Efficient Filtering of XML Documents for Selective Dissemination of Information
Mehmet Altinel, Michael J. Franklin

Discussion Presented by
Yutao Guo
University of Waterloo

Outline

- Overview of the contribution by this paper and novel features of the XFilter system
- Problems with the algorithms of the XFilter system
- System overall evaluation
- Presentation of this paper
- General comments about the paper
Overview of the contribution by this paper and novel features of the XFilter system

- Goal of SDI systems
- Key insight to building efficient and scalable SDI systems
- Choosing of XML and XPath
- Sophisticated index structure and matching algorithms based on a modified FSM

Problems with the algorithms of the XFilter system

- Only supports unordered XML data, but not ordered data.
  --- For example, to query the document tree D (in the right figure):
  T1=/a/b[*]/c]/d,
  T2=/a/b/*[following-sibling::d]/c

  --- Conclusion: unable to tell the difference between the matched documents in their structures
Problems with the algorithms of the XFilter system (cont’d)

- Index on each single element name using a hash table structure.
  - Space cost on the hash table.
  - Keeping track of all instances of partially matched tree patterns before the end of a FSM.
- Dissemination mechanism: simple unicast delivery and entire document sending to each interested user.
  - A bottleneck of the whole system?

System overall evaluation

- Precision
  - Importance of precision evaluation for any filtering system
  - No precision evaluation addressed in this paper
  - Problem of effectiveness in expressing user profile
- Efficiency
  - Drawbacks in the indexing mechanism make XFilter require further efforts to enhance efficiency
  - XFilter’s attempts (i.e., list balancing, pre-filtering) are not significantly helpful
  - How’s the system efficiency when also considering boolean combinations of XPath queries?
System overall evaluation (cont’d)

- **Scalability**
  - not very scalable while documents’ volume, length and tree depth scale up
  - not scalable if queries contain very complicated expressions, e.g., deeply nested path expressions

- **Adaptability**
  - cannot handle non-XML-encoded documents
  - unalterable SAX interface
  - XFilter only performs the batch filtering task, not powerful enough to be an adaptive filtering system

Presentation of this paper

- Points that are not clearly addressed by this paper:
  - Effectiveness of user interests in other SDI systems
  - Why choose XPath rather than UnQ2, Lorel and XML-QL?
  - Any experimental results based on comparison between XFilter and other filtering systems?
General comments about the paper

- pay attention to precision evaluation
- More efforts on better index structure for reducing time and space costs
  --- e.g., index based on decomposing tree patterns into collections of substrings
- Judge the relevance of the retrieved document, and learn a better profile from on-line feedback
- Attempts on broadcast delivery and partial sending of the documents

Open time for discussion…