Data Mining to Support Intensional Answering of Big Data

Mirjana Mazuran, Elisa Quintarelli, Letizia Tanca

DEIB – Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy

Motivations

- **Information overload** was coined 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*
- In the Big Data era, exponential growth, availability and use of information makes this problem even more dramatic!
- A core issue for companies that use new forms of information such as social-network-originated data or biological data.
- The new challenge is *making sense of data*:
    - they are unstructured, irregular
    - they may contain errors, inconsistencies
    - query answers may be so huge that users *don't get the gist* of the resulting dataset

Motivations

- 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*.

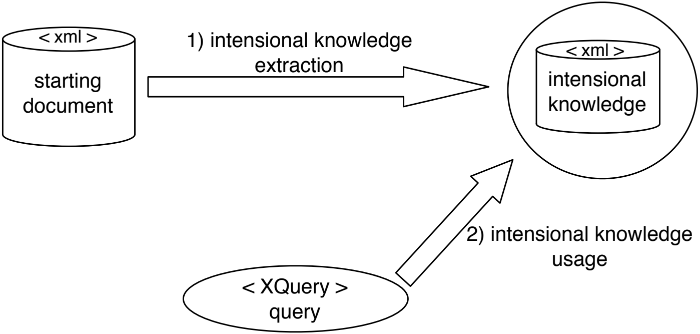
Our proposal

- 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"

We investigate new approaches to support flexible queries in the context of massive, often semistructured, datasets. We focus on relational data and XML data (and RDF data as future work).

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



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"

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

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

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

- 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

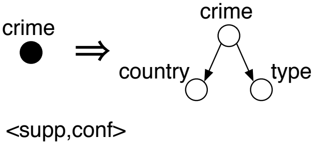
3) Definition of algorithms for querying TARs

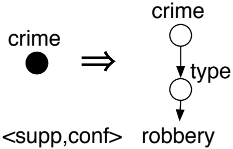
- σ/π-queries
- count-queries
- top-k-queries

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?

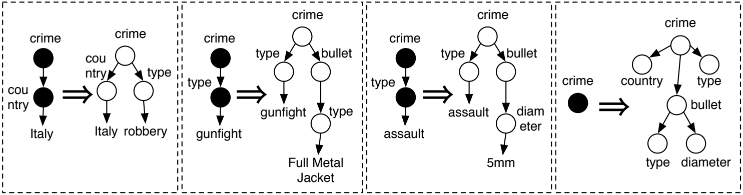
- Given the tree-based representation of an XML document:
  1. *frequent* subtrees are extracted (support above the threshold)
  2. for each frequent subtree, *interesting* rules are computed (confidence above the threshold)

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` ⇒ `(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>
```

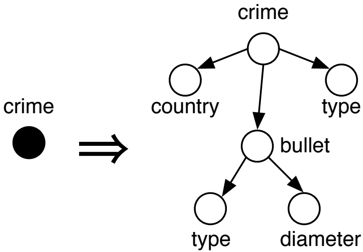
```
<rule id="3" support="0.01"
          confidence="0.73">
  <crime body="true">
    <bullet body="false">
      <diameter body="false">
        5mm
      </diameter>
    </bullet>
    <type body="true">
      assault
    </type>
  </crime>
</rule>
```

```
<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



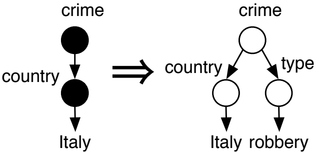
instance Tree-based Association Rules

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

- $\boldsymbol{\sigma}/\boldsymbol{\pi}$**-queries**: “Retrieve all crimes reported in Italy”

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



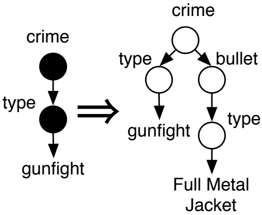
Intensional query answering

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

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

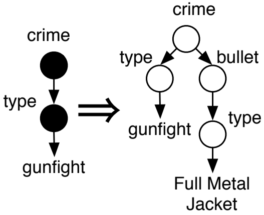
- match in the consequent

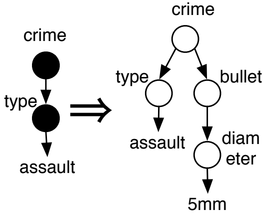
$$frequency(S_H) = supp * cardinality$$

Intensional query answering

Top-k queries

- **top-k-queries**: “Retrieve the 2 most frequent types of crime”





$frequency(S_{H_1}) = 2.968 \approx 3$ $frequency(S_{H_2}) = 0.77 \approx 1$

Theorems

We have proven that the intensional answer constitutes a *representation* of the frequent properties of the extensional one.

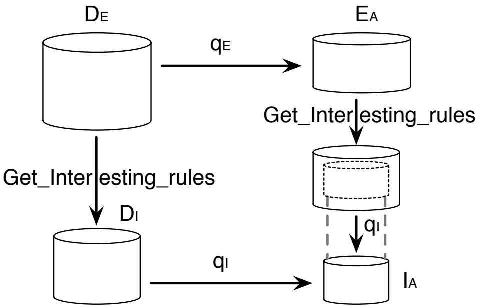
Theorem

Let q_E be a σ/π query on the XML document D_E, q_I the intensional rewriting of q_E, E_A the XML document obtained as result for q_E, and I_A the intensional answer to q_I. The procedure to obtain intensional answers is **sound***, that is, if a TAR $T_r \in I_A$ then $T_r \in q_I(Get - Interesting - Rules(E_A))$.*

Theorem

Let q_E be a σ/π query on the XML document D_E, q_I the intensional rewriting of q_E, E_A the XML document obtained as result for q_E, and I_A the intensional answer. If, in the mining process, the imposed support and confidence thresholds are 0, the procedure to obtain intensional answers is such that $q_I(Get - Interesting - Rules(E_A)) = I_A$, that is, the procedure is both **sound** *and* **complete***.*

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: intentional answer

Theorems

The result of count-queries is exact, up to the approximation introduced by the computation of the support and confidence.

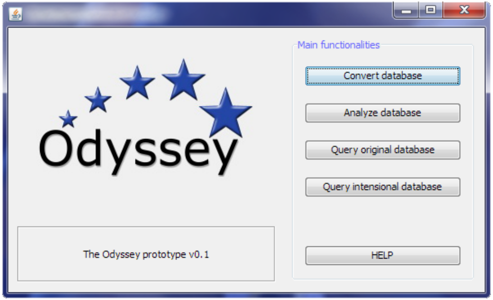
Theorem

Let q_E be a count−query on the XML document D_E, q_I the intensional rewriting of q_E, $\texttt{count}_\texttt{E}$ *the extensional answer, and* $\texttt{count}_\texttt{I}$ *the intensional answer. If we can mine at least a TAR exactly satisfying in the antecedent the constraints in q_E then* $\texttt{count}_\texttt{E} \approx \texttt{count}_\texttt{I}$*, that is, the procedure is* ***sound****.*

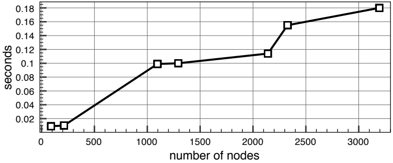
Prototype

TreeRuler

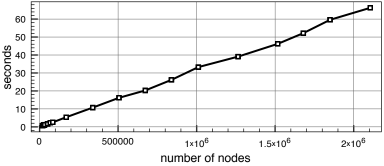
- implemented in Java (and Web)
- manages both XML and relational data
- allows:
  1. intensional knowledge extraction
     - Tree-based association rules from XML documents
     - standard association rules from relational datasets
  2. original dataset querying
  3. intensional knowledge querying



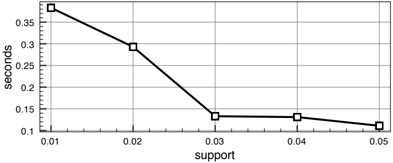
Experimental results



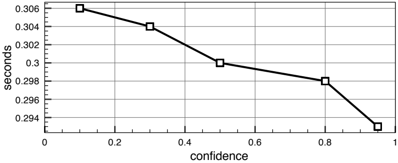
Extraction time real documents



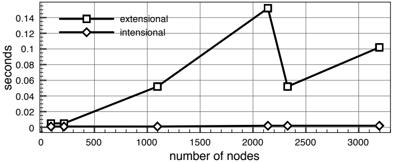
Extraction time synthetic documents



Extraction time confidence=0.95

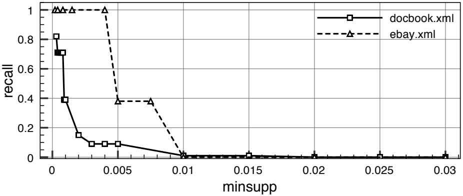


Extraction time support=0.02



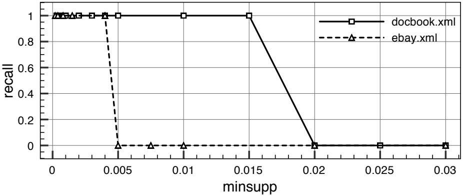
Answer time

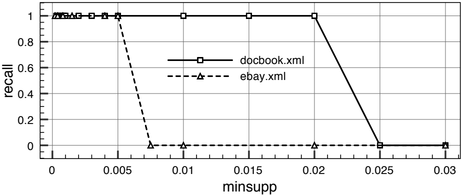
Experimental results



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.

Publications

- Mirjana Mazuran, Elisa Quintarelli, Letizia Tanca: Data mining for XML query-answering support. *IEEE Transactions on Knowledge and Data Engineering, 2012*
- Mirjana Mazuran, Elisa Quintarelli, Letizia Tanca: Mining Tree- Based Frequent Patterns from XML. *Proceedings of the 8th International Conference on Flexible Query Answering Systems, 2009*
- Elena Baralis, Paolo Garza, Elisa Quintarelli, Letizia Tanca: Answering XML queries by means of data summaries. *ACM Transactions on Information Systems, 2007*