Towards Context-Aware Search with Right Click

Aixin Sun
School of Computer Engineering
Nanyang Technological University, Singapore
axsun@ntu.edu.sg

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

ABSTRACT

Many queries are submitted to search engines by right-clicking the marked text (*i.e.,* the query) in Web browsers. Because the document being read by the searcher often provides sufficient contextual information for the query, search engine could provide much more relevant search results if the query is augmented by the contextual information captured from the source document. How to extract the right contextual information from the source document is the main focus of this study. To this end, we evaluate 7 *text component extraction* schemes, and 5 *feature extraction* schemes. The former determines from which text component (*e.g.,* title, meta-data, or paragraphs containing the selected query) to extract contextual information; the latter determines which words or phrases to extract. In total 35 combinations are evaluated and our evaluation results show that noun phrases extracted from all paragraphs that contain the query word is the best option.

Categories and Subject Descriptors

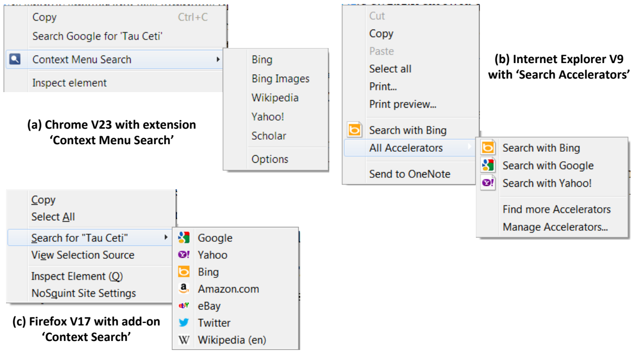
H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation*

Keywords

Right-click query, Context-aware search, Query expansion

1. INTRODUCTION

Search for information online through general or dedicated search engines becomes a part of our daily life. To perform a search, a keyword query is often submitted to a search engine and the latter returns the documents most relevant to the query. A keyword query can be submitted to a search engine through many applications communicating with the search engine. Web browser is one of such applications. Figure 1 captures the pop-up menus of three popular Web browsers (*i.e.,* Chrome, Internet Explorer, and Firefox) when a user right-clicks some selected words in a Web page (*e.g.,* "Tau Ceti" in the figure). The selected words will then be submitted to search engine as a keyword query. Because of the way a query is submitted to a search engine, we refer a query submitted by right-clicking some selected words in Web browser ***right-click query***. Figure 1 also shows that some native or third-party extensions (also known as add-ons or accelerators) that have been developed facilitating user's query submission to her preferred search engine. With this background, we motivate this study with two observations.



(a) Chrome V23 with extension 'Context Menu Search'

(b) Internet Explorer V9 with 'Search Accelerators'

(c) Firefox V17 with add-on 'Context Search'

Figure 1: Right click to search in three popular Web browsers

- First, right-click queries are currently processed in the same way by a search engine as queries submitted through other means (*e.g.,* query box in search engine's Web page). In other words, regardless how a keyword query is submitted to a search engine the same set of results is returned.[1]
- Second, for many queries, particularly those short queries that can be interpreted with multiple semantics (*e.g.,* "apple" and "jaguar"), the source document from which the query is marked for search provides sufficient contextual information to determine the right semantic of the query.

Based on the two observations, we argue that the contextual information available in the *source document* of a right-click query is not utilized by existing search engines. On the other hand, such contextual information can be extracted with limited effort to enable ***Context-Aware Search*** for better user search experiences.

The notion of *context* can be defined in many different ways in Information Retrieval research. In a recent survey on contextual search, Melucci gives two different extremes of the definition: (i) "the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed", and (ii) "the parts of something written or spoken that immediately precede and follow a word or passage and clarify its mean-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609456.

[1] Here we fix the other factors that might affect search results from a search engine, *e.g.,* the location the query is issued, the logon status of a user who submits the query, or the device from which the query is submitted to the search engine.

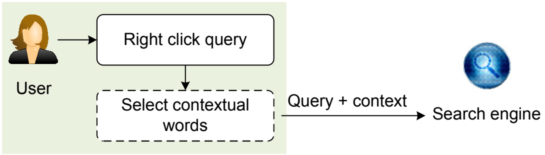


Figure 2: Context-aware search framework

ing" [6]. Our definition of "context", hence the notion of *context-aware search*, is close to the second extreme. That is, in our definition of *context-aware search*, we do not consider any other factors of a user beyond the current document a user is reading, *i.e.,* the source document.

In this paper, we enumerate and compare different approaches to extract contextual information from the source document for a right-click query. We try to answer two main questions: (i) 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? and (ii) What contextual information (*e.g.,* words, nouns, or noun phrases) shall be extracted to augment the query? To answer the first question, we evaluate 7 text extraction schemes to determine the text component from which to extract the contextual information of a right-click query. To answer the second question, we evaluate 5 feature extraction schemes to extract and rank words, nouns, and noun phrases using different weighting schemes. In our experiments, we report the evaluation of 35 context extraction combinations on a collection containing 20 news articles from Yahoo! news.[2] Our evaluation shows that noun phrases extracted from the paragraphs containing the query word of a right-click query provide the best contextual information.

We argue that the extracted contextual information (*e.g.,* in the form of noun phrases with weights) can be easily incorporated into a right-click query for more relevant search results. Figure 2 illustrates a general framework to incorporate the query context. After user marks text (*i.e.,* query) in browser, she explicitly selects the contextual phrases and the search engine through right-click context menu, similar to the menus captured in Figure 1. The selection may also be made implicit if the top-ranked noun phrase is always picked up as context. The query and the selected contextual phrase(s) are then posted to the search engine. Nevertheless, the detailed implementation on utilizing the contextual information is out of the scope of this study.

2. RELATED WORK

In this section, we first discuss the differences and similarities between our contextual search and search personalization. We then survey the works on contextual advertising. Lastly, we distinguish our work from other studies on context-aware search.

Search Personalization. Contextual search is heavily researched in literature and is mainly studied from personalization perspective. A large portion of a recent comprehensive survey on contextual search is devoted to the study of personal interest from interaction, content, social, and geographical variables [6]. As clarified in our earlier discussion, we do not consider user search behavior and click/browsing history. Instead, we only consider the contextual information from the source document of a right-click query. In this setting, user profiling techniques (*e.g.,* by extracting keywords from browsing history) are relevant to our research. Matthijs and Radlinski learn users' long-term interests by summarizing the content of the Web pages in their browsing history. Six different summaries are evaluated in their study by extracting unigrams from full text, title, metadata description, metadata keywords, by extracting noun phrase, and by extracting important keywords (including unigrams and phrases) using linguistic and statistical information. Their study shows that all these summaries are helpful in improving personalization except unigrams from full text. Other than the content of the Web pages, users may also be profiled by their provided social tags to the Web pages [7]. However, in our setting, we do not assume that user provides tags for the Web page she is currently reading.

Contextual Advertising. Contextual advertising is to display advertisement best matching the estimated interests of users. In many settings, the user interests are estimated from the content displayed to user. To minimize latency and communication costs, summarizing the content of the Web page is often done prior to performing a match search on the advertisements. In [1], the role of page components in crafting short but informative page fragment that serve as an alternative description of the entire page is studied. Their experiments show that the combination consisting of page URL, referrer URL, title, meta-data, and headings serves as a good page summary. It uses only 6% of the page text and achieves 97% - 99% of the full-text-based relevance. In [2], Giuliano Armano *et al.* summarize Web page by retrieving relevant blocks of a Web page. They propose and evaluate five summarization techniques, namely Title and First Paragraph (TFP), Title and First Two Paragraphs (TF2P), Title, First and Last Paragraphs (TFLP), Most Titled-words and Keywords (MTK), and *N* most Frequent Keywords (NK). On the BankSearch Dataset consisting of 11,000 Web pages in 11 different categories as the evaluation data set, TFLP provides the best performance in terms of accuracy.

Context-Aware Search. The IntelliZap system reported in [4] enables a user to search the marked text in a document she views as query. The context of the query (*i.e.,* the marked text in the document) is derived from the surrounding text. Semantic keywords are then extracted from the context using a clustering-based approach which are then used to generate augmented queries. While the context is relatively fixed to be the about 50 words surrounding the marked query in [4], we evaluate different approaches of extracting context for a marked query in this paper. This also distinguish our study from [5] where different search strategies (*e.g.,* query rewriting, rank biasing, and iterative filtering meta-search) have been evaluated with explicitly provided query context. Our proposed context-aware search is categorized into the query rewriting strategy. Specifically, query rewriting augments a query using additional words derived from its context, hence can be readily supported by all search engines without any modification to the search engines. It is reported in [4] that "query rewriting performs surprisingly well". Note that, query rewriting is also considered as a form of query expansion [3].

3. CONTEXT EXTRACTION

Given a document d, from which a text string at position p is marked by a user as a right-click query q for search, our task is to identify a few words or phrases from d to serve as the context of q such that the search results matches q with respect to d. Note that, the same text string q may appear multiple times at different positions in the document.

We solve the problem by considering two perspectives: (i) from which text component of the document to extract contextual information; and (ii) which words or phrases to be extracted as contex-

[2] http://sg.news.yahoo.com/

Table 1: The 7 text components T1 – T7 for context extraction

Scheme	Text component
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

Table 2: The 5 feature extraction scheme F1–F5

Scheme	Context words	Weighting scheme
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

tual information to augment the query. For the former, based on the works surveyed in Section 2, we evaluate 7 text components for extracting context from title, full text, paragraphs containing the query, and others, shown in Table 1. For the latter, we evaluate 5 feature extraction options to extract the context word or phrases, shown in Table 2. Nouns and noun phrases are detected using off-the-shelf POS (part-of-speech) tagger. The words or nouns are ranked by 2 weighting schemes, namely, *Frequency-based weighting*, *Proximity-based weighting*. The noun phrases are weighted by *Phrase weighting*. We detail the three weighting schemes.

- Frequency-based weighting of a term t_i, which can be a noun or any other type of word, denoted by $f_w(t_i)$, is computed by using the $TF \cdot IDF$ weighting scheme commonly adopted in IR tasks. The term frequency is determined from the selected text component (*e.g.,* title) in the document.
- Proximity-based weighting scheme factors in the distance between the term and the query in addition to the frequency-based weighting. That is, 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. Specifically, let f_i be the term frequency of term t_i, $dist(t^j, q)$ be the proximity distance (*i.e.,* number of words) between the j-th occurrence of t_i and query q. The proximity-based weighting of term t_i is defined as:

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

  Note that a query string may appear multiple times in a document. The proximity distance is computed based on the nearest query string occurrence in the document (not necessarily the query string marked by the user).
- Let s be a phrase and $t_i \in s$ be a term contained in s. Phrase weighting considers two factors: (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 length of the phrase in number of terms. Specifically, the phrase weighting is defined as follows.

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

  The above weighting scheme considers the importance of the phrase as a whole text unit through $f_w(s)$, and its relevance to the document reflected by the averaged frequency of its contained terms. The second factor is designed based on the observation that many phrases, particularly named entities, are partially repeated in documents. For instance, a person's first name or last name may occur more frequently than her full name in a document.

Table 3: Precision of the 35 combinations. The best result for each text component extraction scheme (T1 - T7) is in boldface, for each feature extraction scheme (F1 - F5) underlined.

	T1	T2	T3	T4	T5	T6	T7	Avg
F1	0.206	0.308	0.262	0.244	0.275	0.269	0.150	0.245
F2	0.119	0.277	0.263	0.219	0.281	0.281	0.094	0.219
F3	0.175	0.374	0.271	0.200	0.300	0.300	0.156	0.253
F4	0.100	0.380	0.271	0.231	0.281	0.331	0.075	0.238
F5	**0.325**	**0.525**	**0.350**	**0.300**	**0.750**	**0.475**	**0.325**	**0.436**
Avg	0.185	0.373	0.283	0.239	0.377	0.331	0.160	–

4. EVALUATION

Dataset and Evaluation Setting. To evaluate the performance of different context extraction and weighting schemes, we conducted a user study using 20 news articles from Yahoo! News (see Table 4 for example articles and queries). These articles/queries in the dataset are 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. The IDF of terms are estimated using Reuters-21578 collection.

From each news article, we first mark the right-click query and then evaluate the top-8 words ranked by each combination of text component/feature extraction schemes (Tables 1 and 2). The top-8 words are selected for evaluation so as to easily compare with Google as the baseline method. Google usually recommends 8 related queries for a given query and these 8 queries are recommended without considering the context as defined in our setting. Note that in our experiments, only one-word queries are evaluated because they are usually vague and ambiguous as compared to multi-word queries. A multi-word query is usually more topic specific making the context less important in search.

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 marked query (*i.e.,* whether the word is helpful *in providing contextual information to the query* with respect to the article). Example relevant words for query "pub" in the article titled "Cameron left 8-year-old daughter in pub" include *drinks*, *friends*, and *mix*. However *flood management* is an example of relevant phrase for the same query in the article titled "PUB to spend S$750 million to improve drainage" where PUB stands for Public Utilities Board, Singapore's National Water Agency.

Because each method is expected to return a fixed number (*i.e.,* top-8) of keywords, we adopt **precision** to be our evaluation metric, which is the ratio of the relevant words. If any scheme returns fewer than 8 words, then precision is computed based on all returned words. For a fair comparison, the top-N phrases are selected such that $\sum_{s \in top-N} |s| \le 8$, *i.e.,* the total number of words in the selected phrases is not larger than 8. In our experiments, we used the Stanford POS Tagger[3] for extracting nouns and noun phrases. The noun phrases are extracted by regular expression: (Adjective | Noun)* (Noun Preposition)? (Adjective | Noun)* Noun.

Evaluation Results. The precisions of the 35 combinations (T1-T7 with F1-F5) are plotted in Figure 3. For easy comparison, the preci-

[3] http://nlp.stanford.edu/software/tagger.shtml

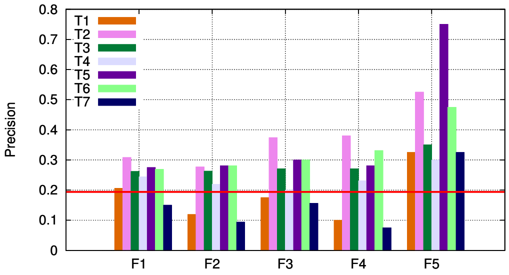


Figure 3: Precision of the 35 combinations against the baseline (the horizontal line in the plot).

sion of the baseline method which is 0.194 is plotted in a horizontal line in the figure. The precision values are also reported in Table 3, with the averaged value over T1-T7 and F1-F5 respectively. From the results, we make the following four observations.

- Noun phrases with phrase weighting (F5) is the best context feature extraction scheme compared to the other four feature extraction scheme F1-F4. In fact, for any given text component extraction scheme (T1-T7), the combination with F5 always leads to the best precision compared to the combination with any of the other four feature extraction schemes. F5 is also the only feature extraction scheme that achieves better precision than the baseline with any text component extraction scheme.
- Paragraphs containing the query words are the best text components for query context extraction compared to the other text components (*e.g.,* title or keywords). The best two combinations among the 35 in terms of precision are T5F5 and T2F5. T2 is the paragraph containing the query word and T5 refers to all paragraphs containing the query word. In fact, for a given feature selection scheme, T2 is the best text component extraction scheme for F1, F3, and F4 respectively; and T5 is the best for F2 and F5 respectively. This observation is consistent with the settings in many applications: the query's surrounding words are used as the query's context. Nevertheless, T5F5 significantly outperforms T2F5 in terms precision based on paired *t*-test. In other words, all paragraphs containing the query word are useful in providing contextual information for the query.
- Proximity-based weighting scheme adversely affects the precision compared with frequency-based weighting scheme. Observe that on average, the 7 text component schemes using F1 (words with frequency-based weighting) have better precision than that with F2 (words with proximity-based weighting); precision with F3 (nouns with frequency-based weighting) is better than that with F4 (nouns with proximity-based weighting). The differences in the precision values, however, are not statistically significant based on paired *t*-test.
- Between nouns and any words, using the same weighting schemes, nouns define better contextual information than any words. With frequency-based weighting, combinations using nouns F3 enjoy better averaged precision over the combinations using any words F1 (0.253 vs 0.245); similarly, with proximity-based weighting, F4 using nouns is better than F2 using any words (0.238 vs 0.219);

To summarize, noun phrases extracted from the paragraphs containing the query word provide better contextual information for right-click query. Other than noun phrases, nouns provide better context than any words and proximity-based weighting schemes adversely affect the quality of context.

Table 4: Example news articles and selected queries

Title of the randomly selected news articles	Query
Apple launches iTunes Store in 12 new Asia markets	apple
Apple Peel Compound May Help Ward Off Obesity	apple
PUB to spend S$750 million to improve drainage	pub
Cameron left 8-year-old daughter in pub	pub
Recent cloud discovery hints at planet formation in Galactic Center	cloud
Cloudonomics: The Business Value of Cloud Computing	cloud
SIA introduces 'quirky' budget carrier, Scoot	scoot
For heart health, fish oil pills not the answer: study	omega

5. CONCLUSION AND FUTURE WORK

In this paper, we focus on right-click query that is submitted to a search engine by marking a text string in a Web page. To extract the contextual information for the right-click query, we evaluate 35 combinations involving 7 text component extraction schemes and 5 feature extraction schemes. Our evaluation shows that paragraphs containing the marked query are the best text component for contextual information extraction. Noun phrases is the best option to extract query context followed by nouns and then any words. However, our experiment also shows that proximity-based weighting scheme adversely affects context extraction. We argue that contextual information can be relatively easily integrated into right-click queries through Web browsers. The integration can be supported natively by the browser or through third-party extensions.

Our evaluation results are based on the relevance judgement of the extracted contextual information with respect to the query and the source document, and not the actual search results from any search engines because of two reasons. First, search results heavily depend on the underline search engines used in evaluation. Second, contextual information augmenting a query only provides a description of the query. Therefore a search result may not necessarily contain these contextual keywords to be a good match. In other words, the search query and the contextual information are not equally important in the search, which is not naturally supported by general search engines through Web interface. It is therefore interesting to study the impact of possible query augmentation schemes for best search results, which is part of our future work.

6. REFERENCES

[1] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel. Web page summarization for just-in-time contextual advertising. *ACM Trans. Intell. Syst. Technol.*, 3(1):14:1–14:32, Oct. 2011.

[2] G. Armano, A. Giuliani, and E. Vargiu. Experimenting text summarization techniques for contextual advertising. In *Proc. Italian Information Retrieval (IIR) Workshop*, 2011.

[3] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, 2012.

[4] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, 2002.

[5] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar. Searching with context. In *WWW*, pages 477–486. ACM, 2006.

[6] M. Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

[7] D. Vallet, I. Cantador, and J. M. Jose. Personalizing web search with folksonomy-based user and document profiles. In *ECIR*, pages 420–431. Springer-Verlag, 2010.