Semantic Knowledge Extraction and Annotation for Web Images

Zhigang Hua, Xiang-Jun Wang, Qingshan Liu, Hanqing Lu
Institute of Automation, Chinese Academy of Sciences
No. 95 East Road, Zhong Guan-Cun
Beijing 100080, P.R. China
(+86-10) 6254-2971
{zghua, xwang, qsliu, luhq}@nlpr.ia.ac.cn

ABSTRACT

Nowadays, images have become widely available on the World Wide Web (WWW). It's essential to develop effective ways for managing and retrieving such abundant images. Advantageously, compared to the traditional images where very little information is provided, the web images contain plentiful context data. This paper introduces a system that can automatically acquire semantic knowledge for web image annotation. By using a page layout analysis method that can precisely assign context to web images, we developed efficient algorithms to extract semantic knowledge for web images, such as description, people, temporal and geographic information. To validate the practicality and efficiency of this system, we applied it to about 6,500 images crawled from Web. Experiments demonstrated that our approach could achieve satisfactory results.

Categories and Subject Descriptors

H.3.m [Information Storage and Retrieval]: Miscellaneous

General Terms: Experimentation

Keywords: Web image mining, web image retrieval, image context, semantic image annotation

1. INTRODUCTION

Nowadays, the digital images have become widely available on World Wide Web (WWW), which has brought about great challenges for organizing and searching a large volume of available images. The traditional image retrieval techniques, such as those content-based image retrieval (CBIR) systems, are usually not scalable for the use in WWW images to handle the vast image amount. Different from the traditional images where very little information is provided, there exist a lot of additional contextual information on the Web like surrounding text and links.

The search engines such as Google's Image Search makes good use of surrounding keywords of web images when available. But according to the study [12], searching images through keywords may be frustrating, because keywords have linguistic and person-dependent components that can make them difficult to use. [12] further pointed out that the primary semantics of an image includes the time when it was taken, the people whom it talks about or it is owned by, and the spot where it was shot, etc. Past studies [9][10] further demonstrated that users often associate their digital photos with event, location, subject, and time. Among them, subject is often defined by combinations of who, what, when, and where [12]. Obviously, in order to refine web image retrieval and understanding, it's quite important and meaningful to mine and extract these kinds of knowledge because they say too much about the semantic content of digital images.

Recently, image annotation techniques have been widely used to annotate semantic information for digital images. Toyama [12] has presented several methods for acquiring textual annotations for digital images. However, such methods are mainly relying on manual or semi-manual entries, which are not always preferred by users. The manual annotation tends to be unscalable and low-proficient, especially for the huge amount of images on the Web.

Advantageously, the thing becomes favorable for the images on the Web, in which plentiful contextual data are ubiquitously available. Mining semantic knowledge from the Web has become a hot topic in recent years. Ding [5] proposes effective algorithms to compute geographic scopes for web resources. TimesMine [11] is a system that can automatically generate timelines from the date-tagged free text corpora. Newsjunkie [6] is a system that is designed to mine named entities to personalize news for users by identifying the novelty of stories. Our previous work [7] can extract and annotate geographical location for the web images. However, none of these studies have involved the extraction of different semantic knowledge for web images.

This leads us to design a system that can automatically extract semantic knowledge for the annotation of web image. By using an efficient DOM-based page layout analysis method, web pages are analyzed to precisely assign each image with its corresponding context. We have also developed effective methods to analyze the contextual information for the extraction of semantic knowledge. Preliminary experiment results demonstrated the effectiveness of this approach. We believe that this solution is effective for the scalable use in the large WWW image collections.

The rest of this paper is organized as follows. Section 2 describes the design of our system, including web image context acquisition and semantic knowledge extraction. In section 4, we provide the experimental results of our system applied to three various image categories on the Web. Section 5 concludes this paper.

2. SYSTEM DESIGN

We have developed a novel system to automatically acquire and annotate semantic knowledge for WWW images. The system framework of our approach is presented in Figure 1. As shown in the

MM'05, November 6–11, 2005, Singapore.

figure, the system mainly comprises two components, namely image context extraction and semantic knowledge acquisition.

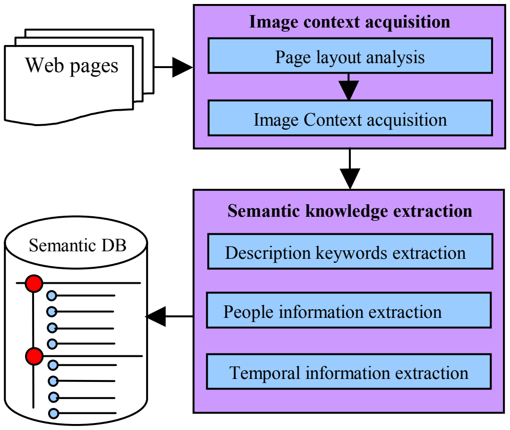

Web pages
Image context acquisition
Page layout analysis
Image Context acquisition
Semantic knowledge extraction
Description keywords extraction
People information extraction
Temporal information extraction
Semantic DB

Figure 1. System framework

2.1 Web Image Context Acquisition

Traditionally, static images include some contextual information in addition to their low-level image features as follows:

- File name. An image's name is sometimes specified by users to indicate its background semantics.
- Caption text. An image's caption usually accompanies with useful information such as person name and capture time, etc.
- Metadata. Most digital cameras follow the EXIF standard defining the format of image header fields. It allows for the insertion of various metadata by the camera processor, such as basic image parameters (e.g. height and width), location/time of capture, and capture settings (e.g. focus distance).
- Annotation. Annotating textual information for digital images is becoming popular to manage the image collections.

However, the metadata and annotation information are not always available for digital images. Additionally, they also require manual maintenance to acquire and manage these contextual data. Fortunately, there naturally exist plentiful context data for WWW images. The problem is left to precisely extract image context from web pages. As shown in [13], the majority of pages contain noise information like ads. Furthermore, it is not appropriate to assign the context of an image as the whole content body of its hosing page, especially when multiple images coexist in a page and each has its own contextual description as shown in Figure 2.

Current web image search engines are mostly relying on text-based search [8]. In such cases, web images are indexed by their surrounding text, which is assumed to be related to the content semantics of that image [1]. Several approaches were proposed to extract surrounding context for web images. [4] proposed a rather simple method that finds a passage consisting of the 20 terms before and after a web image. However, such pure analysis with simple rules without page layout tends to be low-level accuracy. Some researchers proposed to address this problem from the point of image segmentation [13], where a visual based method called VIPS was proposed to analyze the structure of a web page. Chen [2] provides a method that understands the semantic page structure based on detecting visual borders of content objects.

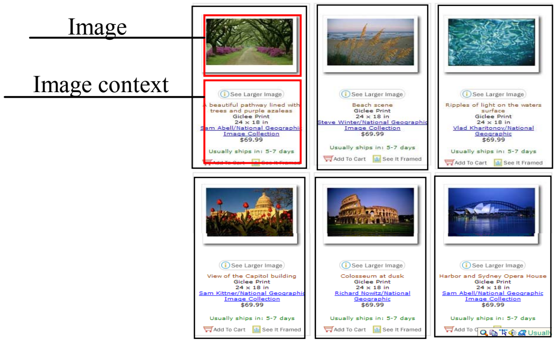
In our system, we use simpler objects as nodes in the DOM (Document Object Model) interface provided by the web browser (e.g. IE). DOM is a type of HTML syntax tree comprising nodes and pointers, by which an HTML node tree can be traversed easily. From bottom up we identify nodes by using simple rules of treating visible objects like image, link or text paragraph as basic element. In this manner, all visible objects in a web page will be elements allocated a tree. Our system mainly processes the image object nodes in the HTML tree. We could apply explicit separator detection to detect their contextual boundaries. The explicit separators can be detected by analyzing the properties of the tags. Several kinds of separators are widely used by web designers:

- The commonly used tags such as <TABLE>, <TD>, <TR> and <DIV> own border properties. When their border properties are set, there would be separators at the corresponding borders.
- <HR> is also the most frequently used explicit separator as a horizontal line for separating different content passages.

By detecting these separators, our algorithm can effectively detect context for images. Figure 2 displays an example for our context acquisition algorithm, in which 6 images (presented as a 3x2 table layout) can be successfully segmented separately. It can be seen from the figure that, web image context includes plentiful data:

- The hierarchical URL of an image indicates its context.
- Surrounding text usually indicates the description of an image.
- Image link indicates the detailed illustration of an image.

For example, Google Image Search matches users' keywords with URL, page title, or alternative text of images. The following subsection describes how to mine contextual semantics from these textual sources. Note that, we don't use the link text in the current implantation because it may cause additional computational load.



A beautiful pathway lined with trees and purple azaleas
Giclee Print
24 x 18 in
Sam Abell/National Geographic Image Collection
$69.99
Usually ships in: 5-7 days

Beach scene
Giclee Print
24 x 18 in
Steve Winter/National Geographic Image Collection
$69.99
Usually ships in: 5-7 days

Ripples of light on the waters surface
Giclee Print
24 x 18 in
Vlad Kharitonov/National Geographic
$69.99
Usually ships in: 5-7 days

View of the Capitol building
Giclee Print
24 x 18 in
Sam Kittner/National Geographic Image Collection
$69.99
Usually ships in: 5-7 days

Colosseum at dusk
Giclee Print
24 x 18 in
Richard Nowitz/National Geographic
$69.99
Usually ships in: 5-7 days

Harbor and Sydney Opera House
Giclee Print
24 x 18 in
Sam Abell/National Geographic Image Collection
$69.99
Usually ships in: 5-7 days

Figure 2. An example for web image context acquisition

2.2 Semantic Knowledge Extraction

As we have described in the introduction section, time and people are important cues for indicating content semantics of an image. This section presents the algorithms to acquire and extract the description, time and people for images from the web context.

2.2.1 Web Image Description Extraction

It is important to extract indicative keywords from an image's context to illustrate the semantics content of that image. Given a text stream, our method parses it and extracts all possible phrases (n-grams) from the contents. We apply stemming to each word using Porter's algorithm. Stop words are then removed from the generated phrases. Thus, the context of an image is represented as a bag of phrases. We explore four properties for each phrase to estimate the importance of a phrase to indicate image content.

Visual weight (*VW*). The visual features such as font, bold, color, hyperlink and position available in web pages indicate important cues for a phase's importance. These visual properties can be combined into one single score according to some heuristic rules, which is described as:

$$\{Font, Bold, Hyperlink, Color, Position\} \rightarrow VW$$

Phrase length (*LEN*) is simply the count of words in a phase. A longer phrase is more preferred by designers for describing images.

$$Len = n$$

Phrase weight (*PW*) in a text passage is heuristically indicated by its frequency (TF) and part of speech (POS). For example, a proper noun is more probable to describe an image.

$$PW = TF \times POS$$

Phrase independence (*IND*). According to [3], a phrase is independent when the entropy of its context is high (i.e. the left and right contexts are random enough). We use *IND* to measure the independence of phrases. The following is the equation for IND_l which is independence value for left context (denoted IND_l) where 0·log0=0 is also defined.

$$IND_l = -\sum_{t\in l(p)} \frac{f(t)}{TF}\log\frac{f(t)}{TF}$$

where f(t) is to measure the times of the term t occurs. The IND_r value for right context could be calculated similarly. The final *IND* value is the average of those two.

Using the four properties (*VW, LEN, PW, IND*), a regression model can be learned to map them into a single salient score. Thus, the textual description of an image is set as the phrases whose salience score are among the top three ranks.

2.2.2 Temporal Information Extraction

The temporal information also indicates important context of an image. The best practice recommended for encoding the date value is defined in a profile of ISO 8601 and follows the YYYY-MM-DD format. However, in practice there exist various forms of time representations. It's hard to recognize all time formats in textual streams. Currently, our system is capable of recognizing three time formats including month, week and date (e.g. Table 1).

Table 1. The time formats recognized by our system

Time format	Examples
Month	January; Jan; February; Feb; etc.
Week	Sunday; Sun.; Monday, Mon.; etc.

Date	1 Jan; Jan 1; Jan 1, 2002;
	Jan 2002; 2002 Jan;
	01-01; 01-01-2002; 2002-01-01

In the future, we plan to extend this function to be able to extract more time formats. Given the variation of the temporal formats used by data providers and the perception that the ability of a standard date is of the measure importance to users, the decision was made to attempt to normalize the temporal information in the date field. The temporal normalization is also a must in future to standardize various time formats to an identical representation.

Moreover, there is also a wide variability in what the time values were used, such as date created, date published or date digitized. It is very difficult to identify which the values are really involved. This problem is expected to be solved by the introduction of the advanced techniques in natural language processing (NLP) field.

2.2.3 Geographical Information Extraction

Previous studies [5][7] have demonstrated that there exist various representation forms for geographical location, e.g. telephone number, postal code and geography place name, and so on.

Currently, our system only recognizes and identifies the name of geographic entities. When a text passage is scanned by the system, such geographic place names will be acquired. Furthermore, it is a common case that multiple geographical names may co-exist in a text passage [5]. It is necessary to determine a final geographical representation for this case of multiple candidate locations. Instead of treating this problem with complicated processing or computing [5][7], we used a simple yet quick method, that is, geographical names with top occurrences are deemed as the final geographic candidates for that passage, which proves effective in practice.

2.2.4 People Information Extraction

People names are widely available in the context of web images. For example, an image with a figure usually accompanies a caption that includes the name of that person. It is quite easy to identify the person names with a prepared name thesaurus. Furthermore, we performed a simple name normalization on the identified names. The name normalization consisted of conflating all persona names with the same last name, and replacing it with the most frequent occurrence, so that:

$$n = \arg\max_{1\le i\le m}\{tf(n_i)\}$$

where n_1, n_2, n_3, … n_m be a series of names that have the common last name that are extracted from an image's context. For example, the names "Timothy McVeigh", "Tim McVeigh", "Timothy James McVeigh" were all replaced with "Timothy McVeigh".

For example, semantic knowledge of the first two images on the top row of Figure 1 is extracted in our case as follows:

Semantics	Image 1	Image 2
Description	Pathway, purple azaleas	Ripples, waters
Time	Oct 1998	July 1999
People	Sam Abell	Vlad Kharitonov
Geography	Montana National Park	Mississippi

3. EXPERIMENTS

3.1 Dataset Preparation

The web images used in our experiments are crawled from Yahoo (http://dir.yahoo.com/) that classifies web pages into hierarchical categories. We selected top 50 web sites from three second-level categories: 1) newspapers in News & Media; 2) shopping in Business & Economy; and 3) history in Arts & Humanities. We crawled 50 pages from each site, and use our page analysis method to extract JPG images and their corresponding context.

Table 2. The experimental web image dataset

Image Category	Image Number	Semantics related ratio (%)		
		Time	People	Geography
News	2,102	92.2%	94.3%	96.8%
Shopping	1,987	38.2%	26.3%	46.2%
History	2,431	87.3%	89.7%	86.4%
Total	6,520	73.9%	71.9%	77.5%

The resulting images crawled from all the pages are listed in Table 2, including 6,520 images totally. Furthermore, we labeled these images related with various kinds of contextual semantics. As shown in Table 1, the image context of the news category includes the maximal percentage of semantics. For all the pages, the average relevant ratio over time, people and geography are about 74%, 72% and 78% respectively, indicating that a high level of semantics are available in the WWW image context.

3.2 Experimental Results

We list in Figure 3 the precision of semantics extraction by using our method. On average, the extraction precisions of description and time are higher than people and location. Such quantitative values indicated by Figure 3 show that our approach can mostly find semantic knowledge available in web image context.

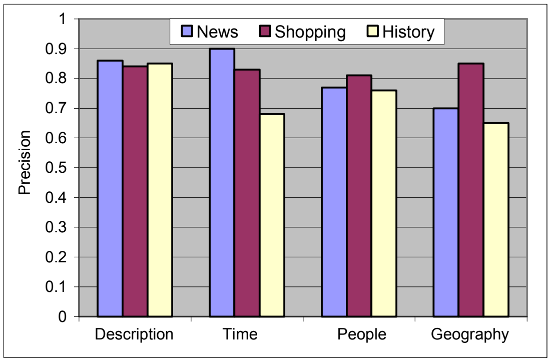


Figure 3. The precision of semantic knowledge extraction

The extraction precision of time ranges from 62% to 90%, due to the availability of a wide variety time representations that are difficult for temporal identification. The low precision in the history category is caused by the temporal representations such as, "the early 18th century", "medieval", "17th c." and so on. As for future work, we plan to use professional techniques to normalize the date field provided the variation of the temporal formats. The reason for lower extraction precisions of people and geography information lies in some existing ambiguities. For example, a person and a geographical entity may share a common appellation like Washington, Mcdonald's is not describing a person name but a snack brand, and Kentucky Chicken is not concerned with the state of Kentucky, etc. It is clear that such appellation confusions can lead to degradation of extraction precision on the people and geography information. In current work, we have not employed sophisticated algorithms to eliminate these ambiguities. Anyway, this problem also remains a great challenge in traditional natural language processing field. In future, we will adopt sophisticated algorithms and heuristic rules to improve this case.

As an initial report of the performance studies, the quantitative precision results verified the effectiveness of our approach. We believe it will achieve higher precision of semantic knowledge extraction by employing more sophisticated algorithms.

4. CONCLUSIONS

This paper introduced an automatic approach that can effectively extract and annotate semantic knowledge for the World Wide Web images. By utilizing the page layout analysis method, surrounding context can be precisely assigned to the web images. We designed different methods to extract semantic knowledge from an image's contextual information. The preliminary experimental results demonstrated the effectiveness of this approach of extracting semantic knowledge for web images. As future work, we plan to apply our solution to refine web image retrieval services to users.

5. ACKNOWLEDGEMENT

This project was supported in part by Natural Sciences Foundation of China under No. 60405005 and 60121302. We also give special thanks to some graduate students on their generous programming helps.

6. REFERENCES

[1] D. Cai, X. He, Z. Li, W.-Y. Ma and J.-R. Wen. Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Analysis. ACM Multimedia Conference 2004, New York, USA, Oct 2004.

[2] Y. Chen, W.-Y. Ma and H.-J. Zhang. Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices. 11th World Wide Web Conference, Budapest, Hungary, May 2003.

[3] L. F. Chien. PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval. 20th ACM SIGIR Conference, Phliadelphia, 1997.

[4] T. A. S. Coelho, P. Calado, and et al. Image Retrieval Using Multiple Evidence Ranking. IEEE Transactions on Knowledge and Data Engineering, 2004.

[5] J. Ding, L. Gravano and N. Shivakumar. Computing Geographical Scopes of Web Resource. The 26th International Conference on VLDB, Cairo, Egypt, Sep 2000.

[6] E. Gabrilovich, S. Dumais and E. Horvitz. Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. New York, USA. May 2004.

[7] Z. Hua, C. Wang, X. Xie, H. Lu and W.-Y. Ma. Automatic Annotation of Location Information for WWW Images. International Conference on Multimedia and Expo (ICME) 2005, Amsterdam, Netherlands, July 2005.

[8] G. Lu and B. Willam. An Integrated WWW Image Retrieval System. 5th Australian World Wide Web Conference, 2004.

[9] M. Naaman, Y. J. Song, A. Paepcke and H. Garcia-Molina. Automatic Organization for Digital Photographs with Geographic Coordinates. 4th ACM and IEEE-CS Joint Conference on Digital Libraries, Tuscon, USA, June 2004.

[10] K. Rodden and K. R. Wood. How Do People Manage Their Digital Photographs? The CHI 2003 Conference on Human Factors in Computing Systems, Florida, USA, Apr 2003.

[11] R. Swan and J. Allan: TimeMine: Visualizing Automatically Constructed Timelines. 23th ACM Conference on SIGIR. New Orleans, USA, July 2000.

[12] K. Toyama, R. Logan, A. Roseway and P. Anandan. Geographic Location Tags on Digital Images. ACM Multimedia 2003 Conference, Berkeley, California, USA, Nov 2003.

[13] S. Yu, D. Cai, J.-R.Wen and W.-Y. Ma. Improving Pseudo Relevance Feedback in Web Information Retrieval Using Web Page Segmentation. 12th World Wide Web Conference, Budapest, Hungary, May 2003.