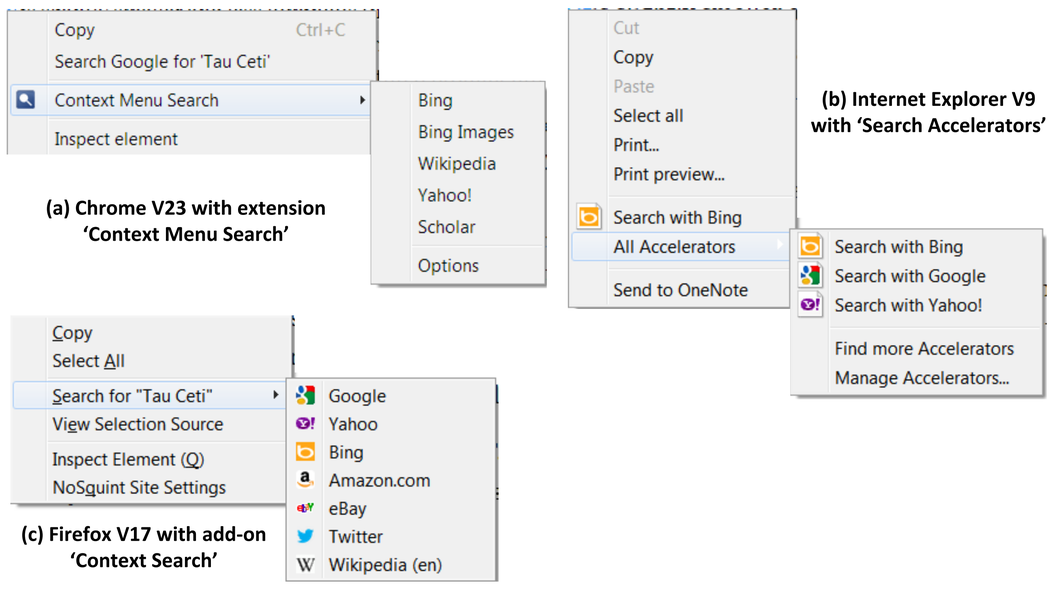
Towards Context-Aware Search with Right Click

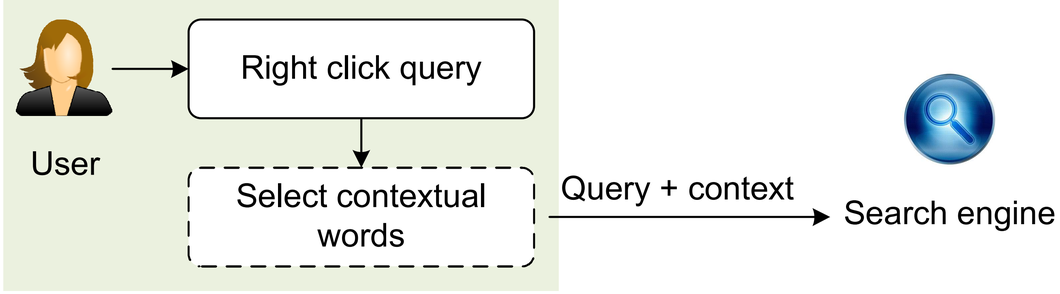
Aixin Sun **Chii-Hian Lou**
School of Computer Engineering, Nanyang Technological University, Singapore
axsun@ntu.edu.sg louc0001@e.ntu.edu.sg

Right-click Query



- Many queries are submitted to search engines by **right-clicking** some selected words in Web browser.
- Right-click queries are currently processed in the same way by a search engine as queries submitted through other means.
- The source document from which the query is marked for search provides sufficient contextual information to determine the right semantic of the query.

Context-Aware Search Framework



- Contextual information can be extracted to enable ***Context-Aware Search*** for better user search experiences
- **Two main research questions:**

→ Given the source document of a right-click query, which component of the document (*e.g.,* title, full text, paragraph containing the right-click query) is best in providing contextual information for the query?

→ What contextual information (*e.g.,* words, nouns, or noun phrases) shall be extracted to augment the query?

Context Extraction

- **Seven text components T1 – T7**

T1	Full text of the page
T2	Paragraph of the selected query word(s)
T3	Title of the page
T4	Title, the first and last paragraph
T5	Paragraphs containing the query word(s)
T6	Meta description and keyword of the page
T7	Full text of the current and referenced articles

- **Five feature extraction schemes F1–F5**

F1	Words	Frequency-based Weighting
F2	Words	Proximity-based Weighting
F3	Nouns	Frequency-based Weighting
F4	Nouns	Proximity-based Weighting
F5	Noun Phrases	Phrase Weighting

Weighting Scheme

- **Frequency-based weighting**: $TF \cdot IDF$ weighting scheme
- **Proximity-based weighting**: A term is more important if (i) its $TF \cdot IDF$ score is large, (ii) it occurs for multiple times in the selected text component, and (iii) the occurrences are close to the query in terms of proximity distance.

$$p_w(t_i) = \sum_{j=1}^{f_i} \frac{f_w(t_i)}{dist(t^j, q)}$$

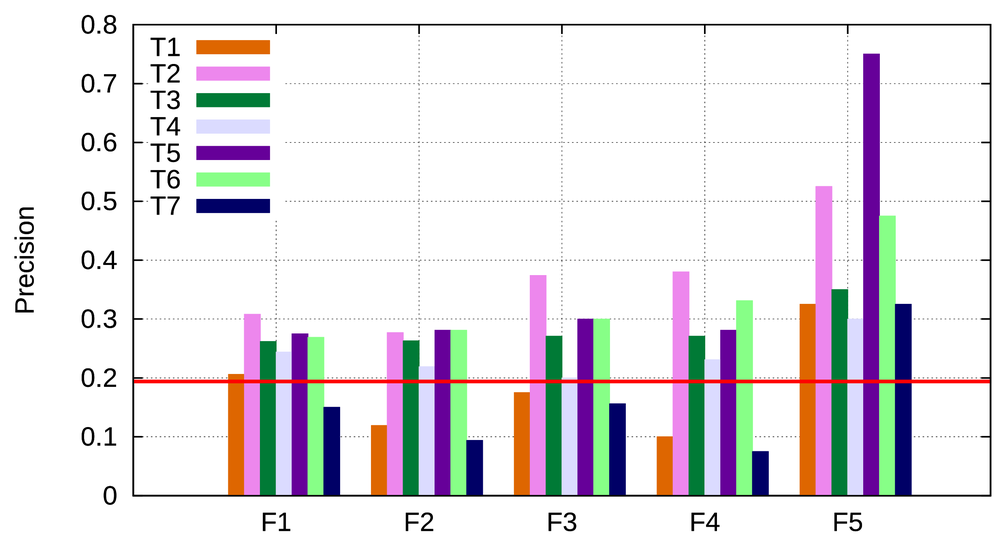
- **Phrase weighting**: (i) the phrases's $TF \cdot IDF$ score $f_w(s)$ by treating each phrase as a token, and (ii) the average frequency of all terms contained in the phrase $\sum_{t_i \in s} f_i/|s|$, where $|s|$ is the number of terms in phrase s.

$$s_w(s) = f_w(s)\frac{\sum_{t_i \in s} f_i}{|s|}$$

Evaluation

- **Data collection:** A user study using 20 news articles from Yahoo! News selected mainly based on two criteria: (i) article contains an ambiguous query term; (ii) two or more articles contain the same query term but with different semantics.
- **Baseline method:** Top-8 keywords recommended by Google search engine for the right-click query.
- **Our methods**: Top-8 keywords (or at most 8 words from top-N phrases) ranked by the text component T_x and feature extraction scheme F_x
- **Evaluation metric:** Each of the top-8 ranked words is manually judged to be relevant or irrelevant based on the content of the news article and the right-click query. Precision is used to evaluate the methods.

Result



Observations:

- Noun phrases with phrase weighting (F5) is the best context feature extraction scheme
- Paragraphs containing the query words (T5) are the best text components for query context extraction
- Proximity-based weighting scheme *adversely affects* the precision compared with frequency-based weighting scheme.
- Between nouns and any words, using the same weighting schemes, nouns define better contextual information than any words.