

Extracting Logical Hierarchical Structure of HTML Documents Based on Headings

Tomohiro Manabe and Keishi Tajima

Graduate School of Informatics, Kyoto Univ.

Sakyo, Kyoto 606-8501 Japan

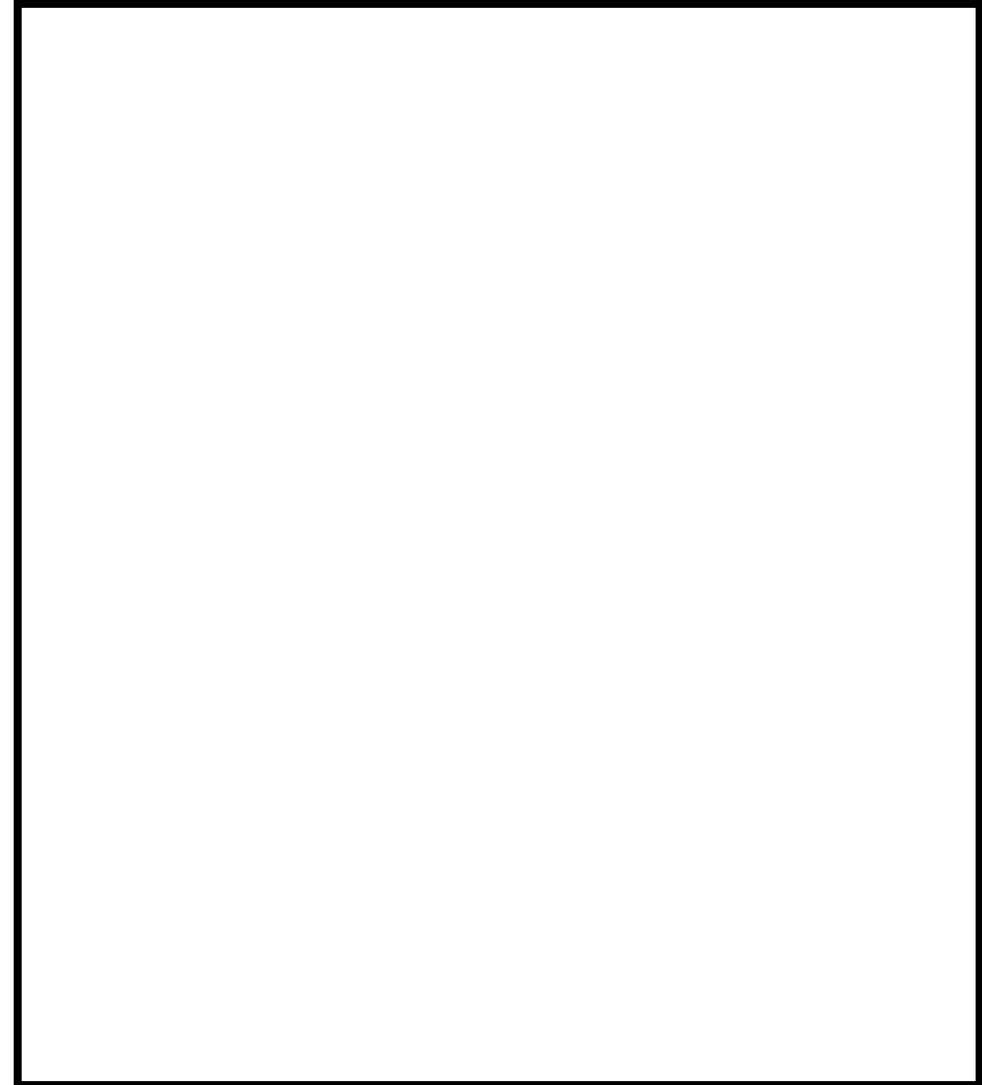
{manabe@dl.kuis, tajima@i}.kyoto-u.ac.jp

Background

- **Understanding of structure in web pages is important** for many applications
 - Web search
 - Automatic summarization of web pages
 - Web information extraction

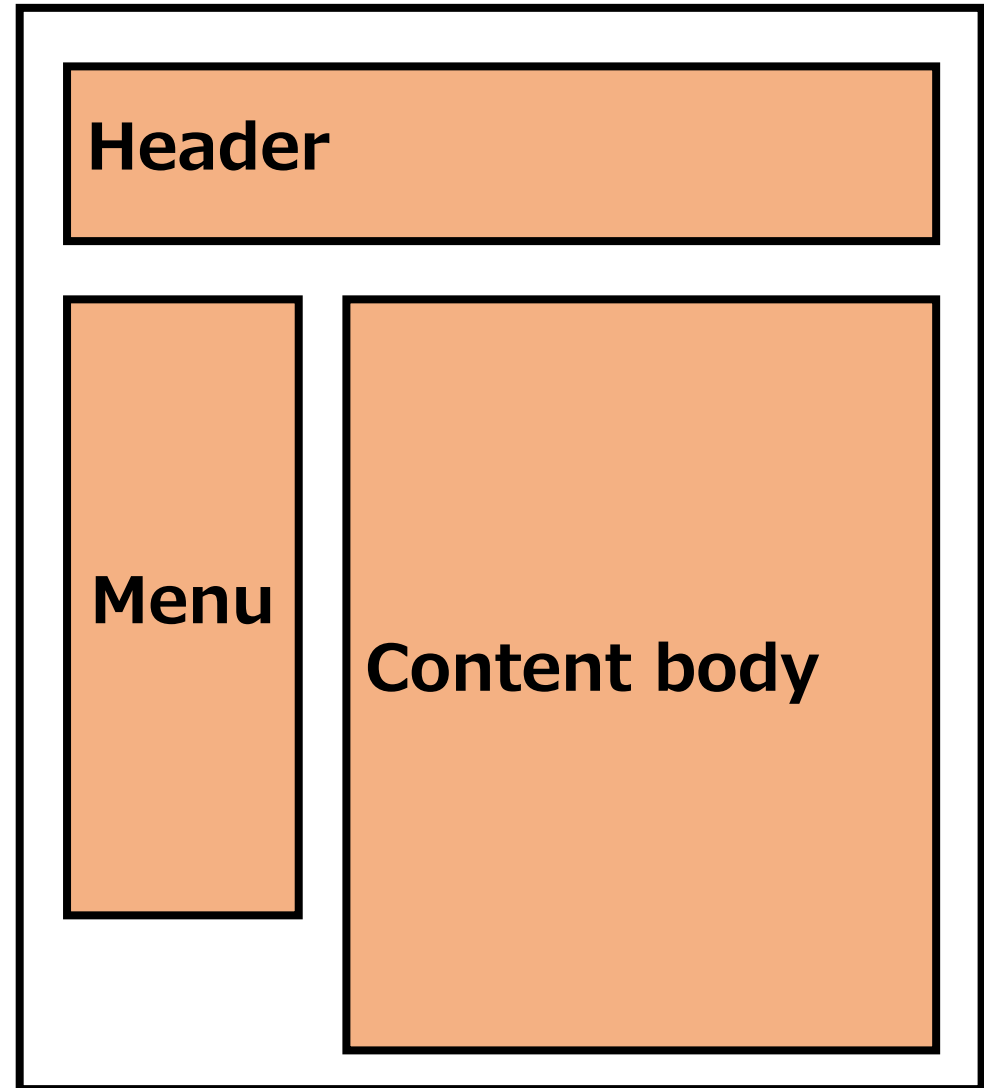
Structure in web pages

- Web pages contain various types of structures



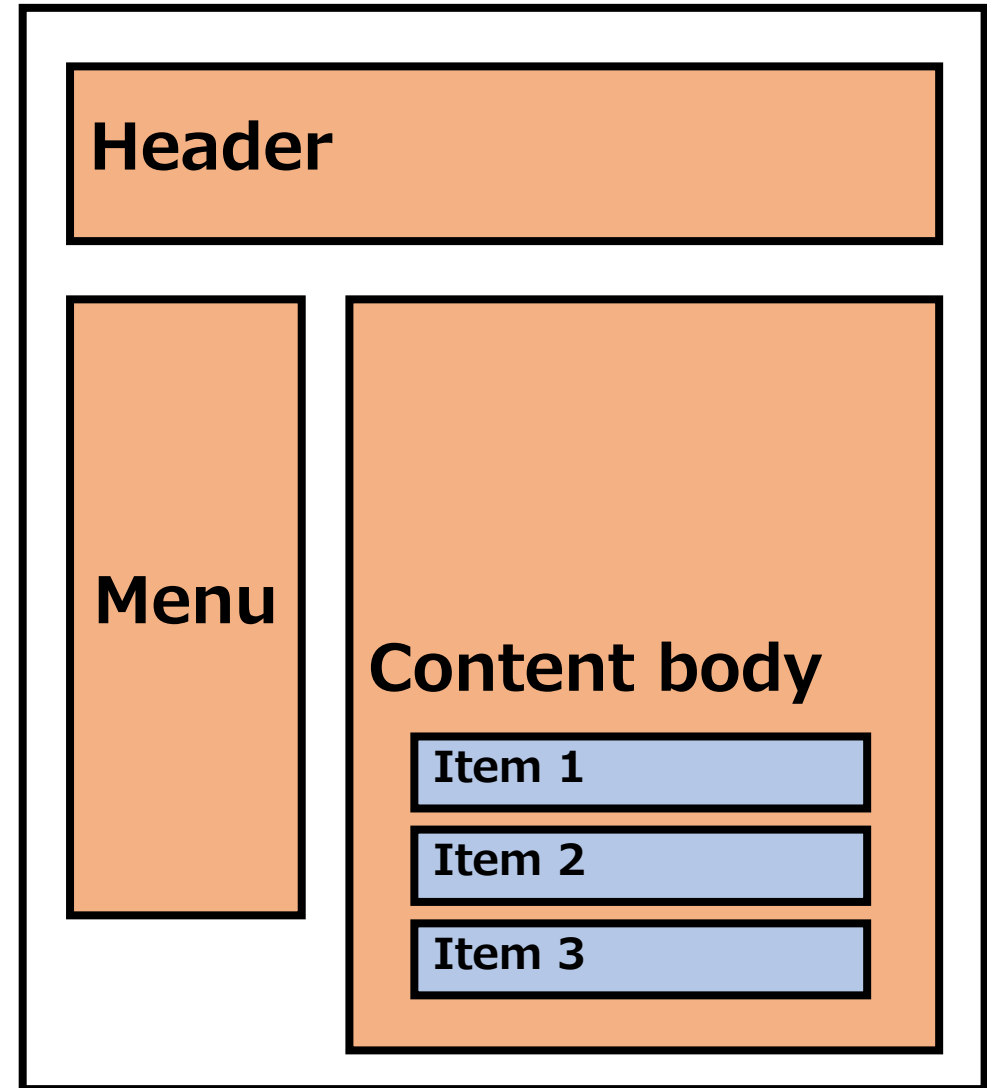
Structure in web pages

- Web pages contain various types of structures
 - Layout structure,



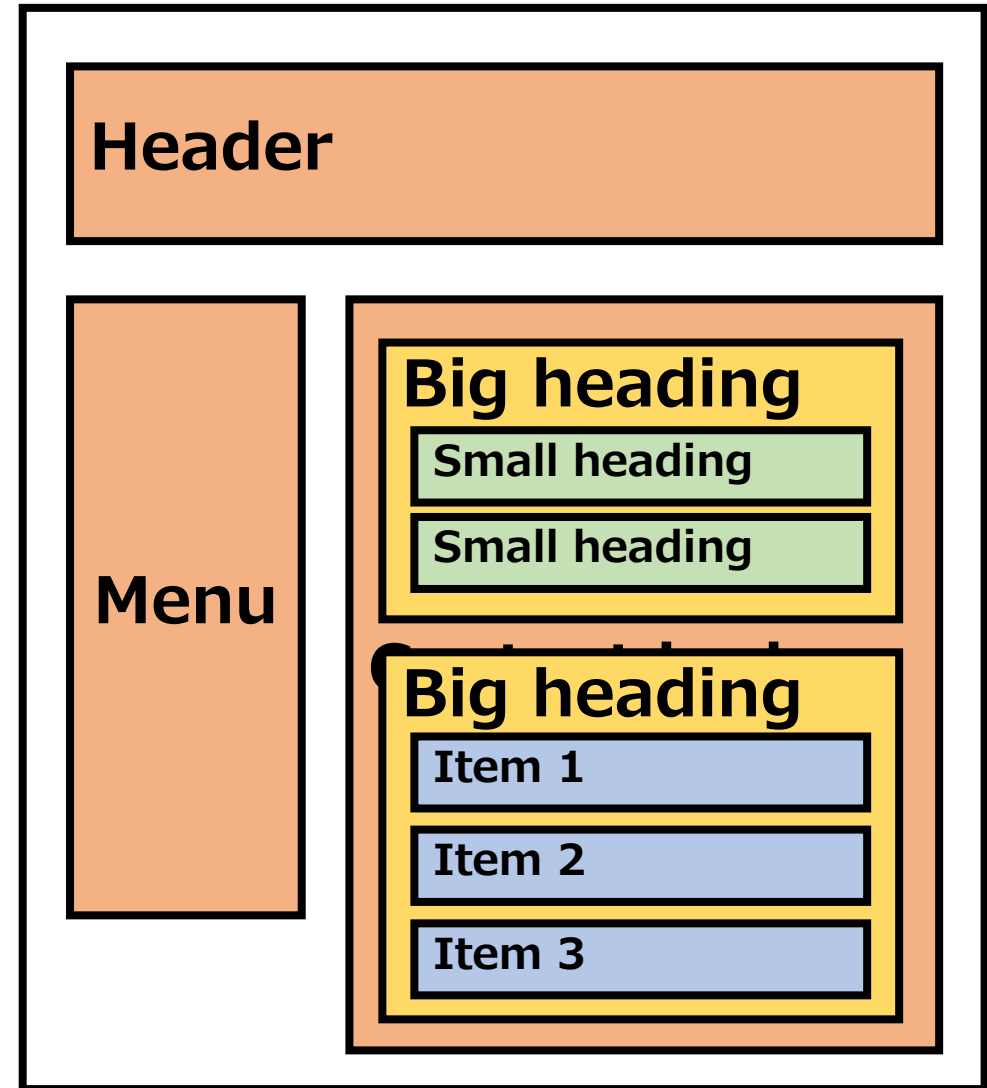
Structure in web pages

- Web pages contain various types of structures
 - Layout structure, list or table structure, ...



Structure in web pages

- Web pages contain various types of structures
 - Layout structure, list or table structure, ...
- We focus on **hierarchical heading structure**
 - 78% of pages contain it



Hierarchical heading structure

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment
- *Block*
 - A segment with its heading
 - may contain each other

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment
- *Block*
 - A segment with its heading
 - may contain each other

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment
- *Block*
 - A segment with its heading
 - may contain each other

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Hierarchical heading structure

- *Heading*
 - Topic description of a segment
- *Block*
 - A segment with its heading
 - may contain each other



- *Hierarchical heading structure*
 - composed of these
- 15 headings and blocks

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Importance of heading structure

2010 Mar

Search

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Importance of heading structure

- Traditional search engines:
 - This page contains both words
 - return this page incorrectly

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

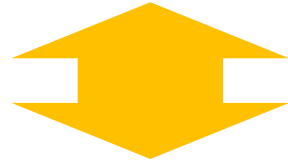
- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Importance of heading structure

2010 Mar

Search

- Traditional search engines:
 - This page contains both words
 - return this page incorrectly



- Heading-aware engines:
 - “Mar.” occurs under “2012”, not “2010”
 - **Will not return this page**

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

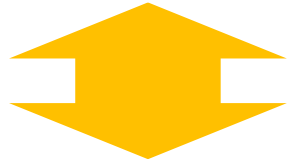
- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Problem to be solved

- Hierarchical heading structure is useful
- It seems easy to extract the structure

Problem to be solved

- Hierarchical heading structure is useful
- It seems easy to extract the structure



- In fact, it's NOT easy

Our research problem:

Extraction of hierarchical heading structure

Hierarchical heading structure extraction is NOT easy

- HTML has tags for describing headings
 - H1 to H6 and DT tags

Hierarchical heading structure extraction is NOT easy

- HTML has tags for describing headings
 - H1 to H6 and DT tags
- These tags are not always used or used incorrectly
 - In our data set:
 - Only 32% of headings were tagged by these tags
 - Only 67% of components tagged by these tags were headings

Hierarchical heading structure extraction is NOT easy

- HTML has tags for describing headings
 - H1 to H6 and DT tags
- These tags are not always used or used incorrectly
 - In our data set:
 - Only 32% of headings were tagged by these tags
 - Only 67% of components tagged by these tags were headings
- **More sophisticated extraction method is necessary**

Humans use visual style

- How do humans extract hierarchical heading structure?

About 597,000 results (0.37 seconds)

[List of F5 and EF5 tornadoes - Wikipedia, the free ...](#)

en.wikipedia.org/wiki/List_of_F5_and_EF5_tornadoes ▾ Wikipedia ▾

Among the most violent known meteorological events are **tornadoes**. Each year, more than 2,000 **tornadoes** occur worldwide, with the vast majority occurring in ...

[2013 El Reno tornado](#) - [2011 Joplin tornado](#) - [2013 Moore tornado](#) - [TORRO scale](#)

[Category:F5 tornadoes - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Category:F5_tornadoes ▾ Wikipedia ▾

Wikimedia Commons has media related to **F5 tornadoes**. These tornado outbreaks had their strongest tornado rate as an F5 on the Fujita scale or an EF5 on the ...

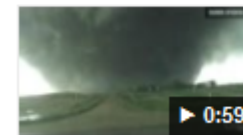
[Images for F5 tornado](#)

[Report images](#)



[More images for F5 tornado](#)

[MASSIVE F5 TORNADO CAUGHT ON CAMERA! - YouTube](#)

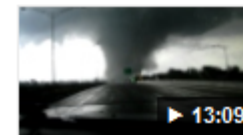


www.youtube.com/watch?v=oaDmpcG0Nw0 ▾

Apr 29, 2012 - Uploaded by ben rumford

STARTING TODAY 32 MILLION FARMERS, HILLBILLIES, AND REDNECKS WILL FACE TORNADOES. I CALL ...

[F5 Tuscaloosa tornado - YouTube](#)




www.youtube.com/watch?v=Tlx26tN6pCk ▾

Apr 29, 2011 - Uploaded by Ryne Chandler

Nate Hughett and Ryne Chandler chasing the **F5 tornado** in Tuscaloosa AL. This storm was like nothing else ...

Humans use visual style

- How do humans extract hierarchical heading structure?
- 
- They use *visual style*
 - consists of various visual attributes of components
 - e.g. font-size, color

About 597,000 results (0.37 seconds)

[List of F5 and EF5 tornadoes - Wikipedia, the free ...](#)

en.wikipedia.org/wiki/List_of_F5_and_EF5_tornadoes ▾ Wikipedia ▾

Among the most violent known meteorological events are **tornadoes**. Each year, more than 2,000 **tornadoes** occur worldwide, with the vast majority occurring in ...

[2013 El Reno tornado](#) - [2011 Joplin tornado](#) - [2013 Moore tornado](#) - [TORRO scale](#)

[Category:F5 tornadoes - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Category:F5_tornadoes ▾ Wikipedia ▾

Wikimedia Commons has media related to **F5 tornadoes**. These tornado outbreaks had their strongest tornado rate as an F5 on the Fujita scale or an EF5 on the ...

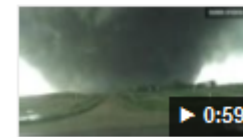
[Images for F5 tornado](#)

[Report images](#)



[More images for F5 tornado](#)

[MASSIVE F5 TORNADO CAUGHT ON CAMERA! - YouTube](#)

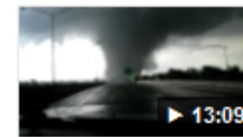


www.youtube.com/watch?v=oaDmpcGONw0 ▾

Apr 29, 2012 - Uploaded by ben rumford

STARTING TODAY 32 MILLION FARMERS, HILLBILLIES, AND REDNECKS WILL FACE TORNADOES. I CALL ...

[F5 Tuscaloosa tornado - YouTube](#)




www.youtube.com/watch?v=Tlx26tN6pCk ▾

Apr 29, 2011 - Uploaded by Ryne Chandler

Nate Hughett and Ryne Chandler chasing the **F5 tornado** in Tuscaloosa AL. This storm was like nothing else ...

Humans use visual style


- How do humans extract hierarchical heading structure?
- 
- They use *visual style*
 - consists of various visual attributes of components
 - e.g. font-size, color

About 597,000 results (0.37 seconds)

[List of F5 and EF5 tornadoes - Wikipedia, the free ...](#)
en.wikipedia.org/wiki/List_of_F5_and_EF5_tornadoes ▾ Wikipedia ▾
Among the most violent known meteorological events are **tornadoes**. Each year, more than 2,000 **tornadoes** occur worldwide, with the vast majority occurring in ...
[2013 El Reno tornado](#) - [2011 Joplin tornado](#) - [2013 Moore tornado](#) - [TORRO scale](#)

[Category:F5 tornadoes - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Category:F5_tornadoes ▾ Wikipedia ▾
Wikimedia Commons has media related to **F5 tornadoes**. These tornado outbreaks had their strongest tornado rate as an F5 on the Fujita scale or an EF5 on the ...

[Images for F5 tornado](#) Report images



[More images for F5 tornado](#)

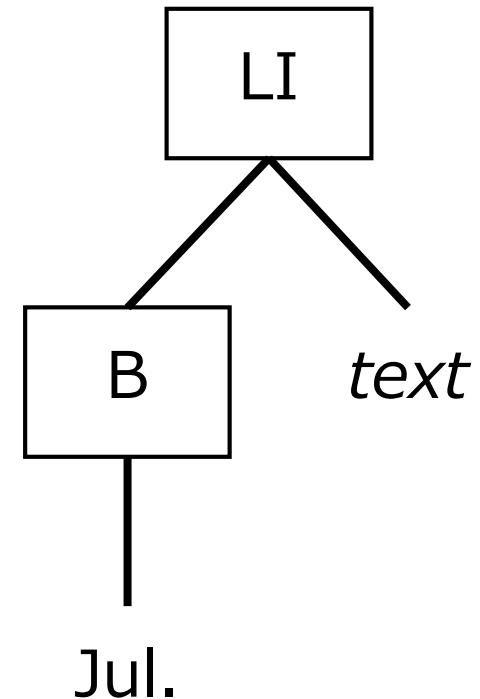
[MASSIVE F5 TORNADO CAUGHT ON CAMERA! - YouTube](#)
www.youtube.com/watch?v=oaDmpcGONwU ▾
Apr 29, 2012 - Uploaded by ben rumford
STARTING TODAY 32 MILLION FARMERS, HILLBILLIES, AND REDNECKS WILL FACE TORNADOES. I CALL ...
▶ 0:59

[F5 Tuscaloosa tornado - YouTube](#)
www.youtube.com/watch?v=Tlx26tN6pCk ▾
Apr 29, 2011 - Uploaded by Ryne Chandler
Nate Hughett and Ryne Chandler chasing the **F5 tornado** in Tuscaloosa AL. This storm was like nothing else ...
▶ 13:09

Visual style can be easily detected

- Visual style is assigned to each DOM node
 - DOM node is a pair of tags or a text fragment split by tags

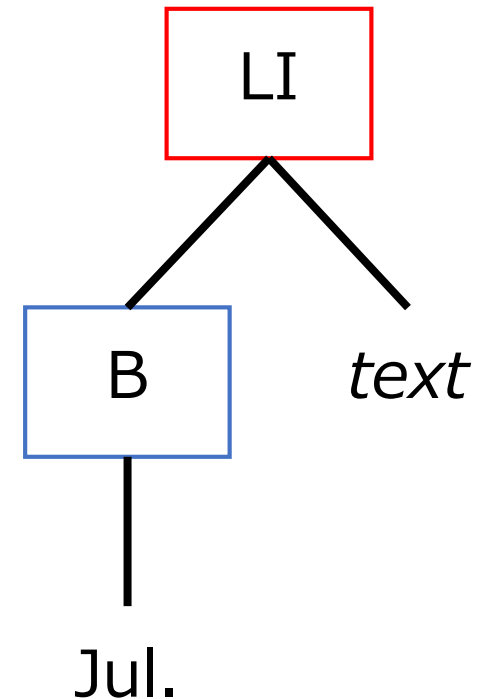
```
<LI>  
  <B>  
    Jul.  
  </B>  
  Construction started.  
</LI>
```



Visual style can be easily detected

- Visual style is assigned to each DOM node
 - DOM node is a pair of tags or a text fragment split by tags

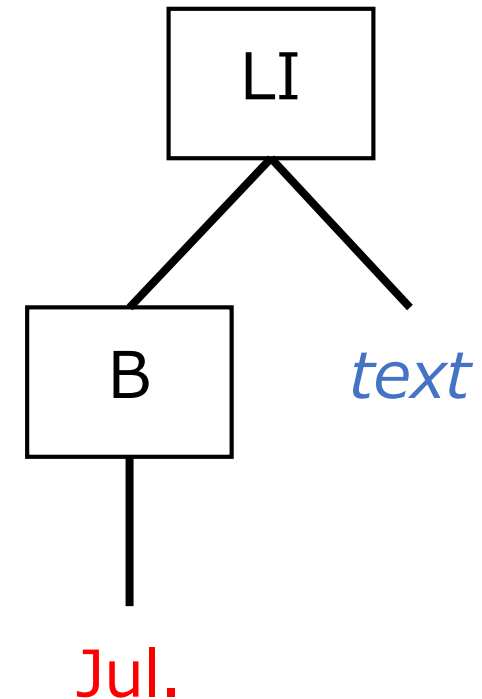
```
<LI>  
  <B>  
    Jul.  
  </B>  
  Construction started.  
</LI>
```



Visual style can be easily detected

- Visual style is assigned to each DOM node
 - DOM node is a pair of tags or **a text fragment split by tags**

```
<LI>  
  <B>  
    Jul.  
  </B>  
  Construction started.  
</LI>
```



Visual style can be easily detected

- Visual style is assigned to each DOM node
 - DOM node is a pair of tags or a text fragment split by tags
- Visual style can be easily detected by computers
- We use it to extract hierarchical heading structure

Disadvantages of existing methods

- There exists some methods that use visual style of nodes

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Disadvantages of existing methods

- There exists some methods that use visual style of nodes
- Existing methods
 - check nodes one-by-one [Okada, Arakawa]

[Okada, Arakawa] H. Okada and H. Arakawa. Automated extraction of non <h>-tagged headers in webpages by decision trees. In *Proc. of SICE Annual Conf.*, pages 2117–2120, 2011.

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Disadvantages of existing methods

- There exists some methods that use visual style of nodes
- Existing methods
 - check nodes one-by-one
 - compare two nodes and judge which one is more likely to be a heading [Pembe, Güngör]

[Pembe, Güngör] F. C. Pembe and T. Güngör. A tree learning Approach to web document sectional hierarchy extraction. In Proc. of ICAART, pages 447–450, 2010.

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Disadvantages of existing methods

- There exists some methods that use visual style of nodes
- Existing methods
 - check nodes one-by-one
 - compare two nodes and judge which one is more likely to be a heading
- They do not use global information within given page

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Our idea

- To use more information, our method
 - groups nodes by visual style into *node sets*
 - judges if each node set is a set of actual headings
- Each node set is
 - a set of headings of same level
 - or a set of non-headings

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Example node sets

- Node sets indicated by color

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Example node sets

- Node sets indicated by color

An example set of actual headings

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Example node sets

- Node sets indicated by color

An example set of non-heading components.

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

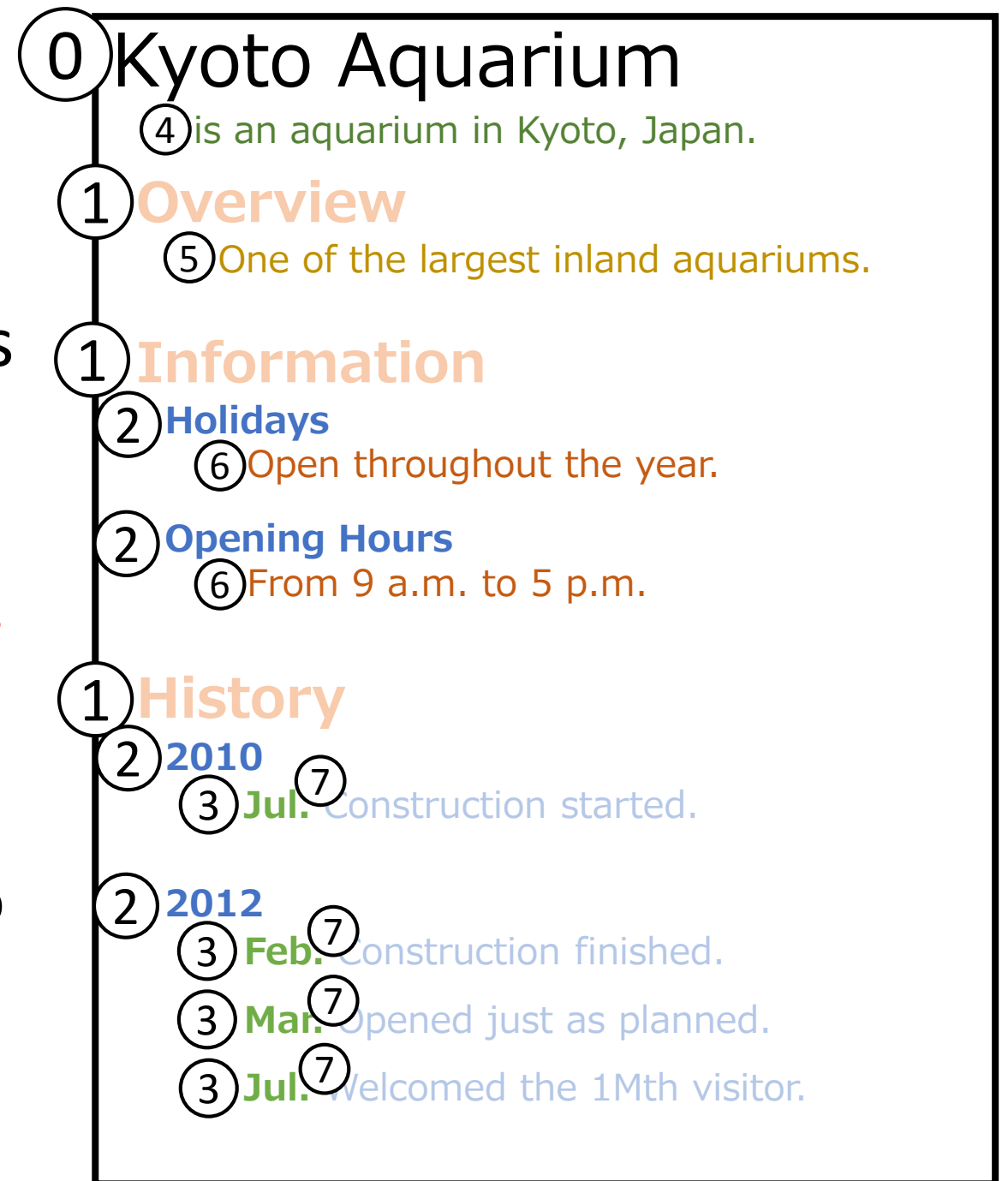
- Jul. Construction started.

2012

- Feb. Construction finished.
- Mar. Opened just as planned.
- Jul. Welcomed the 1Mth visitor.

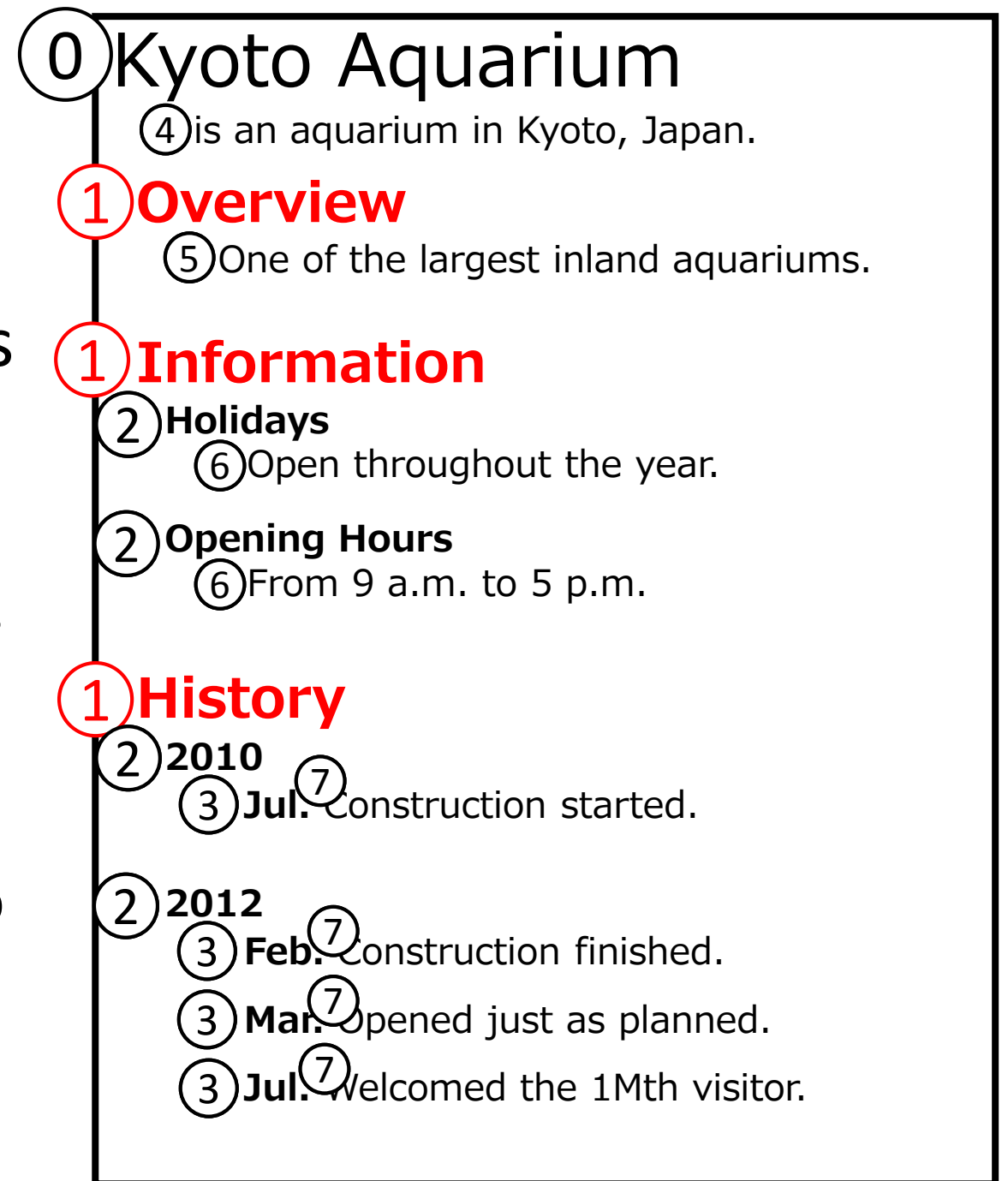
Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks



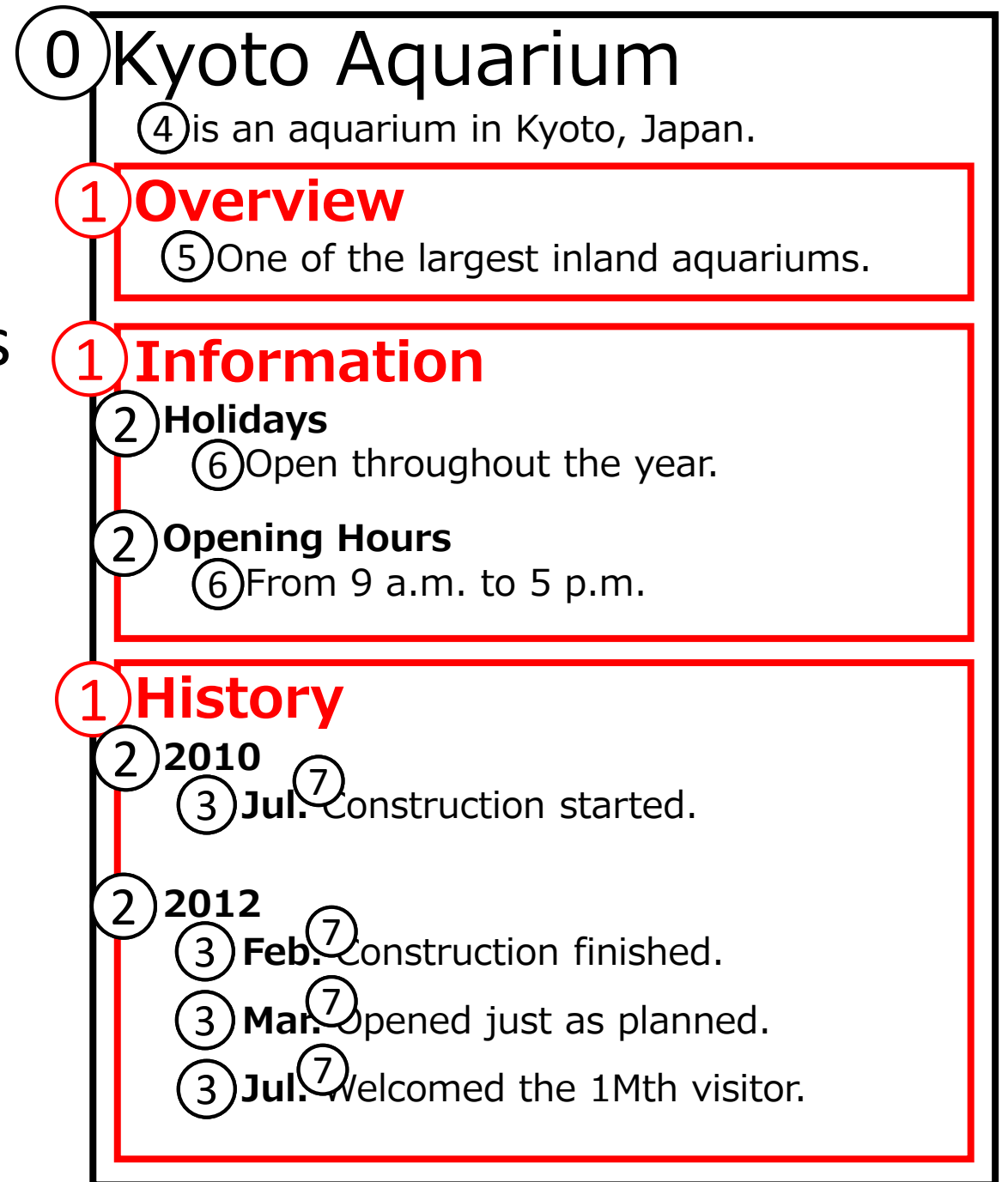
Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks



Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks



Outline of our method

1. Group candidate headings
2. Sort node sets by significance of their style
3. For each node set in desc. order of significance
 - 3.1 Judge if the node set is a set of actual headings
 - 3.2 For actual headings, also extract corresponding blocks

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

2012

- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

Step 1. Group candidate headings

- Candidate heading nodes: a single text or image node
- Group candidates with exactly the same attribute values

Three types of attributes for grouping

1. Visual attribute values computed by web browsers
 - Font-size, font-style, font-weight, text-decoration, and color
2. Tag path
 - Sequence of node names between a node and the root
 - e.g. /HTML/BODY/TABLE/TR/TD/UL/LI/text()
3. Height of images

Step 2.

Sort node sets by significance of their style

Four sort keys in this priority order

1. Depth of corresponding *blocks* in hierarchy
 - because blocks never include blocks at upper levels
2. Font-size
3. Font-weight
4. Document order
 - because a heading of a parent block usually appear earlier than that of a child block

Step 3.

Scan node sets in order of significance

- Our method
 - recursively scans node sets in the descending order of their significance
 - When an actual heading set is found, extracts the blocks corresponding to the headings
- Two sub-steps
 - 3.1 Judge if a node set is an actual heading set
 - 3.2 Detect the corresponding blocks from headings

Step 3.1 Judging if a node set is actual heading set

5 heuristic rules

- e.g. all headings in one parent block are unique

Kyoto Aquarium

is an aquarium in Kyoto, Japan.

Overview

One of the largest inland aquariums.

Information

Holidays

Open throughout the year.

Opening Hours

From 9 a.m. to 5 p.m.

History

2010

- **Jul.** Construction started.

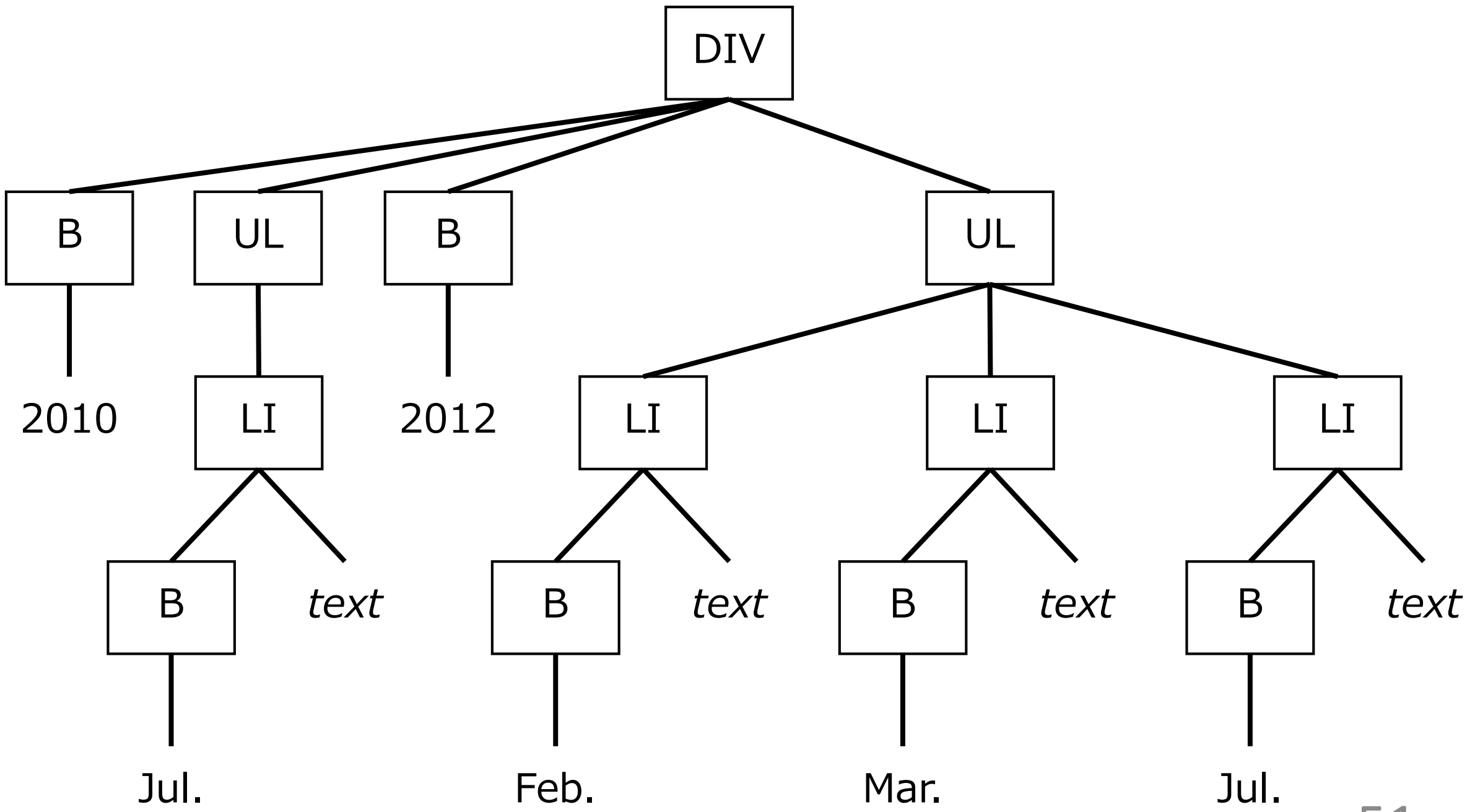
2012

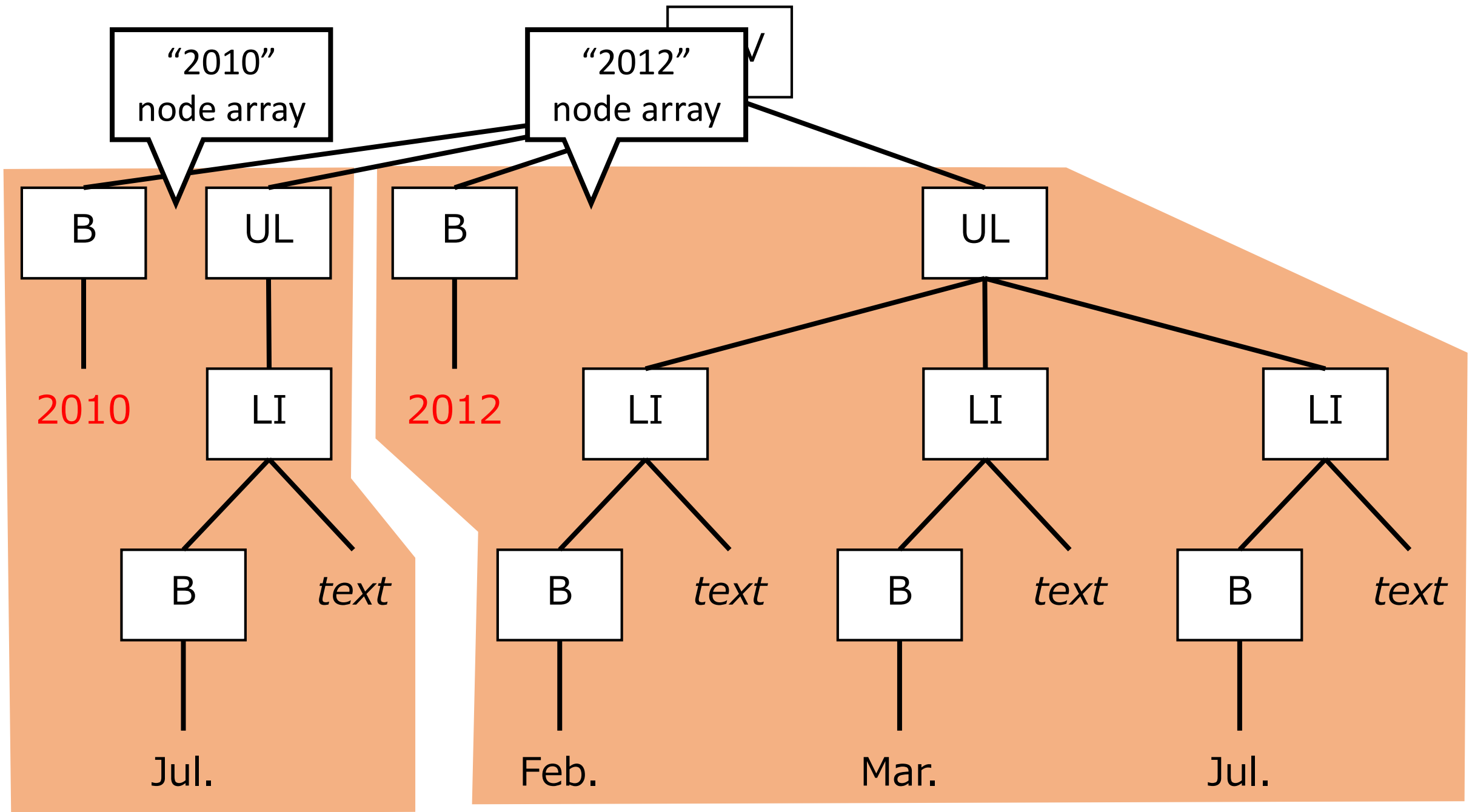
- **Feb.** Construction finished.
- **Mar.** Opened just as planned.
- **Jul.** Welcomed the 1Mth visitor.

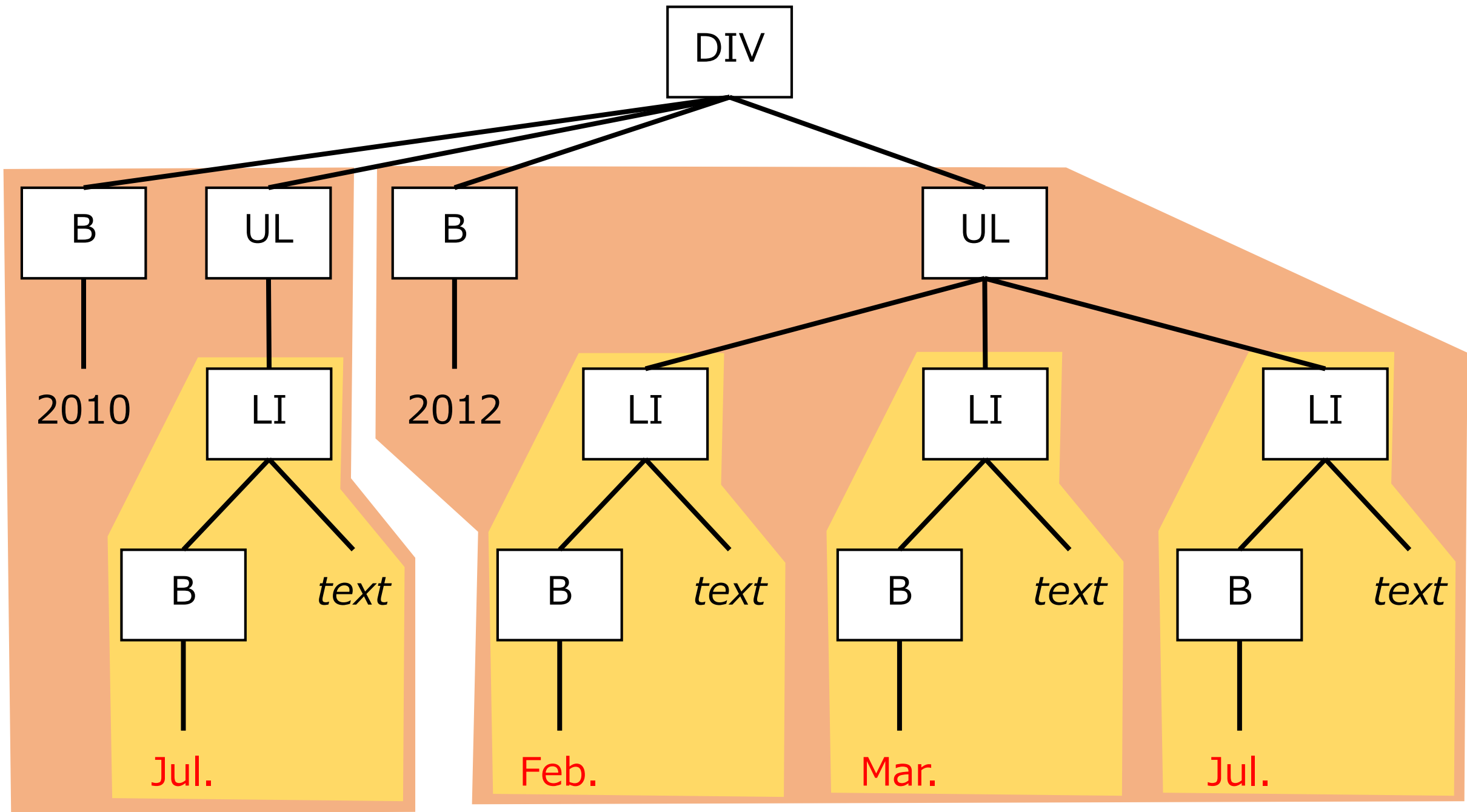
Step 3.2

Detecting corresponding blocks from headings

- When a node set passed all the rules, our method
 - regards it is an actual heading set
 - detects blocks corresponding to the headings
- To determine blocks from headings, our method use correspondence between them and DOM sub-tree
 - A heading corresponds to a single text or image node
 - A block corresponds to a *node array*, a set of adjoining sibling nodes and their descendants







DIV

B

UL

B

UL

2010

LI

2012

LI

LI

LI

B

text

Jul.

B

text

Feb.

B

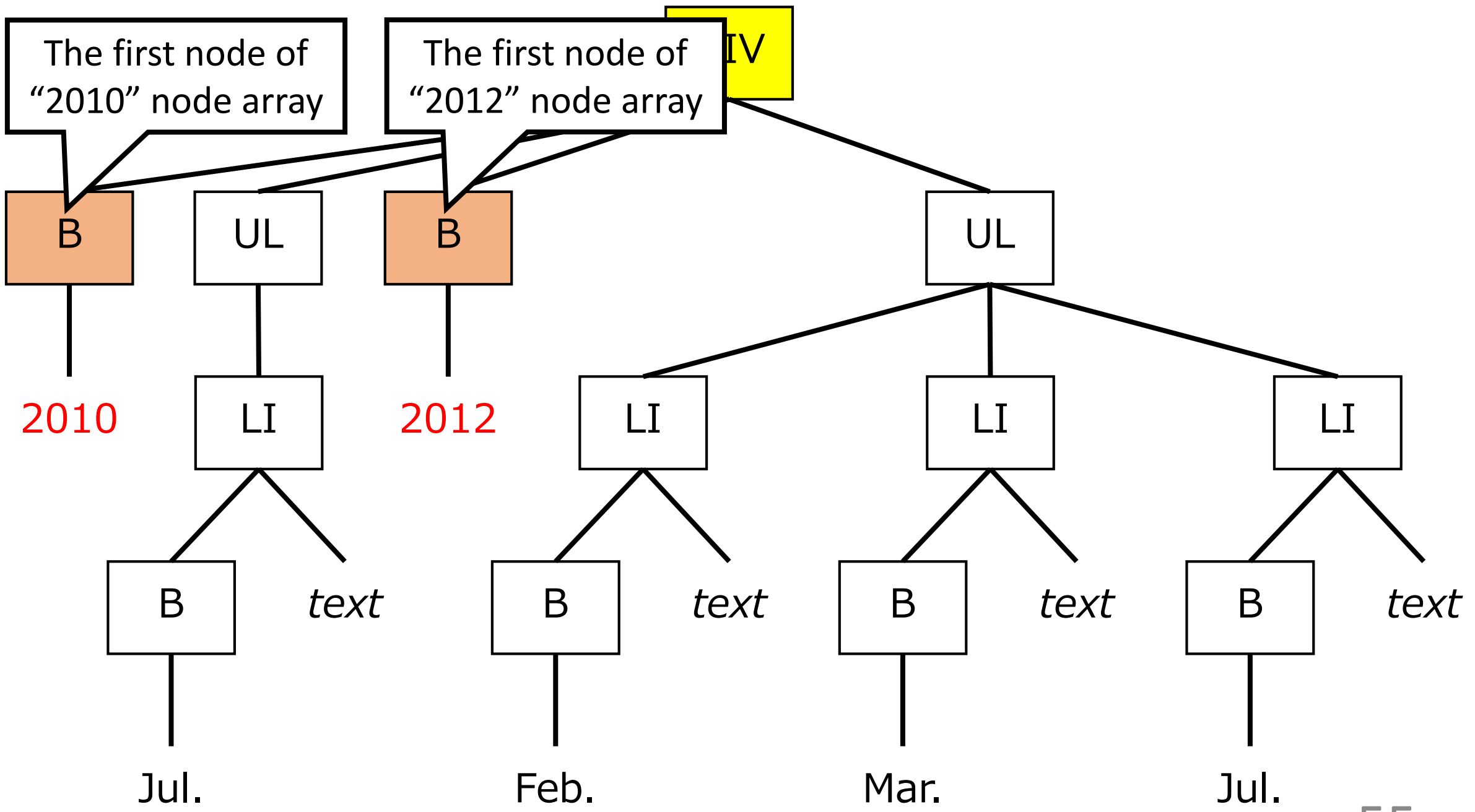
text

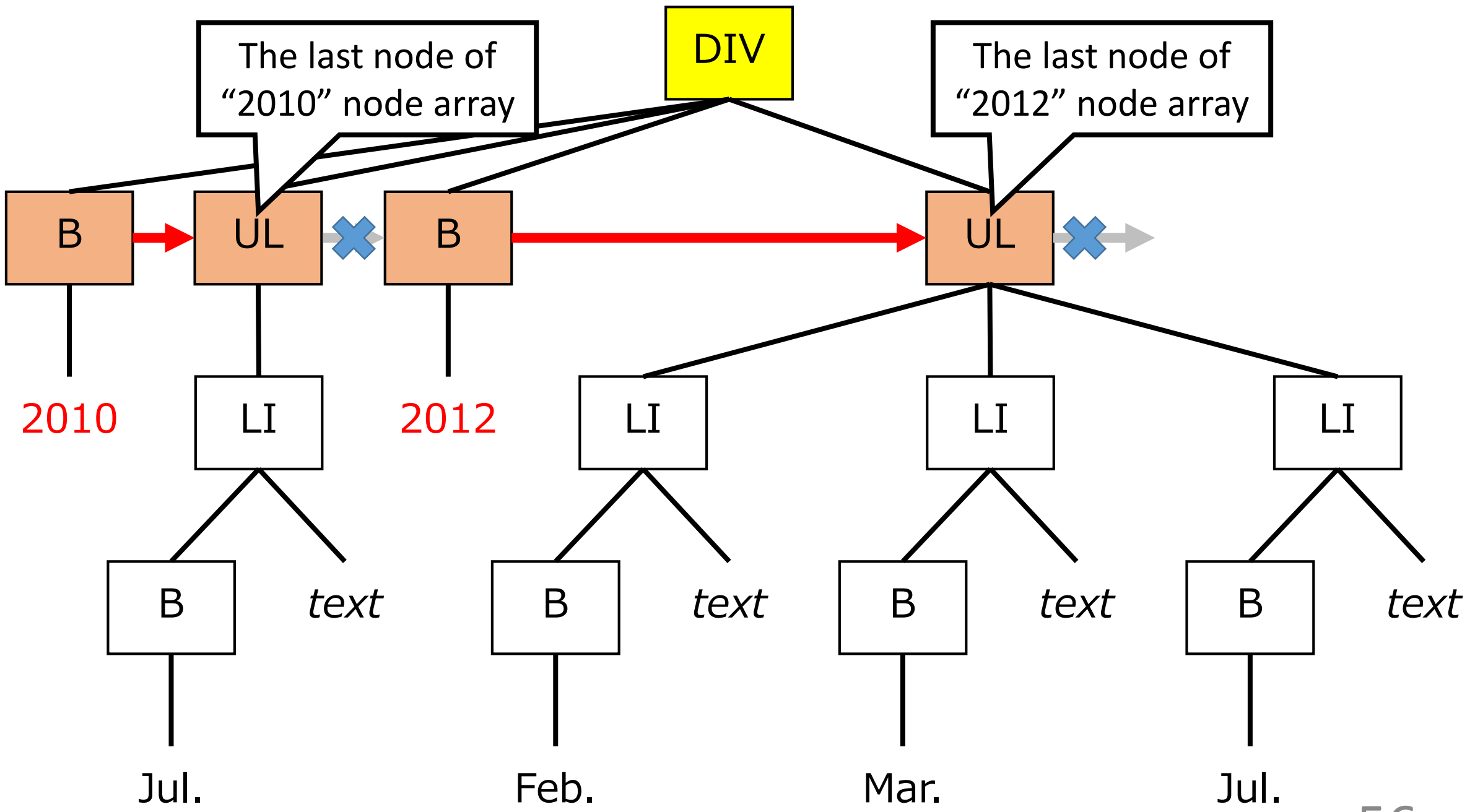
Mar.

B

text

Jul.





Experimental setting

To evaluate our method

- Random 803 pages from ClueWeb09
 - For excluding spam pages, only pages relevant to some intents in TREC Web track were collected
- For each page, 1 of 5 annotators hand-annotated hierarchical heading structure in its content body
 - Fleiss' Kappa: .693 for headings and .583 for blocks

Evaluation result (heading extraction)

Method	Precision	Recall	F1-score
Decision tree learning [Okada, Arakawa]	.084	.884	.154
Naïve method based on tag names	.668	.320	.433
Our method	.638	.569	.602

Evaluation result (heading extraction)

Method	Precision	Recall	F1-score
Decision tree learning [Okada, Arakawa]	.084	.884	.154
Naïve method based on tag names	.668	.320	.433
Our method	.638	.569	.602

- The decision tree learning method did not work well
 - Most test pages did not share visual style with training pages

Evaluation result (heading extraction)

Method	Precision	Recall	F1-score
Decision tree learning [Okada, Arakawa]	.084	.884	.154
Naïve method based on tag names	.668	.320	.433
Our method	.638	.569	.602

- The decision tree learning method did not work well
 - Most test pages did not share visual style with training pages
- Our method achieved a far better recall retaining about same precision as the naïve method

Evaluation result (block extraction)

Method	Precision	Recall	F1-score
VIPS [Cai+]	.215	.070	.106
Our method	.586	.563	.574

Evaluation result (block extraction)

Method	Precision	Recall	F1-score
VIPS [Cai+]	.215	.070	.106
Our method	.586	.563	.574

- VIPS did not work well
 - because its extraction target is layout structure
 - VIPS is complementary to our method

Evaluation result (block extraction)

Method	Precision	Recall	F1-score
VIPS [Cai+]	.215	.070	.106
Our method	.586	.563	.574

- VIPS did not work well
 - because its extraction target is layout structure
 - VIPS is complementary to our method
- Our method: in accuracy close to heading extraction
 - Extracted blocks from actual headings by precision of .769

Conclusion

- Extraction of hierarchical heading structure is important for various applications of the web
- We proposed a method based on an idea that headings of the same level share their visual style
- Our method achieved high recall and satisfactory precision
- Our code and data sets will be available online
 - <https://github.com/tmanabe>