18th Italian Symposium on Advanced Database Systems (SEBD '10)
Rimini, Italy, June 20-23, 2010

Clustering XML Documents: a Distributed Collaborative Approach

Sergio Greco
Francesco Gullo
Giovanni Ponti
Andrea Tagarelli
Giuseppe Agapito
DEIS – University of Calabria

UNIVERSITA DELLA CALABRIA
Dipartimento di ELETTRONICA, INFORMATICA E SISTEMISTICA

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

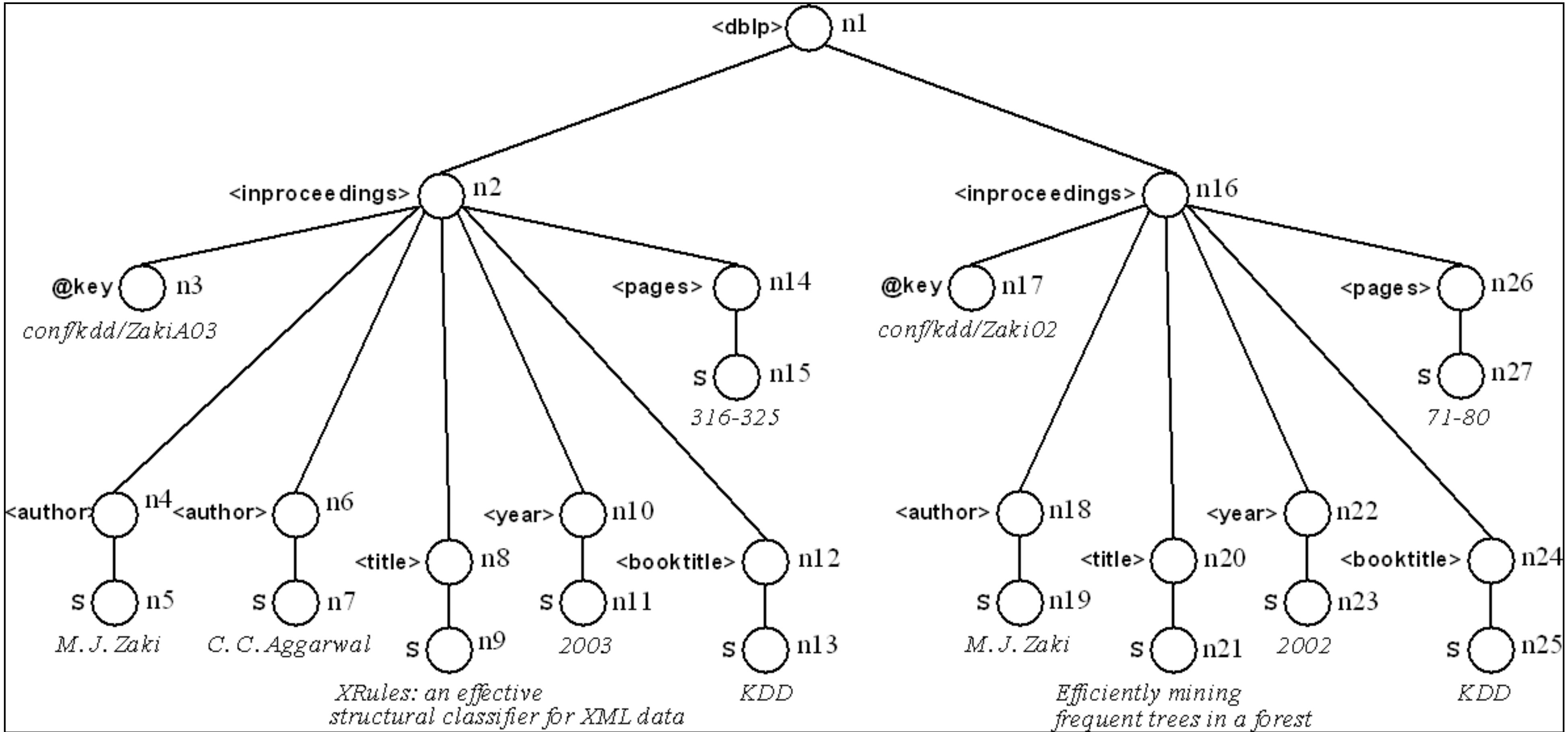
Clustering XML documents: the core method

[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

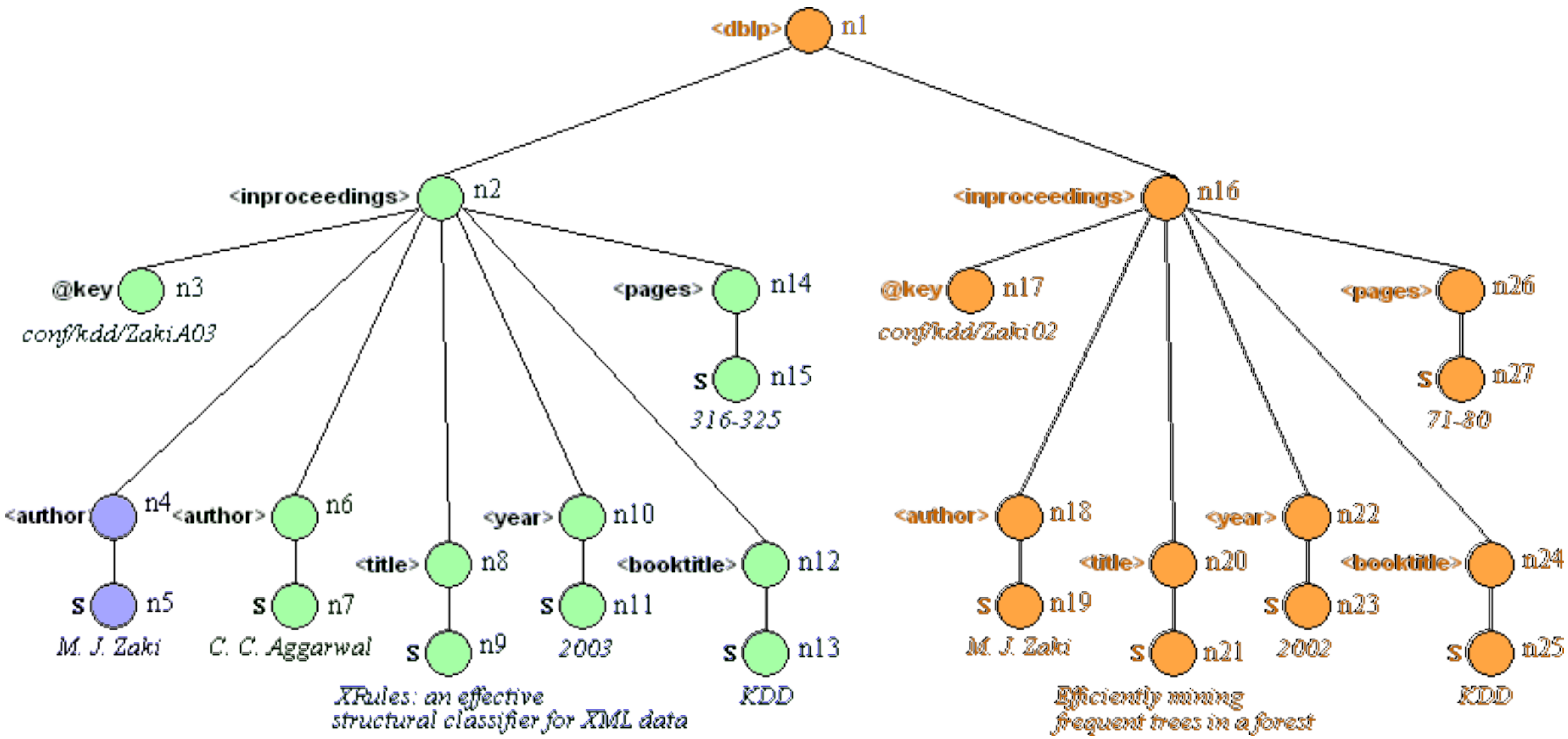


<booktitle>KDD**</booktitle>**

</inproceedings>

</dblp>

Extracting XML tree tuples:
The DBLP Example



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_{19}
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}

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

$$sim(e_i, e_j) = f \times sim_S(e_i, e_j) + (1 - f) \times sim_C(e_i, e_j)$$

- Tolerance-aware matching
  - Notion of γ-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-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

- **CXK-means**: process N_i

receive $(\{Z_1, \ldots, Z_m\}, k, \gamma)$ from N_0;
let $Z_i = \{j_1, \ldots, j_{q_i}\}$, with $0 \le q_i \le k$, $\sum_{i=1}^{m} q_i = k$;
/ selects q_i initial global clusters */*
select $\{tr_1, \ldots, tr_{q_i}\}$ from $\mathcal{S}^i$ 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 */*
    $C_{k+1}^i = \{tr \in S^i | sim_J^\gamma(tr, \ell_j^i) = 0, \forall j \in [1..k]\}$;
    for each $j \in [1..k]$ **do**
      $C_j^i = \{tr \in S^i \setminus C_{k+1}^i | sim_J^\gamma(tr, \ell_j^i) \ge sim_J^\gamma(tr, \ell_t^i), \forall t \in [1..k]\}$;
      $\ell_j^i = \mathsf{ComputeLocalRepresentative}(C_j^i)$;
    end for
  until no transaction is relocated;
  if ℓ_j^i does not change, $\forall j \in [1..k]$ **then**
    send (broadcast) $(\{\}, V_i = done)$;
  else
    send $(\{(\ell_j^i, |C_j^i|) | j \in Z_h\}, V_i = continue)$ 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. $V_h = continue)$ **then**
    $g_j = \mathsf{ComputeGlobalRepresentative}(\{(\ell_j^1, |C_j^1|), \ldots, (\ell_j^m, |C_j^m|)\}), \forall j \in Z_i$;
until $V_1 = \cdots = V_m = done$;

Collaborative Clustering of XML Transactions: Local XML Cluster Representative

Compute the set of γ-shared items among all the transactions within cluster *C*

1. for each transaction in *C*, compute the union of the γ-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 γ-match

Collaborative Clustering of XML Transactions: Complexity

The picture can't be displayed.

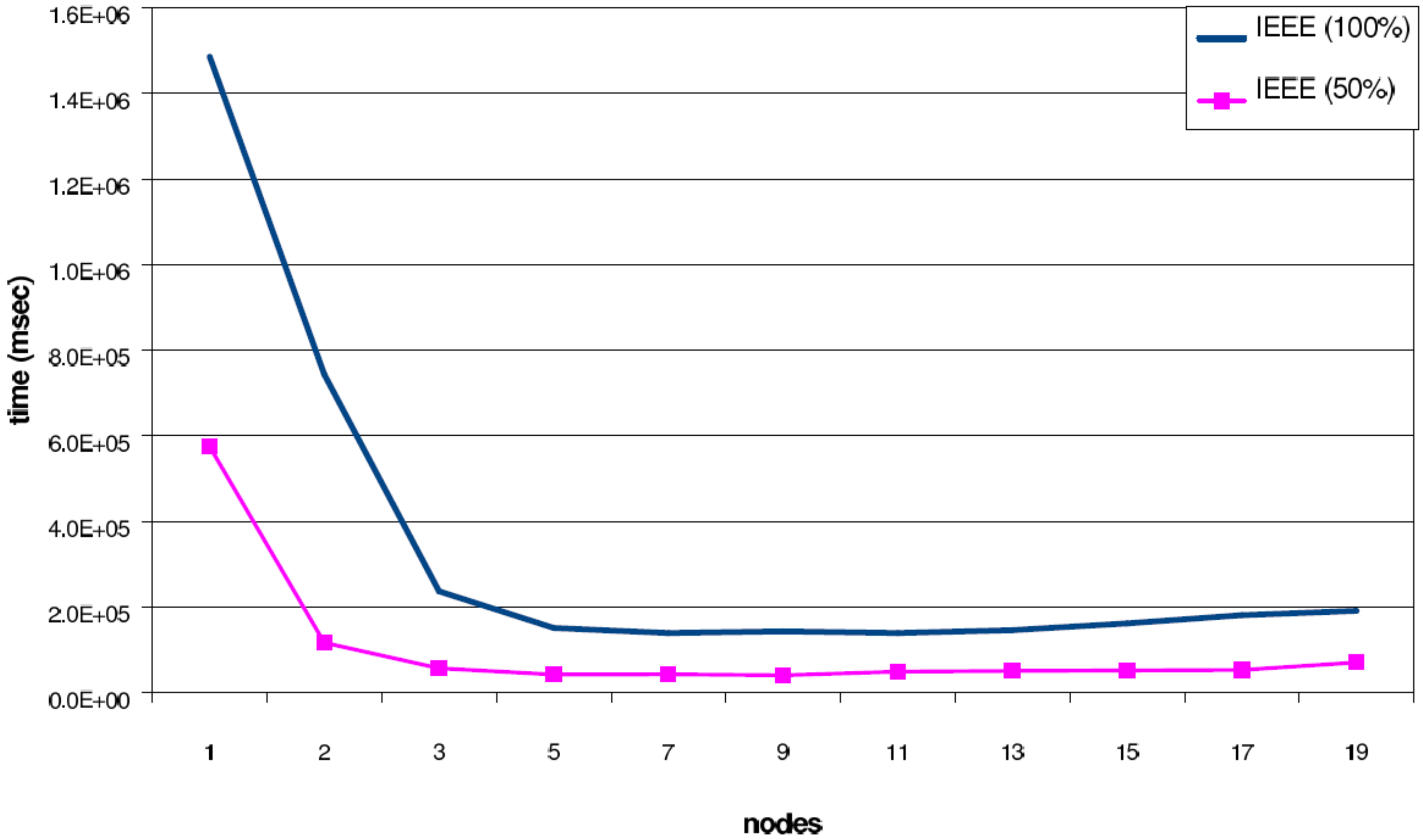
- m — number of nodes
- k — number of clusters
- $|S^i| = |S| / m$ — number of transactions node i
- $|tr|$ — max size transaction
- $|V|$ — vocabulary size
- c_1 — cost main memory operation
- c_2 — communication cost
- $1 \leq h \leq k$ — transactions distribution over clusters

Experimental evaluation: Data description

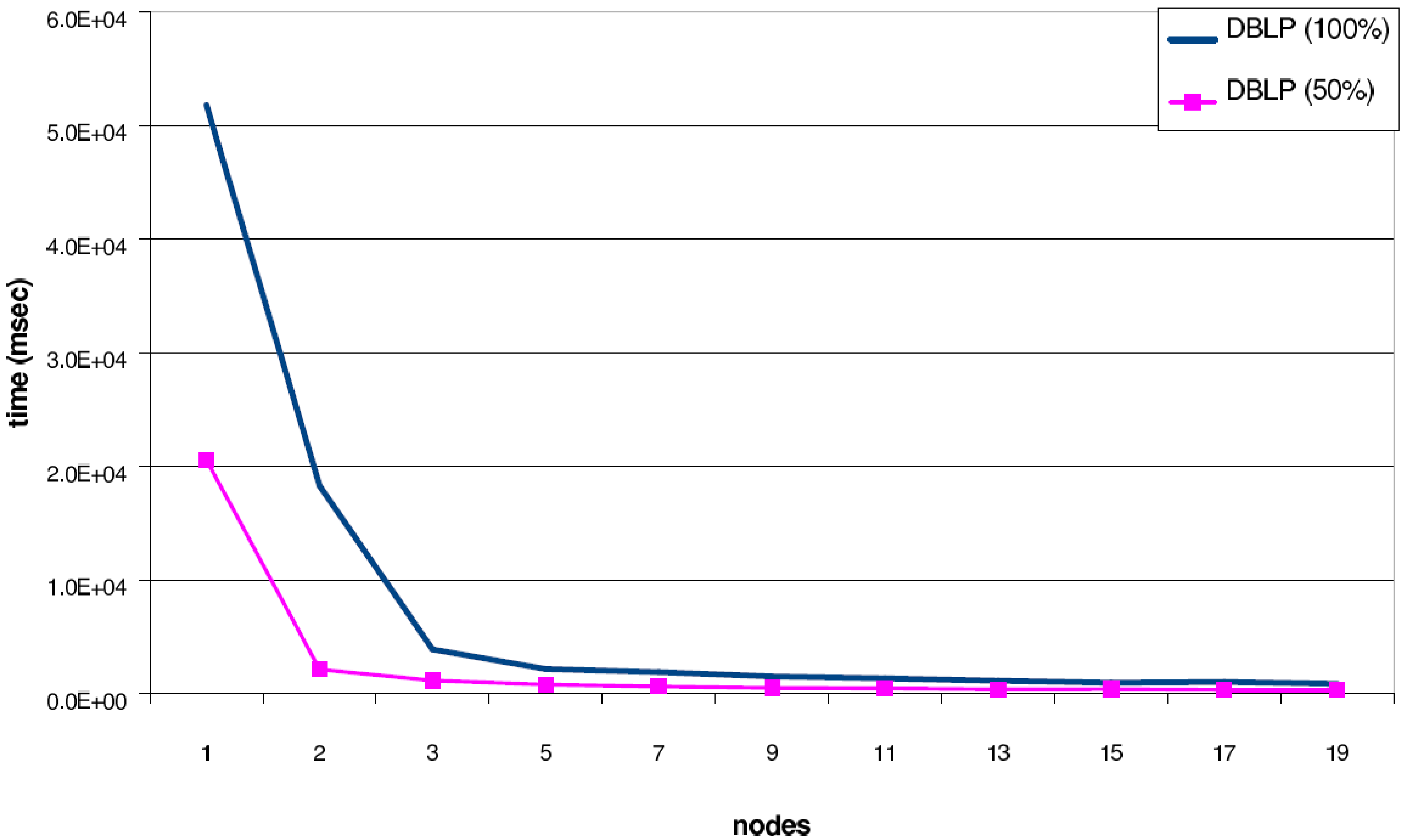
- Real XML data sources

data	# docs	# trans.	# items	max fan out	avg depth
IEEE	4,874	211,909	135,869	43	5
DBLP	3,000	5,884	8,231	20	3

Experimental evaluation: Efficiency results



Experimental evaluation: Efficiency results



Experimental evaluation: Accuracy results

dataset	*# of clusters*	*# of nodes*	*F-measure (avg)*
IEEE	8	1	0.593
		3	0.523
		5	0.485
		7	0.421
		9	0.376
DBLP	6	1	0.764
		3	0.702
		5	0.662
		7	0.612
		9	0.547

TABLE I
CLUSTERING RESULTS WITH $f \in [0..0.3]$
(CONTENT-DRIVEN SIMILARITY)

dataset	*# of clusters*	*# of nodes*	*F-measure (avg)*
IEEE	14	1	0.564
		3	0.497
		5	0.451
		7	0.404
		9	0.356
DBLP	16	1	0.772
		3	0.721
		5	0.676
		7	0.614
		9	0.558

TABLE II
CLUSTERING RESULTS WITH $f \in [0.4..0.6]$
(STRUCTURE/CONTENT-DRIVEN SIMILARITY)

dataset	*# of clusters*	*# of nodes*	*F-measure (avg)*
IEEE	2	1	0.618
		3	0.542
		5	0.497
		7	0.433
		9	0.386
DBLP	4	1	0.988
		3	0.934
		5	0.882
		7	0.819
		9	0.716

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