Enterprise Track: Thread-based Email Retrieval

Yejun Wu and Douglas W. Oard

College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park

Goal

- Explore
  - -- document expansion using threads,
  - -- inclusion or exclusion of quoted text,
  - -- normalization of date and version expression.
- Jointly creating first email retrieval test collection.

Email Search

- **Known Item (KI) Search**: find a specific email that the user knows to exist.
- **Discussion Search (DS)**: search for emails that contribute at least one pro or con related to a specified topic in new (not quoted) text.

W3C Collection

- NIST Crawled from w3c.org in June 2004
- XML/HTML markup
- Parsed 174,313 email messages.
- 22,252 multi-message threads, average length = 4.7 messages, median = 3.
- 68,899 single-message threads.
- Threading based on subject line reply chains.

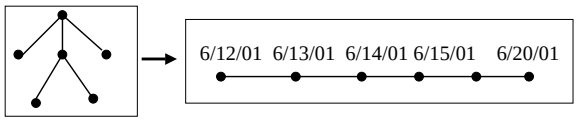
Relevance Assessment

- NIST pool: top 50 docs.
- Average of 529 messages/topic across all the 59 topics.
- Relevant: on topic + pro/con; Partially relevant: on topic, no pro/con.

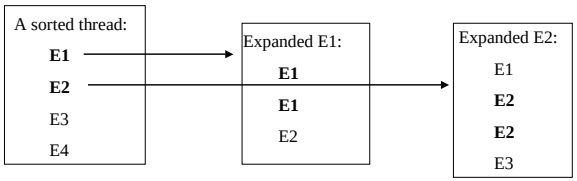
Conclusion

- Helpful: normalization of date and version expression.
- Not Helpful:
  - -- suppression of quoted text;
  - -- document expansion using linearized threads.

Linearization of Threads



Document Expansion with Adjacent Messages



Date and Version Normalization

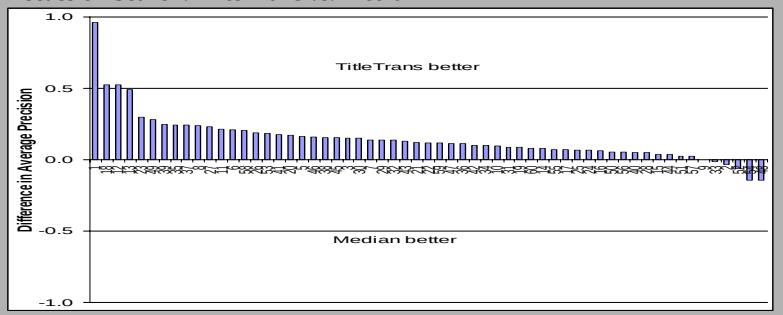
Before transformation	After transformation
12-Jan-2000	**day12 January** 2000
Jan 6-12 or Jan 6 - 12	January day6 - day12
6~12 January or 6-12 January	day6 - day12 January
6 January 2001	day6 January 2001
Jan 6 2001 or Jan. 6 2001	January day6 2001
2001-01-06 (Not in URL)	January day6 2001
HTML-4.1	HTML four **pointx** one
HTML5.1c1	HTML five pointx one c1
J2SDK1.4.1	J2SDK one pointx four pointx one
gcc-make 3.7	gcc-make three pointx seven
gcc-12.7.2.1p1	gcc twelve pointx seven pointx two pointx one p1

Automatic Runs

- **TitleDefault (DS)** **KIDefault (KI):**
  Baseline, messages indexed in original form.
  Queries were constructed from all words in "query/title" field.
- **ThrQuot (DS)** **KIThrQuot (KI):**
  Document expansion: thread with quoted text. "Query/title" field.
- **ThrNoQuot (DS)** **KIThrNoQuot (KI):**
  Document expansion: thread without quoted text. "Query/title" field.
- **ThrNoQNarr (DS):**
  Document expansion: thread without quoted text.
  Queries were constructed from all words from "query" and "narrative".
- **TitleTrans (DS)** **KITrans (KI):**
  Document transformation. "Query" field transformed correspondingly.
- **TitleNewText (DS)** **KITitleNewText (KI):**
  Messages without quoted text. "Query" field. Not submitted.

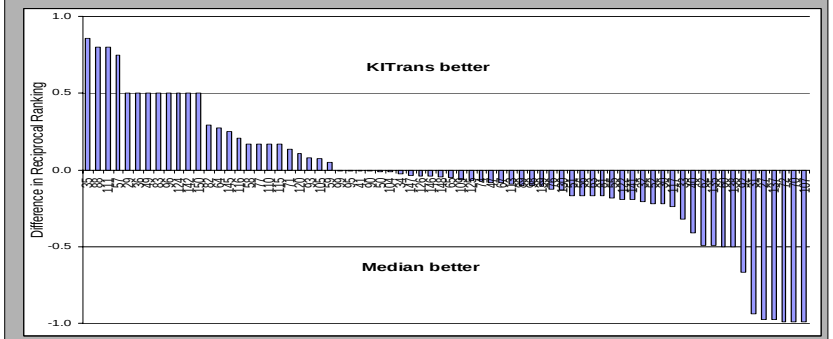
Results (evaluated on topicality)

Discussion Search: TilteTrans vs. Median



Wilcoxon signed-rank test for paired samples: statistically significant at p>0.05.

Known Item Search: KITrans vs. Median



- 125 topics, 67 runs: 25 better, 53 identical, 47 worse. Top 100 documents submitted.
- KI70: Target has a short new text body with long quoted text. TitleNewText ranked top 1.
- KI107: Query: "access tables on the web". Target: subject="Access tables on the web." Phrase query?

Discussion Search