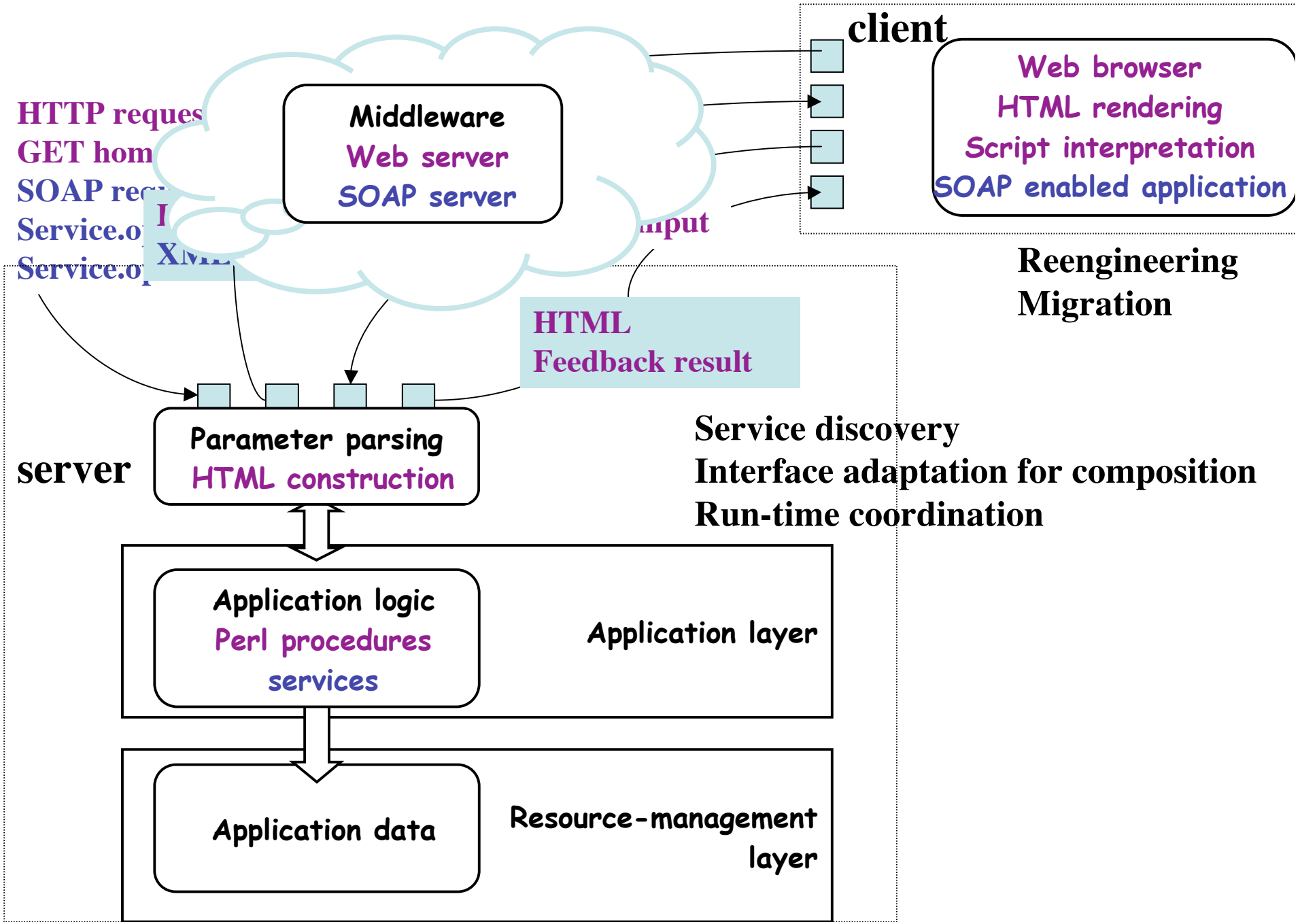
“Change minder”: Towards a general web-change notification system, based on HTML differencing

Eleni Stroulia and Rimon Mikhaeil
Computing Science department
University of Alberta


client
Web browser
HTML rendering
Script interpretation
SOAP enabled application
Middleware
Web server
SOAP server
HTTP reques
GET hom
SOAP req
Service.o
Service.o
I
XML
nput
HTML
Feedback result
Reengineering
Migration
server
Parameter parsing
HTML construction
Service discovery
Interface adaptation for composition
Run-time coordination
Application logic
Perl procedures
services
Application layer
Application data
Resource-management
layer

Objective

- To build an HTML Comparison tool that detects HTML editing operations
- To recognize the HTML elements (content and structure) that were “modified” , “deleted”, “inserted” or “moved”.

- Typical uses cases
  - To monitor the evolution of HTML Pages in CVS to resolve collisions/conflicts between edits of multiple developers
  - To recognize updates of web pages of interest (mind-it!)

(Some of) The state of the art

```
<html>
<body>
<b>test text</b>
</body>
</html>
```

⇒

```
<html>
<body>
<a href="">test text</a>
</body>
</html>
```

```
1:  <html>
2:  <body>
3:  <a href="">test text</a> </body>
4:  <b>
5:  test text
6:  </b>
7:  </body>
8:  </html>
```

HTMLMatch

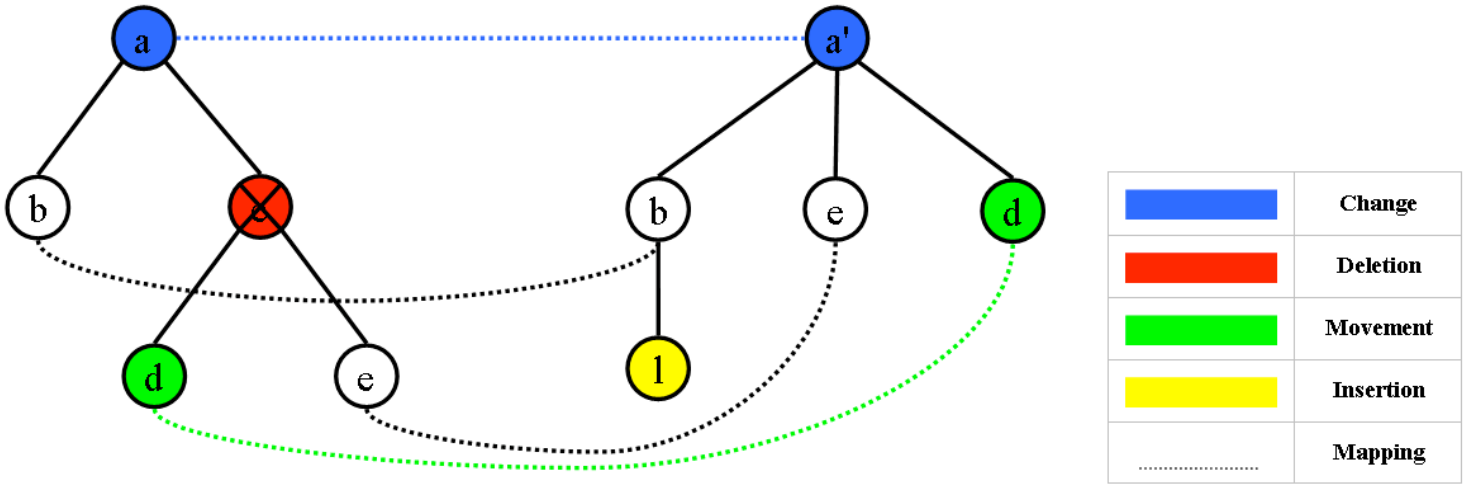
Diff Doc

```
<html>
<body>
<i>test text</i>
</body>
</html>
```

```
<html>
<body>
<b>test text</b>
</body>
</html>
```

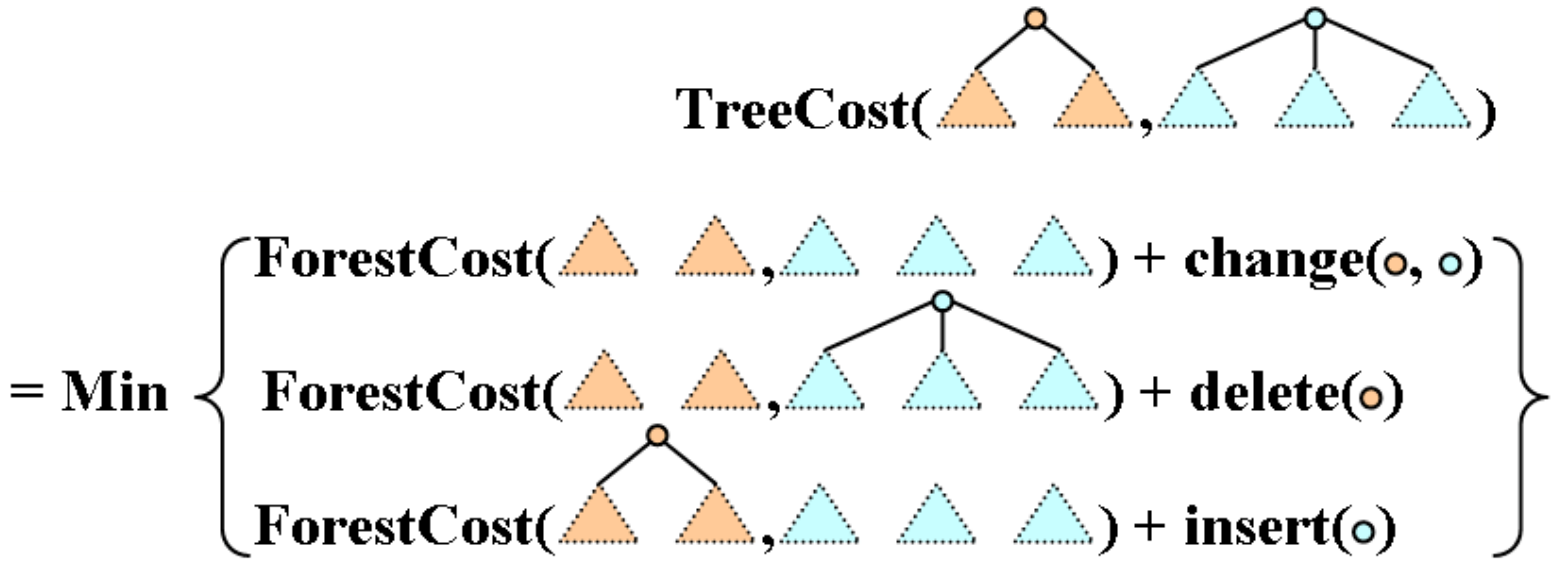
The Research Problem

- To analyze changes to the order of elements as well as their content in HTML documents
  - To compare **ordered labeled trees**
- To develop a tree-alignment algorithm to identify the minimum-cost editing operations that transform the tree corresponding to the original version to the one corresponding to the subsequent version

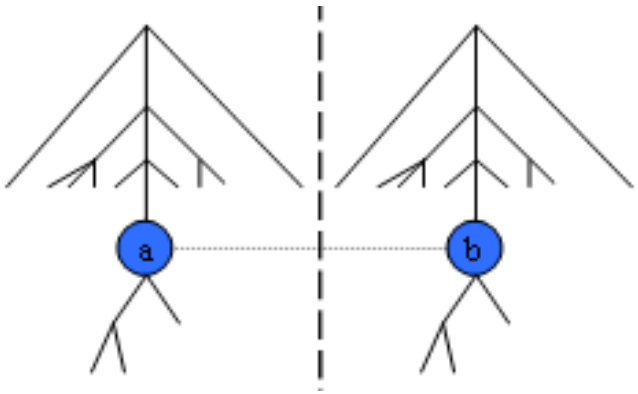


Tree-Editing Distance Algorithm [Zang-Shasha]

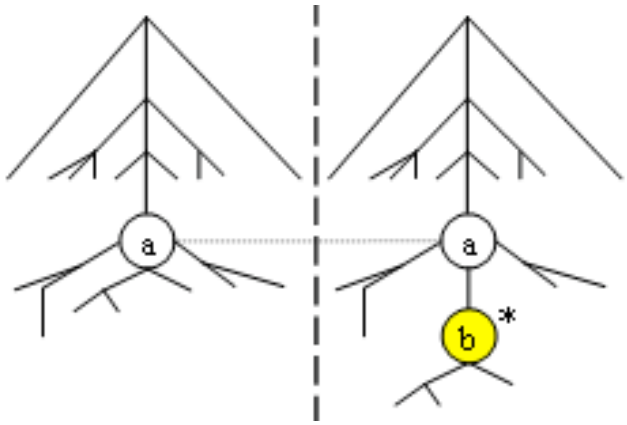
- Given
  1. two trees and
  2. a *suitable* cost function
- Calculate the least expensive transformational editing sequence of operations (insertion, deletion, and modification)



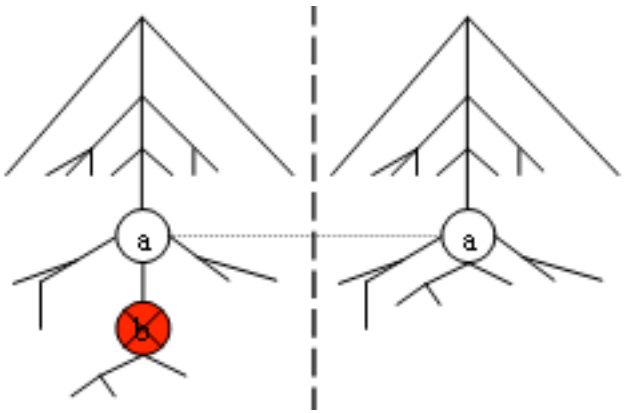
Operations and their costs



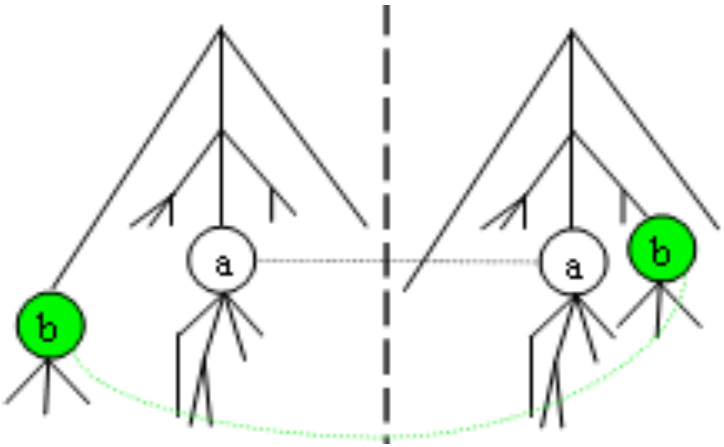
modification



insertion



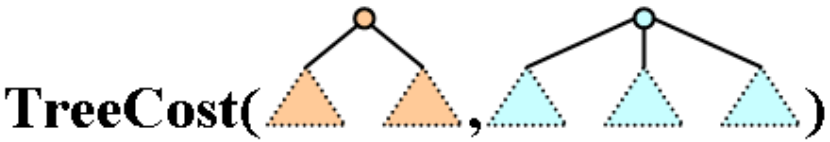
deletion

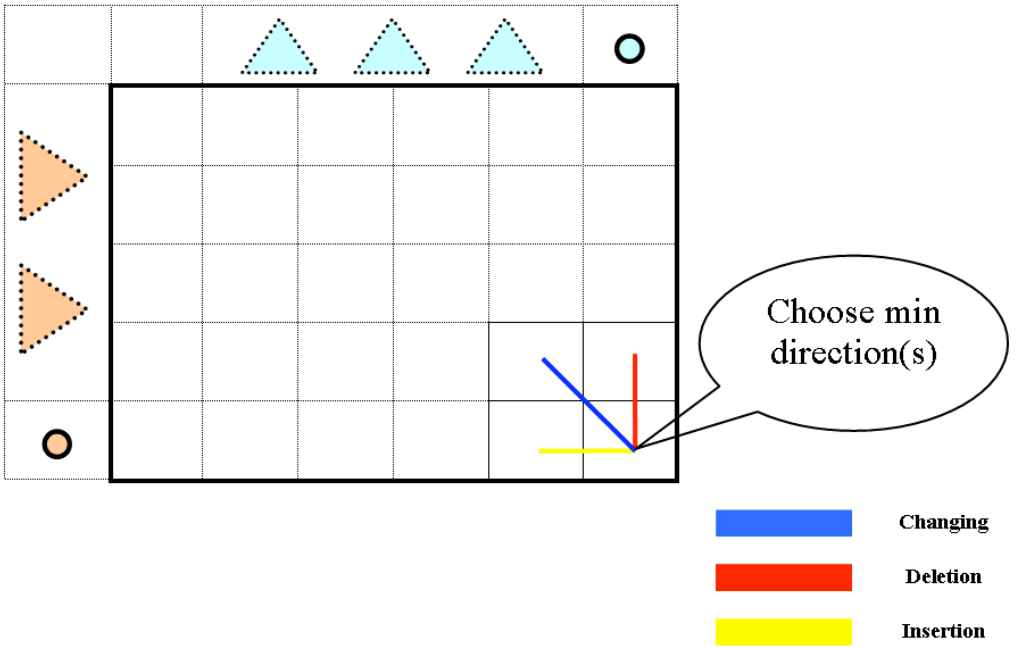


move

Editing-operation sequence Construction

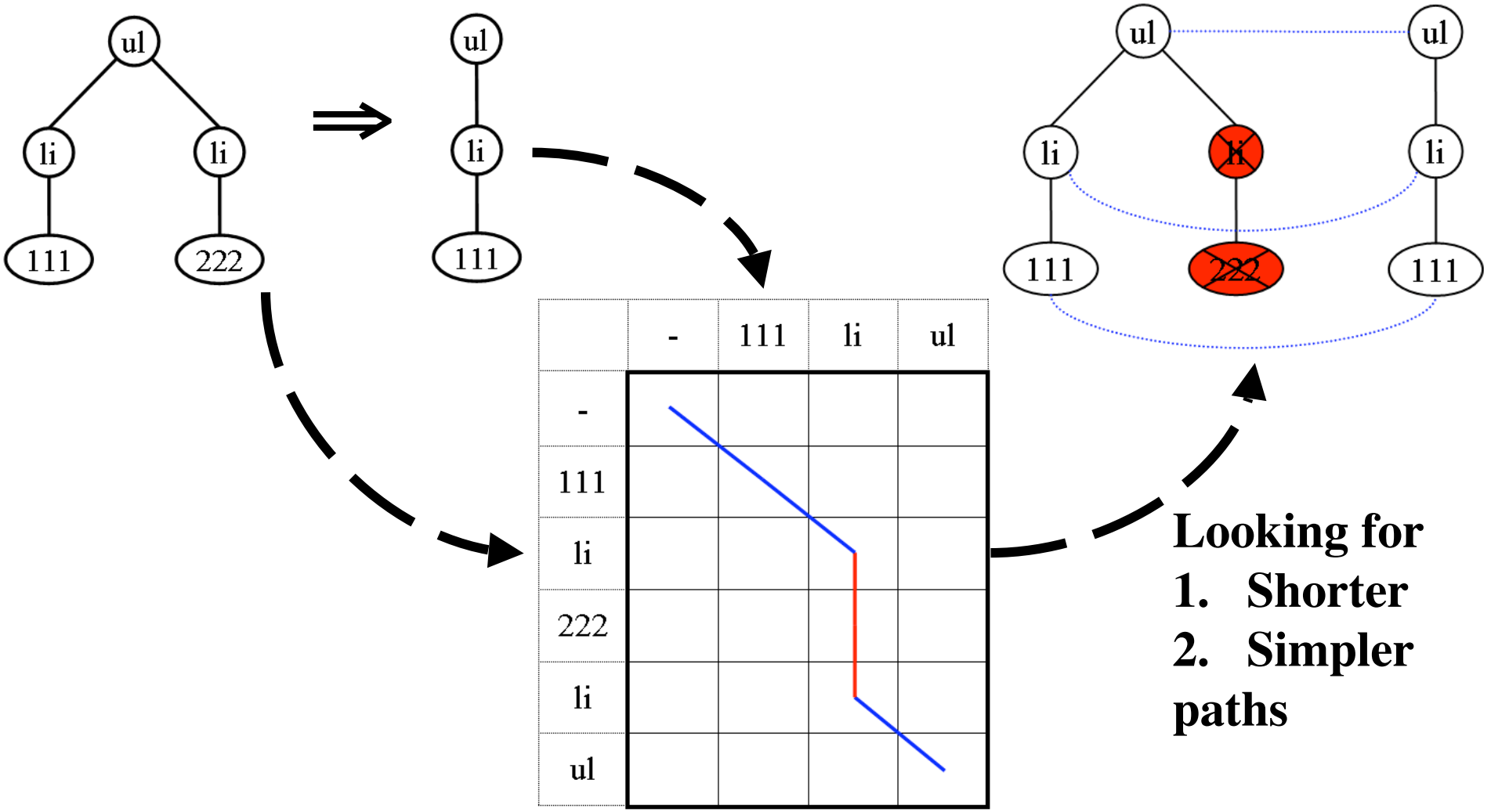
- Develop a matrix:
  - Row(i) represents the nodes of T_1 (the tree before).
  - Column(j) represents the nodes of T_2 (the tree after).
  - M[i, j] is the cost(T_1[l(i)..i] ⇒ T_2[l(j)..j]).
- Using a dynamic-programming approach
  - diagonal: T_1(i) is mapped to T_2(j)
  - up: T_1(i) is deleted
  - left: T_2(j) is newly inserted





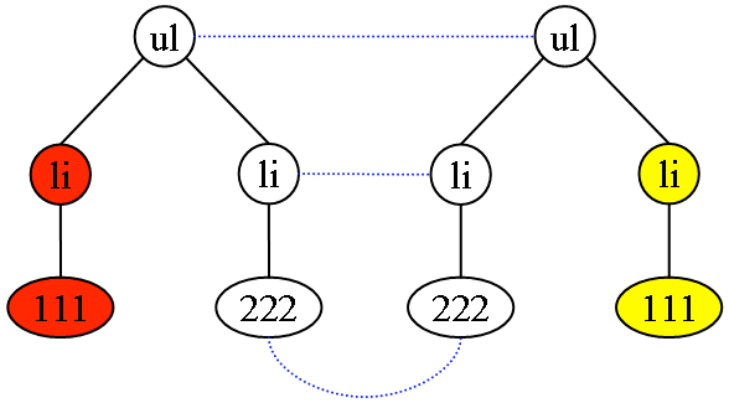
Editing-operation sequence Construction

- Recursively calculate the best editing sequence (Path) from the lower-right corner up to the upper-left one.

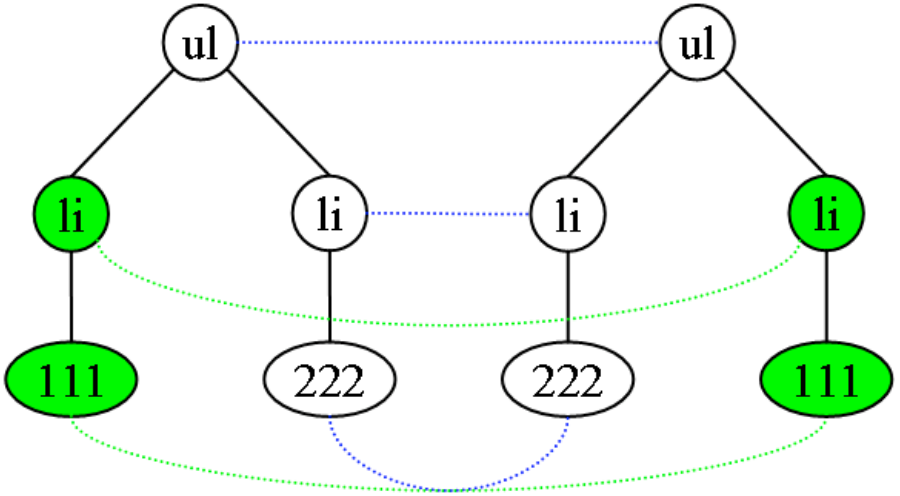


Post-processing: Movement Recognition

- This comparison may result in the depicted editing operations:
  - identical sub-trees are deleted from the first and inserted in the second tree.



- To recognize such movement operations, we compare the deleted sub-trees against the inserted ones:
  - if a deleted sub-tree is mapped to an inserted one, this sub-tree is reported as “moved”.



Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

Computer Architecture
Aritificial Intelligence
Database Systems
Graphics, Vision and Imaging
General Courses

Communication Networks
Software Engineering and Programming Languages
Software Systems
Computational Theory

Example#1

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

- Computer Architecture
- Aritificial Intelligence
- Database Systems
- Graphics, Vision and Imaging
- General Courses

- Communication Networks
- Software Engineering and Programming Languages
- Software Systems
- Computational Theory

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

- Computer Architecture
- Aritificial Intelligence
- Database Systems
- Graphics, Vision and Imaging
- General Courses

- Communication Networks
- Software Engineering and Programming Languages
- Software Systems
- Computational Theory

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

Computer Architecture
Aritificial Intelligence
Database Systems
Graphics, Vision and Imaging
General Courses
Communication Networks
Software Engineering and Programming Languages
Software Systems
Computational Theory

Example#2

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

Computer Architecture and organization
Aritificial Intelligence
Advanced Hi-Tech
Communication Networks
General Courses
Graphics, Vision and Imaging
Software Engineering and Programming Languages
Software Systems

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

Computer Architecture and organization
Aritificial Intelligence
Advanced Hi-Tech ~~Database Systems~~
Communication Networks
General Courses
Graphics, Vision and Imaging
Software Engineering and Programming Langua
Software Systems ~~Computational Theory~~

	Change
	Deletion
	Movement
	Insertion
	Structure Change

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

Computer Architecture
Aritificial Intelligence
Database Systems
Graphics, Vision and Imaging
General Courses

Communication Networks
Software Engineering and Programming Languages
Software Systems
Computational Theory

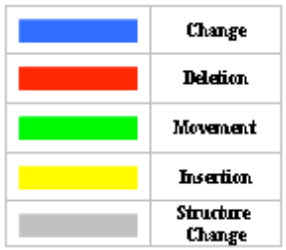
Example#3

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.

1. Computer Architecture
2. Aritificial Intelligence
3. Graphics, Vision and Imaging
4. General Courses
5. Adavced Courses

1. Communication Networks
2. Software Engineering and Programming Languages
3. Computational Theory
4. Software Systems (new)

Our Department offers a wide variety of graduate courses. Please be advised that only a selection of these graduate courses are offered each year. Offerings depend on the interest and the availability of the Faculty members. The course schedule is usually finalized in the month of June.



1. Computer Architecture
2. Aritificial Intelligence
3. Graphics, Vision and Imaging
4. General Courses
5. Adavced Courses

1. Communication Networks
2. Software Engineering and Programming Languages
3. Computational Theory
4. Software Systems (new)

Summary

- HTML Minder compares regular ordered trees
- It recognizes movement operations.
- Building suitable cost functions (to reflect their domain-specific semantics), the same algorithm could be used to compare other structured entities (XML schemas and documents, UML diagrams, …)

- Visualization of the results for human consumption is a challenge