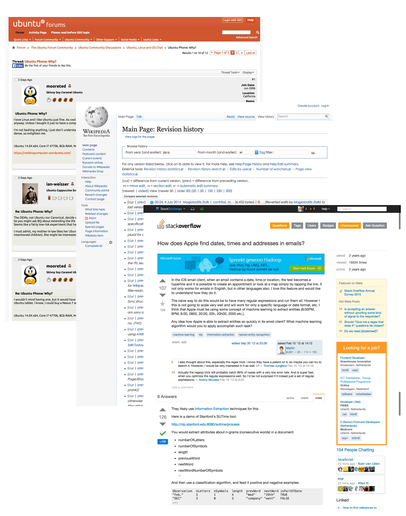
Sub-document Timestamping of Web Documents

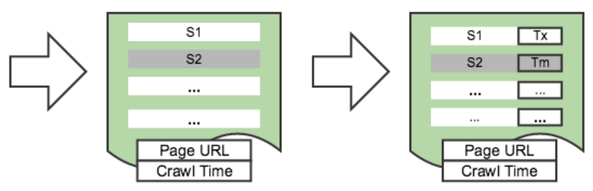
Yue Zhao and Claudia Hauff
Web Information Systems Group, TU Delft
{Y.Zhao-1, C.Hauff}@tudelft.nl

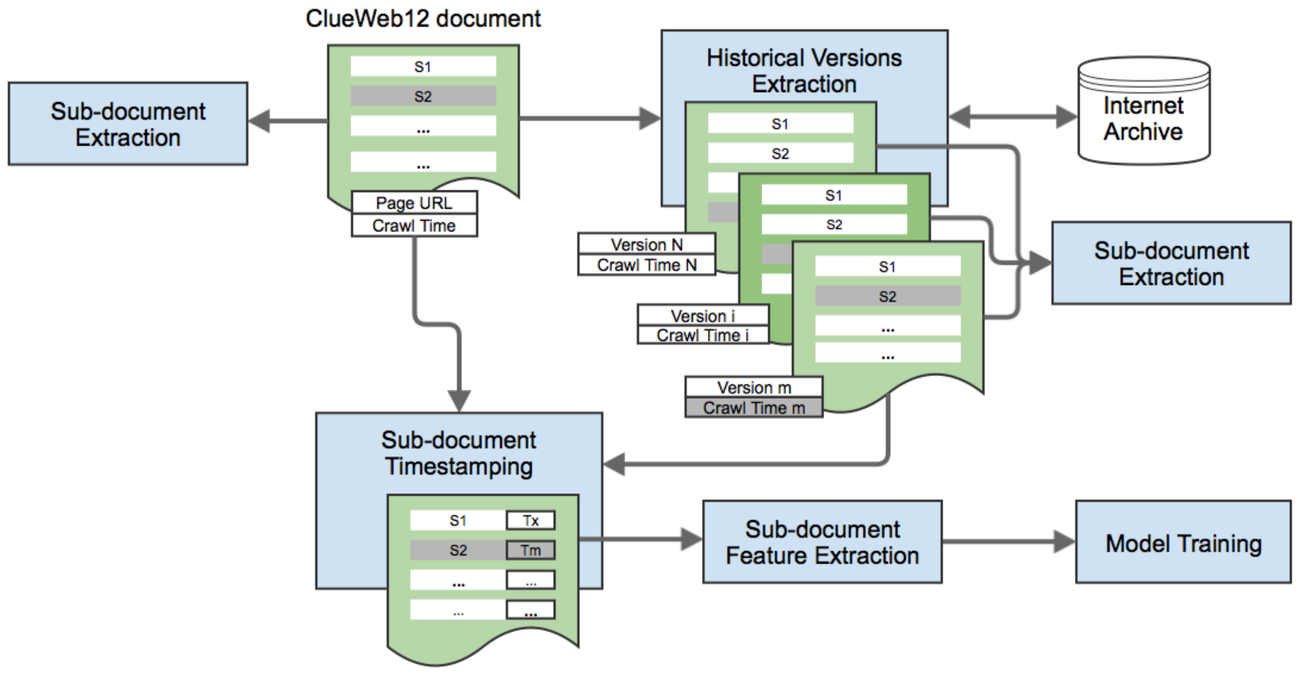
Goal
To estimate the creation time of individual Web documents' components (so-called sub-documents).

Current Situation
1. Creation time is the important information for building timelines, generating snippets and clustering pages.
2. Existing work is all based on the assumption that each Web page has only one creation time.
3. We know that for many pages not all the blocks on a Web page are created at once, like blogs, Q&A, forums…

Timestamping Pipeline: A pipeline which is used to timestamp sub-documents of Web documents based on Internet Archive and to extract features for inferring timestamps of sub-documents which are not crawled by Internet Archive.

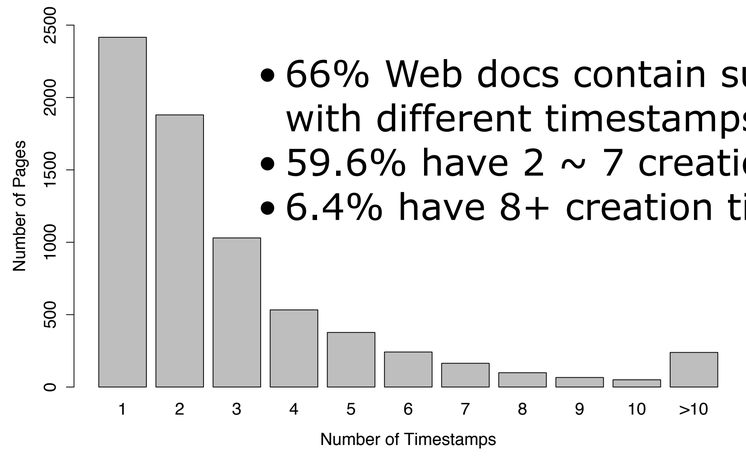






Research Question 1:

1.1 To what extent do Web documents consist of sub-documents created at different times?

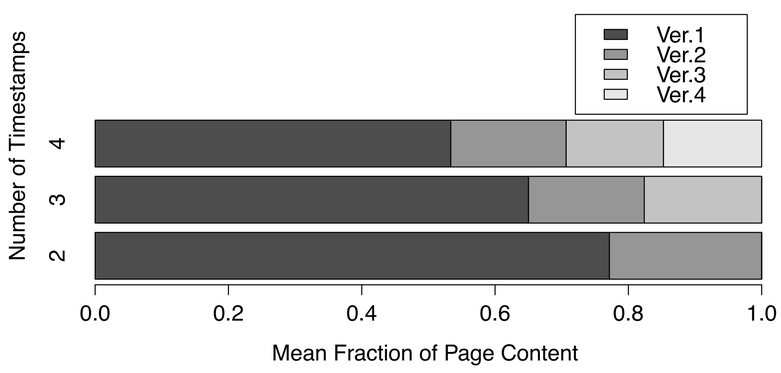


- 66% Web docs contain sub-documents with different timestamps in total;
- 59.6% have 2 ~ 7 creation timestamps;
- 6.4% have 8+ creation timestamps.

1.2 What is the timespan between the oldest and most recent sub-document of a document?

- 1079 days on average, 813 days on median as well

1.3 What fraction of the current document has been created in each version (a version corresponding to a particular timestamp)?



- Docs with 2, 3, 4 creation timestamps has 77%, 65%, 53% of the content created in the first version on average respectively.

1.4 For a specific topic, what is the distribution of its relevant documents in terms of different numbers of timestamps?

- For most topics a relatively large percentage of relevant documents contain 2 or more creation timestamps.

Research Question 2:

2.1 To what extent are we able to classify each sub-document as either having been created within the past month (relative to the document crawl time), within the past year, or more than *m* years ago?

	#Instances	Method	Misclassified	F-Measure / Class A	B	C	D	E
Entire Data Set	556,243	RF	23.85%	0.72	0.70	0.73	0.79	0.86
Data Set with TEs only	145,038	RF	27.54%	0.72	0.67	0.66	0.73	0.86
Data Set with TEs only	145,038	BL: earliest TE	64.60%	0.43	0.32	0.22	0.20	0.39
Data Set with TEs only	145,038	BL: latest TE	64.68%	0.43	0.32	0.21	0.20	0.41

2.2 What document features are used in the classification?

- Features about the structure of sub-documents
- Features about the structures and values of temporal expressions (TEs) in sub-documents

Future Work:

1. Combine linguistic features with the structural features of the sub-documents
2. Answer the next question “What is updated?”

TU Delft Delft University of Technology

SIKS