Description-Based Multimedia Clipart Retrieval in WWW

Hiongun KIM, Bong-Kee SIN, and Ju-Won SONG
Multimedia Research Laboratories, Korea Telecom
Wumyun-dong 17, Socho-gu, Seoul 137-140 Korea
{hiongun@ktweb, bkshin@pine, jwsong@ktweb}.kotel.co.kr

Abstract

The Internet today is teemed with not only text data but also other media such as sound, still and moving images in a variety of formats. Unlike text, however, that can be retrieved easily with the help of numerous search engines, there has been few way to access data of other media unless the exact location or the URL is known.

Multimedia data in the WWW are contained in or linked via anchors in the hyper-documents. They can most reliably be retrieved by analyzing the binary data content, which is far from being practical yet by the current state of the art. Instead we present another technique of searching based on textual descriptions which are found at or around the multimedia objects. The textual description used in this research includes file name (URL), anchor text and its context, alternative descriptions found in ALT HTML tags. These are actually the clues assumedly relevant to the content.

Although not without a possibility of missing or misinterpreting images and sounds, the description-based search is highly practical in terms of computation. The prototype search engine will soon be deployed to the public service through the prestige search engine, InfoDetective, in Korea.

1. Introduction (Previous research)

Over the past several years a growing number of researchers have focused on creating descriptions from digital images and sounds. When the work is combined with the retrieval based on the description, it is commonly called *content-based retrieval*. Content-based retrieval is based on the analysis of binary data content, that recognizes automatically the important features contained in an image without human intervention. In the case of images, the work usually focuses on the description of color, texture, shape, spatial location, regions of interest, facial characteristics, and specifically for moving images, key frames and scene change.

The basic idea of content-based image retrieval is that, when the user provides a description of some of the prominent visual features of an image, the system can search the archive and return the images that best match the patterns in the description. At present, most of the research mainly deals with such visual features as color, texture, and shape. Color indexing, first studies by Swain and Ballard [1], is computationally simple and fast. But it suffers from instability--prone to produce false positives and negative--and inadequate expression power. Many researchers still consider other low-level features like texture and shape features.

Typically sound data is described in features like pitch, loudness, duration, timbre. Researchers report the utility of neural networks in indexing sound data with some success [5]. According to Wold [4] there are several possible methods of accessing sounds using simile, acoustical/perceptual features, subjective features, or onomatopoeia. Those features can be parameterized and fed to statistical analyzer. In retrieval applications, the above features may be used with traditional keyword or text-based queries. It's very recent that researchers start studying about indexing and searching the moving images, and they rely on the scene change detection and key frame extraction.

With the advent of Internet, the amount of multimedia data is overwhelmingly exploding . Yet however, technology for analyzing multimedia data is far from being practical to categorize or classify the content itself, in a way we can easily retrieve what we need. The current state of art content-based technology, while very impressive, has yet to overcome/develop a series of basic technologies and establish the generalized methods that are needed for wide acceptance.

In this paper, we present a practical approach of indexing and retrieval of multimedia data found in WWW. Documents found in WWW contain lots of images and sounds, and the contents of those sounds and images are roughly described by the text around them. We suggest a simple but effective method to utilize those surrounding texts.

2. Cliparts in WWW

By *cliparts* we mean those data such as image, sounds and moving pictures that are found in the hyper-documents of WWW. Following table shows some key-facts about the cliparts in WWW.

<table>
<tr><td colspan="2">Number of Sample Documents</td><td colspan="4">28,100</td></tr>
<tr><td colspan="2">Number of Cliparts Extracted</td><td colspan="4">131,400</td></tr>
<tr><td rowspan="2">Images</td><td rowspan="2">GIF, JPG, JPEG, BMP, TIF, TIFF, XBM, MAP, ICO, IMG, MAP</td><td colspan="4">Total: 130,400</td></tr>
<tr><td>GIF</td><td>JPG</td><td colspan="2">Etc.</td></tr>
<tr><td></td><td></td><td>97,000</td><td>31,100</td><td colspan="2">2,300</td></tr>
<tr><td rowspan="3">Sounds</td><td rowspan="3">AU, M3U, MID, WAV</td><td colspan="4">Total: 2,700</td></tr>
<tr><td colspan="2">AU</td><td colspan="2">Etc.</td></tr>
<tr><td colspan="2">1,200</td><td colspan="2">1,500</td></tr>
<tr><td>Moving Pictures</td><td>AVI, MP3, MOV, MPEG, MPG</td><td colspan="4">200</td></tr>
</table>

Table 1. Number of clipart data in WWW documents.

Roughly speaking, an average Web document contains about three pictures and most of the pictures are in GIF and JPG format and the number of moving pictures and sounds are very small.

Web documents in HTML format contain some sort of structural information like <head>, <title> or <body> tags. Unlike in strict SGML format, those tags are not strictly hierarchical and many tags are used for specifying the attributes of text, not for specifying the structure of a document. Fortunately however, all the filenames (or URL) of the multimedia data are easily extractable by analyzing the HTML tags. Most of the pictures are found in <img> tag or <a> tag and many sounds are found in <embed> tag. Moving pictures are found in between <a> and . It's recommended to use "ALT" feature within <img> to describe the content of the <img> tag, but few people use ALT tags in their HTML files. Many HTML pages are produced by "*copy-and-paste*", in result, lots of pages have wrong titles. Anchor text in between <a> tag and tag seems to play the same role as the "ALT" feature in <img> tag.

3. Automatic Description Generation

The main point of this study is to achieve an effective way of getting descriptive texts (string/sentence/phrase) from HTML documents for each multimedia item. Once the descriptive texts are found, it's easy to apply traditional text-based techniques of information retrieval. Figure 1 depicts a Web document and each part marked in the picture is to be used in automatic generation of description. Description will be used for two different purposes: one for searching the multimedia clipart, and the other for displaying the summary about the data in the search result.

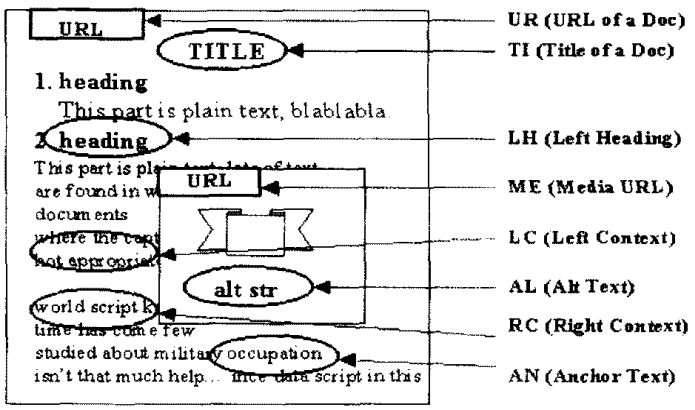

Figure 1. Surrounding Source of Description Keywords

As you see in Figure 1, a picture is surrounded by many text items and we extracted eight kinds of texts. Here are those eight fields:

Description (X) = {AL, AN, LC, LH, ME, RC, TI, UR}
where
AL: Alt string of X within <img> tag
AN: Anchor text of X in between <a> and tag
LC: Left Context, text (up to 10 words) just before the <img> or <embed> tag
LH: Left Heading, the heading item just before the <img> or <embed> tag
ME: URL string of the clipart
RC: Right Context, text (up to 10 words) after the <img> or <embed> tag
TI: Title of the HTML file, which contains this clipart
UR: URL of the HTML file, which contains this clipart

We can simply gather all the eight fields as the clues of searching, but simple-minded search is very imprecise and the text gathered from those eight fields are very noisy to be used as the summary string.

Let us see the characteristics of each fields, first. In many cases, HTML documents are not long enough and the position of the embedded data within an HTML may vary, so it is not always possible to extract all the eight surrounding information from HTML texts. Following table indicates the availability of each fields.

The Number of Embedded Clip-art	131,400
AL (Alt Text)	33,800
AN (Anchor Text)	5,400
LC (Left Context)	128,700
LH (Left Heading)	126,200
ME (Media URL)	131,400
RC (Right Context)	58,100
TI (Title String)	127,200
UR (URL of the Document)	131,400

Table 2. Availability of Surrounding Information

Guidebooks on HTML recommend to use ALT text field to describe the content of an image or sound, and ALT field is presumably the best candidate for the description of a multimedia data. It turned out, however, that only one forth of the total cliparts have ALT field and even the ALT field are given, they are very noisy to be used as a certain clue for searching or to be used as the summary string in search result. Rather, it proved that the filename or directory name found in UR and ME field are much more useful for the retrieval clues.

4. Experiments

Following table shows the retrieval precision for each field for 10 query string over 28,100 HTML documents.

AN	AL	ME	LH	LC	RC	TI	UR
0.78	0.48	0.59	0.43	0.36	0.46	0.44	0.25

Table 3. Precision of each description field

According to this result, Anchor Text (AN) and Media URL (ME, Filename of the clipart) is the best clue for searching pictures and sounds. During this experiment, we have identified following to two factors that cause significant amount of errors. And we simply *filtered out* these two factors to get a significantly increased precision.

a. Iconic pictures are the main cause of errors. Normally Web documents contain links to some relevant pages and the links are displayed in some iconic pictures, indicating "return to homepage", "previous page", "next page", etc. Those iconic pictures are of no use in most retrieval. This error is very easily solved by adopting a short list of "stop words", i.e., do not retrieve items that contain following words in the ME field (in the media file name).

   Stop word list = {home, prev, next, return, back, line, ball, bullet, point, new, link, mail, button, last}

   This stop word list was very effective to block out all the "iconic figures" from retrieval. This increases the precision remarkably because those "iconic figures" is the largest part of pictures in WWW.

b. Hostnames found in ME field and UR field are very often meaningful words, which cause unexpected retrieval. For example, if somebody wants to search all the pictures of "Jupiter" in WWW, certainly he is not searching the pictures on all the hosts named, "jupiter.domain.name". So we can simply rule out the address part of ME and UR field from being searched to get more precise result.

Following graph shows the effectiveness of filtering these two factors.

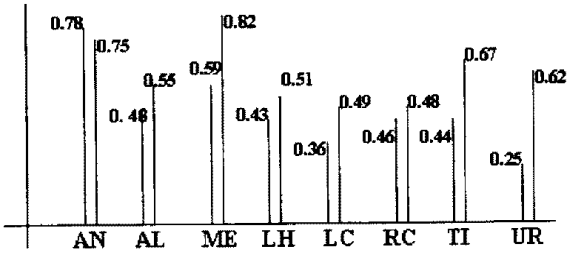

Figure 2. The Effectiveness Filtering.

This filtering increased the average precision from 0.51 to 0.72, which is about 41% gain of precision. Three fields ME, UR, and TI gain the precision after the filtering. According to this result, we can conclude that the URL strings (both from ME and UR fields) are the most effective clues for searching. The AN field is also a precise clue, but it is very low in availability according to the Table 2.

5. Issues and Discussion

We are applying the result of this study to the public service, *InfoDetective*, which is widely used in Korean domain. During the application to the public service, we have found that following factors are also important to the public users.

a. Search Clues vs. Summary Text: According to the result in Figure 2, the most useful search clue is the URL string found in UR and ME field. But when we are to display the search result to the users, we need some descriptive text as the summary, in order to let users decide whether to 'click' (really jump into the multimedia data, which is very time-consuming for the network). Users prefer to click some better explained items in the result. Originally we thought that automatically extracted AL field or AN field are useful for this summary string, but it turned out that AN field is very low in availability and AL field is very noisy, to be used as the summary string. Rather the TI (Text of the HTML file) is more useful for that purpose. So we need some technique to summarize the context around the picture or sound data, which evidently requires higher level of natural language understanding/processing.

b. Automatic filtering of Useless Data: We suggested a simple filtering in this paper, but during the development we found that the size of a media file, and the filename would also be useful for removing duplicated pictures. Because the Internet is vastly open to anyone and every body publishes his own Web pages, using freely gathered pictures, the amount of duplication is tremendous and the detection of the duplication would be easily done by filenames and sizes.

c. Automatic Classification: In many cases, people search a group of pictures at the same time because the freely achieved picture is more inappropriate for his own use. So if the search engine provides a classified search result, users may easily compare the classified data to select the best item for his own use.

6. Conclusion

In this paper we presented a practical approach of indexing and searching cliparts found in Web documents, based on the automatically gathered textual description of each clip-art. We also suggested a simple-minded filtering technique that effectively increases the retrieval precision.

We have seen that by this relatively simple processing, we can achieve relatively high precision. However, we haven't yet found a good way of getting the summary string of a clipart, which is essential for public service. In order to get summary string of each clipart, a more sophisticated, linguistic analysis would be necessary.

Refrence

[1] M. J. Swain and D. H. Ballard, "*Color Indexing*," International Journal of Computer Vision, Vol. 7, No. 1, pp. 11-32, 1991.

[2] M. K. Mandal, T. Aboulnast and S. Panchanathan, "*Image Indexing Using Movements and Wavelets*," IEEE Transactions on Consumer Electronics, Vol. 42, No.3, pp.557-565, August, 1996.

[3] M.-S. Park, B.-H. Song, S.-H. Lee, "*Design and Implementation of WWW Image Search Engine*", Proc. Fall Conf. KISS, Vol.23, No.2, pp.155-158, 1996.

[4] E. Wold, T. Blum, D. Keislar, J. Wheaton, "*Content-based Classification, Search and Retrieval of Audio*," IEEE Multimedia, Vol.3, No.3, pp.27-36, Fall 1996.

[5] B. Feiten and S. Gunszel, "*Automatic Indexing of a Sound Database Using Self-Organizing Neural Nets*," Computer Music J., Vol.18, No.3, pp.53-65, Summer 1994.