

Extracting Visual Snippets for Query Suggestion in Collaborative Web Search

Hannarin Kruajirayu, Teerapong Leelanupab

Knowledge Management and Knowledge Engineering Laboratory
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang (KMITL)
Bangkok, Thailand. 10520

hannarin_k@hotmail.com, teerapong@it.kmitl.ac.th

Abstract: This paper proposes an approach to generate query suggestions by employing information from user-created visual snippets. In order to generate query suggestions, we apply the optical character recognition (OCR) technique to extract a set of words presented in the visual snippet. The natural language processing (NLP) is used to identify the words that might be relevant to users' information need. The goal of this paper is to compare the accuracy of terms extracted by our approach with that by a Lemur search toolkit. In our experiment, we use 1,000 English web pages from a dataset in category B of ClueWeb09. The visual snippets are created from the web pages. Then, their important terms are extracted by our approach. Outcomes of extracting terms for query suggestions are evaluated by using Precision, Recall, and F-measure. The experimental results indicate that our approach effectively identifies and generates useful terms for query suggestion.

Keywords: Query Suggestion, Visual Snippet, ClueWeb09, Lemur Search Toolkit.

1 INTRODUCTION

The visual summarization of web pages has become a useful feature for searching information from the web. Typically, to achieve search and re-finding tasks, users might rely on common features representing previously visited web pages. These web pages are often represented in the forms of page titles in a user's history, a bookmark of URLs, and thumbnails of some part of web pages. These different representations have their own advantages for each task. For example, in different contexts users may require to visit the same web page with different information needs. Therefore, users cannot recognize the thumbnails that are previously captured from some parts of visited web pages. As a result, one form of the common page representations may not be suitable for other tasks. However, the visual snippet is proposed and designed to bring together benefits of typical representations. It combines both textual snippets and thumbnails for a better recognition.

We presented a new feature, "Snapboard", in our collaborative web search system [1], called "CoZpace." Figure 1 demonstrates the Snapboard that is a shared space in the form of a display board that allows a user to share "Visual Snippets" among collaborators. The visual snippet is a snapshot of focused and salient information, captured by the user or collaborators. It acts as a concise and quick human-generated summary, which is potentially relevant to the users' information needs. These users are those who

have the same interest and collaboratively search for particular information. The visual snippet in CoZpace is not only useful for emphasizing important information for team collaborations, but also valuable for search and revisitation tasks. By means of visual snippets, users can easily navigate to the same web page having previously viewed, or read information on them and pose new queries based on given information. To support the finding of new information, we propose the use of visual representations of web pages that call out relevant texts for generating query suggestions. This can be regarded as implicit collaboration, since one user unknowingly employs information generated from other users' efforts in an automatic way.

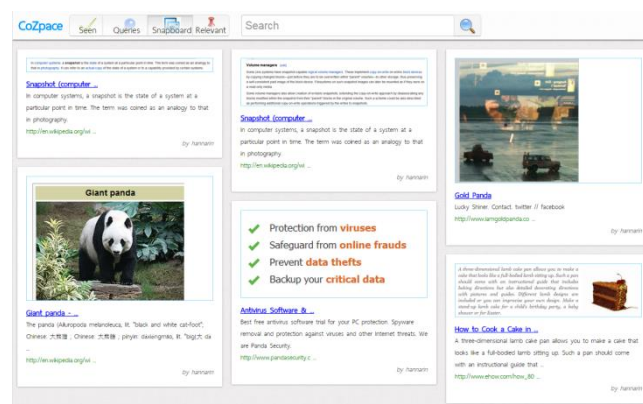


Fig. 1. Example Visual Snippets of CoZpace.

The contributions of this paper are structured as follows. We first describe a brief overview of related works and an

approach to generate query suggestions in the next section. Section 4 explains experiment design to evaluate the performance of our approach, followed by the results discussion. We then describe conclusion of our evaluation in section 6.

2 RELATED WORK

In 2009, Teevan et al. [2] presented visual snippet that is a summarization of web pages for search and re-finding tasks. In the experiment, they compared the performance of visual and textual snippet in three main types of search tasks: homepage finding, shopping and medical tasks. Their outcomes demonstrated the visual and textual snippet not much of a difference in speed for searching new web pages, which have never been seen by the users, and the visual snippet is faster than textual snippet for re-finding the previously seen web pages. In 2010, Jiao et al. [3] continued to explore visual snippet of web pages. They described about generating the internal image for visual snippet, which is a representative of the content in the web pages. It is difficult work to create an internal image when the internal image of a web page is unavailable or the web page has a complex layout and rich content. They proposed a new strategy, which retrieves the representative image from the external entire Internet to summarize web page, called external images. Moreover, their study investigated not only how different internal and external image summarizations perform in search and revisitation, but also how different internal and external image summarizations perform for different kinds of web pages. Their results demonstrated that the internal images could provide a reliable summarization on web pages with dominant images and web pages with simple structure and the external images were a useful information to complement the internal images and very useful in helping users understanding new web pages. However, in 2010, Aula et al. [4] compared the efficacy and effectiveness of visual and textual information summaries for web page previews. Their results suggested that the visual summarizations perform worse than the textual summarization. Moreover, they found that the textual summarizations tended to make the users systematically overestimate the helpfulness of web pages. The best performance was obtained by combining textual and visual information summarizations. The combination by placing title and URL below the visual summarization makes users focus on both visual and textual information.

Optical Character Recognition (OCR) [5] handles a problem of recognizing optically processed characters.

OCR is the conversion of images of typewritten or printed text into machine-encoded/computer-readable text. The performance of OCR is directly dependent upon the quality of the input image. Recently, OCR tools became commercially available in the markets, but only few of them are open source. The Tesseract [6] is one of the popular open sources OCR software [7]. It was developed by HP Labs Bristol in between 1984 to 1994. The Tesseract was modified with greater accuracy and concentrated more on improving rejection efficiency than on base-level accuracy.

Most existing search engines are able to automatically suggest a list of textual query terms based on users' current query input called Textual Query Suggestion. In 2009, Zha et al. [8] proposed a new query suggestion scheme named Visual Query Suggestion (VQS) which was dedicated to image search. They demonstrated that VQS was able to more precisely and more quickly help users specify and deliver their search intents. Denney et al. [9] (2009) introduced a concept of related search suggestion for searching video. When a user puts a query for video searches, their system not only displays the video search results but also shows related video search of the query. In 2012, Bian et al. [10] also studied Visual Query Suggestion. They proposed a new query suggestion scheme named Visual Query Attributes Suggestion (VQAS) for image search. VQAS suggests a list of informative attributes based on users' query image. These attributes reflect the visual properties and key components of the query, which is able to help users specify and deliver their search intention in a more precise way.

3 APPROACH

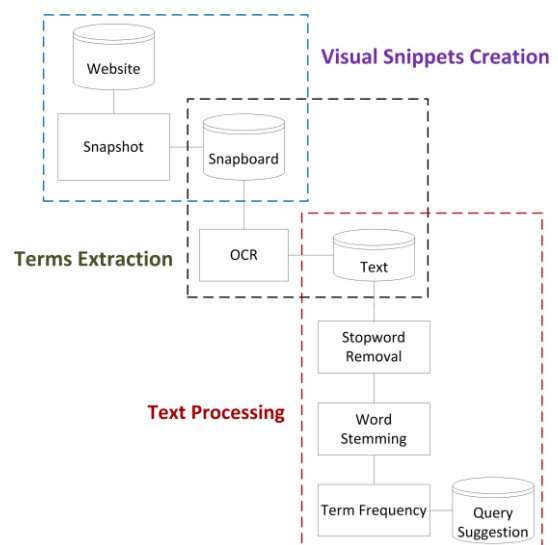


Fig. 2. The framework of our approach.

As shown in Figure 2, our approach is composed of three main steps: i) Visual snippets creation, ii) Terms extraction, and iii) Text processing.

3.1 Visual Snippets Creation

The first stage of visual snippets creation is to retrieve the English web pages. Then, we generate quality visual snippets in PNG image format by using an open source Python tool called Webkit2png [11].

3.2 Terms Extraction

In order to get more exploits on visual snippet, we explore important terms in the visual snippet by employing the OCR tool called Tesseract to convert text image in visual snippet into computer-readable text. However, it is usually difficult to extract text from visual snippet because of the different size, different style, and complex background of visual snippet. Therefore, the effectiveness of OCR depends on image pre-processing. We adjust the quality of visual snippet by increasing image contrast, increasing image size, and changing image color to black & white, as an example in Figure 3.

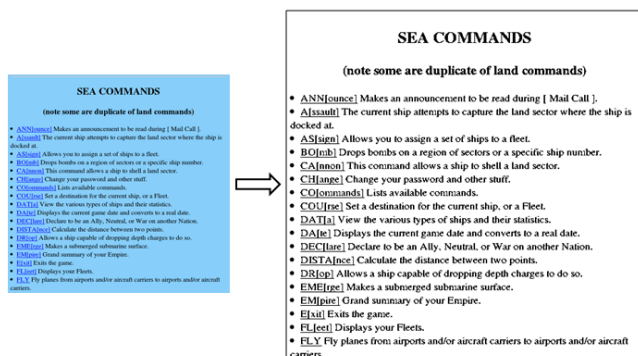


Fig. 3. Example adapting visual snippet.

3.3 Text Processing

First of all, frequently occurring and insignificant words, called stopwords, e.g., a, an, and the, are removed from the extracted text. Then, we remove inflectional and derivational suffixes in order to combine word variants into the same stem or root, called words stemming, e.g., processes, processed, and process can be shorten to process. Finally, we also use the term frequency (TF) as the number of times that a term occurs in visual snippet. If a term has high TF, the term might be relevant to users' information need.

4 EXPERIMENT DESIGN

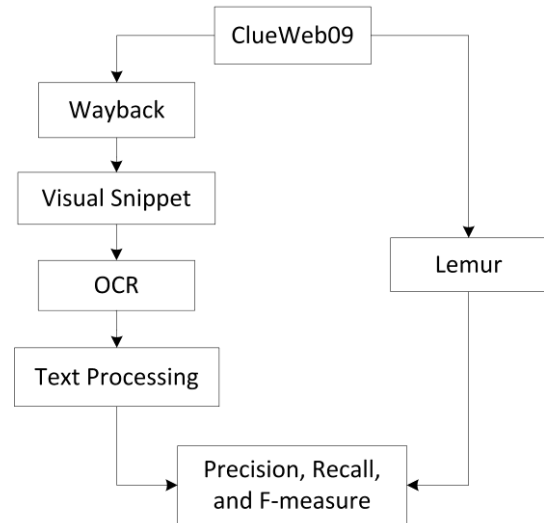


Fig. 4. The framework of experiment design.

We conduct an experiment to compare the performance of our approach with that of Lemur search toolkit. In our experiment, we only focus on the English Web pages from ClueWeb09 dataset. We use Wayback tool to access ClueWeb09 and Webkit2png to create a visual snippet. We then employ OCR technique to recognize terms in the visual snippet followed by text processing. In order to assess effectiveness of our approach, we use the most prevailing measures that are precision, recall, and F-measure, as shown in Figure 4.

4.1 ClueWeb09 Dataset

The ClueWeb09 [12] dataset was created to support research on information retrieval and related human language technologies. It consists of about 1 billion web pages in ten languages that were collected in January and February 2009. The dataset is used by several tracks of Text REtrieval Conference (TREC) [13] which is a series of workshops focusing on a list of different information retrieval (IR) research areas. As we found out that all dataset in Category B contains only English web pages, We randomly select one dataset in categories B of ClueWeb09 consisting of about 30,000 English web pages. As English web pages in our experiment are stored as the Web ARchive (WARC) file format [14] in ClueWeb09 dataset, the first stage of visual snippets creation is to retrieve the English web pages by using Wayback tool. A simple logical overview of the Wayback Web application in Figure 5 shows that Wayback tool is used to index the WARC file and keep the index in the CDX file [15].

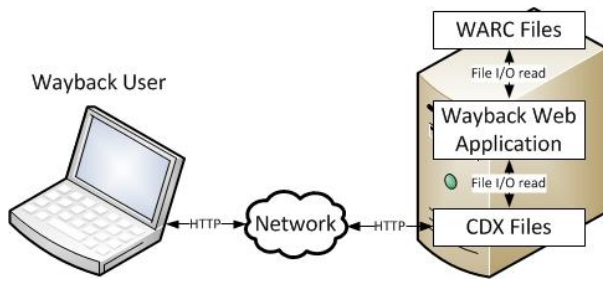


Fig. 5. A logical overview of a Wayback Web application.

To evaluate the overall performance, we use randomly selected 1,000 web pages. However, we found that the number of terms in the majority of web pages was varied from about 100 to 400 terms. Consequently, in order to check whether the number of terms impacts the performance, we also check performance separately from 4 different groups of web pages that are classified by the number of terms: i) 50-200 terms, ii) 201-400 terms, iii) 401-600 terms, and iv) more than 600 terms. Each group contains randomly selected 250 web pages.

4.2 Lemur Search Toolkit

Lemur is a toolkit [16] designed to facilitate research in language modeling and information retrieval. The Lemur toolkit [17] is effective to support the construction of basic text retrieval systems, it comes with all the source code and make files necessary to build the libraries for indexing and retrieval. Accordingly, we index the English web pages by using the Lemur toolkit and take its result as a baseline to compare with the number of terms extracted by our approach. As we assume that the result from the Lemur toolkit is a correct result, precision, recall and F-measure are used as measures for the comparison. Table 1 shows the number of terms and the number unique terms from the indexing by the Lemur toolkit.

Table 1. Statistics of experimental collections.

Name	Docs	Terms	Uniq. Terms
ClueWeb09	1,000	1,307,646	439,227

4.3 Precision, Recall and F-measure

Precision, recall, and F-measure are the basic measures used in evaluating search strategies. We apply precision, recall, and F-measure to evaluate relevant terms extracted from visual snippet. Figure 6 illustrates that we divide the extracted terms to 3 sets for the evaluation as follow:

T1: The set of terms that are extracted by our approach and also indexed by Lemur toolkit.

T2: The set of terms that are indexed by Lemur search toolkit but not by our approach.

T3: The set of terms that are extracted by only our approach.

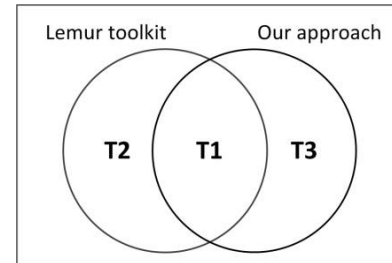


Fig. 6. The set of T1, T2, and T3 for evaluation.

Precision is the ratio of the number of terms that are extracted by our approach and also indexed by Lemur toolkit (T1) to the total number of terms that are extracted by our approach (T1+T3).

$$Precision = \frac{N_{T1}}{N_{T1} + N_{T3}} \quad (1)$$

Recall is the ratio of the number of terms that are extracted by our approach and also indexed by Lemur toolkit (T1) to the total number of terms that are indexed by Lemur toolkit (T1+T2).

$$Recall = \frac{N_{T1}}{N_{T1} + N_{T2}} \quad (2)$$

F-measure is useful in giving a single numeric measure of overall performance combining both precision and recall measure.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

5 EXPERIMENTAL RESULTS AND DISCUSSION

The experimental results in Table 2 shows that the terms extracted by our approach match with 81.44% of the terms in baseline (T1). It means that the performance of our approach to retrieve terms is very high similar to Lemur search toolkit. Even though, there are a large number of terms in T3 which is 28.63% of baseline, we cannot judge that they are incorrect. Because, Lemur may not be comprehensive to retrieve all the terms from ClueWeb09. Still, the terms extracted by our approach may possibly be useful for generating query suggestion. Moreover, it can be observed that our approach can achieve the precision, recall

and F-measure rate up to 73.9%, 81.4%, and 77.5%, respectively.

Table 2. The result of 1,000 Web pages

T1	357721 terms (81.44%)
T2	81,506 terms (18.56%)
T3	125,765 terms (28.63%)
Precision	0.739
Recall	0.814
F-measure	0.775

The precision, recall and F-measure in Table 3 shows that our approach performs well in every range of the number of terms. Moreover, there is no signal of impact from the different of the number of terms to the performance of our approach because the precision, recall and F-measure in the four groups are only a bit different.

Table 3. Precision, recall and f-measure of the web pages was classified to four group

Group of terms	Precision	Recall	F-measure
50-200 terms/page	0.709	0.836	0.767
201-400 terms/page	0.719	0.837	0.774
401-600 terms/page	0.717	0.828	0.768
>600 terms/page	0.767	0.796	0.781

6 CONCLUSIONS

In this work, we proposed an approach to generate query suggestion by exploiting information in visual snippet. We conducted an experiment to evaluate our approach by using ClueWeb09 dataset, Wayback, Webkit2png, Lemur search toolkit, precision, recall, and F-measure. The experiment results demonstrate that our approach has a high precision, recall, and F-measure. In other words, our approach is capable of generating satisfactory and useful terms for query suggestion, which is relevant to users' search intention.

REFERENCES

[1] H. Kruajirayu, A. Tangsomboon, and T. Leelanupab. Cozpace: A proposal for collaborative web search for sharing search records and interactions. In *The 2014 Third ICT International Student Project Conference, ICT-ISPC '14*, 2014.

[2] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. Andr e, and C. Hu. Visual snippets: Summarizing web pages for search and revisitation. In *Proc. of the*

SIGCHI Conference on Human Factors in Computing Systems, CHI '09, 2009.

[3] B. Jiao, L. Yang, J. Xu, and F. Wu. Visual summarization of web pages. *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 499–506, 2010.

[4] A. Aula, R. M. Khan, Z. Guan, P. Fontes, and P. Hong. A comparison of visual and textual page previews in judging the helpfulness of web pages. In *Proc. of the 19th International Conference, WWW '10*, 2010.

[5] S. Mori, H. Nishida, and H. Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

[6] GOOGLE. Tesseract-ocr. <http://code.google.com/p/tesseract-ocr/>, 2012.

[7] R. Smith. An overview of the tesseract ocr engine. In *Proc. of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07*, 2007.

[8] Z. J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proc. of the 17th ACM International Conference on Multimedia, MM'09*, pages 15–24, 2009.

[9] J. Denney, J. Hamilton, T. C. Hoad, K. R. McDonald, and R. J. Qian. Visual query suggestions, June 27 2008. US Patent App. 12/163,655.

[10] J. Bian, Z. J. Zha, H. Zhang, Q. Tian, and T.-S. Chua. Visual query attributes suggestion. In *Proc. of the 20th ACM International Conference on Multimedia, MM '12*, 2012.

[11] P. Hammond. Webkit2png. www.paulhammond.org/webkit2png/, 2013.

[12] J. Callan, M. Hoy, C. Yoo, and L. Zhao. Clueweb09 data set, 2009.

[13] D. Harman. Overview of the first trec conference. In *Proc. of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 1993.

[14] I. ISO. 28500: 2009 information and documentation-warc file format. *International Organization for Standardization*, 2009.

[15] B. Tofel. Wayback for accessing web archives. In *Proc. of the 7th International Web Archiving Workshop*, 2007.

[16] P. Ogilvie and J. P. Callan. Experiments using the lemur toolkit. In *TREC*, volume 10, pages 103–108, 2001.

[17] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. N. Truong, P. Ogilvie, et al. The lemur toolkit for language modeling and information retrieval. 2003.