Improving Relevance Judgment of Web Search Results with Image Excerpts

Zhiwei Li
Microsoft Research Asia
Sigma Center, Beijing, China
zli@microsoft.com

Shuming Shi
Microsoft Research Asia
Sigma Center, Beijing, China
shumings@microsoft.com

Lei Zhang
Microsoft Research Asia
Sigma Center, Beijing, China
leizhang@microsoft.com

ABSTRACT
Current web search engines return result pages containing mostly text summary even though the matched web pages may contain informative pictures. A text excerpt (i.e. snippet) is generated by selecting keywords around the matched query terms for each returned page to provide context for user's relevance judgment. However, in many scenarios, we found that the pictures in web pages, if selected properly, could be added into search result pages and provide richer contextual description because a picture is worth a thousand words. Such new summary is named as image excerpts. By well designed user study, we demonstrate image excerpts can help users make much quicker relevance judgment of search results for a wide range of query types. To implement this idea, we propose a practicable approach to automatically generate image excerpts in the result pages by considering the dominance of each picture in each web page and the relevance of the picture to the query. We also outline an efficient way to incorporate image excerpts in web search engines. Web search engines can adopt our approach by slightly modifying their index and inserting a few low cost operations in their workflow. Our experiments on a large web dataset indicate the performance of the proposed approach is very promising.

Categories and Subject Descriptors
H.3.3[**Information Systems**]: Information Search and Retrieval, I.2.6 [**Computing Methodologies**]: Artificial Intelligence

General Terms
Design, Algorithms

Keywords
Image Excerpts, Dominant Image, Web Search, Usability, User Interface

1. INTRODUCTION
The web search engines have been indispensable tools to find information from the Internet. They answer the user's query by a ranked list. Each item of the list is a web page, but only text summary of the page is displayed in result pages, which contains only page title and some keywords around the query terms. The purpose of providing a text summary for each result page is to enable the user to quickly judge whether it is what he or she needs. Providing such a simple interface has been philosophy of many search engines because it is quick but informative.

However, such a user interface misses very valuable information in web pages, say images. Usually, a web page may contain some informative images, and these images are indispensable components to present the ideas of the page. For example, we cannot imagine a news site will be if all news images are removed. Why we place some images in web pages when we make them? The reason is very straightforward: we must think images are useful to present our ideas. Thus, intuitively, showing some informative images in search results may be helpful for users to quickly understand what the page is taking about, as well as make better relevance judgment. Figure 1 illustrates the idea of showing some important images in search results of web search engines. Those images displayed in search results are extracted from corresponding web pages. It is obvious that the search results with image are more vivid and informative than traditional search results, in which only text summaries are provided. We define such search results are *image excerpts*, and these informative images are *dominant images*.

Figure 1: Search results of query "the White House". The upper figure is text summary, the lower figure is image excerpts.

From the aspect of designers, we often think a web page consists of two indispensable components, say text contents and images (or other multimedia contents). The two components should be regarded as elements of an "atom". However, current search companies build web search engine to search pages and build image search engine to search images. The two components are not utilized together in search engines to exert their combinational values. Actually, some web search engines have realized this problem, and began to use images to improve their usability. Search engines, like Live.com and Google, will insert a few images got from their image search engines on the top of the search result page for some queries (e.g. the query "David Beckham"). Obviously, such interface is far from enough to embody the value of web images. Such a user interface only can improve the overall usability of web search engines, but cannot help users to make quicker relevance judgment.

WWW 2008, April 21-25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

In this paper we do not deal with problems on how to generate better text snippets [20], while we only focus on extracting dominant images from web pages to generate image excerpts along with existing text snippets. However, extracting dominant images is non-trivial, there are two difficulties:

1. For most web pages, there are lots of images embedding in them, but not all of these images are dominant images (e.g. advertisement images and decoration pictures).

2. A web page may have many dominant images, but not all these images are relevant with the user's query. For example, the web page illustrated in Figure 2 has three dominant images, but the three images represent different digital cameras, respectively.

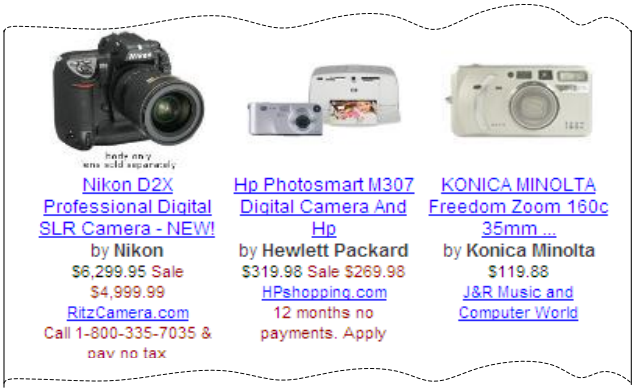

Figure 2: A web page may contain lots of images, and each image may have different meanings.

To address the two problems, we propose an approach consisting of two consecutive steps. In the first step, we train a classifier to classify images to dominant images vs. non-dominant images. But different from a common classifier, we optimize our classifier to assign a dominant score to each dominant image. This score will be used in the next step to select the best images. The first step can be performed off-line. In the second step, we combine the user's query and the dominant score got in the first step to select the most important and relevant image to generate image excerpts. This step has to be performed on-line, but the cost of this step is very low if we have indexed images according to their annotation text (i.e. file name and surrounding text).

This paper is organized as follows. In section 2, we review previous work. The framework and details of the proposed approach are given in section 3, 4 and 5. Experiments to evaluate this approach are reported in section 6. The user study is given in section 7. At last, we conclude this paper and point our future work in section 8.

2. PREVIOUS WORK

Previous studies have used different methods to summarize web documents. Some works are focused on extracting most representative sentences or phases [15, 16, 20, 28]. Ocelot [15] is a system for summarizing web pages using probabilistic models to generate the gist of a web page. Buyukkokten et al. [3] introduce five methods for summarizing parts of Web pages on handheld devices. Delort el al. [20] exploit the effect of context in web page summarization. Shen et al. [28] propose a new web summarization algorithm, which extracts the main topic of a web page through a page-layout analysis to enhance the accuracy of classification. In the web search tasks, the summarization needs consideration of search queries. Current web search engines like Google or Live most set the summaries as the texts in which search terms appear in the documents. However, presenting text summaries to users has proven to be less effective than graphical summaries in some search tasks [21, 13].

A number of studies have involved the design of graphical interfaces for presenting documents. Ayers and Stasko's thumbnails [14] consist of a reduced view of the left upper corner of a document, which is assumed to be most representative part in the document. Dziadosz and Chandraseka [21] claimed that graphical thumbnails can greatly improve the efficiency by which users to find out relevant documents from list of documents in search results. Kopetzky and Mühlhuser [24] describe a system in which links from a web page are represented by corresponding thumbnail of the document that appears temporarily when users move a mouse over the hyperlink. If the user has previously seen the page, the visual representation may aid in recognizing or classifying it [19, 23], which is usually not true in web search tasks where users are unlikely to have seen many of the documents before.

As demonstrated in previous studies [21, 13], although thumbnails are perceived as images, people usually need to read textual information presented in thumbnail previews, which causes additional time cost and reading difficulty due to poor accessibility of textual information on thumbnails. Thus, Woodruff et al [13] designed a new kind of textually-enhanced thumbnail that enforces readability of certain parts of the document within thumbnail and displays highlighted terms transparently overlaid on the reduced document. However, experiments in this study also showed that most of users were highly relying on the highlighted keywords for identifying document relevance, which again, to some extension, falls into the inefficiency suffered from text summaries.

Using a thumbnail of the "whole" page as an indication of layout of the page and all other methods in previous work leads us to ask: whether there are other more informative methods for summarizing web documents. Previous study [29] is most similar to our work, which produces web page "caricatures", containing selected features of a page often rendered in an abstract form: title, representative image, number of images, abstract, etc. In this work, the representative images in a document are selected as that can best convey the content of that document. Thus, a web document may contain multiple representative images with different contextual indication. However, in the web search tasks, the extraction of representative images needs to comprehensively consider consistence of an image with users' search queries.

We believe such indicative images are more suitable for indicating document content than thumbnails. However, as page thumbnail can give hints about the style as well as the layout of the page, one may argue that it can also present the included images to the users. However, this is untenable due to the poor accessibility of images on the thumbnails usually rendered as limited size at search results. Moreover, the desired image may not be contained on the reduced version of thumbnails [14].

Google news search [22] makes good use of images in its search results. The presence of images on the news search results is helpful to let users identify whether the news are relevant to the information need. However, it can only provide the heading or logo images on the site or newspapers of a news result, consequently resulting in an inconsistence of the displayed images with the news content. Moreover, we found that images are also available and useful in general web search tasks.

3. FRAMEWORK OF DOMINANT IMAGE EXTRACTION ALGORITHM

As a value-added component to web search engines, the proposed approach adopts the same workflow as web search engines. The workflow of a web search engine can be divided into two phases: off-line indexing and on-line searching. In the first phase, it crawl web pages and build index for them. In the second phase, it matches query terms in its index, and ranks pages as some criterion [7, 27]. Accordingly, our dominant image extraction algorithm also consists of two consecutive steps: an off-line dominant image detection step and an on-line dominant image selection step. Figure 3 illustrates the workflow of our approach.

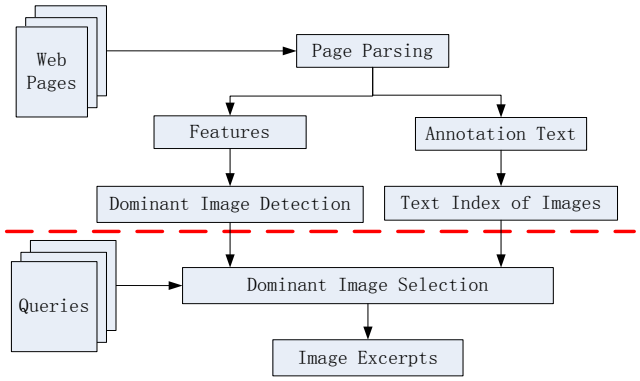

Figure 3: The workflow of the proposed dominant image extraction approach. The upper part is the off-line dominant image detection and index process, and the lower part is the on-line dominant image selection process.

Dominant Image Detection At first, we extract features both from web pages and images, and then use a classifier to determine which images are dominant images for their hosting web pages. At the same time, we compute a dominant score *d(p)* for each dominant image *p*. This real valued score reflects how important the image *p* is for its hosting web page. In feature extraction module, we also extract text annotations of dominant images from their hosting pages, and these annotations are indexed to enable quick relevance measure in the next step.

Dominant Image Selection After receiving the user's query *q*, we first retrieve the most relevant web pages (this is the work of web search engines), and then select the most relevant dominant image for each page in the search results. Since some pages have more than one dominant image, which may have different meanings, we use the annotation text automatically extracted in the first step to compute a relevance score *r(p,q)* for each dominant image *p*. Actually, this is exactly the work what web image search engines are doing. This relevance score can be used to select the most relevant dominant images.

Because in our interface design, for each item (a web page) in search results, only its "best" dominant image will be used to generate its image excerpt, we combine the two evidences together to determine which image is the best one. The final score *s(p,q)* of an image *p* is computed by

$$s(p,q) = \beta \cdot d(p) + (1-\beta) \cdot r(p,q) \qquad (1)$$

where $\beta \in [0,1]$ is a coefficient determined experimentally, and q is the query. This simple combination enables a lot of fast approximate ranking algorithms. Moreover, because our approach can be fit in workflow of web search engines, and can adopt the same distributed computing architecture, its scalability is not a problem.

4. DOMINANT IMAGE DETECTION

Usually, even in one web page, there are lots of images, but most of them are advertisement images, logo and decoration images. We stat the image numbers of pages used in our experiments, there are 13.47 images on average. Thus, we need a classifier to discriminate dominant images from non-dominant images. It is worth noticing that a page may have no dominant images, even if there are lots of images in it.

4.1 Features for Dominant Image Detection

For practical considerations, we carefully select some low-cost features to train the classifier. These features can be categorized into three groups according to their properties.

4.1.1 Image Level Features

This group of features is extracted by analyzing image content. Different from traditional usages of visual features, we do utilize some middle-level features instead of utilizing low-level features directly. Usually, dominant images tend to have better qualities than non-dominant images. Thus this group of features focuses on measuring the qualities of images.

Image Size is computed by $width \times height$, where $width$ and $height$ denote the width and height of an image, respectively. Dominant images lean to be bigger than non-dominant images.

Aspect Ratio of an image is simply computed by

$$\frac{\min(\text{width}, \text{height})}{\max(\text{width}, \text{height})}$$

Dominant images lean to be with bigger aspect ratios than non-dominant images.

Image Quality features consist of three kinds of image quality metrics: sharpness, contrast and colorfulness. We adapt the image quality measurement methods proposed in [9, 11]. Sharpness of an image is assessed by computing the ratio of the number of "clear" edges to the number of all edges. Contrast is defined as the ratio of the brightness of foreground to the brightness of background. Colorfulness of an image is defined as how many colors in this image, but we quantized it into 10 levels.

Image Categorizations consists of two kinds of image taxonomies, named as *photo* vs. *graphics* [5] and *with human faces* vs. *without human face* [12]. These features are Boolean valued. Dominant images lean to be photo and contain human faces in them.

Image Format feature reflects whether an image is an animation or not. Dominant images lean to be static images, while advertisement images, logos and banners are often animation images.

4.1.2 Page Level Features

Dominant images are defined as the most important and informative images of their hosting web pages. They are often placed in the attention attractive areas of web pages. Thus, the layout of web page is an important evidence to determine which images are dominant images. According to these observations, we extract of a group of features which can reflect the importance of an image in a web page.

Position consists of the *x* and *y* coordinates of an image in its hosting page. Dominant images lean to locate at the center or at the top of a web page.