A Multi-Language Search Scheme using a Multithread Processing for Yahoo Image Search

Anucha Tungkasthan, Sarayut Intarasema, and Wichian Premchaisawadi

***Abstract*— This paper presents a multi-languages search scheme for Yahoo Image Search. The methodology for searching is to retrieve a wide variety of images annotated in different languages websites from yahoo image search engine. The proposed methodology consists of two main parts as follows: 1) multi-language translation and 2) multi-crawling, retrieving results with images annotated in different languages. The first part is used to accept user's query and to perform cross-language translation into different languages. The second part, a given query is sent to Yahoo BOSS API to perform image retrieval task. The multithreading processes were designed and implemented to resolve the problem of crossing language queries and retrieving images processes. It can greatly reduce lot of time and effort for the search. The experiments on diverse queries on Yahoo Images search have shown that the proposed scheme can improve the images results for Non-English keyword effectively.**

I. Introduction

There is high competition in the image search engine developers, such as Yahoo, Google and the latest one, namely Bing [1], which is introduced by Microsoft. Bing can now grab market share and move in the top two of the search engine developers. There are several popular image search engines in the market. Some of them use the shared image database from free-access social networking websites such as Facebook and Flickr.The image search engine developer emphasis more on a variety of searching techniques in order to search images that match the needs of users. For example, Google has released the Google Similar Images [2] that allows user to search for images using pictures rather than words. A user can click on "Similar images" link under an image to find other images that are similar it. It is an experimental service from Google Labs [3] that lets user find images that are similar to an image he or she selected.

Basically, Text-based image search engines such as Google Images have achieved great success on exploiting text information to index and retrieve large-scale online image collections [4]. There are several techniques to improve the efficiency of the search. For example, Yahoo uses the most frequently searched images for re-ranking of the results. The existing techniques of image search may not be enough to satisfy the user needs. The number of pictures on the internet increase rapidly due to higher efficiency technology of hardware and software for managing storages. There is the number of pictures denoted in different languages from various sources on the internet that are not retrieved. Some of them may exactly match with the user needs. The user needs to search the pictures from on-line sources or internet because it is fast and convenient. Generally, most users use only one specific language keyword for the search. The image search engines typically use the text around the picture to create the indexes for that picture. Thus, there are the hundreds or thousands pictures that have similar meaning to that keyword but they are annotated in different languages. Therefore, these pictures are not retrieved. Practically, users need only the pictures that match exactly with the keyword they search. The user mostly focuses on the pictures but not the contents or text around the pictures or in the webpage. Thus this paper presents an automated multi-language search engine that helps user search the variety of pictures in different languages using just one keyword in any languages. The search engine transforms the keyword to other provided languages and searches for the picture automatically. One of the problems of this technique is that it takes a lot of time for crossing language queries and retrieving images. This paper overcomes the problem by using a multithread processing in crossing language queries and retrieving images processes. It can greatly reduce lot of time and effort for the search.

A. Tungkasthan, Graduate School of Information Technology in Business, Siam University, Bangkok 10163, Thailand (email: aimdala@hotmail.com).

S. Intarasema, Graduate School of Information Technology in Business, Siam University, Bangkok 10163, Thailand (email: sarayut100@gmail.com).

W. Premchaisawadi, Graduate School of Information Technology in Business, Siam University, Bangkok 10163, Thailand (email: wichian@siam.edu)

II. Related work

In this section, we review the most relevant work according to key issues for online multi-languages image search as follow:

FlickLing [5], a multilingual search interface for Flickr designed and implemented for the CLEF 2008 interactive task. It is the latest version of CLEF (Cross-Language Evaluation Forum) community that was developed and published between 2004 and 2007 [6-9], an annual evaluation exercise for Multilingual Information Access systems. FlickLing consists of two search modes (mono and multilingual) which allow retrieving Flickr images annotated in different languages. From a given query, FlickLing is able to automatically translate it into several languages (remembering the user's preferred term translations) and offer the user mechanisms to refine the query and improve the translations provided by the system. FlickLing very similar with our work, however, the goal of FlickLing was to

collect large usage logs that reflect users' behavior when facing a multilingual search task rather than focus on the result of image. It support only five languages

PanImages [10] are developed by the University of Washington that allows users to search for images in their native language and receive far more results than with traditional search, supports over 300 languages. Users simply type in the query and the language they are performed and see a result set that includes translations. Clicking on a result returns Google Image and Flickr search results for that term.

III. Supporting Technology

1) Google Translation API [11]: Google has released API for language detection and translation. The API helps developers automatically translate content in their applications. Users on these sites will have an easier time communicating across lingual boundaries. Google translation API consists of automatic language detection and translation service. The "Detect language" option automatically determines the language of the text users are translating. The accuracy of the automatic language detection increases with the amount of text entered. To do a translation, the user simply specify the text he or she would like to translate, the language he or she is translating from, the language he or she is translating to, and then click on a translation button.

2) BOSS API (Build your Own Search Service) [12]: is Yahoo!'s open search web services platform. The goal of BOSS is simple: to foster innovation in the search industry. Developers, start-ups, and large Internet companies can use BOSS to build and launch web-scale search products that utilize the entire Yahoo! Search index. BOSS gives you access to Yahoo!'s investments in crawling and indexing, ranking and relevancy algorithms, and powerful infrastructure. By combining your unique assets and ideas with our search technology assets, BOSS is a platform for the next generation of search innovation, serving hundreds of millions of users across the Web. The figure 2 shows the XML result of query by "apple".

3) *Multithreading [13]:* A thread is a sequence of instructions that can execute in parallel with other threads. This paper uses a multi-thread processing to help increase the performance in language translation and downloading images processes. To measure the performance of multithreading, the speedup is calculated. The formula is shown below.

$$S_p = \frac{T_1}{T_p}$$

, when S_p = speedup, p = number of the threads, T_1 = time in single thread, T_p = Total time of all threads

IV. Design of Multilingual Image Search

Multilingual search web application using a multithread processing for Yahoo Image Search has the following two main components. Figure 1 presents the scheme of multi-language of image search, multi-language translation and multi-crawling.

1) Multi-language Translation: This paper uses Google's translation API to translate the keyword. The keyword is automatically examined and translated into four different languages (Arabic, Korean, Spanish and English) by default. The Google's translation API allows the user to translate only one specific language at a time. In order to translate into different languages, there should be several requests to the Google's translation API. It causes a slowdown in the multi-language translation process. Therefore, multithreading is considered to resolve this problem. Multiple threads can be run in parallel for sending requests to API and receiving the responses.

2) Multi-crawling: The keyword is translated to different languages from previous step. The keywords in different languages are now used for querying images form yahoo BOSS API. BOSS gives access to Yahoo!'s investments in crawling and indexing, ranking and relevancy algorithms, and powerful infrastructure. However, BOSS returns only XML results containing with necessary information of all images such as date, file name, format and links for downloading the pictures (shows in figure 2). The program reads the returned XML file and starts downloading from original image sources. The most important problem we encounter is that access time of each source is different. It causes the bottle neck from the source that response slowly. Thus multithreading is also considered to resolve this problem in this part. The main thread is used to control and distribute jobs to other threads. All of them run in parallel for downloading images.

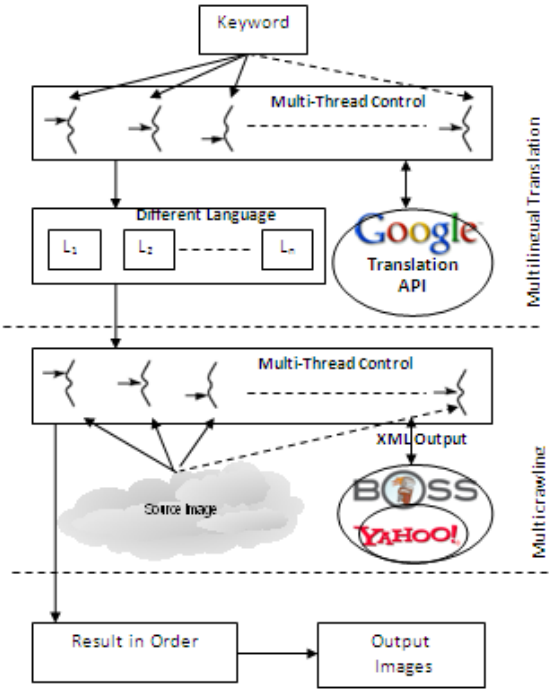

Fig. 1. The scheme of multi-language of image search

<ysearchresponse responsecode="200">
<prevpage>
/ysearch/images/v1/apple?format=xml&count=1&appid=PRG2QiflkY1cA3fLzhgwlVlr48rjlVUmaiWa&start=0
</prevpage>
<nextpage>
/ysearch/images/v1/apple?format=xml&count=1&appid=PRG2QiflkY1cA3fLzhgwlVlr48rjlVUmaiWa&start=2
</nextpage>
<resultset_images count="1" start="1" totalhits="16836085" deephits="16836085">
<result>
<abstract>
Tras muchos años de disputas y acuerdos de concordia Apple Computer y Apple Corps la discografía que gestiona los derechos de Los Beatles podrían estar a punto de llegar a un acuerdo
</abstract>
<clickurl>
http://lrd.yahooapis.com/
_ylc=X3oDMTRjMHJ2aTlsBF9TAzIwMjMxNTI3MDIEYXBwaWQDUFJHMlFpZklrWTFjQTNmTHpoZ3dsVmxyN
DhyamxWVW1haVdhBGNsaWVudANib3NzBHNlcnZpY2UDQk9TUwRzbGsDdGl0bGUEc3JjcHZpZANCYm43
SjBnZUF1MTdvY3JoeGZKLndlQVpkeTQzU2tvellHVUFDZG5k/SIG=1235rt557/**http%3A//
es.appleweblog.com/wp-content/uploads/2006/11/apple.jpg
</clickurl>
<date>2006/11/27</date>
<filename>apple jpg</filename>
<format>jpeg</format>
<height>330</height>
<mimetype>image/jpeg</mimetype>
<refererclickurl>
http://lrd.yahooapis.com/
_ylc=X3oDMTRjMHJ2aTlsBF9TAzIwMjMxNTI3MDIEYXBwaWQDUFJHMlFpZklrWTFjQTNmTHpoZ3dsVmxyN
DhyamxWVW1haVdhBGNsaWVudANib3NzBHNlcnZpY2UDQk9TUwRzbGsDdGl0bGUEc3JjcHZpZANCYm43
SjBnZUF1MTdvY3JoeGZKLndlQVpkeTQzU2tvellHVUFDZG5k/SIG=12sasnpj5/**http%3A//
es.appleweblog.com/2006/11/27/apple-y-los-beatles-a-punto-de-llegar-a-un-acuerdo
</refererclickurl>
<refererurl>
http://es.appleweblog.com/2006/11/27/apple-y-los-beatles-a-punto-de-llegar-a-un-acuerdo
</refererurl>
<size>19100</size>
<thumbnail_height>102</thumbnail_height>
<thumbnail_url>http://thm-a01.yimg.com/image/2f31a32a15522fac</thumbnail_url>
<thumbnail_width>140</thumbnail_width>
<title>apple jpg</title>
<url>
http://es.appleweblog.com/wp-content/uploads/2006/11/apple.jpg
</url>
<width>450</width>
</result>
</resultset_images>
</ysearchresponse>

Fig. 2. XML file results from yahoo boss

V. An Implementation and Experimental Result

The techniques mentioned in earlier chapters are developed. The experimental testing is set on the Windows environment with 100Mbps LAN, Pentium IV 2.8 GH_Z, and 1 GB RAM system. The scheme is developed by using Microsoft .NET and implemented based on Yahoo' API and Google' API. The multithreading processes substantially increase the efficiency of the tool. The results of the image search are shown in figure 3, 4, and 5.

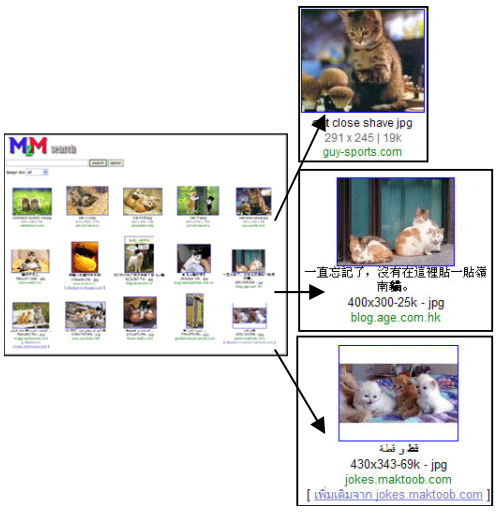

Fig. 3. Our online system for multilingual images search from Yahoo Images, where the keyword "cat" is used and image results (languages in order)

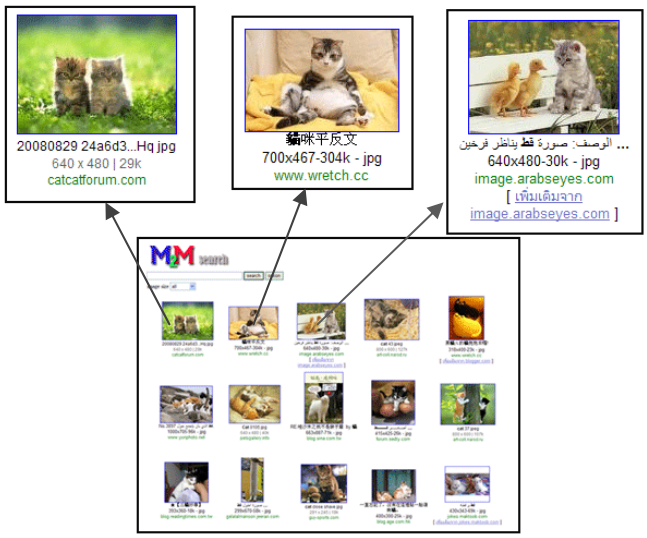

Fig. 4. Image results (languages not in order)

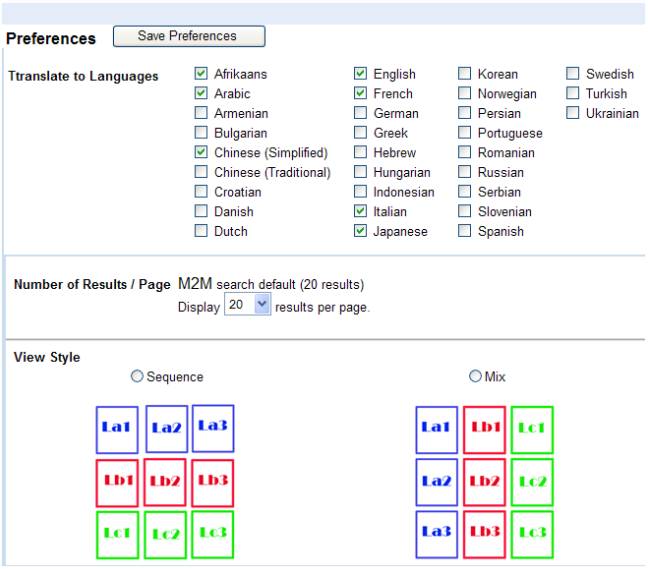

Fig. 5. Options for advanced search

To evaluate the performance of the proposed scheme, experiments were carried out using online database of Yahoo with various keywords in different languages. The system performance was tested in term of both retrieval accuracy and processing time. The first one was measured in terms of the relevant image of retrieving results. Let n_c and n_m be the number of correct and missing results, respectively. The retrieval accuracy for keyword query is applied from [14] as shows in equation (1).

$$R_q = \frac{n_c}{n_c+n_m} \qquad (1)$$

where k is the number of keyword for query. Then, the average rank is modified as (2).

$$AV\ R_q = \frac{1}{k}\sum_{i=1}^{k}\frac{n_c}{n_c + n_m} \qquad (2)$$

The retrieval accuracy on both single language and multi-languages is considered. Ten independent keywords in four languages (Arabic, Korean, Spanish and English) were used in this experiment, some results as shown in table 1. In term of processing time, a single-thread and multi-thread running results are measured. Even though there are large number of images and several languages in querying, speed is close to linear growth.

From the experimental results, one can observe that our proposed system can show a wide variety of images. We exploit the available APIs such as Google translation and yahoo BOSS for multi-languages and multi-crawling. The multithreading processes designed and implemented in the application to overcome the problem in crossing language queries and retrieving images processes. It can greatly reduce lot of time and effort for the search. The experiments on diverse queries on Yahoo images have shown that our proposed scheme can improve the image's results for Non-English keyword effectively.

TABLE 1
Retrieval Accuracy Rate of the First 20 Searched Images

No.	English		Spanish		Arabic		Korean		R_{q2}
	w	R_{q1}	w	R_{q1}	w	R_{q1}	w	R_{q1}	
1	grape	1	uva	0.7	عنب	0.5	포도	0.8	0.8
2	telephone	0.95	teléfono	0.65	الهاتف	0.5	전화	0.75	0.8
3	snake	0.85	serpiente	0.7	ثعبان	0.8	뱀	0.55	0.9
4	snooker	0.95	billar	0.85	السنوكر	0.4	스누커	0.85	0.85
5	turtle	0.95	tortuga	0.7	سلحفاة	0.7	거북이	0.35	0.9
6	watermelon	1	melón	0.05	البطيخ	0.45	수박	0.7	0.7
7	monkey	0.9	mono	0.25	القرد	0.5	원숭이	0.9	0.8
8	boat	0.95	barco	0.7	قارب	0.9	보트	0.65	0.9
9	key	0.95	llave	0.45	مفتاح	0.4	열쇠	0.65	0.75
10	cat	1	gato	1	قط	0.9	고양이	0.6	1
Avg. R_q		0.95		0.6		0.6		0.68	0.84

w= keyword, R_{q1} = Retrieval Accuracy of single language R_{q2}= Retrieval Accuracy of multi-language

Time-oriented performances metrics will be tested by calculations determine the speedup comparison to working between single thread and multi-thread. The results as shown in table 2 and 3.

TABLE 2
Average Time (in second.) and Speedup of Translations

Thread Type	Number of Languages		
	3	5	10
Single Thread	1.75	2.76	4.48
Multithread	0.61	0.65	0.68
Speedup	2.87	4.25	6.59

.

TABLE 3
Average Time (in second.) and speedup of Image Downloading

Thread Type	Image Quantity		
	20	50	100
Single Thread	92.5	223.17	480.02
Multithread	4.21	8.9	18.51
Speedup	21.97	25.1	25.93

VI. Conclusion

A Multi-languages search scheme using a multithread processing for Yahoo Image Search is purposed. For a given text-based image query, our system can automatically translate to different language using multithread processing.

References

[1] http://www.bing.com/
[2] http://similar-images.googlelabs.com/
[3] http://www.googlelabs.com/
[4] Y. Gao, J. Fan, H. Luo, and S. Satoh, "A Novel Approach for Filtering Junk Images from Google Search Results," *Springer Advances in Multimedia Modeling*, Volume 4903, 2008, pp. 1-12.
[5] V. Peinado, J. Artiles, J. Gonzalo, E. Barker, F. Lopez-Ostenero, (2008). "FlickLing: a Multilingual Search Interface for Flickr," Interactive track of CLEF (Cross-Language Evaluation Forum). Available: http://www.clef-campaign.org/.
[6] F.Florea, A.Rogozan, V.Cornea, A.Bensrhair and S.Darmoni, "MedIC at ImageCLEF 2006: Automatic Image Categorization and Annotation Using Combined Visual Representations ," *Springer Evaluation of Multilingual and Multi-modal Information Retrieval*, Volume 4730, 2007, pp. 670–677.
[7] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. *Springer Lecture Notes in Computer Science*, Volume 3491, 2005, pp. 371–391.
[8] M. Sanderson, P. Clough, C. Paterson, and W. T. Lo, "Measuring a Cross Language Image Retrieval System," *Springer Advances in Information Retrieval*, Volume 2997, 2004, pp. 353–363.
[9] T. Deselaers, T. M. Deserno, H. Müller "Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion," *Elsevier Science Inc., Pattern Recognition Letters*, Volume 29, Issue 15, 2008, pp 1988-1995.
[10] S. M. Colowick, (2008, March) "Multilingual search with PanImages," MultiLingual Computing, Volume 19 Issue 2, Available: www.multilingual.com.
[11] http://code.google.com/intl/th/apis/ajaxlanguage/
[12] http://developer.yahoo.com/search/boss/

[13] K. Manjunathachri, "A Parallel Processing Approach to Image Processing Application Using simultaneous Multi Threading," *Asian Journal of Information Technology*, 2006, pp. 1137-1141.

[14] H. Young Lee, H. K. Lee, and Y. H. Ha, *Senior Member, IEEE,* "Spatial Color Descriptor for Image Retrieval and Video Segmentation," IEEE Trans. on Multimedia, VIL. 5, NO. 3, 2003, pp.358-367.