Web Image Retrieval Systems with Automatic Web Image Annotating Techniques

Hsien-Tsung Chang
Computer Science and Information Engineering
Chang Gung University
No.259, Wenhua 1st Rd., Guishan Shiang, Taoyuan County 333
Taiwan (R.O.C.)
smallpig@widelab.org
http://www.widelab.org/

Abstract: - Due to the popularity of digital cameras and web authors' enriching the visual aesthetics, the number of web images is growing in an uncontrolled speed. The images in the World Wide Web are becoming a large image library for browsing. It is an important issue that how to retrieve the images accurately on the World Wide Web. In this paper we describe the architecture of the web image retrieval systems with automatic image annotation techniques. And we propose four methods to generate the annotation automatically for every image from its hosted web page, by analyzing the structural blocks, collecting anchor text of link structures, and gathering shared annotation with other images with the same visual signature.

Key-Words: - Image Annotation, Image Retrieval

1 Introduction

After the invention of the World Wide Web, the most important invention of the network applications, by Tim Berners-Lee on November 12, 1990, the population on the Internet is growing in a rapid and uncontrolled speed. The population brings commercial potential onto the World Wide Web. In order to attract users to stay for more time and return to the web site more often, the looks of webs pages are more and more elegant. A large number of images are used to enrich the visual aesthetics by the authors of the web sites. On the other hand, more and more digital images are posted onto the web space with the popularity of digital cameras. Lots of free web album services are announced for users put their images taken by their digital cameras for sharing. For the two reasons we mentioned above, the number of images on the World Wide Web is increasing in a fast way.

As we know, the colossal number of text content in the web pages is a large free text library for web users to surf. Such a huge number of images created recently are posted on the World Wide Web, the images are also formed another library as the text web pages. Therefore, it is becoming an important issue to retrieve images with a quick and accurate method on the World Wide Web. Unlike the web pages search, a lots of text retrieval technique have already developed for a long time. There are many difficult problems for a web image retrieval system.

In the traditional images retrieval systems, they assign annotations to each image manually. Although it is a good methodology to retrieve images through text retrieval technologies, it is gradually becoming impossible to annotate images manually one by one due the huge and rapid growing number of web images. Some researches [1] [2] [3] [4] [5] [6] [7] create image retrieval systems by analyzing the image content. They find out the size, shapes, patterns, colors, and distribution of hues [8] [9] [10] [11] of each image, and transfer the properties into a signature for the image. User can specify the properties as the query signature, or submit a similar image and generate the query signature with the same method by the image retrieval system. However, how to select the properties for images is a difficult task. In the past research paper, we can notice that the accuracy is not good enough by this method. The properties of the image with "Tokyo Tower" or one with "Eiffel Tower", as an example, may be similar, but the semantic meanings of the two images are poles apart. Therefore, even we can develop an accuracy method to select the best properties for an image retrieval system, it is still hard to recognize the semantic difference only by signatures generating. On the other hand, it is also hard for a user to submit a query by specifying the properties or submitting a similar image as a query. Therefore, in some other researches [12] [13], they allow users to submit a text query, and then use the retrieved images by text based retrieval method to generate the signatures. The generated signatures will use as the input for the second phase image retrieval. However, it is easy to understand that the signatures generating is a time consuming process, it will need a lot of hardware and time to obtain the results. It is becoming hard to accomplish through the same reason with annotating image manually.

Large image retrieval systems [14] [15] [16], like Google Image Search [17], Yahoo Image Search [18], and MSN Image Search [19], nowadays on the World Wide Web allow users submitting keyword queries similar to web page search. Those web image retrieval systems use automatic image annotation and text inverted index techniques to match the users' queries, and then return the images that relevant to query keywords. In this paper, we will sketch the architecture of the image retrieval system and how it works with the automatic image annotation techniques.

2 Architecture of the Image Retrieval System

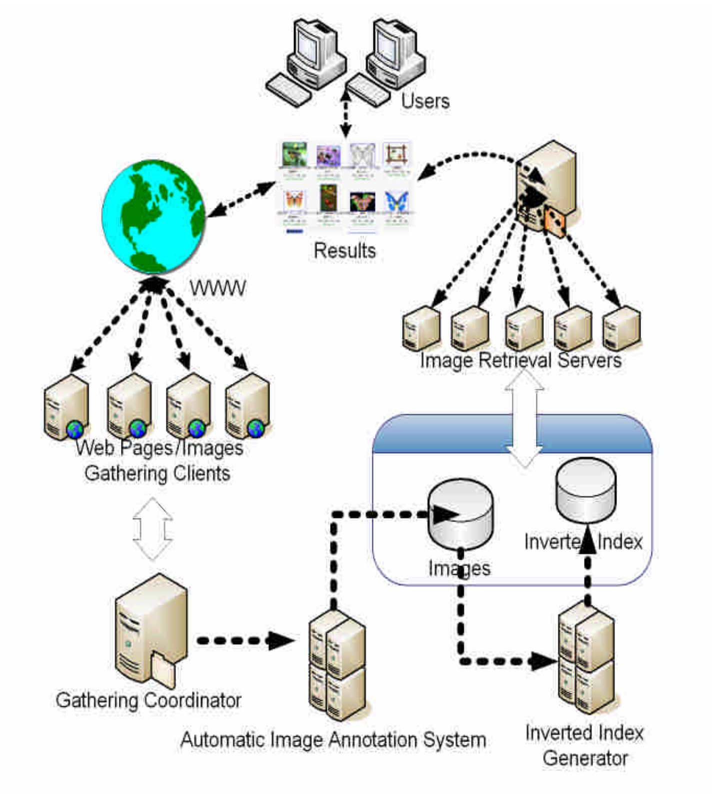

Figure 1 The architecture of the image retrieval system on the World Wide Web

We can model the World Wide Web as a large database. There is a main difference between the traditional information retrieval systems and web information retrieval systems, that is, we should generate the results to the users in a short time (usually in one second). If users need to wait the results from a web information system, they will give up the system and choose another for instead. Therefore, it is impossible for a web image retrieval system generating the results by accessing the remote web pages and images through the network when the system receives a query keyword from a web user, because it is too slow to bear. Therefore, in order to response in a short time, a web image retrieval system needs to store a special copy of the web images in advanced. For saving the storage space, we only need to store the thumbnails from the original ones. The stored thumbnails and data can be utilized for generating the inverted index and providing the search function..

There are many tasks that a web image retrieval system needs to accomplish. A web image retrieval system can be generally separated into four sub-systems. They are data gathering sub-system, automatic image annotation sub-system, index generator sub-system, and image retrieval sub-system. Figure 1 shows the general architecture of the web image retrieval systems.

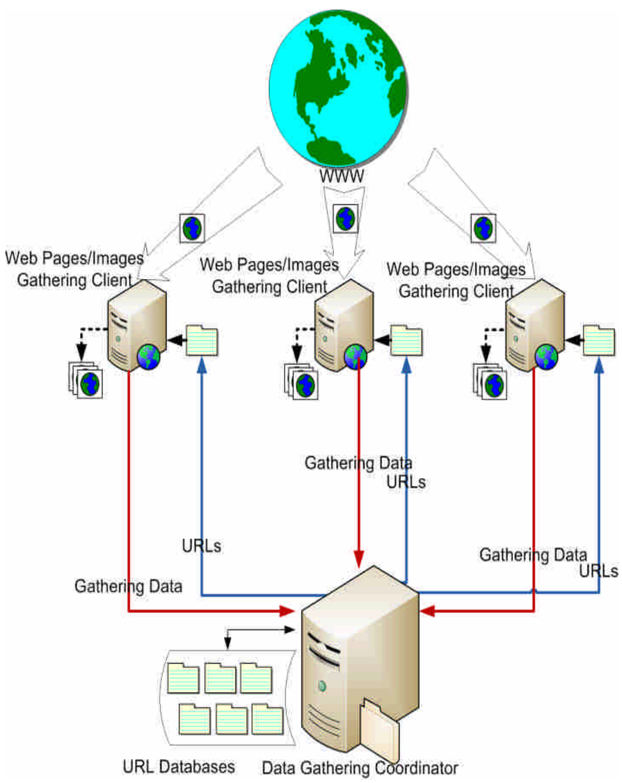

Figure 2 The flow chart of a data gathering sub-system

2.1 Data Gathering Sub-System

The purpose of a data gathering sub-system [20] [21] is to gather the data for the rest of the sub-systems. A simple data gathering sub-system starts with a set of URLs, we call them as seed URLs, through the protocol HTTP. It then extracts the URLs from the starting pages, following the new URLs recursively by the method of breadth-first or depth-first tracing. We can collect the portal sites as the seed URLs. It is because the data gathering sub-system can access a lot of new URLs from portal sites. After gathering the web pages on the World Wide Web, we need to extract the image URLs from the pages. And then dispatches the image URLs to the gathering clients for gathering the images. Generally speaking, we design the data gathering sub-system in a distributed model to increase the concurrency and gathering speed. One problem is how to coordinate the gathering robots to avoid traversing a URL many times. Traversing one same URL many times wastes the bandwidth to fetch and time to process. We can design the data gathering sub-system in a centralized model, which is putting one centralized gathering coordinator to arrange the URLs for the gathering clients. Figure 2 shows the flow chart of a data gathering sub-system. Besides the unique issue, we also need to limit the access number to a single server in the same time. It is because the computing power and network bandwidth for a data gathering sub-system is definitely large. It will crash the web servers of small companies easily. This behavior will be treated as DOS attacks. The data gathering sub-system need to follow the robots.txt rules set the web authors due to the web politeness.

2.2 Automatic Image Annotation Sub-System

After the gathering phase, we now have the images and pages in our database for the next phase automatic image annotation. We will generate the thumbnails and collect the basic information, like image width, height, file format of the image… for all the images and transfer them into our internal format. And automatic image annotation sub-system will generate the annotation for each image automatically from the database. The detail of automatically generating annotation will be described in the section 3.

2.3 Index Generator Sub-System

The automatically generated annotation for all gathered images can be used for building the inverted index through the traditional information retrieval techniques, and we call it as index generator sub-system. The produced inverted index is a word-oriented structure for indexing the pages collection. Therefore the text needs to break into word-based form. We call it the page index phase. Collecting the words information and generating the inversion indices for each word is the inverted index phase. Figure 3 demonstrates the data structures of traditional inverted index.

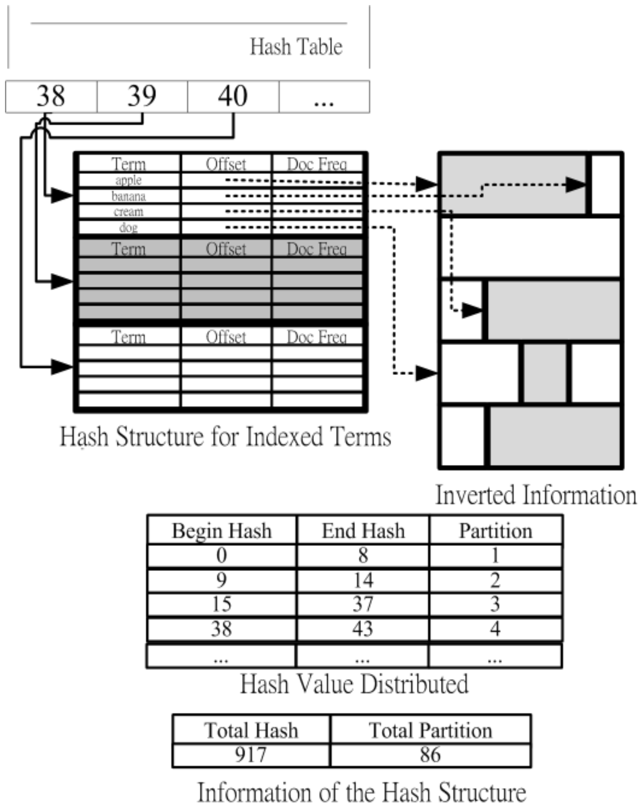

Figure 3 The data structures for the inverted index

2.4 Image Retrieval Sub-System

The generated inverted index, the thumbnails, and the basic information from the images can be utilized for the image retrieval sub-system. When a query keyword is submitted to the image retrieval sub-system, it will prepare the result images according to the query keyword. Figure 4 shows the framework of the query processing from the image retrieval sub-system. We will generally put a powerful server [22] in the front of image retrieval servers which contains a small part of the entire images and relative data. The powerful front server will first issue the query task to the first of query servers. And it will decide whether to issue the query task to the rest query servers that base on the results returned from the first server. The image retrieval server will look up the inverted index to matching the images with the automatic generated annotation which is relative to the keyword. And then show the image thumbnails to the user according to some pre-defined ranking methods.

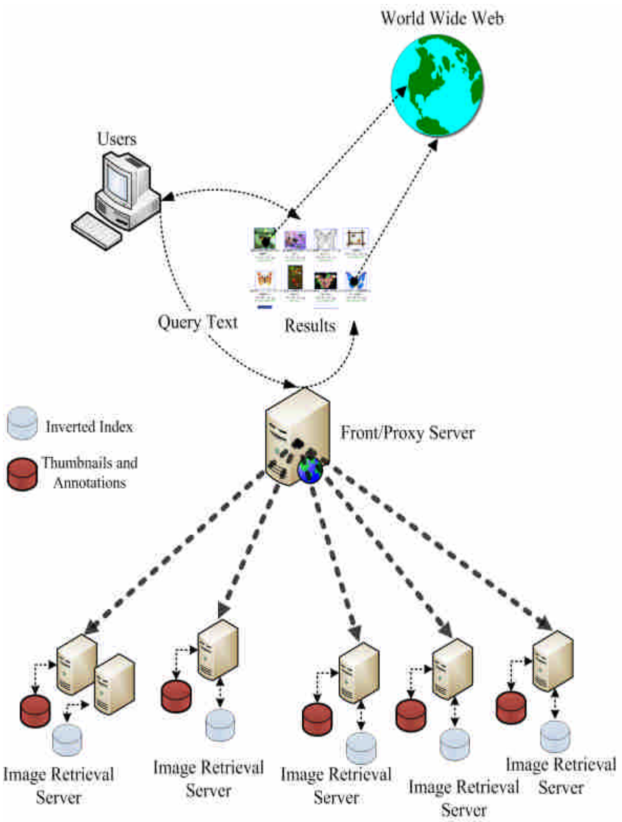

Figure 4 Framework of the image retrieval sub-system

3 Automatic Web Image Annotation

Web pages are composed by HTML. There are a lot of clues can be used for automatically generating image annotation, especially the links characteristics and structures. We will utilize the special properties of the World Wide Web to generate the image annotation automatically. The four methods we used are annotation from the hosted web pages, annotation from structural blocks, anchor text, and the shared annotation. We will describe the detail by the following sub-sections.

3.1 Annotation from the Hosted Web Page

Web page authors use the ***<img>*** tag to show an image in a HTML page, the value of the property "***alt***" is a statement composed by some words defined by the author. It is generally to tell page consumers what the image represents when the browser fails to load it. Therefore, it is the best annotation candidate for an image. It is direct assigned by the web page authors who recognize the image most.

Generally speaking, the text around the ***<img>*** tag is used for describing the image. We can also say that the image is used for emphasizing the text around it. It is impossible for a web page author putting some irrelevant images to the text in the web page. Therefore, the text around the image can also be treated as an important annotation for an image. We can select 10 to 20 words before and after the <***img***> tag as the annotation.

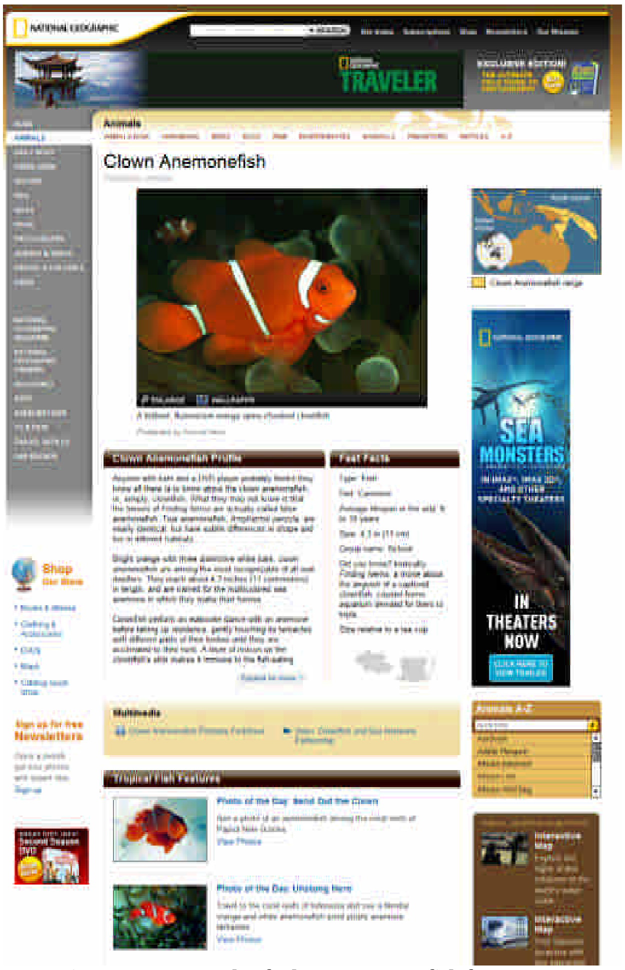

Figure 5 An example of Clown Anemonefish from URL: http://animals.nationalgeographic.com/animals/fish/clown-anemonefish.html

The author of the web page may specify a meaningful title to the page. The specified title can be treated as the name of the page. Generally speaking, the name contains the essential meaning of the page. The specified title can also be the annotation to images within the page. By observation, we can find out that the image with the largest area within the page can share the meaning of the page title. Figure 5 shows that, the web page with title ***"Clown Anemonefish, Clown Anemonefish Profile, Facts, Information, Photos, Pictures, Sounds, Habitats, Reports, News - National Geographic"*** from the National Geography official web site, and it is designed for describing clown anemonefish. Although there are many images within the page, we can easily notice that the image with largest area can just share the meaning of the specified title. Generally speaking, the precision of the relationship between the annotation and the image from this property is very high. If the query keyword matches the annotation from this property, we can higher the rank position for the image.

3.2 Annotation from Structural Blocks

The structural blocks are very common in a web page. Generally speaking, the authors of web pages utilize the <***table***> to present a structural blocks. They usually use structural blocks to present a sort of items. For example, if you want to display the products that produce by your company, you may use a structural blocks to place the product images and relevant annotation.

To analyze and discover the structural blocks can gather the annotation for the image accurately. Figure 6 is an example also from the National Geography official web site, which is a sort of images of the annotation with the structural blocks. Each block contains an image in the left, and the detail annotation written by the authors of the web page for the image was placed in the right. It is clear that if we can find out the mapping relation between images and annotations, we can easily get the accurate annotation from the authors of the web pages. Figure 7 is another example of structural blocks from the URL http://animals.nationalgeographic.com/animals/mammals/cheetah.html.

In fact, we can easily build the DOM tree for a web page through open source tools, the DOM tree will systematically represents the web into a tree view. By counting the repeating sequence of HTML tags, we can easily find out the repeated structural blocks. And we can find out the annotation for each image in the structural blocks easily by referencing the DOM tree.

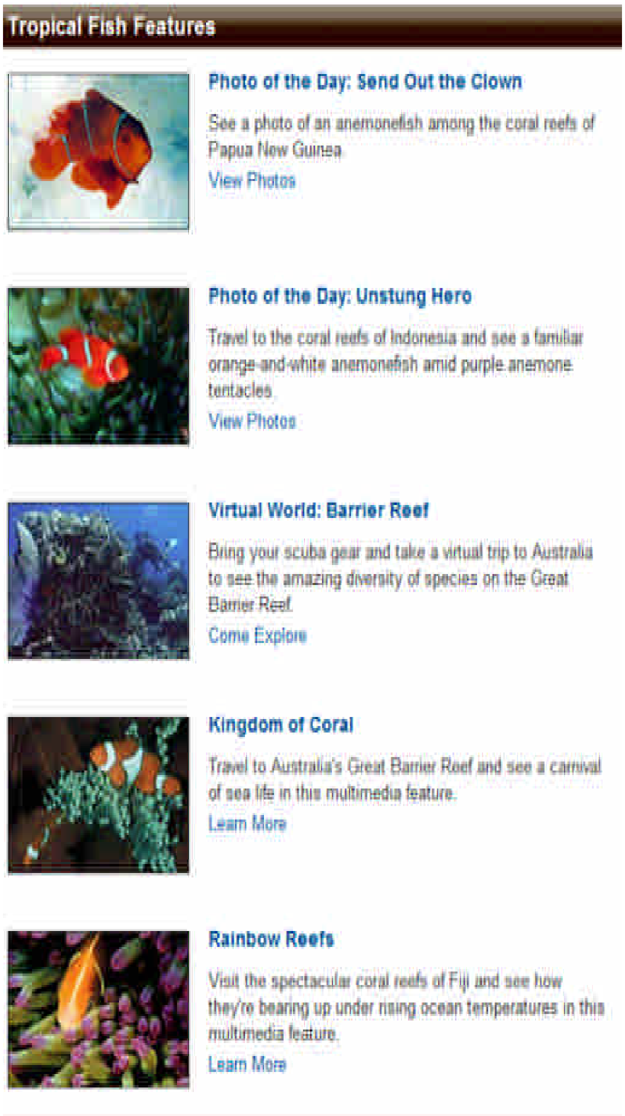

Figure 6 An example of structural blocks annotation in URL: http://animals.nationalgeographic.com/animals/fish/clown-anemonefish.html

Figure 8 is a small portion of the DOM tree from Figure 6. We can notice that the text and the image are always located in the leaves of the DOM tree. In this example, we can also notice that the text portion of the structural blocks is within the second <***td***> tag, and the image information is located within the first <***td***> tag. The text portion "***Photo of the Day: Send Out the Clown***", "***See a photo of an anemonefish among the coral reefs of Papua New Guinea***", and "***View Photos***" can be the annotation for the image. If we count the repetitions of the text within the page, we will find that "***View Photos***" repeats many times. It can be treated as noise, and remove it out from the annotation.

Figure 7 Another example of structural blocks in URL http://animals.nationalgeographic.com/animals/mammals/cheetah.html

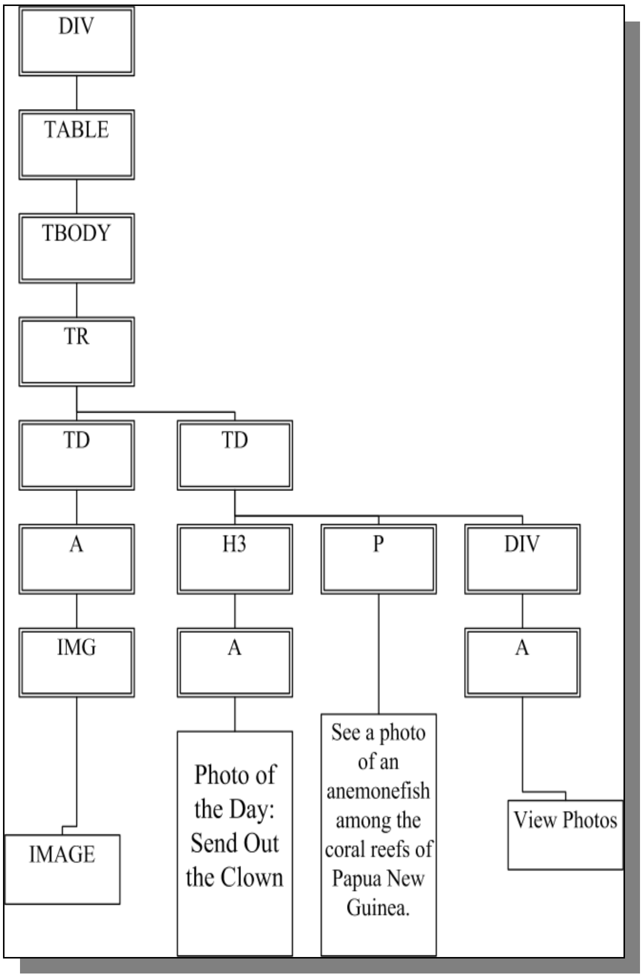

Figure 8 The DOM tree example

3.3 Anchor Text

A link is represented as *** Anchor Text ***. The anchor text of a link assigned by the author usually is a good annotation for the linked URL. In other words, the anchor text is the name that others like to call. It is because the anchor text is the text for page consumers moving their cursors to click and surf to another page. Page consumers decide if surfing to the URL only according to the anchor text. Therefore, authors of the web pages generally assign the anchor text with more attentions.

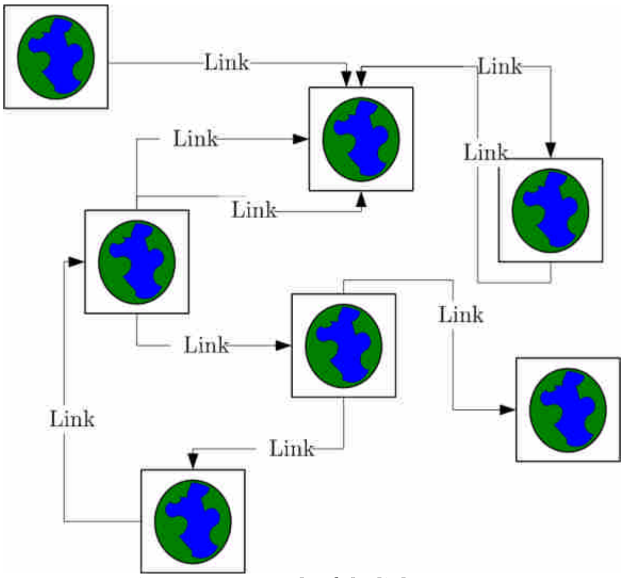

Figure 9 An example of the link structure

We can gather all kinds of annotation from different authors to a single image by collecting the anchor text that link to that image. Even we can collect the annotation for the image in different languages, because the anchor text of the links could be assigned by the people from any countries. Therefore, user can retrieve the images by all kinds of languages by assigning the anchor text as the annotation to the image.

The anchor text is widely used for web page search engine. Besides the anchor text for the image, the anchor text for the URL of the hosted web page is also acting a good role for the image annotation. As we mention in section 3.1, since the image with the largest area can share the meaning of the page title. The anchor text associated to the hosted page can be treated as the alias titles of the page, because they are the titles other authors of the web pages like to call. Therefore the image with the largest area in the hosted page can also share the anchor text linked to the page.

A link can provide anchor text for automatically web image annotating, the structure of links can also be used for ranking the image like popular ranking algorithms, PageRank [23], and HITS [24]. Figure 9 displays the concept of the link structure. Although the characteristics of web pages and web images are not totally the same, the link based ranking algorithms can also be an important clues for ranking the web images. However, because of the rampant drowsing spam techniques, the first positions of results ranking by links structure will be occupied by spam sites if you do not apply any spam filtering technologies.

3.4 Shared Annotation

In the traditional images retrieval system, we treat two images as different ones if they have different URLs. However, web page authors often duplicate images from others sites, and put the original duplicated image or resized one into their web sites. In fact that if the two images are visually the same, even the size or brightness were different, we still can treat these two images the same in the web image retrieval system. It is because that, image retrieval system users do not concern about which image is the original one, they only concern about if they can get the images that are relevant to their submitted query keywords. Therefore, one image can share the annotation generated from sections 3.1 to 3.4 from visually the same images.

In order to share the annotation for visually the same images, we need to generate a visual signature for each image. However, how to generate the visual signatures perfectly is a very difficult image processing problem. It is not the main issue in this paper. We will not discuss the detail about generating the visual signature in this paper.

In order to proof this concept, we can just use a simple method for the visual signatures by resizing every image to a pre-defined dimension, and then use the MD5 digest method to generate the signature for each image. The visual signatures of duplicated images and original one will be the same. Even the duplicated images are resized, and the generated visual signatures will also be the same. It is because we force all the images transfer into the same size. As our definition, images with the same visual signature could share the annotation to each others.

By sharing annotation from images with the same visual signature can substantially increase the source of annotation for every image. Although the share annotation cannot increase the result number of retrieved images with different visual signatures (It is because you always can retrieve one of the visually the same images, even they do not share the annotation to each other.), it can help the ranking algorithm higher the score if one image can gather more shared annotation from others. It is because one image might be good if many authors duplicate it and put it in their own sites. On the other sides, if we combine the links count for the visually the same images, it will highly increase the total links count for the images. It will influence the link based ranking algorithms we mentioned in section 3.3, and higher the accuracy for the web image retrieval systems.

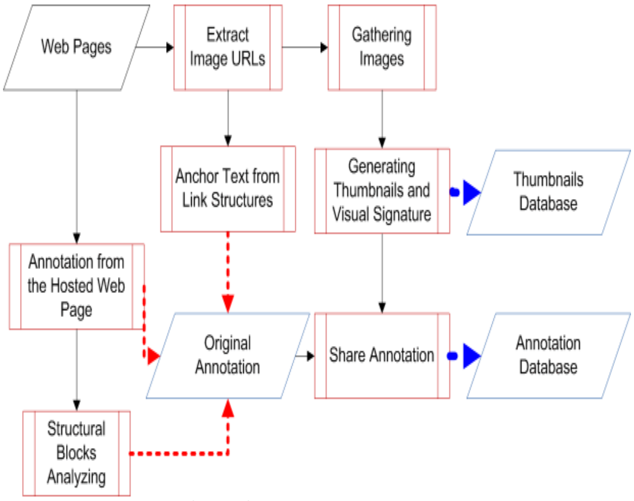

Figure 10 The data flow of automatic image annotation generation.

Figure 10 displays the data flow of automatic image annotation generation for the automatic image annotation sub-system. We will generate the annotation for each image through the methods we described in the section 3.1 to 3.3. And then generate the thumbnails and visual signature for each image. The shared annotation can be generated the by sorting the internal format according to their visual signatures. The images with the same visual signature will be put in the continuous location. It will be very easy to collect all the shared annotation in linear time.

The importance of annotations from difference sources should be different. We can assign different vectors of weights to each annotation from different sources. For example, the title of the hosted web page for the max area image is much more important than the annotation from the text surrounding the image. If the query keyword matches the title, we can score it as 1000, for example. And if the query keyword matches the text surrounding the image, we can score it only as 100, for example.

4 Experimentation

We make an experimentation to gather some free web albums and e-commerce sites in Taiwan. In the Figure 11, it is the snapshot of the portal site http://www.yahoo.com/. We can notice that there are many functional images in the web page. The functional images are used for navigation, separator, or advertisement banner. Generally speaking, the functional images are not good for a web image retrieval system. It is because that only few users will try to find out the advertisement banners through a web image retrieval system. Therefore, we should first mark this kind of images for further processing.

Figure 11 An example of http://www.yahoo.com/

In fact, we can notice that the size of the functional images is relative small to the normal ones. And the images for separator can be filtered out by calculating their ratio of width and height. On the other hands, the functional images could be appeared in one page more than ones, like the marks for the items. We can mark the images as functional ones by the following rules.

A. Width/Height > 4
B. Height/Width > 4
C. Width * Height < 2500
D. Image in GIF format with more than one layer
E. The image appears in the single page more than once

We will take the marked functional images out from the results, because users are not interesting in these functional ones. If users need to retrieve those functional images, they can use advanced search function to retrieve them.

# of Web Pages	20,219,113
# of total images	352,295,297
# of functional images	303,190,414
# of images for retrieval	49,104,883

Table 1 The information of gathering data

Table 1 shows the number of web pages, functional images, and normal images we gathered for this experimentation. We will create two set of annotation for the experiment of the influence on shared annotation. The first set is the annotation generated by the methods described in section 3.1 to 3.3, and the second set is adding the shared annotation described in the section 3.4. We select the top 10 query keywords from our query logs, and use blind test as shown in Figure 12 to 20 users. The retrieved results from two set will randomly put in the two sides of the results page. Testing users will assign one value from 1 to 5 for the better side. We call it as blind test.

It is a very difficult problem to calculate the recall and precision of a web image retrieval system according to the definitions of the traditional information retrieval. The first problem to calculate recall and precision is what the exact number of all relative images is? In fact, we even have no idea about the exact number of all the images. It is nearly impossible to find out the exact number of relevant images, because the size of web images is definitely huge. Since the scientific statistic cannot be calculated, we may utilize the user feedbacks or user evaluation to evaluate the quality of image retrieval systems. Generally speaking, the use experience of users is more important to a web service than the traditional recall and precision notations. It will be more practical to design a measuring system for users to evaluate good or bad than just calculate the value of recall and precision.

Users can judge the two sides of results to which side is better, or a draw. The evaluation system will record the value from the users. After analyzing the judgment data, we can notice the overall comparison between the two sets. We can also discover the inferior results in specific terms in our system. The information can be used to improve our system. It is the reasons that why we choose blind test for this experiment.

The average value to the first set is 0.8, and the value to the second set is 1.65. We can say that shared annotation can improve the results in an image retrieval system.

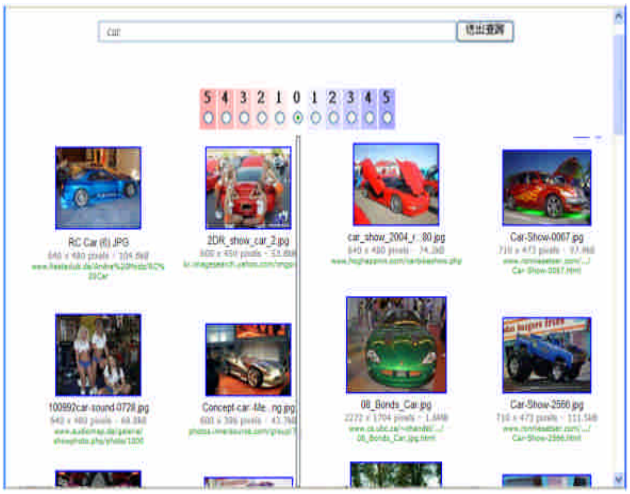

Figure 12 Blind test for the experimentation

5 Conclusion

In this paper, we describe the architecture of the image retrieval system with automatic image annotation. It can be generally separated into four sub-systems; they are data gathering sub-system, automatic image annotation sub-system, index generator sub-system, and image retrieval sub-system. In order to generate the annotation automatically for the image retrieval systems, we proposed four methodologies to gather the annotation for every image. We can first gather the annotation from hosted web page, analyze the structural blocks for accuracy annotation, collect the anchor text through the link structure, and last share the annotation from images with the same visual signature.

How to generate the annotation for every image automatically is an important issue in the context-based image retrieval system. The amount of annotation can be widely increased by our proposed methods. Especially the concept of shared annotation can substantially increase the source of annotation for every image. Although the share annotation cannot increase the result number of retrieved images with different visual signatures, it can help the ranking algorithm for generating more accuracy results for the users.

References:

[1] V.N. Gudivada and V.V. Raghavan, "Content based image retrieval systems", Journal of Computer, Volume 28, Issue 9, Sep 1995

[2] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22, Issue 12, Dec 2000

[3] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, “Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval,” IEEE Trans. Circuits and Systems for Video Technology, 1998

[4] J. R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system", Proceedings of the fourth ACM international conference on Multimedia, Nov. 1997, Massachusetts, United States

[5] L. Zhang, L. Xi, B. Zhou, "Image Retrieval Method Based on Entropy and Fractal Coding.", WSEAS TRANSACTIONS on SYSTEMS Volume 7, pp. 332-341, 2008

[6] H. Kim, T.-K. Kim, S. R. Kim, W. Hwang, S. C. Kee, "A Recursive Matching Method for Content-based Image Retrieval.", WSEAS TRANSACTIONS on COMPUTERS Issue 4, Volume 2, pp. 1035-1038, October 2003

[7] R. Dobrescu, M. Dobrescu, L. Ichim, "Improved Content Based Image Retrieval Technique with Fractal Discriminator.", WSEAS TRANSACTIONS on SYSTEMS Issue 12, Volume 5, pp. 2781-2788, December 2006

[8] T.A.S. Coelho, P.P. Calado, L.V. Souza, B. Ribeiro-Neto, and R. Muntz, “Image Retrieval Using Multiple Evidence Ranking,” IEEE Trans. Knowledge and Data Eng., vol. 16, no. 4, pp. 406-417 Apr. 2004.

[9] A.K. Jain and A. Vailaya, "Image Retrieval Using Color and Shape", Pattern Recognition, vol. 29, no. 8, pp. 1,233-1,244, 1996

[10] A.K. Jain and A. Vailaya, "Shape-Based Retrieval: A Case Study with Trademark Image Databases", Pattern Recognition, vol. 31, no. 9, pp. 1,369-1,390, 1998

[11] D. Niyomua, S. Supratid, C. Kimpan, "Improving Image Retrieval By Fuzzy C-Means Initialized by Fixed Threshold Clustering : Case Studies Relating to a Color Temperature Histogram And a Color Histogram.", WSEAS TRANSACTIONS on SYSTEMS Issue 9, Volume 5, pp. 2182-2188, September 2006

[12] S.-F. Chang, J.R. Smith, M. Beigi and A. Benitez, "Visual Information Retrieval from Large Distributed Online Repositories", Comm. ACM, vol. 40, no. 12, pp. 63-71, 1997

[13] C. Frankel, M.J. Swain and V. Athitsos, "WebSeer: An Image Search Engine for the World Wide Web," Technical Report TR-96-14, Univ. of Chicago, Aug. 1996
[14] Z. Chen, W. Liu, F. Zhang, M. Li, H. J. Zhang, "Web Mining for Web Image Retrieval", Journal of the American Society for Information Science andTechnology 52 (10) (2001) pp. 831-839
[15] H. J. Zhang, W. Liu, C. Hu, "iFind – A System for Semantics and Feature BasedImage Retrieval over Internet" , Proc. 8thACM Multimedia Conf., Los Angeles,USA, 2000, pp. 477-478. 24. Z. Chen, W. Liu, F. Zhang, M. Li, H. J. Zhang, Web Mining for Web ImageRetrieval, Journal of the American Society for Information Science andTechnology 52 (10) (2001) pp. 831-839
[16] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab and M.G. Strintzis, “Semantic annotation of images and videos for multimedia analysis”, Proceedings of the 2nd European Semantic Web Conference (ESWC 2005) 29 May–1 June 2005, Heraklion, Greece
[17] Google Image Search, http://images.google.com/
[18] Yahoo Image Search, http://images.search.yahoo.com/images
[19] MSN Image Search, http://search.msn.com/images/
[20] Junghoo Cho , Hector Garcia-Molina, The Evolution of the Web and Implications for an Incremental Crawler, Proceedings of the 26th International Conference on Very Large Data Bases, p.200-209, September 10-14, 2000
[21] T. -C. Hsu. Design and implementation of a data gathering system for super scale search engines. Doctoral dissertation. 2007.
[22] H.-T. Chang, T.-C. Hsu, and S. Wu, “A Switching Search Proxy”, Proceedings of the Web Technologies, Applications, and Services, 2006
[23] The Google Search Engine: Commercial search engine founded by the originators of PageRank. http://www.google.com/.
[24] J. Kleinberg, “Authoritative sources in a hyperlinked environment”, Proceedings of the ACM- SIAM Symposium on Discrete Algorithms, 1998.