Extracting Logical Hierarchical Structure of HTML Documents Based on Headings

Tomohiro Manabe and Keishi Tajima {manabe@dl.kuis, tajima@i}.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto Univ.

Motivation: Extraction of hierarchical heading structure (HHS) seems easy, but is NOT

- Only 32% of headings are tagged by heading tags
- Only 67% of heading tags are actual headings

Definitions: HHS consists of nested blocks with headings

- Heading: Topic description of a segment
- Block: A segment with its heading

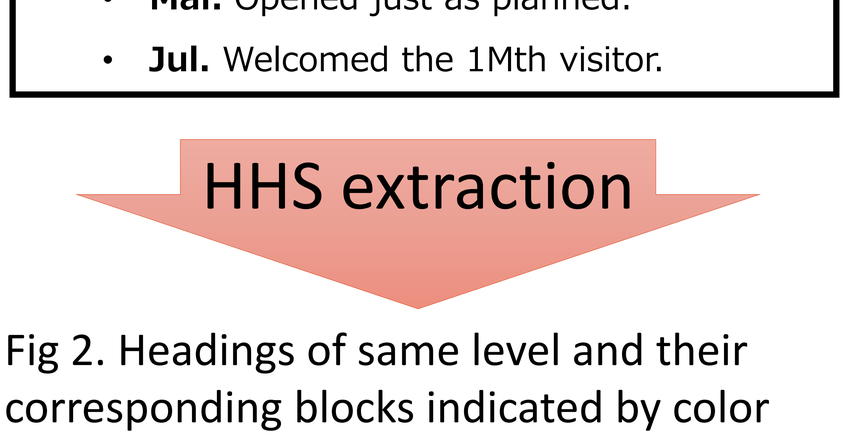
Our idea: Headings of same level share visual style

- which is easily detected by computers based on attributes computed by browsers (e.g. font-size) and tag path (e.g. /html/body/ul/li/b/text)

Our method HEPS:

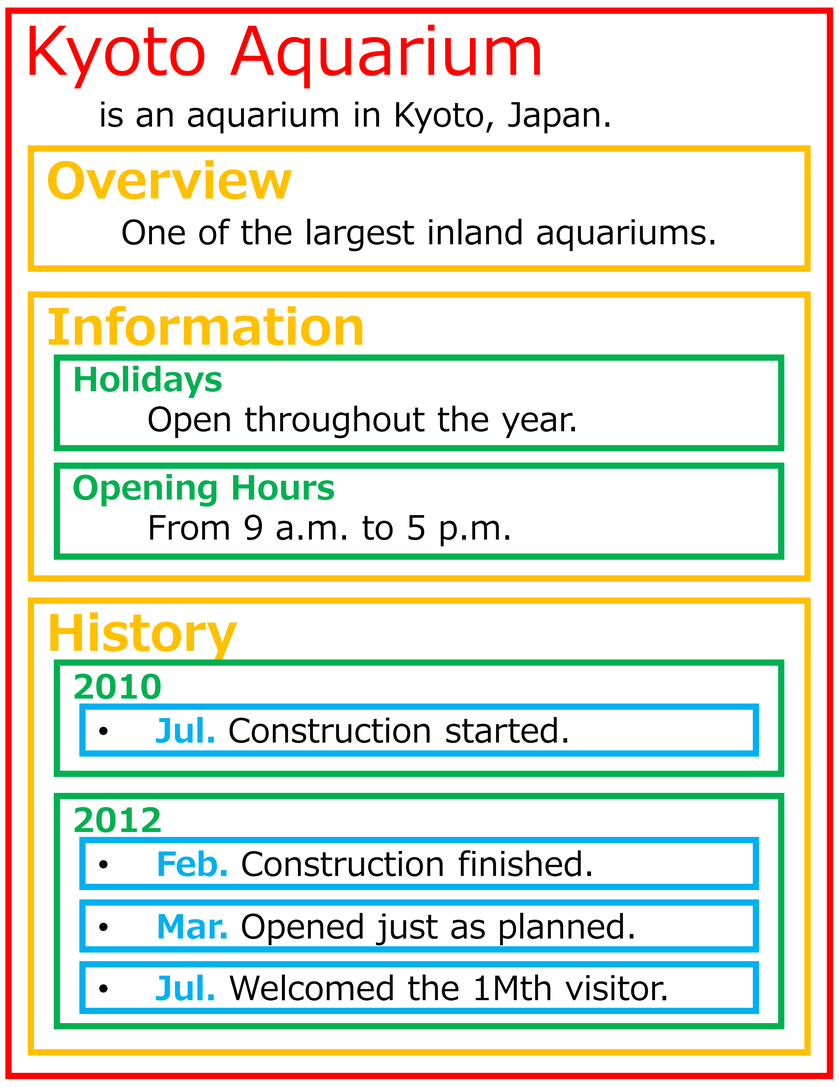
1. Groups candidate headings into sets by style
2. Sorts sets by significance of style
3. Scans all sets in descending order of significance
   - 3.1 Judges if the set is an actual heading set
     - Point: Set by set, not node by node
   - 3.2 On finding headings, also extracts blocks

Fig 1. Example page



HHS extraction

Fig 2. Headings of same level and their corresponding blocks indicated by color



Evaluation results:

Heading extraction

	P	R	F
Learning[1]	.084	.884	.154
Naïve	.668	.320	.433
HEPS	**.638**	**.569**	**.602**

Block extraction

	P	R	F
VIPS[2]	.215	.070	.106
HEPS	**.586**	**.563**	**.574**

- HEPS extracted many headings retaining same precision as naïve method that uses tag names
- HEPS extracted blocks in the precision and recall close to heading extraction

Resources: Our code and data sets will be available at https://github.com/tmanabe

[1] H. Okada and H. Arakawa. Automated extraction of non <h>-tagged headers in webpages by decision trees. In *SICE*, 2011.
[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: A vision-based page segmentation algorithm. MSR–TR–2003–79, 2003.