Intensional Answers and Exploratory Computing to tackle the "Big Data" challenge

Letizia Tanca

DEIB, Politecnico di Milano

co-authors:

Nicoletta Di Blas, Mirjana Mazuran,Paolo Paolini, Elisa Quintarelli, Gianni Mecca

Motivation

- Organizations grow and generate more information: they capture billions of bytes of information about their customers, suppliers and operations
- The pervasiveness of digital technologies has changed the way individuals interact with the external world (sensor technologies) and with one another (social media), generating a huge mass of content

Data has become as a torrent that flows through all possible digital channels.

Managing data abundance

- The term *Information overload* was already used by Alvin Toffler in his book Future Shock, back in 1970.
- It refers to the difficulty to understand and make decisions when too much information is available
- The interpretation of the obtained answers may be non-trivial, since the dataset returned as answer may be too big to be easily human-readable

State of the Art

Many approaches have been developed

They attack one specific side of the problem:

- efficient querying
- analysis techniques that summarize data or reduce its dimensionality
- data visualization

Data analysis

A process of inspecting, cleaning, transforming, and modelling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Not only data analysis

- *Data exploration*: a preliminary exploration of the data to better understand its characteristics.
- *Exploratory Data Analysis* (Tukey, J. W. (1977)): a field of statistics and data analysis where the role of the researcher is to explore the data in many possible ways keeping an open mind instead of starting with a-priori hypotheses.
- *Data mining* (Agrawal, Srikant, 1994): focuses on modeling and knowledge discovery for predictive or descriptive purposes.
- *Business intelligence*: relies heavily on aggregation: On Line Analytical Processing (OLAP), typically performed over a Data Warehouse (Golfarelli and Rizzi, 2009, Kimball, 1996).
- *Faceted search*, also called faceted navigation, (Tunkelang, 2009, Yee et al., 2003, Ranganathan, 1964)
- *Intensional answers* (Pirotte, Goldin, Kanellakis, ...)
- *Approximate intensional answers* (Mazuran, Quintarelli and Tanca, 2012)
- *Personalization and context-awareness:* can eliminate *information noise* reducing the available data only to the part that is appropriate for the current user and context (Tanca et al., 2008, etc.)

Intensional answers

- Ever since Frege and Russell's doctrine on the foundations of Mathematics, the term **intension** suggests the *idea of denoting objects by means of their properties rather than by exhibiting them*
- Intensional characterization replaces a lengthy list of items with a succinct description.
- In real life we use intensional knowledge very often, since our brain is much more apt to capturing (and reasoning over) properties of objects, than to memorizing long lists of them
- The egg of Columbus? Intensional definitions will allow us to *make sense of Big Data*.

Intensional answers: the problem

- Very seldom an intensional definition is possible, since finding a minimal and complete set of properties that precisely characterize a collection of data is easier in mathematics than in real life!
- Often, reality can be (partially) described by means of succinct, but approximate, intensional properties.
- "80% of crimes are robberies"

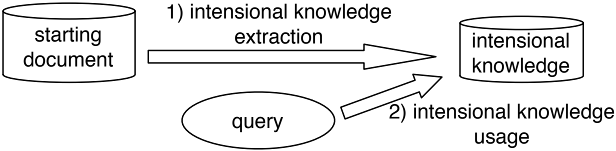
Idea: investigate new approaches to support flexible queries in the context of massive, possibly semistructured, datasets.

Approximate intensional answers: example

- **Database:** crimes in the EU
- **approximate intensional knowledge:**
  - "80% of crimes carried out in Italy are robberies"
  - "in 65% of gunfights Full Metal Jacket bullets were used"
  - "in 73% of assaults bullets with 5mm diameter were used"
  - "78% of crimes carried out in the UK involve blue Fords"
  - ...
- **query:** "retrieve the crimes carried out in Italy"
  - **extensional answer:** *list* of all crimes carried out in Italy
  - **(a possible) intensional answer:** "80% of crimes carried out in Italy are robberies"

Vision

- Given a huge document, D
- Provide a way of:
  1. extracting intensional, approximate knowledge from D
  2. using this intensional knowledge in order to:
     - provide quick, approximate information on both the structure and the content of D
     - provide approximate answers to queries over D


starting document
1) intensional knowledge extraction
intensional knowledge
query
2) intensional knowledge usage

Association rules

- “Implications” extracted with data mining techniques from a database D

$$\{\texttt{country="Italy"}\} \Rightarrow \{\texttt{crime\_type="robbery"}\}$$

- They quantify the correlation between the elements in Body and those in Head

$$support = \frac{frequency(\{\texttt{Italy}, \texttt{robbery}\}, D)}{cardinality(D)} = 0.2$$

$$confidence = \frac{frequency(\{\texttt{Italy}, \texttt{robbery}\}, D)}{frequency(\{\texttt{Italy}\}, D)} = 0.8$$

- They are used to extract approximate knowledge

XML data

Motivations

- XML data is growing fast because XML is a flexible model to represent and share semistructured information
- We have experienced the growth of huge XML documents which are hard to manage because XML is very verbose:
    - a lot of storage space is needed
    - query response time is high
- Analysis of an XML document in order to extract *approximate intensional knowledge*

XML data

Contributions

1) Definition of Tree-based Association Rules (TARs)

- a new way of representing approximate intensional knowledge
- based on the association rule paradigm

2) Definition of methods for managing TARs

- extraction
- storage
- usage

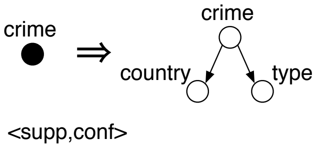
3) Definition of algorithms for querying TARs

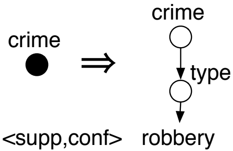
- σ/π-queries
- count-queries
- top-k-queries

Tree-based Association Rules

What are they?

- They are both trees and association rules
  - structure TARs
  - instance TARs





Support and confidence

$$support(S_B \Rightarrow S_H) = \frac{frequency(S_H, D)}{cardinality(D)}$$

$$confidence(S_B \Rightarrow S_H) = \frac{frequency(S_H, D)}{frequency(S_B, D)}$$

- They preserve the structure of the extracted information

Tree-based Association Rules

How are they extracted?

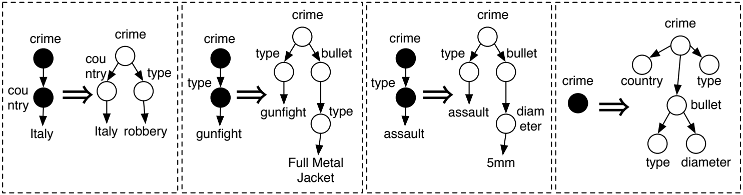
- Given the tree-based representation of an XML document:
    1. *frequent* subtrees are extracted (support above the threshold)
    2. for each frequent subtree, *interesting* rules are computed (confidence above the threshold)

1. There are many algorithms for frequent subtree extraction. This work is based on the use of CMTreeMiner (Y. Chi, Y. Yang, Y. Xia, R. R. Muntz, 2003)
2. Given a frequent subtree S:
    - all possible, not empty, node subsets B are generated
    - the rule `B` $\Rightarrow$ `(S - B)` is generated
    - the rule is considered "interesting" if its confidence is above the threshold

Tree-based Association Rules

How are they stored?

- Graphically rules are represented as trees



- Phisically rules are stored in an XML file

```
<rule id="1" support="0.01"
        confidence="0.8">
  <crime body="true">
    <country body="true">
      Italy
    </country>
    <type body="false">
      robbery
    </type>
  </crime>
</rule>
```

```
<rule id="2" support="0.01"
        confidence="0.65">
  <crime body="true">
    <bullet body="false">
      <type body="false">
        Full Metal Jacket
      </type>
    </bullet>
    <type body="true">
      gunfight
    </type>
  </crime>
</rule>
```

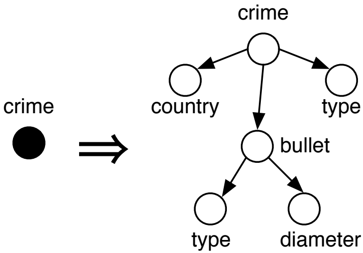
```
<rule id="3" support="0.01"
        confidence="0.73">
  <crime body="true">
    <bullet body="false">
      <diameter body="false">
        5mm
      </diameter>
    </bullet>
    <type body="true">
      assault
    </type>
  </crime>
</rule>
```

```
<rule id="4" support="0.03"
          confidence="0.9">
  <crime body="true">
    <country body="false">
    </country>
    <type body="false"></type>
    <bullet body="false">
      <type body="false"></type>
      <diameter body="false">
      </diameter>
    </bullet>
  </crime>
</rule>
```

structure Tree-based Association Rules

What are they used for?

- They provide information about the structure of the XML document:
    - useful when the XML document does not have an explicit DTD
    - can be used as a DataGuide (Goldman, Widom, 1997) to allow queries which are consistent with the data contained in the XML document



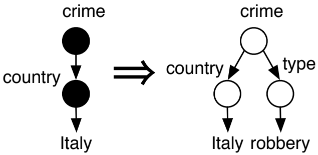
instance Tree-based Association Rules

What are they used for?

- They provide approximate knowledge about the content of the XML document and can be used for query answering:
  - queries that are too specific may not return results
  - we allow three classes of queries

- ***σ*/*π*-queries**: "Retrieve all crimes reported in Italy"

  we look for a match in both the antecedent and consequent of the extracted TARs



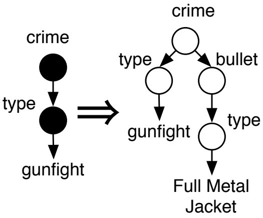
Intensional query answering

Count queries

- **count-queries**: "Retrieve the number of gunfights"

$$supp = \frac{frequency(S_H)}{cardinality} = 0.01$$

$$conf = \frac{frequency(S_H)}{frequency(S_B)} = 0.65$$



- match in the antecedent

$$frequency(S_B) = \frac{supp * cardinality}{conf} = 2.968 \approx 3$$

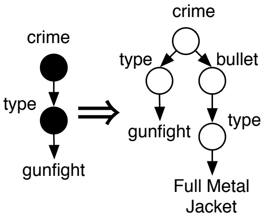
- match in the consequent

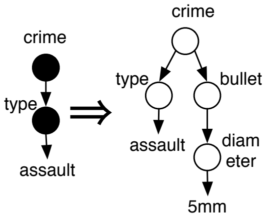
$$frequency(S_H) = supp * cardinality$$

Intensional query answering

Top-k queries

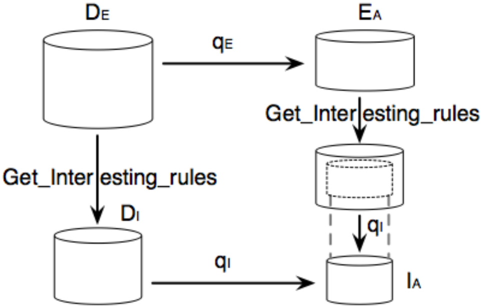
- **top-k-queries**: “Retrieve the 3 most frequent types of crime”





$frequency(S_{B_1}) = 2.968 \approx 3$ $\qquad$ $frequency(S_{B_2}) = 0.77 \approx 1$

Intensional query answering commutative diagram

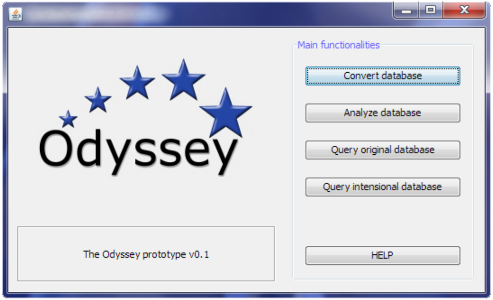


- D_E: original document
- D_I: intensional knowledge
- q_E: query over extensional knowledge
- q_I: query over intensional knowledge
- E_A: extensional answer
- I_A: intensional answer

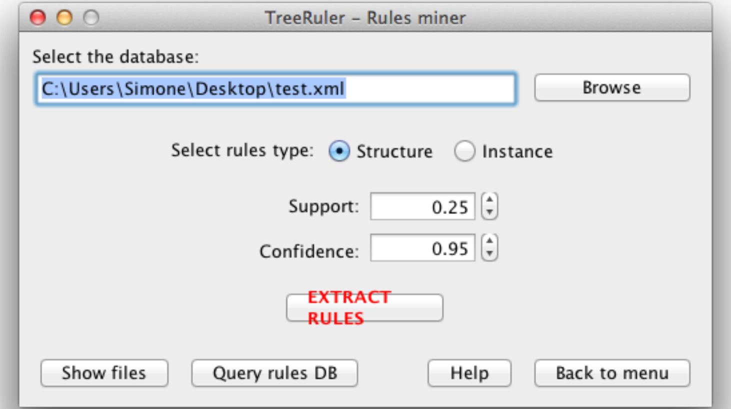
Prototype

TreeRuler

- implemented in Java (and Web)
- manages both XML and relational data
- allows:
  1. intensional knowledge extraction
     - Tree-based association rules from XML documents
     - standard association rules from relational datasets
  2. original dataset querying
  3. intensional knowledge querying



TreeRuler



TreeRuler – Rules miner

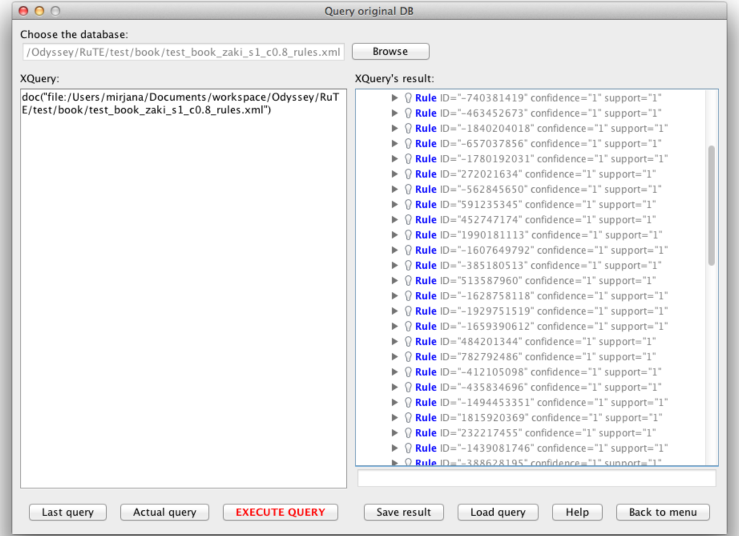
Select the database: C:\Users\Simone\Desktop\test.xml

Select rules type: ● Structure ○ Instance

Support: 0.25

Confidence: 0.95

TreeRuler



Choose the database:

/Odyssey/RuTE/test/book/test_book_zaki_s1_c0.8_rules.xml

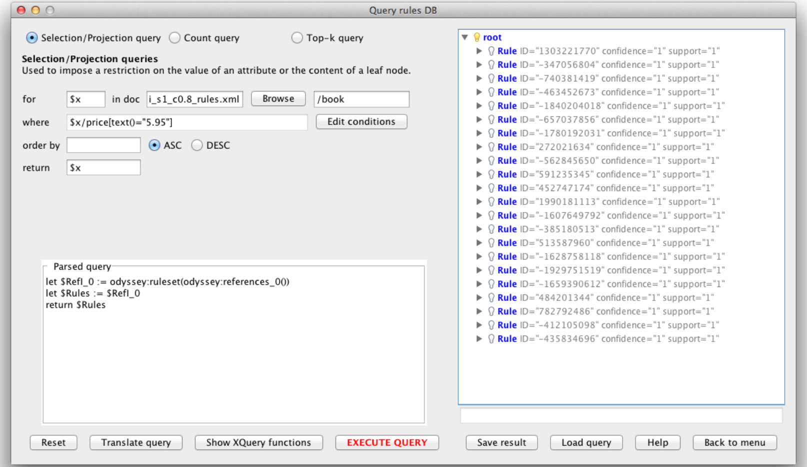
XQuery:

```
doc("file:/Users/mirjana/Documents/workspace/Odyssey/RuT
E/test/book/test_book_zaki_s1_c0.8_rules.xml")
```

XQuery's result:

```
▶ Rule ID="-740381419" confidence="1" support="1"
▶ Rule ID="-463452673" confidence="1" support="1"
▶ Rule ID="-1840204018" confidence="1" support="1"
▶ Rule ID="-657037856" confidence="1" support="1"
▶ Rule ID="-1780192031" confidence="1" support="1"
▶ Rule ID="272021634" confidence="1" support="1"
▶ Rule ID="-562845650" confidence="1" support="1"
▶ Rule ID="591235345" confidence="1" support="1"
▶ Rule ID="452747174" confidence="1" support="1"
▶ Rule ID="1990181113" confidence="1" support="1"
▶ Rule ID="-1607649792" confidence="1" support="1"
▶ Rule ID="-385180513" confidence="1" support="1"
▶ Rule ID="513587960" confidence="1" support="1"
▶ Rule ID="-1628758118" confidence="1" support="1"
▶ Rule ID="-1929751519" confidence="1" support="1"
▶ Rule ID="-1659390612" confidence="1" support="1"
▶ Rule ID="484201344" confidence="1" support="1"
▶ Rule ID="782792486" confidence="1" support="1"
▶ Rule ID="-412105098" confidence="1" support="1"
▶ Rule ID="-435834696" confidence="1" support="1"
▶ Rule ID="-1494453351" confidence="1" support="1"
▶ Rule ID="1815920369" confidence="1" support="1"
▶ Rule ID="232217455" confidence="1" support="1"
▶ Rule ID="-1439081746" confidence="1" support="1"
▶ Rule ID="-388628195" confidence="1" support="1"
```

TreeRuler



● Selection/Projection query ○ Count query ○ Top-k query

Selection/Projection queries
Used to impose a restriction on the value of an attribute or the content of a leaf node.

for: $x in doc: i_s1_c0.8_rules.xml /book

where: $x/price[text()="5.95"]

order by: ● ASC ○ DESC

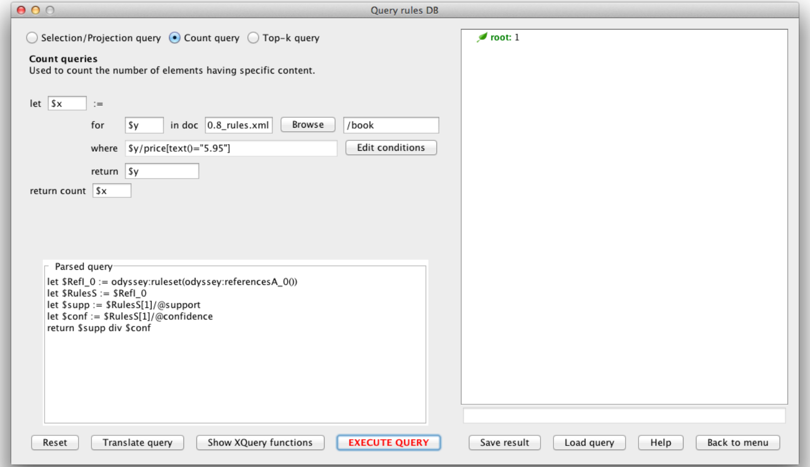
return: $x

Parsed query

```
let $Refl_0 := odyssey:ruleset(odyssey:references_0())
let $Rules := $Refl_0
return $Rules
```

```
▼ root
  ▶ Rule ID="1303221770" confidence="1" support="1"
  ▶ Rule ID="-347056804" confidence="1" support="1"
  ▶ Rule ID="-740381419" confidence="1" support="1"
  ▶ Rule ID="-463452673" confidence="1" support="1"
  ▶ Rule ID="-1840204018" confidence="1" support="1"
  ▶ Rule ID="-657037856" confidence="1" support="1"
  ▶ Rule ID="-1780192031" confidence="1" support="1"
  ▶ Rule ID="272021634" confidence="1" support="1"
  ▶ Rule ID="-562845650" confidence="1" support="1"
  ▶ Rule ID="591235345" confidence="1" support="1"
  ▶ Rule ID="452747174" confidence="1" support="1"
  ▶ Rule ID="1990181113" confidence="1" support="1"
  ▶ Rule ID="-1607649792" confidence="1" support="1"
  ▶ Rule ID="-385180513" confidence="1" support="1"
  ▶ Rule ID="513587960" confidence="1" support="1"
  ▶ Rule ID="-1628758118" confidence="1" support="1"
  ▶ Rule ID="-1929751519" confidence="1" support="1"
  ▶ Rule ID="-1659390612" confidence="1" support="1"
  ▶ Rule ID="484201344" confidence="1" support="1"
  ▶ Rule ID="782792486" confidence="1" support="1"
  ▶ Rule ID="-412105098" confidence="1" support="1"
  ▶ Rule ID="-435834696" confidence="1" support="1"
```

TreeRuler



○ Selection/Projection query ● Count query ○ Top-k query

Count queries
Used to count the number of elements having specific content.

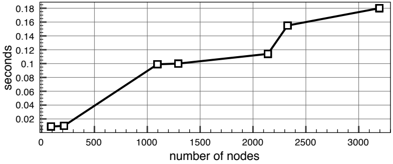
```
let $x :=
        for $y in doc 0.8_rules.xml  /book
        where $y/price[text()="5.95"]
        return $y
return count $x
```

Parsed query

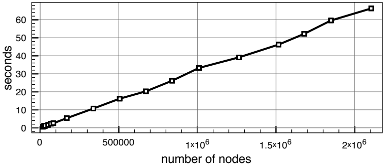
```
let $Refl_0 := odyssey:ruleset(odyssey:referencesA_0())
let $RulesS := $Refl_0
let $supp := $RulesS[1]/@support
let $conf := $RulesS[1]/@confidence
return $supp div $conf
```

root: 1

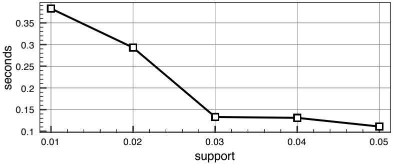
Experimental results



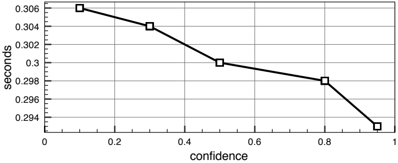
Extraction time real documents



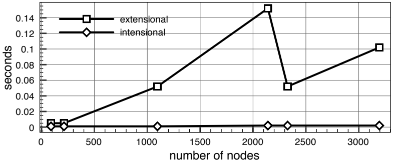
Extraction time synthetic documents



Extraction time confidence=0.95

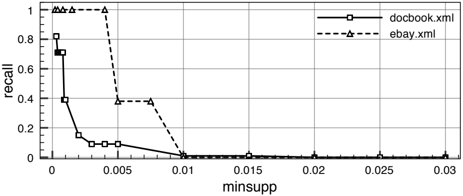


Extraction time support=0.02



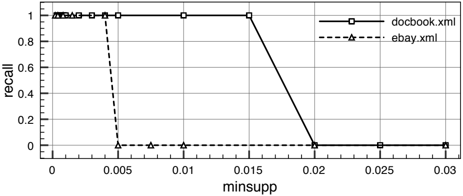
Answer time

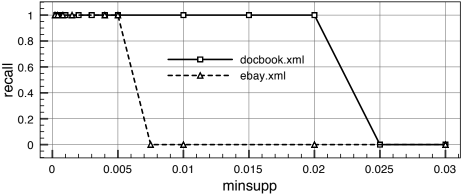
Experimental results



σ/π-queries recall

count-queries recall





top-k-queries recall

Conclusions

- New ways for dealing with the Big Data problem
- We considered both relational and tree-based data whose growth has been significant in recent years
- The TreeRuler tool allows us to analyze tree-shaped data and query both the data itself and its frequent properties
- We plan to investigate the problem also for graph-based data
- Long-term goal: a formal framework for manipulating intensionally-defined datasets.

The Exploratory Computing paradigm

Taking a comprehensive view

- supporting a variety of user experiences, like investigation, inspiration seeking, monitoring, comparison, decision-making, research
- taking into account that human exploration should be an iterative and multi-step process, in a sort of "dialogue" between the user and the system.
- dealing with "semantically rich" datasets

A step-by-step progress

Exploratory Computing (EC) is a process supported by the system which, at each user request, gives her feedback by emphasising the interesting properties of the current result and possibly suggesting one or more possible actions that can be taken in her next exploration step.

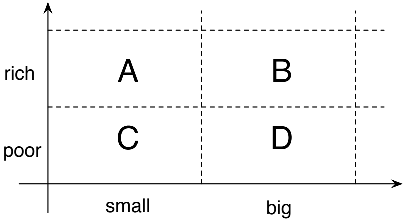
An exploratory experience...

...basically consists in making sense of data by creating, refining, modifying, comparing various datasets

Rich vs. poor...

Two ample categories of Big Data

- *semantically poor*: time series, data from industrial plants, sensor readings in general
- *semantically rich*: business data, data about health, and in fact most of the data that must be directly examined by users to the purpose, for instance, of taking a decision.



Rich vs. poor (2)

Poor Data

poor data lend themselves to be summarised and compressed in many ways, providing the users with a concise idea of a dataset contents

Rich Data

obtaining exhaustive, deep, if necessary detailed knowledge of rich, faceted data needs a step-by-step process where the result of a first query should be thoroughly understood before progressing towards issuing the next one

Accordingly...

...the name "Exploratory Computing" underlines the fact that the various steps of the same exploratory experience may entail a wide range of different operations, not confined to simple search or querying but including the use of statistics and data analysis, exact and approximate query-answering, support for intensional answers, data exploration, query suggestion, graphical and design tools, personalised answers, recommendations, etc. , all with the aim of providing users with the necessary feedback to progress towards the next step

Preliminary experiences

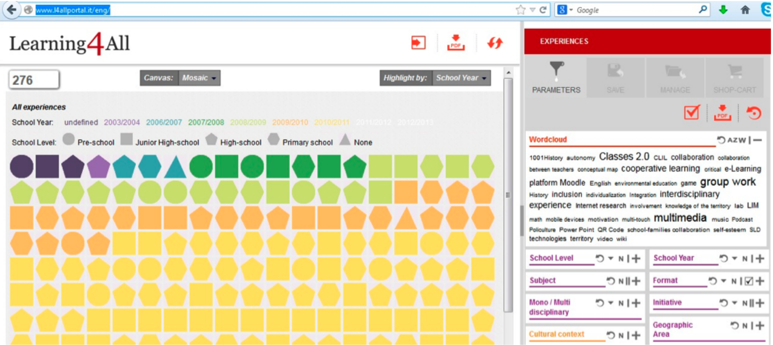
The Learning4All portal `http://www.l4allportal.it/eng/` is based on a dataset of reports about technology-based educational experiences at school

Each item in this dataset includes rich information about the pedagogical implementation, like the subject dealt with, the benefits achieved, the technology used, etc.

The educational experiences are tagged according to a sophisticated taxonomy of more than 30 facets (e.g. "subject matter") and 200 values (e.g. "humanities").

Portals like Learning4Al are located in square A: the added value of EC will be fully expressed when the involved technological background and tools are mature enough to move this paradigm to square B.

The user interface of the Learning4All portal



276

Canvas: Mosaic

Highlight by: School Year

All experiences

School Year: undefined 2003/2004 2006/2007 2007/2008 2008/2009 2009/2010 2010/2011 2011/2012 2012/2013

School Level: ● Pre-school ■ Junior High-school ⬟ High-school ⬢ Primary school ▲ None

Wordcloud AZW

1001History autonomy Classes 2.0 CLIL collaboration collaboration between teachers conceptual map cooperative learning critical e-Learning platform Moodle English environmental education game group work History inclusion individualization Integration interdisciplinary experience Internet research involvement knowledge of the territory lab LIM math mobile devices motivation multi-touch multimedia music Podcast Policulture Power Point QR Code school-families collaboration self-esteem SLD technologies territory video wiki

School Level N

School Year N

Subject N

Format N ☑

Mono / Multi disciplinary N

Initiative N

Cultural context N

Geographic Area N

To the right, part of the classification taxonomy

Aims and characterizing aspects

1. Investigation/understanding/making sense: the user wants to understand a phenomenon and/or has an ill-defined idea about what she may be interested in; as the exploration of the dataset moves on, she either refines, focuses, expands or changes her initial attitude.
2. Inspiration seeking: the user has a quite generic interest and she is willing to be offered suggestion.
3. Supervision: the user needs to understand how things are going about something.
4. Comparison: the user needs to compare two phenomena, under various perspectives.
5. Decision-making: the user has to take a well-informed decision about something.
6. Research: the user wants to refine or verify some research hypothesis, or she is looking for research hypothesis.
7. Leisure browsing: the user just wants to stroll around playing with data.

Example 1

Investigation

Janet, a teacher at junior high school, would like to introduce tablets in her class. Examples of "Facets" :

- what benefits they brought about
- what technology was used
- what kind of pedagogical implementation was adopted

Janet selects "tablets" within the facet about technology:

- what changes has this selection brought about within the other facets? Too bad!! relational benefits drop down dramatically!
- maybe a more collaboration-supporting technology could be a better choice, since children of Janet's class are not very close with one another
- Janet changes her mind and starts a new exploration, selecting "relational benefits" to see what she could use to support them...

Example 2

Inspiration-seeking

Jack, a teacher at primary school level, is told that his class will be given a tablet for each pupil.

- Jack selects “primary school” as school level and “tablet” as technology
- He takes a look at how the other facets have reacted and he sees that tablets do not seem to promote group-work but rather individual work.
- Jack then explores the benefits and sees that cognitive benefits seem quite good
- He therefore concentrates on those experiences and saves all the related material, including the teachers’ reports, interviews, detailed descriptions of the activities’ implementation, for finding inspiration about how to use the tablets with her pupils.

Example 3

Supervision

Karen, an educational director at district level wants to know how things are going with Interactive White Boards at junior high-school level.

- She accesses the Learning4All portal, selects "junior high school" and also the geographical area of her interest and then she checks how values in all the facets have changed
- She makes a number of positive discoveries: relational benefits are improving, students are more motivated, media literacy is increasing
- Karen collects all this information to prepare her report for the ministry.

Example 4

Comparison

Let's get back to Janet, the teacher of Example 1. She is wondering whether there is any difference according to the school level in the way storytelling is implemented within the class.

- Janet creates further subsets according to the school levels. She now compares them
- Janet sees that, from the benefits point of view, cognitive benefits are higher at high-school level with respect to lower grades
- She also discovers that heterogeneous group-work (i.e. groups where students of different performances work together) is more used at lower grades.
- She goes on comparing the sets for further discoveries.

Example 5

Decision-making

Karen, the same educational director of Example 3, has to decide whether to extend the adoption of Interactive White Board to primary school level. She makes the same exploration described above and then checks closely the way activities with the IWBs are implemented within the class:

- are they suitable for primary school too?
- She saves all the experiences where implementation strategies seem appropriate for younger pupils,
- she goes carefully through them and concludes that IWBs could be a very good resource for that school level too.

Examples 6 and 7

Research

Michael, a scholar in educational technology would like to understand whether digital storytelling in the class can support group work.

- Michael accesses the Learning4All portal and selects “digital storytelling” as technology-based activity
- he discovers that group-work is the most widely adopted pedagogical implementation
- Michael saves all the experiences characterized by “digital storytelling” and “group-work” for further investigation.

Leisure-browsing

Teacher Christine,, who has just heard about the Learning4All portal, accesses it and browses around at leisure starting from the class organization. She bumps into interesting facts, like for example that in order to support inclusion homogeneous group work can be recommendable and so on. She keeps browsing and making serendipitous discoveries.

First conclusions

- The interactive process matters more than single queries; an exploratory experience builds upon previous turn-takings and does not start from scratch at every new step.
- In dialogue terms, the feedbacks are as relevant as the queries; it means that new turns of events depend on how the system reacts to the user's choices and requests.
- The emphasis of the feedback is on the properties of the sets of items rather than on their sheer list; in mathematical words, the intension of the set the system provides as feedback matters more than its extension.

Requirements

In human-to-human dialogue people *do not start from scratch at every turn-taking*: on the contrary, what one of the interlocutors says is crucial to steer the conversion towards sometimes unexpected turns, and where the meaning of what is being said is generally more relevant than sheer lists of items.

- An EC user typically reaches her goal through a number of sessions, each one consisting of *several tasks, often unplanned*, in the sense that the user refines her needs and wishes along the way.
- Each session displays an iterative process: an initial dataset gets refined into smaller datasets. For example, the teacher from Example 2 refines the initial dataset by selecting the experiences related to primary school; then, he refines the dataset selecting only those were tablets are used, and so on.

Technical requirements

Data modeling:

Data are modeled according to "facets" (i.e. categories); facets can be flat or hierarchically organized. For example, a facet describing the technology used at school can be represented by a flat attribute "Technology", while a facet describing the geographical position of schools can be represented by a hierarchy of attributes "City→ Region → State". Facets can be modelled as classical attributes or using Boolean attributes. For example, the facet "school level" can be represented as an attribute whose domain is {primary, junior, high-school,...}, or as a set of Boolean attributes where each attribute represents one value of the domain.

Experience modeling:

The exploratory experience can be *modelled as a graph*, where each node represents a step in the exploratory process.

Graph traversal:

- The experience starts by accessing an initial portion of the data, represented by a node of the graph
- The set is iteratively refined, moving to one of the accessible nodes, each representing the current dataset.
- Each refinement brings the user into a new step of the process.
- As the process continues, a *path in the graph is created* where each node represents a subset of the initial data and each edge represents a refinement action that brings from one set of data to the other one.

The system should be able to:

- suggest data refinements to the user
- allow the possibility of backtracking from each iterative step.

the first feature means that the EC system should suggest, at each step of the exploratory process, one or more *interesting refinements* to the user

Relevance

A definition of *"interestingness"*, or *"relevance"*, is needed to understand which facets should be suggested.

At each iteration

Summarization:

Being able to compute some properties that describe in a compact way the values taken by the facets in the query results These properties might be:

1. simple statistical measures (e.g. mean, median, etc.);
2. more complex statistical properties , like the distribution of each facet;
3. other (intensional, approximate) knowledge about the facets such as frequent approximate properties of a single facet or correlations between two different facets within a query result.

Since we might deal with huge amounts of data, *developing efficient computation algorithms, that summarize data in various ways, plays a key role in the development of EC systems.*

At each iteration

Intensional comparisons:

- The result of an exploration step gains more importance if it is compared w.r.t. the user's "expectations"
- Given a facet (or a set of facets), the user's expectation of a statistical property (e.g. the mean, the distribution) for that facet, evaluated on a set of data, *might dynamically change from one refinement step to the other* and can be expressed in several ways.
- E.g., if the use of tablets at primary school in Italy is 20% on average, the system should help the user to detect that in a specific region the use is 35% instead, and it should also help to identify the implications on the other facets.

The user expectation might be synthetically expressed:

- as a constant specified by the user or chosen for the case study (e.g. the mean value of a facet in a data set is expected to be equal to 0);
- as a particular distribution (e.g. the uniform one)
- as the distribution of the values obtained in a previous exploration step;

A digression on the concept of *relevance*

Different possible meanings of *relevant facet* :

- If it *differs with respect to the user's expectations*: E.g. the teacher, as a first step, searches the benefits of the experiences without the use of tablets. As a second step, she expects that the benefits, when using the tablet will increase, still keeping a similar distribution. If the distribution of the benefit facet is very different in the two steps of the exploration and in particular if it does not increase, the benefit facet should be suggested by the system as a relevant aspect to be investigated.
- If it is *close to the user's expectations:* user's expectations may have different origins: previous background, common knowledge, exploration on previous similar datasets, exploration of other portions of this datasets, etc.
- If it *differs with respect to the previous (or initial) dataset*: E.g. the teacher searches the benefits of the experiences without the use of tablets. If the distribution of the benefit facet is very different from its distribution over the whole dataset then the benefit facet may be considered as a relevant aspect.

Intensional descriptions may help:

- Defining the common set operations, such as union, intersection and difference, also for *intensionally defined sets*.
- What is the actual meaning of applying the difference to two sets of *relevant properties* of two sets of data instead of applying it to the two sets? For instance, in the Examples some kind of *intensional difference* between the experiences S_1, that have involved the use of tablets in Northern Italy, and those of Southern Italy (S_2) must be applied, computing those properties that "make the difference" between the two sets S_1 and S_2
- The answer might be that the wealth distribution of Southern Italy is different from that of Northern Italy, and this should be pointed out by the system for the user to take a decision about the next exploratory step.

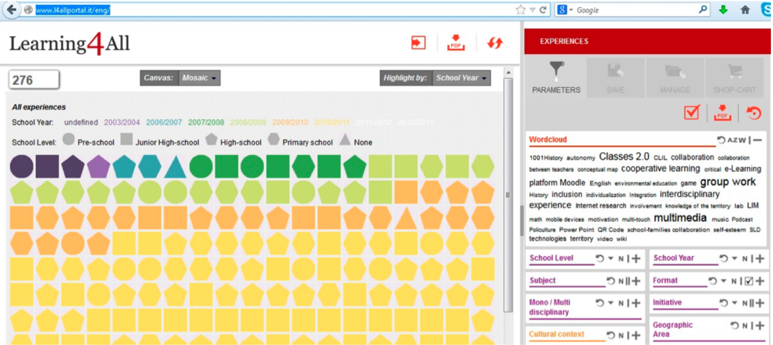
At each iteration

Visualization:

- The properties determined by the EC system should be shown to users in a comprehensive way
- Thus efficient and effective visualizations are needed.
- The figure shows the use of different colors and shapes to emphasize data facets
- Here different colors stand for different temporal spans while different shapes denote different school levels

Research on visualization carried out in the area of Exploratory Data Analysis can come to the rescue in this task.

The user interface of the Learning4All portal



www.l4allportal.it/eng/

276

Canvas: Mosaic

Highlight by: School Year

Google
Learning4All
All experiences
School Year: undefined 2003/2004 2006/2007 2007/2008 2008/2009 2009/2010 2010/2011 2011/2012 2012/2013
School Level: Pre-school Junior High-school High-school Primary school None
EXPERIENCES
PARAMETERS
SAVE
MANAGE
SHOP-CART
PDF
Wordcloud
AZW
1001History autonomy Classes 2.0 CLIL collaboration collaboration between teachers conceptual map cooperative learning critical e-Learning platform Moodle English environmental education game group work History inclusion individualization Integration interdisciplinary experience Internet research involvement knowledge of the territory lab LIM math mobile devices motivation multi-touch multimedia music Podcast Policulture Power Point QR Code school-families collaboration self-esteem SLD technologies territory video wiki
School Level N
School Year N
Subject N
Format N
Mono / Multi disciplinary N
Initiative N
Cultural context N
Geographic Area N

To the right, part of the classification taxonomy

Conclusions

Exploratory Computing provides a new paradigm for accessing big and rich datasets

Basic Constituents:

- The interactive aspects of the paradigm, like canvases to display objects, widgets etc. are basic constituents
- The feedback from the system is as important as the requests by the users
- The exploration is not about "which objects belong to" a given dataset, but more about "which are the relevant properties" of these objects
- The feedback can take different flavors, for example highlighting the facets' distribution or how a facet influences other facets.
- The richness of the taxonomy should go beyond the user's expectations leading to *serendipitous discoveries*

Future work:

- Deeper understanding of the internal modeling, introducing at a larger extent the notion of relevance in its various interpretations;
- Better understanding of the properties of a dataset in terms of properties' distribution;
- Implementation of the operations for detecting facets' influences and comparison among different various datasets;
- Tackling a new challenge and extending EC to big (not just rich) datasets.