Clustering XML Documents: a Distributed Collaborative Approach

Sergio Greco
Francesco Gullo
Giovanni Ponti
Andrea Tagarelli
Giuseppe Agapito

DEIS – University of Calabria
Motivations

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

**Our proposal**: the first collaborative distributed framework for efficiently clustering XML documents.
Centroid-based partitional clustering in a collaborative distributed framework

- **Centroid-based partitional clustering**
  - Partition a set of objects into $k$ clusters
  - Object-to-cluster assignment is driven by similarity of data to cluster representatives (cluster centroids)

- Cluster centroids can efficiently be exchanged through the network
  - Each peer computes
    - a “local” clustering solution
    - and a subset of the “global” clustering solution
  - Global centroids are used to update local solution
Clustering XML documents: the core method

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

Main steps

1. Extracting XML tree tuples
2. Modeling XML tree tuples as transactions
   - XML feature generation
3. Clustering XML transactions
Extracting XML tree tuples: The DBLP Example

```xml
<dblp>
  <inproceedings key="conf/kdd/ZakiA03">
    <author>M. J. Zaki</author>
    <author>C. C. Aggarwal</author>
    <title>XRules: an effective structural classifier for XML data</title>
    <pages>316-325</pages>
    <year>2003</year>
    <booktitle>KDD</booktitle>
  </inproceedings>

  <inproceedings key="conf/kdd/Zaki02">
    <author>M. J. Zaki</author>
    <title>Efficiently mining frequent trees in a forest</title>
    <pages>71-80</pages>
    <year>2002</year>
    <booktitle>KDD</booktitle>
  </inproceedings>
</dblp>
```
Extracting XML tree tuples: The DBLP Example
Modeling XML transactions: 
The DBLP Example

<table>
<thead>
<tr>
<th>path (p)</th>
<th>τ_{1.p}</th>
<th>node ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>dblp.inproceedings.@key</td>
<td>&quot;conf/kdd/ZakiA03&quot;</td>
<td>n_3</td>
</tr>
<tr>
<td>dblp.inproceedings.author.S</td>
<td>&quot;M. J. Zaki&quot;</td>
<td>n_5</td>
</tr>
<tr>
<td>dblp.inproceedings.title.S</td>
<td>&quot;XRules: an effective ...&quot;</td>
<td>n_9</td>
</tr>
<tr>
<td>dblp.inproceedings.year.S</td>
<td>&quot;2003&quot;</td>
<td>n_11</td>
</tr>
<tr>
<td>dblp.inproceedings.booktitle.S</td>
<td>&quot;KDD&quot;</td>
<td>n_13</td>
</tr>
<tr>
<td>dblp.inproceedings.pages.S</td>
<td>&quot;316-325&quot;</td>
<td>n_15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>path (p)</th>
<th>τ_{2.p}</th>
<th>node ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>dblp.inproceedings.@key</td>
<td>&quot;conf/kdd/ZakiA03&quot;</td>
<td>n_3</td>
</tr>
<tr>
<td>dblp.inproceedings.author.S</td>
<td>&quot;C. C. Aggarwal&quot;</td>
<td>n_7</td>
</tr>
<tr>
<td>dblp.inproceedings.title.S</td>
<td>&quot;XRules: an effective ...&quot;</td>
<td>n_9</td>
</tr>
<tr>
<td>dblp.inproceedings.year.S</td>
<td>&quot;2003&quot;</td>
<td>n_11</td>
</tr>
<tr>
<td>dblp.inproceedings.booktitle.S</td>
<td>&quot;KDD&quot;</td>
<td>n_13</td>
</tr>
<tr>
<td>dblp.inproceedings.pages.S</td>
<td>&quot;316-325&quot;</td>
<td>n_15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>path (p)</th>
<th>τ_{3.p}</th>
<th>node ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>dblp.inproceedings.@key</td>
<td>&quot;conf/kdd/Zaki02&quot;</td>
<td>n_17</td>
</tr>
<tr>
<td>dblp.inproceedings.author.S</td>
<td>&quot;M. J. Zaki&quot;</td>
<td>n_19</td>
</tr>
<tr>
<td>dblp.inproceedings.title.S</td>
<td>&quot;Efficiently mining ...&quot;</td>
<td>n_21</td>
</tr>
<tr>
<td>dblp.inproceedings.year.S</td>
<td>&quot;2002&quot;</td>
<td>n_23</td>
</tr>
<tr>
<td>dblp.inproceedings.booktitle.S</td>
<td>&quot;KDD&quot;</td>
<td>n_25</td>
</tr>
<tr>
<td>dblp.inproceedings.pages.S</td>
<td>&quot;71-80&quot;</td>
<td>n_27</td>
</tr>
</tbody>
</table>

item ID | corresponding node IDs
---|---------------------
\(e_1\) | \(n_3\)
\(e_2\) | \(n_5, n_{10}\)
\(e_3\) | \(n_9\)
\(e_4\) | \(n_{11}\)
\(e_5\) | \(n_{13, n_{25}}\)
\(e_6\) | \(n_{15}\)
\(e_7\) | \(n_7\)
\(e_8\) | \(n_{17}\)
\(e_9\) | \(n_{21}\)
\(e_{10}\) | \(n_{23}\)
\(e_{11}\) | \(n_{27}\)
Clustering XML transactions: XML tree tuple item similarity

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

- Tolerance-aware matching
  - Notion of \( \gamma \)-matched items

- Similarity by structure
  - computed by comparing tag paths

- Similarity by content
  - cosine similarity between TCUs
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-mean**: 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 $c_{i1}, \ldots, c_{i_qi}$, using the local representatives computed by all nodes
receive \( \{Z_1, \ldots, Z_m\}, k, \gamma \) from \( N_0 \);
let \( Z_i = \{j_1, \ldots, j_{q_i}\} \), with \( 0 \leq q_i \leq k \), \( \sum_{i=1}^{m} q_i = k \);
/* selects \( q_i \) initial global clusters */
select \( \{tr_1, \ldots, tr_{q_i}\} \) from \( S^i \) coming from distinct original trees;
\( g_{j_s} = tr_s, \forall s \in [1..q_i]; \)
\( C_j^i = \{\}, \forall j \in [1..k]; \)
repeat
    send (broadcast) \( \{g_j | j \in Z_i\} \) to \( N_1, \ldots, N_m \);
    receive \( \{g_j | j \in Z_h\} \) from \( N_h, \forall h \in [1..m]; \)
    \( \ell_j^i = g_j, \forall j \in [1..k]; \)
    repeat /* transaction relocation */
    \( C_{k+1}^i = \{tr \in S^i | \text{sim}_j^\gamma(tr, \ell_j^i) = 0, \forall j \in [1..k]\}; \)
    for each \( j \in [1..k] \) do
        \( C_j^i = \{tr \in S^i \setminus C_{k+1}^i | \text{sim}_j^\gamma(tr, \ell_j^i) \geq \text{sim}_j^\gamma(tr, \ell_t^i), \forall t \in [1..k]\}; \)
        \( \ell_j^i = \text{ComputeLocalRepresentative}(C_j^i); \)
    end for
    until no transaction is relocated;
    if \( \ell_j^i \) does not change, \( \forall j \in [1..k] \) then
    send (broadcast) \( \{\}, V_i = \text{done} \); 
else
    send \( \{(\ell_j^i, |C_j^i|) | j \in Z_h\}, V_i = \text{continue} \) to \( N_h, \forall h \in [1..m]; \)
    receive \( \{(\ell_j^h, |C_j^h|) | j \in Z_i\}, V_h \) from \( N_h, \forall h \in [1..m]; \)
    if \( \exists h \in [1..m] \) s.t. \( V_h = \text{continue} \) then
    \( g_j = \text{ComputeGlobalRepresentative}((\ell_j^1, |C_j^1|), \ldots, (\ell_j^m, |C_j^m|)), \forall j \in Z_i; \)
    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
  - The structural rank $g_{rank}$ associated with an item considers the rank associated with each item (instead of the number of items) having a $\gamma$-match
Collaborative Clustering of XML Transactions: Complexity

- $m$ number of nodes
- $k$ number of clusters
- $|S^i| = |S| / m$ number of transactions node $i$
- $|tr|$ max size transaction
- $|V|$ vocabulary size
- $c_1$ cost main memory operation
- $c_2$ communication cost
- $1 \leq h \leq k$ transactions distribution over clusters
Experimental evaluation: Data description

- Real XML data sources

<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:

Efficiency results
Experimental evaluation: Efficiency results

![Graph showing efficiency results with two lines representing DBLP (100%) and DBLP (50%) with time (msec) on the y-axis and nodes on the x-axis.]
**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)

<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)
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
Thanks