Motivation
- Problem
- Applications
- Metrics
- Related Work

TSearch
- Intuition
- Sequencing
- Transformation

Experiments

Conclusion

Efficient Similarity Search for Tree-Structured Data

Guoliang Li Xuhui Liu Jianhua Feng Lizhu Zhou

(SSDBM 2008)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

July 10, 2008

RoadMap

1. Motivation
   - Problem
   - Applications
   - Similarity Metrics
   - Existing Methods
2. Sequence-Based Tree-Structured Data Similarity Search
   - Intuition
   - Sequencing
   - Edit Distance Transformation
3. Experiments
4. Conclusion

RoadMap

1. Motivation
   - Problem
   - Applications
   - Similarity Metrics
   - Existing Methods
2. Sequence-Based Tree-Structured Data Similarity Search
   - Intuition
   - Sequencing
   - Edit Distance Transformation
3. Experiments
4. Conclusion

Problem

Problem

- Given a set of trees and a query tree
- Find all the trees that are ***similar*** to the query tree

Applications

Applications

- Comparison of hierarchically structured data
- Alignment of RNA secondary structures in computational biology
- Approximate XML document match
- Schema mapping of tree-structured data

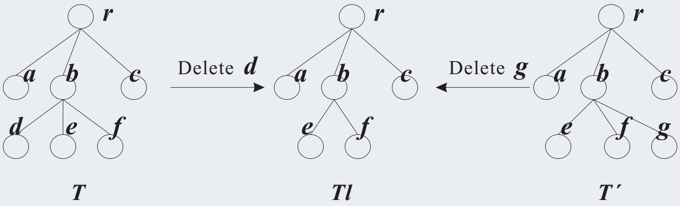
Tree Similarity Metrics

Tree Similarity Metrics

- Largest Common Subtree
- Smallest Common Super-tree
- Tree Edit Distance
- ...

Largest Common Subtree Distance

Largest Common Subtree

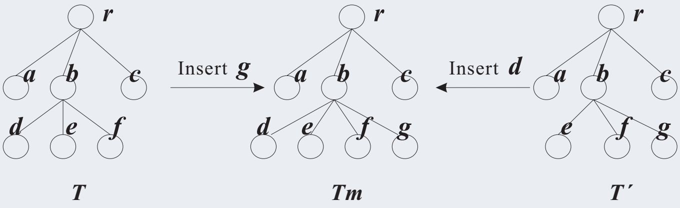


Largest Common Subtree Distance (LCST)

- the sum of # of operations to transfer the two trees into the largest common subtree

Smallest Common Super-tree Distance

Smallest Common Super-tree



Smallest Common Super-tree Distance (SCST)

- the sum of # of operations to transfer the two trees into the smallest common super-tree

Tree Edit Distance

Tree Edit Operations

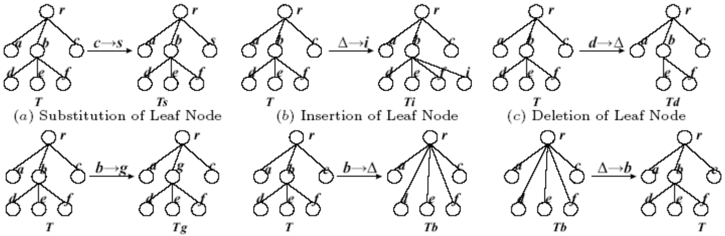
- Insert Node
- Delete Node
- Substitute Node

Tree Edit Distance

- # of tree edit operations

Edit Operations

Examples



(a) Substitution of Leaf Node (b) Insertion of Leaf Node (c) Deletion of Leaf Node

(d) Substitution of Internal Node (e) Deletion of Internal Node (f) Insertion of Internal Node

Edit Operations

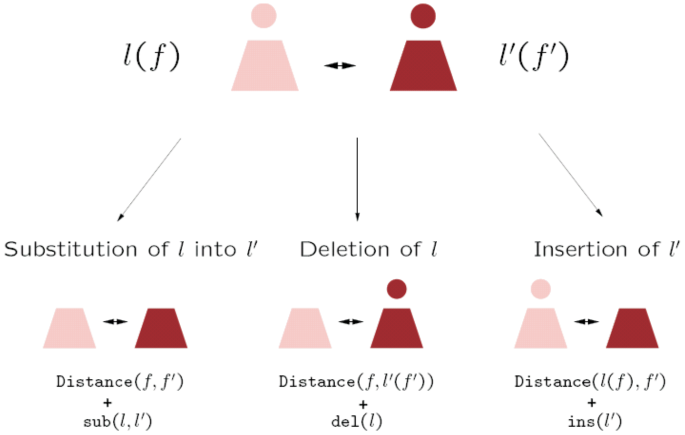
Edit Operations

$$\lambda_s(v) = \texttt{cSize}(v)$$

$$\lambda_d(v) = \begin{cases} 2 & \textit{if } v \textit{ is a leaf node} \\ 1 & \textit{if } v \textit{ is an internal node and } \texttt{parent}(v) = v \\ \texttt{cSize}(v) & \textit{if } v \textit{ is an internal node and } \texttt{parent}(v) \neq v \end{cases} \quad (1)$$

$$\lambda_i(v) = \begin{cases} 2 & \textit{if } v \textit{ is a leaf node} \\ 1 & \textit{if } v \textit{ is an internal node and } \texttt{parent}(v) = v \\ \texttt{cSize}(v) & \textit{if } v \textit{ is an internal node and } \texttt{parent}(v) \neq v \end{cases}$$

Tree Edit Distance

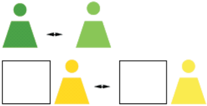


Forest Edit Distance

$l(f) \circ t$  $l'(f') \circ t'$

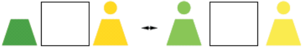
Leftmost decomposition

Substitution of l into l'



$\texttt{Distance}(l(f), l'(f'))$ $+$ $\texttt{Distance}(t, t')$

Deletion of l



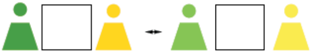
$\texttt{Distance}(f \circ t, l'(f') \circ t') + \texttt{del}(l)$

Insertion of l'

$\texttt{Distance}(l(f) \circ t, f' \circ t') + \texttt{ins}(l')$

Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

Forest Edit Distance

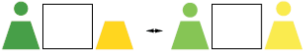
$t \circ l(f)$  $t' \circ l'(f')$

Rightmost decomposition

Substitution of l into l'

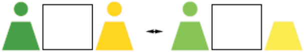
$\texttt{Distance}(l(f), l'(f'))$
$+$
$\texttt{Distance}(t, t')$

Deletion of l



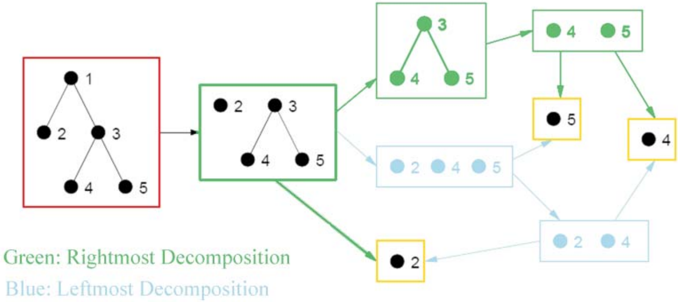
$\texttt{Distance}(t \circ f, t' \circ l'(f')) + \texttt{del}(l)$

Insertion of l'



$\texttt{Distance}(t \circ l(f), t' \circ f') + \texttt{ins}(l')$

Forest Edit Distance



Tree Edit Distance

Edit Distance

- Different decomposition strategies
- Dynamic programming
- The costs of a commonly used algorithm
  - Space: $|T_1| * |T_2|$
  - Time: $|T_1|^*|T_2|^*min(|depth(T_1)|, |leaves(T_1)|)^*min(|depth(T_2)|, |leaves(T_2)|)$
  - Worst Case: $|T_1|^2 * |T_2|^2$
- High CPU and IO costs!

Tree Edit Distance

Edit Distance

- Different decomposition strategies
- Dynamic programming
- The costs of a commonly used algorithm
  - Space: $|T_1| * |T_2|$
  - Time: $|T_1|{*}|T_2|{*}min(|depth(T_1)|, |leaves(T_1)|){*}min(|depth(T_2)|, |leaves(T_2)|)$
  - Worst Case: $|T_1|^2 * |T_2|^2$
- High CPU and IO costs!

Complexity of Tree Similarity Metrics

Complexity

Tree edit distance

variant	type	time	space	reference
general	O	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$	[43]
general	O	$O(\|T_1\|\|T_2\|\min(L_1, D_1)\min(L_2, D_2))$	$O(\|T_1\|\|T_2\|)$	[55]
general	O	$O(\|T_1\|^2\|T_2\|\log\|T_2\|)$	$O(\|T_1\|\|T_2\|)$	[25]
general	O	$O(\|T_1\|\|T_2\| + L_1^2\|T_2\| + L_1^{2.5}L_2)$	$O((\|T_1\| + L_1^2)\min(L_2, D_2) + \|T_2\|)$	[8]
general	U	MAX SNP-hard		[54]
constrained	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$	[51]
constrained	O	$O(\|T_1\|\|T_2\|I_1I_2)$	$O(\|T_1\|\|D_2I_2)$	[37]
constrained	U	$O(\|T_1\|\|T_2\|(I_1 + I_2)\log(I_1 + I_2))$	$O(\|T_1\|\|T_2\|)$	[52]
less-constrained	O	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1 + I_2))$	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1 + I_2))$	[29]
less-constrained	U	MAX SNP-hard		[29]
unit-cost	O	$O(u^2\min(\|T_1\|, \|T_2\|)\min(L_1, L_2))$	$O(\|T_1\|\|T_2\|)$	[41]
1-degree	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$	[38]

Tree alignment distance

general	O	$O(\|T_1\|\|T_2\|(I_1 + I_2)^2)$	$O(\|T_1\|\|T_2\|(I_1 + I_2)^2)$	[18]
general	U	MAX SNP-hard		[18]
similar	O	$O((\|T_1\| + \|T_2\|)\log(\|T_1\| + \|T_2\|)(I_1 + I_2)^4s^2)$	$O((\|T_1\| + \|T_2\|)\log(\|T_1\| + \|T_2\|)(I_1 + I_2)^4s^2)$	[17]

Tree inclusion

general	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\min(D_2L_2))$	[21]
general	O	$O(\|\Sigma_{T_1}\|\|T_2\| + m_{T_1,T_2}D_2)$	$O(\|\Sigma_{T_1}\|\|T_2\| + m_{T_1,T_2})$	[36]
general	O	$O(L_1\|T_2\|)$	$O(L_1\min(D_2L_2))$	[7]
general	U	NP-hard		[22, 32]

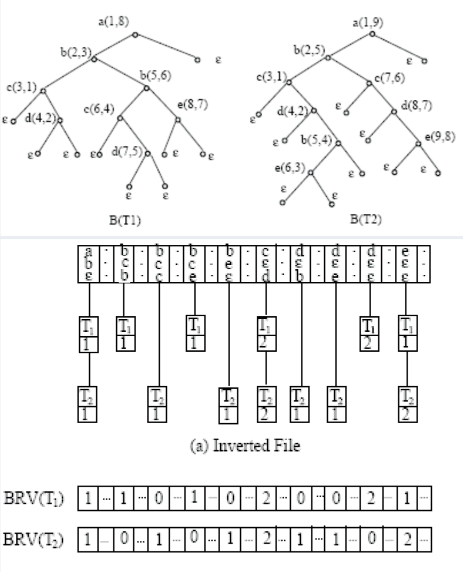
Existing Methods

Existing Methods

- PQ-Gram based method
- Binary tree based method
- Filter and Refine

Binary tree based similarity search [SIGMOD05]

Binary tree

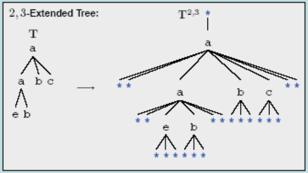


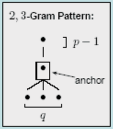
pq-gram based similarity search [VLDB05]

pq-gram

- **Extended Tree $\mathbf{T}^{pq}$:**
  Patch boundaries by adding null nodes (*):
  - $p - 1$ ancestors to the root
  - $q - 1$ nodes before the first and after the last child of each non-leaf node
  - q children to each leaf
- pq**-Gram $\mathbf{G}$:** Subtree of $\mathbf{T}^{pq}$.
  - Anchor node
  - with $p - 1$ ancestors
  - and q children.

  Contiguous siblings in $\mathbf{G}$ are contiguous siblings in $\mathbf{T}^{pq}$.



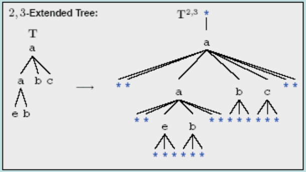


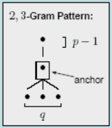
pq-gram based Similarity Join [VLDB05]

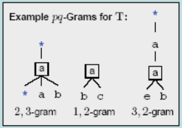
pq-gram

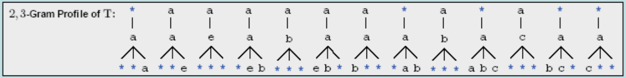
- **Extended Tree** $\mathbf{T}^{pq}$:
  Patch boundaries by adding null nodes (*):
  - $p-1$ ancestors to the root
  - $q-1$ nodes before the first and after the last child of each non-leaf node
  - q children to each leaf
- pq-**Gram** $\mathbf{G}$: Subtree of $\mathbf{T}^{pq}$.
  - Anchor node
  - with $p-1$ ancestors
  - and q children.

  Contiguous siblings in $\mathbf{G}$ are contiguous siblings in $\mathbf{T}^{pq}$.
- pq-**gram Profile** $\mathbf{P}^{p,q}(\mathbf{T})$:
  - Bag of all pq-grams of $\mathbf{T}$.









RoadMap

1. Motivation
   - Problem
   - Applications
   - Similarity Metrics
   - Existing Methods
2. Sequence-Based Tree-Structured Data Similarity Search
   - Intuition
   - Sequencing
   - Edit Distance Transformation
3. Experiments
4. Conclusion

TSearch

Tree Search
Guoliang Li
Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

Intuition

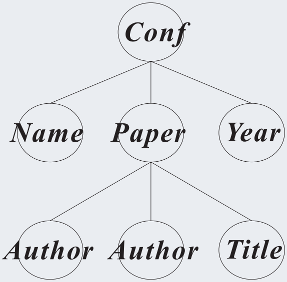
- Efficiency
- Both Structural and Textual features

Intuition

- Trees -> Strings
- Using approximate string search for approximate tree search

Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

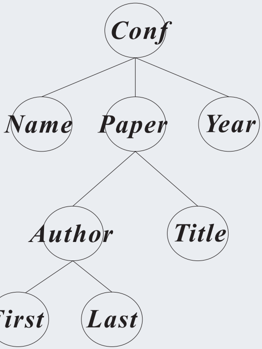
Sequencing



Name Conf Author Paper Author Paper Title Paper Conf Year Conf

Sequence-Based TSearch

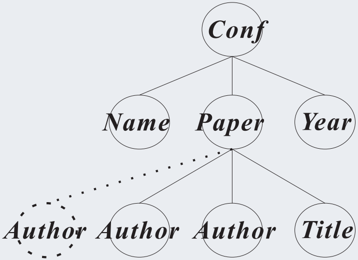
Sequencing



Name Conf First Author Last Author Paper Title Paper Conf Year Conf

Tree Search
Guoliang Li
Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

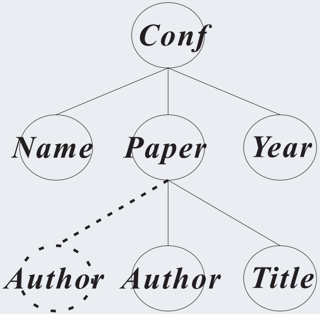
Insert Leaf Node



Name Conf Author Paper Author Paper Author Paper Title Paper Conf Year Conf

Sequence-Based TSearch

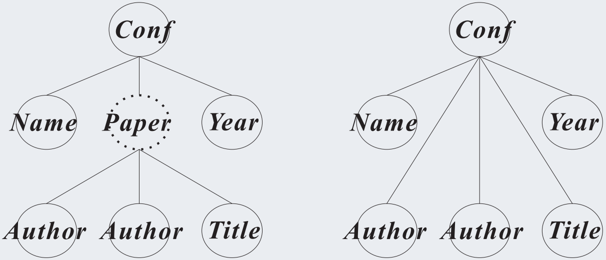
Delete Leaf Node



Name Conf <u>*Author Paper*</u> *Author Paper Title Paper Conf Year Conf*

Sequence-Based TSearch

Delete/Insert Internal Node



Name Conf Author Paper <u>*Conf*</u> *Author* Paper <u>*Conf*</u> *Title* Paper *Conf Year Conf*

Edit Distance Transformation

Edit Distance Transformation

$$ted(T, T') \leq ed(S, S') = ed(T, T') \leq C_{max} * ted(T, T')$$

LCST Distance Transformation

LCST Distance Transformation

$$lcstd(T, T') \geq ted(T, T')$$

$$lcstd(T, T') \geq ted(T, T') \geq \frac{1}{C_{max}} * ed(S, S')$$

SCST Distance Transformation

SCST Distance Transformation

$$scstd(T, T') \geq ted(T, T')$$

$$scstd(T, T') \geq ted(T, T') \geq \frac{1}{C_{max}} * ed(S, S')$$

RoadMap

Tree Search

Guoliang Li

Motivation
Problem
Applications
Metrics
Related Work

TSearch
Intuition
Sequencing
Transformation

Experiments

Conclusion

1. Motivation
   - Problem
   - Applications
   - Similarity Metrics
   - Existing Methods
2. Sequence-Based Tree-Structured Data Similarity Search
   - Intuition
   - Sequencing
   - Edit Distance Transformation
3. Experiments
4. Conclusion

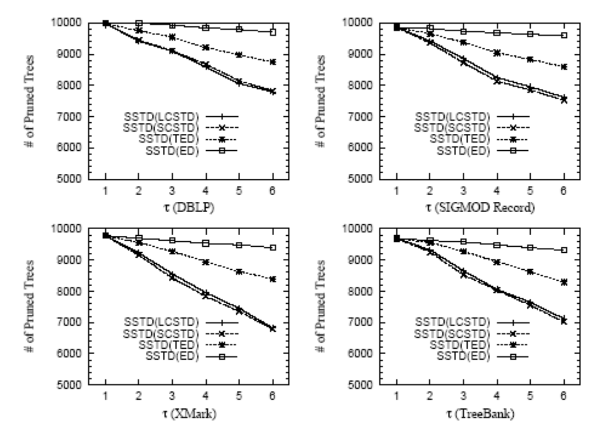
Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

Data Sets

Datasets	Average # of elements	Maximal depth	Maximal fan-out
DBLP	13.6	4	8
SIGMOD Record	12.2	4	8
XMark	10.4	6	6
Treebank	11.8	8	6

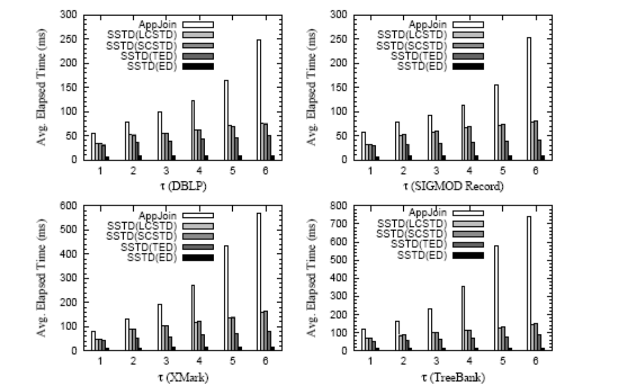
Tree Search
Guoliang Li
Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

Average # of Pruned Trees vs. Different Values of τ



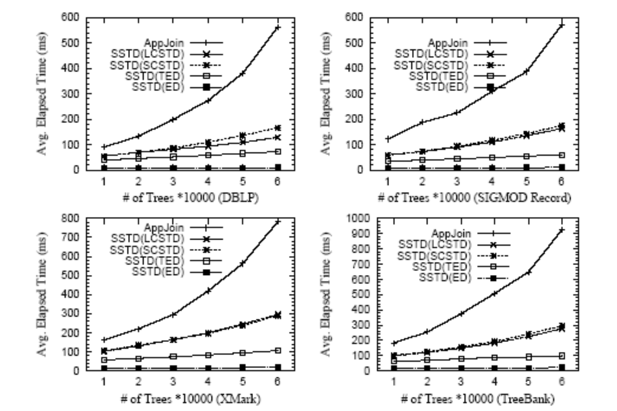
Performance

Average Elapsed Time vs. Different Values of τ



Tree Search
Guoliang Li
Motivation
Problem
Applications
Metrics
Related Work
TSearch
Intuition
Sequencing
Transformation
Experiments
Conclusion

Average Elapsed Time vs. Different Numbers of Trees (τ=3)



RoadMap

1 Motivation
- Problem
- Applications
- Similarity Metrics
- Existing Methods

2 Sequence-Based Tree-Structured Data Similarity Search
- Intuition
- Sequencing
- Edit Distance Transformation

3 Experiments

4 Conclusion

Sequence Based Similarity Search

Sequence Based Similarity Search for Tree-Structured Data

- Trees -> Strings: Sequencing method
- Using approximate string search for approximate tree search : Edit Distance Transformation
- Efficient methods for different metrics

Motivation
- Problem
- Applications
- Metrics
- Related Work

TSearch
- Intuition
- Sequencing
- Transformation

Experiments

Conclusion

Thanks!!

Questions?