Collaborative Clustering of XML Documents

Sergio Greco
Francesco Gullo
Giovanni Ponti
Andrea Tagarelli

DEIS – University of Calabria
Outline

- Introduction
  - Motivations
- Our proposal:
  - distributed collaborative approach to XML document clustering
- Experimental evaluation
- Conclusion
Motivations

- In a nutshell, **XML**
  - The extensible, self-describing de-facto standard for data representation and exchange on the Web

- Rapid increase of the volume and heterogeneity of XML sources
  - Documents exhibit too diverse structure and contents
    - may encode related semantics
  - Documents are often schema-less

- **XML data management and XML mining**
  - Web source integration, Querying semistructured data, Document classification
  - Change detection, schema matching
Motivations

- The size of collections of XML documents is often **huge** and inherently distributed
- Classical centralized approaches may be not efficient

Our proposal: a distributed framework for efficiently clustering XML documents
- Peer-to-peer network
- Each peer has access to a portion of the whole document collection
- Centroid-based partitional clustering
- Each peer computes “local” centroids and a subset of “global” centroids
Clustering semantically related XML documents

[Tagarelli and Greco, SDM’06]
[Tagarelli and Greco, TOIS’09]

- **XML features**
  - Structure information (from tag paths)
  - Content information (from textual elements)

- **XML transactional model**
  - based on the notion of XML tree tuple
    - identifies semantically cohesive substructures
    - enables relational-like representation of XML data
Preliminaries

- **XML tree path**
  - A sequence \( p = [s_1, \ldots, s_m] \) of symbols in \( Tag \cup Att \cup \{\$\} \)
  - **Tag Path**: last symbol is a tag name
  - **Complete Path**: last symbol is either an attribute name or a textual element content

- **Path answer:**
  - A set of node identifiers (Tag path case)
  - A set of string values (Complete path case)
Extracting XML tree tuples

**Definition:**

- Given an XML tree $XT$, a *tree tuple* $\tau$ is a maximal subtree of $XT$ such that, for every path $p$ that can be applied to $XT$, the answer $A_\tau(p)$ contains at most one element.

**Meaning in the XML context:**

- A (sub)tree representation of a complete set of distinct concepts that are correlated according to the structure semantics of the original tree.
Extracting XML tree tuples: The DBLP Example
Extracting XML tree tuples: The DBLP Example
Modeling XML transactions

- Decomposition of each tree tuple into a set of tree tuple items
  - Tree tuple item is a pair \((p, A_{\tau}(p))\), such that:
    - \(p\) is a complete path on \(\tau\)
    - \(A_{\tau}(p)\) is the (string) answer of \(p\) applied to \(\tau\)

- **Item**: a tree tuple item

- **Item domain**: the union of the tree tuple item sets over all the tree tuples extracted from a target collection

- **Transaction**: a tree tuple, represented by its set of tree tuple items
  - Each path applied to a tree tuple yields a unique answer \(\Rightarrow\) each item in a transaction refers to a distinct information
## Modeling XML transactions: The DBLP Example

### Path (p) | $\tau_1.p$ | node ID
---|---|---
`dblp.inproceedings.@key` | “conf/kdd/ZakiA03” | $n_3$
`dblp.inproceedings.author.s` | “M. J. Zaki” | $n_5$
`dblp.inproceedings.title.s` | “XRules: an effective ...” | $n_9$
`dblp.inproceedings.year.s` | “2003” | $n_{11}$
`dblp.inproceedings.booktitle.s` | “KDD” | $n_{13}$
`dblp.inproceedings.pages.s` | “316-325” | $n_{15}$

### Path (p) | $\tau_2.p$ | node ID
---|---|---
`dblp.inproceedings.@key` | “conf/kdd/ZakiA03” | $n_3$
`dblp.inproceedings.author.s` | “C. C. Aggarwal” | $n_7$
`dblp.inproceedings.title.s` | “XRules: an effective ...” | $n_9$
`dblp.inproceedings.year.s` | “2003” | $n_{11}$
`dblp.inproceedings.booktitle.s` | “KDD” | $n_{13}$
`dblp.inproceedings.pages.s` | “316-325” | $n_{15}$

### Path (p) | $\tau_3.p$ | node ID
---|---|---
`dblp.inproceedings.@key` | “conf/kdd/Zaki02” | $n_{17}$
`dblp.inproceedings.author.s` | “M. J. Zaki” | $n_{19}$
`dblp.inproceedings.title.s` | “Efficiently mining ...” | $n_{21}$
`dblp.inproceedings.year.s` | “2002” | $n_{23}$
`dblp.inproceedings.booktitle.s` | “KDD” | $n_{25}$
`dblp.inproceedings.pages.s` | “71-80” | $n_{27}$

### Item ID | corresponding node IDs
---|---
$e_1$ | $n_3$
$e_2$ | $n_5, n_{10}$
$e_3$ | $n_9$
$e_4$ | $n_{11}$
$e_5$ | $n_{13}, n_{25}$
$e_6$ | $n_{15}$
$e_7$ | $n_7$
$e_8$ | $n_{17}$
$e_9$ | $n_{21}$
$e_{10}$ | $n_{23}$
$e_{11}$ | $n_{27}$

### Transactions

- $tr_1$: $e_1 \ e_2 \ e_3 \ e_4 \ e_5 \ e_6$
- $tr_2$: $e_1 \ e_7 \ e_3 \ e_4 \ e_5 \ e_6$
- $tr_3$: $e_8 \ e_2 \ e_9 \ e_{10} \ e_5 \ e_{11}$
Clustering XML transactions: XML tree tuple item similarity

- Function of structure and content features
  \[ \text{sim}(e_i, e_j) = f \times \text{sim}_S(e_i, e_j) + (1 - f) \times \text{sim}_C(e_i, e_j) \]

- Match at a degree not below a threshold \( \gamma \)
  - Notion of \( \gamma \)-matched items

- Similarity by structure
  - computed by comparing tag paths

- Similarity by content
  - cosine similarity between TCUs
  - terms in TCUs are weighted by a syntactic relevance function
Clustering XML transactions: XML tree tuple item similarity

- **Structure similarity**
  - Comparison of tag paths by resorting to a simple case of string matching of their respective element names

The *structural similarity* between the XML tree tuple items $e_i$ and $e_j$, having $p_i = t_{i1}.t_{i2}. \ldots .t_{in}$ and $p_j = t_{j1}.t_{j2}. \ldots .t_{jn}$ as their respective tag paths, is:

$$
\text{sim}_S(e_i, e_j) = \frac{1}{n + m} \left( \sum_{t \in p_i} \text{sim}(t, p_j) + \sum_{t \in p_j} \text{sim}(t, p_i) \right)
$$

where

$$
\text{sim}(t_{ih}, p_j) = \text{avg}_{t_{jk} \in p_j} \left\{ \frac{1}{1 + |h - k|} \times \delta(t_{ih}, t_{jk}) \right\}
$$
Clustering XML transactions: XML tree tuple item similarity

Content similarity

- Syntactic relevance function: TF-IDF
  - Proportional to the term density (number of occurrences) in a TCU
  - Proportional to the informativeness of term (its rarity across the whole collection of TCUs)

- Tree tuple Term Frequency – Inverse Tree tuple Frequency: TTF-ITF
  - Proportional to the term frequency within the local TCU
  - Proportional to the term popularity across the TCUs of the local tree tuple and the TCUs of the local document tree
  - Proportional to the term rarity across the whole collection of TCU
Clustering XML transactions:
XML tree tuple item similarity

- Content similarity

\[ ttf.\text{itf}(w_j, u_i|\tau) = \text{tf}(w_j, u_i) \times \exp\left(\frac{n_{j,\tau}}{N_\tau}\right) \times \frac{n_{j,XT}}{N_{XT}} \times \ln\left(\frac{N_T}{n_{j,T}}\right) \]

- \( \text{tf}(w_j, u_i) \) is the number of occurrences of \( w_j \) in \( u_i \),
- \( n_{j,\tau} \) is the number of TCUs in \( \tau \) that contain \( w_j \),
- \( N_\tau \) is the number of TCUs in \( \tau \),
- \( n_{j,XT} \) is the number of TCUs in XT that contain \( w_j \),
- \( N_{XT} \) is the number of TCUs in XT,
- \( n_{j,T} \) is the number of TCUs in \( T \) that contain \( w_j \),
Clustering XML transactions: XML tree tuple item similarity

Content similarity

- A TCU $u_i$ is modeled with a vector $\vec{u}_i$ whose $j$-th component corresponds to an index term $w_j$ and contains the $ttf.itf$ relevance weight.
- The well-known cosine similarity is used to measure the similarity between TCU vectors:

$$sim_C(e_i, e_j) = \frac{\vec{u}_i \cdot \vec{u}_j}{\|\vec{u}_i\| \times \|\vec{u}_j\|}$$
Clustering XML Transactions: XML Transaction Similarity

- Search for shared items, when comparing two transactions
  - Enhance the notion of standard intersection to capture even minimal similarities between XML elements

- Set of $\gamma$-shared items:
  - Intuitively, the union of best $\gamma$-matched items between two XML transactions

- XML transaction similarity:
  $$sim_\gamma(tr_1, tr_2) = \frac{|match_\gamma(tr_1, tr_2)|}{|tr_1 \cup tr_2|}$$
Collaborative Clustering of XML transactions

**CXK-Means**

- Centroid based partitional
  - notion of *representative* of cluster of XML transactions
- Transaction-centric
  - pair-wise similarity between transactions guides the construction of clusters
- Suitable for a collaborative distributed environment
  - peer network: each peer node is responsible of “local” and “global” choices

Define three main notions:

- XML transaction similarity
- XML local cluster representative
- XML global cluster representative
Collaborative Clustering of XML transactions

- **CXK-means**: process $N_0$

  - Data are distributed over $m$ peer nodes
  - Each node communicates with all the other ones sending local representatives and receiving global representatives
  - An initial process corresponding to a node $N_0$ defines a partition of the $k$ clusters into $m$ subsets $Z_j$:

    **Process $N_0$**

    **Method:**

    define a partition of \{1..$k$\} into $m$ subsets $Z_1, \ldots, Z_m$;

    for $i = 1$ to $m$ do

    send ($\{Z_1, \ldots, Z_m\}, k, \gamma$) to $N_i$;
Collaborative Clustering of XML transactions

**CXK-means**: process \( N_i \)

- Each node \( N_i \) computes:
  - Local clusters \( C_1^i, \ldots, C_k^i \)
  - Local representatives \( c_1^i, \ldots, c_k^i \)
  - (A subset of) global representatives \( c_{i_1}, \ldots, c_{i_{q_i}} \), using the local representatives computed by all nodes
Collaborative Clustering of XML transactions

**CXK-means:**

**process** $N_i$

Method:

1. receive $(\{Z_1, \ldots, Z_m\}, k, \gamma)$ from $N_0$;
2. let $Z_i = \{i_1, \ldots, i_{q_i}\}$, with $0 \leq q_i \leq k$;
3. /* selects $q_i$ initial global clusters */
4. select $\{c_{i_1}, \ldots, c_{i_{q_i}}\}$ transactions coming from distinct original trees;
5. $C_{j}^{i} = \{\}$, $\forall j \in [1..k]$;
6. repeat
   - send (broadcast) $\{c_{i_1}, \ldots, c_{i_{q_i}}\}$ to $N_1, \ldots, N_m$;
   - receive $c_j$ from $N_h$ with $h \in [1..m]$ and $j \in Z_h$;
   - repeat /* computes local clusters */
     - $C_{j}^{i} := \{tr \mid tr \in S^i \land \text{sim}_j^\gamma(tr, c_{j}^{i}) > \text{sim}_l^\gamma(tr, c_{l}^{i}), l \in [1..k]\}$, $\forall j \in [1..k]$;
     - $C_{k+1}^{i} := \{tr \mid \text{sim}_j^\gamma(tr, c_{j}^{i}) = 0\}$, $\forall j \in [1..k]$;
     - $c_{j}^{i} := \text{computeLocalRepresentative}(C_{j}^{i})$, $\forall j \in [1..k]$;
   - until $Q(C^i)$ is maximized;
7. if $c_j^i$ does not change, $\forall j \in [1..k]$ then
   - send (broadcast) $([], \text{done})$;
8. else
   - send $(\{c_j^i, |C_j^i|\} | j \in Z_h\}$, continue) to $N_h, \forall h \in [1..m]$;
9. receive $(\{c_j^h\} | j \in Z_h\}, V_h)$ from $N_h, \forall h \in [1..m]$;
10. if $(\exists h \in [1..m] s.t. V_h = \text{continue})$ then
    - for $j \in Z_i$ do $c_j = \text{ComputeGlobalRepresentative}(\{c_j^1, \ldots, c_j^m\})$
    - until $V_1 = \cdots = V_m = \text{done}$;
Collaborative Clustering of XML Transactions: Local XML Cluster Representative

Compute the set of $\gamma$-shared items among all the transactions within cluster $C$

1. for each transaction in $C$, compute the union of the $\gamma$-shared item sets w.r.t. all the other transactions in $C$

2. compute a raw representative
   - by selecting the items with the highest frequency from the previously obtained union sets
   - possibly conflate those items sharing the same path

3. perform a greedy heuristic to refine the raw representative
   - by iteratively adding the remaining most frequent items until the sum of pair-wise similarities between transactions and representative cannot be further maximized
Collaborative Clustering of XML Transactions: Global XML Cluster Representative

- The global representative of a cluster \( C \) is computed by considering the \( m \) local representatives \( c^1, \ldots, c^m \)
  - Procedure similar to that used for computing local representatives
  - Only a difference: the structural rank \( g_{\text{rank}} \) associated with an item \( e \) considers the rank associated with each item (instead of the number of items) having a \( \gamma \)-matching
Collaborative Clustering of XML transactions: CXK–Means - other features

- Trash cluster
  - Contains only transactions having zero-similarity when compared with each cluster representative

- Cluster initialization
  - Tree tuples selected as initial cluster centroids are constrained to come from different XML documents
    - favoring the construction of clusters with low intersimilarity
Experimental evaluation: Data description

- Real XML data sources
  - the IEEE collection version 2.2
    - benchmark in the INEX document data mining track 2008
    - complex article schemas: front matter, back matter, section headings, text formatting tags, mathematical formulas, ...
  - the DBLP digital bibliography
    - variety of structures, small average depth
    - short text descriptions (paper titles, event topics, author names)

<table>
<thead>
<tr>
<th>data</th>
<th># docs</th>
<th># trans.</th>
<th># items</th>
<th>max fan out</th>
<th>avg depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>4,874</td>
<td>211,909</td>
<td>135,869</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td>DBLP</td>
<td>3,000</td>
<td>5,884</td>
<td>8,231</td>
<td>20</td>
<td>3</td>
</tr>
</tbody>
</table>
Experimental evaluation: Methodology and goals

- **Structure-driven** clustering
- **Content-driven** clustering
- **Structure/Content-driven** clustering
  - detecting common structures across different topics
  - identifying classes that both cover common topics and share structure type
Experimental evaluation: Methodology and goals

- **Evaluation**

  - Clustering quality (*F-Measure)*:
    
    \[
    P_{ij} = \frac{|C_j \cap \Gamma_i|}{|C_j|}, \quad R_{ij} = \frac{|C_j \cap \Gamma_i|}{|\Gamma_i|}, \quad F_{ij} = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}}
    \]

    \[
    F(C, \Gamma) = \frac{1}{|S|} \sum_{i=1}^{H} |\Gamma_i| \max_{j \in [1..K]} F_{ij}
    \]

  - Time performances
Experimental evaluation:

Accuracy results

<table>
<thead>
<tr>
<th>dataset</th>
<th># of clusters</th>
<th># of nodes</th>
<th>$F$-measure (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>8</td>
<td>1</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>0.421</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.376</td>
</tr>
<tr>
<td>DBLP</td>
<td>6</td>
<td>1</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.547</td>
</tr>
</tbody>
</table>

**TABLE I**

Clustering results with $f \in [0..0.3]$ (content-driven similarity)
Experimental evaluation: Accuracy results

<table>
<thead>
<tr>
<th>dataset</th>
<th># of clusters</th>
<th># of nodes</th>
<th>F-measure (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>14</td>
<td>1, 3, 5, 7, 9</td>
<td>0.564, 0.497, 0.451, 0.404, 0.356</td>
</tr>
<tr>
<td>DBLP</td>
<td>16</td>
<td>1, 3, 5, 7, 9</td>
<td>0.772, 0.721, 0.676, 0.614, 0.558</td>
</tr>
</tbody>
</table>

**TABLE II**

Clustering results with $f \in [0.4..0.6]$ (structure/content-driven similarity)
Experimental evaluation: Accuracy results

<table>
<thead>
<tr>
<th>dataset</th>
<th># of clusters</th>
<th># of nodes</th>
<th>$F$-measure (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE</td>
<td>2</td>
<td>1</td>
<td>0.618</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.542</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>0.433</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.386</td>
</tr>
<tr>
<td>DBLP</td>
<td>4</td>
<td>1</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>0.819</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.716</td>
</tr>
</tbody>
</table>

**TABLE III**
Clustering results with $f \in [0.7..1]$
(structure-driven similarity)
Experimental evaluation: Efficiency results

![Graph](image)

- **IEEE (100%)**
- **IEEE (50%)**
Experimental evaluation: Efficiency results

- DBLP (100%)
- DBLP (50%)
Conclusion

- **Collaborative distributed framework for clustering XML documents**
  - CXK-means: a distributed, centroid-based partitional clustering algorithm
  - Peer-to-peer network
  - Local and global decisions for each peer

- **XML documents modeled in a transactional domain**
  - Modeling of XML transactions starting from the notion of tree tuple
  - Similarity between transaction computed according to both structure and content features
Thank you