8-2010

Benchmarking Web-Based Image Retrieval

Ryan Fendley
Wright State Research Institute, ryan.fendley@wright.edu

Phani Kidambi
Wright State Research Institute & Department of Biomedical, Industrial & Human Factors Engineering, phani.kidambi@wright.edu

Benchmarking Web-Based Image Retrieval

Phani Kidambi
Wright State Research Institute & Department of Biomedical, Industrial & Human Factors Engineering
phani.kidambi@wright.edu

Dr S. Narayanan
Wright State Research Institute & Department of Biomedical, Industrial & Human Factors Engineering
s.narayanan@wright.edu

Ryan Fendley
Wright State Research Institute
ryan.fendley@wright.edu

ABSTRACT (REQUIRED)

An explosion of digital photography technologies that permit quick and easy uploading of any image to the web, coupled with the proliferation of personal, recreational users of the internet over the past several years have resulted in millions of images being uploaded on the World Wide Web every day. Most of the uploaded images are not readily accessible as they are not organized so as to allow efficient searching, retrieval, and ultimately browsing. Currently major commercial search engines utilize a process known as Annotation Based Image Retrieval to execute search requests focused on retrieving an image. Despite the fact that the information sought is an image, the ABIR technique primarily relies on textual information associated with an image to complete the search and retrieval process. Using the game of cricket as the domain, this article compares the performance of three commonly used search engines for image retrieval: Google, Yahoo and MSN Live. Factors used for the evaluation of these search engines include query types, number of images retrieved, and the type of search engine. Results of the empirical evaluation show that while the Google search engine performed better than Yahoo and MSN Live in situations where there is no refiner, the performance of all three search engines dropped drastically when a refiner was added. Further research is needed to overcome the problems of manual annotation embodied in the annotation-based image retrieval problem.

Keywords (Required)

Image Search Engine, Performance Evaluation, Benchmark, Query Types, Annotation Based Image Retrieval (ABIR), Precision, Retrieval Length.

INTRODUCTION

The Internet contains a nearly limitless archive of images. These images are incredibly diverse in their content and include representations of nearly every global social interest, from the popular types such as politics, sports, and entertainment to the specialized issues such as animal breeds, family genealogies, and restored vintage automobiles. In this vast archive, retrieving a specific image representative of a certain person, place, or event can be challenging.

Image retrieval is the process of searching for an image or a particular set of images from the database of images. As the number of images on the internet increase, those conducting image searches continue to face the problem of image overload. Image overload is the condition where the available images far exceed the user's ability to review them. That is, the user is inundated with a deluge of images – only some of which are relevant to the user (Kidambi and Narayanan, 2008).

Currently major commercial search engines utilize a process known as Annotation Based Image Retrieval (ABIR) to execute search requests focused on retrieving an image. Despite the fact that the information sought is an image, the ABIR technique primarily relies on textual information associated with an image to complete the search and retrieval process. This text-based approach to image retrieval can be traced back to the late1970's, when both the number of images and the number of users

posting and searching for images were much smaller than today. Then as now, the images are typically manually annotated by a human analyst. It is this annotation that is evaluated during a search.

To complete a search, an ABIR driven engine employs a number of standard steps (Yates and Neto, 1999). Images are retrieved by evaluating the vector of word frequencies in their annotations and returning the images with the closest vectors. A relevancy ranking is calculated by evaluating the degree of the match of the order and separation of the words that exists between the search terms and the annotation of each individual image (Witten et al., 1999). Thus, even though the user is searching for images, the images that are retrieved are actually determined by the textual annotation. This annotation usually consists of the manually assigned keywords or the text associated with the images such as captions.

While a significant body of research exists evaluating textual information retrieval processes (Kuralenok and Nekrestyanov, 2002; TREC, 1992), few research efforts have focused on evaluating image retrieval on the Internet. This article describes a benchmarking study which evaluates the effectiveness of three popular search engines in executing image based searches. The domain selected to assess their performance is the game of cricket.

RELATED WORK

Evaluating the effectiveness of information retrieval is important but challenging. Ironically, even though "evaluation" is used often in the literature of information and image retrieval, there is not a single accepted definition for this term in this context. Most researchers' adopt the definition of evaluation posited by Hernon et al (Hernon et al., 1990) which states that evaluation is

> " the process of identifying and collecting data about specific services
> or activities, establishing criteria by which their success can be assessed,
> and determining both the quality of the service or activity and the
> degree to which the service or activity accomplishes stated goals
> and objectives."

Meadow et al (Meadow et al., 1999) classified information retrieval measures into two categories: evaluation of performance (descriptive of what happens during the use of the information retrieval system) and evaluation of outcome (descriptive of the results obtained). Hersh (Hersh, 1995) also classified evaluation into two categories, although different from those proposed by Meadow, Hersh identifies: macro evaluation (investigates information retrieval system as a whole and its overall benefit) and micro evaluation (investigates different components of the system and their impact on the performance in a controlled setting). Lancaster et al (Lancaster, 1993) defined three levels of evaluation. The first level evaluates the effectiveness of the system, the second level evaluates the cost effectiveness and the third level evaluates cost benefits of the system. While Smith (Smith, 1998) proposed several measures for image retrieval evaluation including precision, recall, fallout and F-measure (F-measure = 2 (Precision * Recall) / (Precision + Recall)). Finally, Cooper (Cooper, 1968) suggests Expected Search Length (ESL) as an alternative to recall and precision. ESL measures the number of unwanted documents the user can expect to examine before finding the desired number of relevant documents.

Though the research literature contains numerous studies on the evaluation metrics for information and image retrieval, there is a dearth of literature benchmarking the performance of image search engines. This paper begins to fill this information gap, employing a systematic approach to evaluate search engines based on a number of independent factors including query types, number of images retrieved and the type of search engine. The remainder of this paper discusses the experiment that has been performed, the results, and their implications.

METHODOLOGY

Research studies (Cakir et al., 2008; Smith, 1998), have showed that the quality of images retrieved are typically a function of query formulation, type of search engine and the retrieval level. These independent factors and their levels are illustrated in the Ishikawa diagram illustrated in Figure 1. In this experiment three search engines: Google, Yahoo, and MSN Live are evaluated for varying query types and retrieval levels. Details on the query types and query levels are provided below.

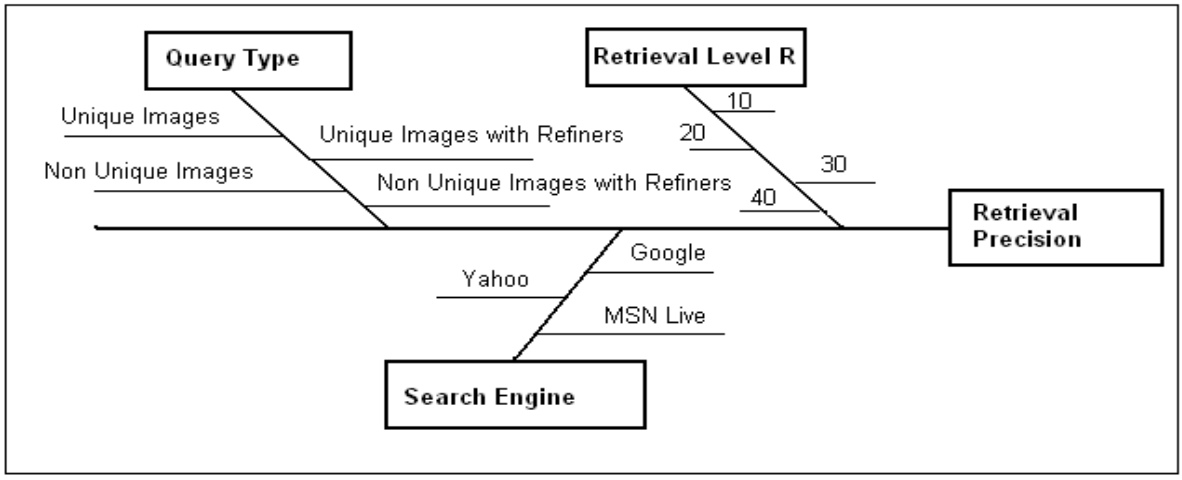

Figure 1 - Ishikawa diagram of independent factors used in this study

Query Types

Broder (Broder, 2002), proposed a three-pronged approach of web searching types for text retrieval: navigational, informational and transactional. Navigational searches are those where the user intends to find a specific website. Informational searches intend to find some information assumed to be present on one or more web pages. Transactional searches perform some web mediated activity, i.e., the purpose is to reach a site where further interactions will happen. Unfortunately, Broder's query types cannot be easily extended to image retrieval solutions since the end goal of the user in these two scenarios varies significantly.

Ensor & McGregor (Ensor and McGregor, 1993) summarized that the user search requests for images fall into four different categories:

i. Search for unique images – The property of uniqueness is a request for the visual representation of an entity where the desired entity (image) can be differentiated from every other occurrence of the same entity type. An example is – "find the image of Sachin Tendulkar".
ii. Search for unique images with refiners – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – "find the image of Sachin Tendulkar in 2004".
iii. Search for non – unique images – The property of non – uniqueness is a request for the visual representation of an entity where the desired entity (image) cannot be differentiated from every other occurrence of the same entity type. An example is – "find the images of Indian cricketers".
iv. Search for non – unique images with identifiers – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – "find images of Indians waving the Indian flag".

Search Engines

According to Nielsen's May 2009 ratings (Nielsen Search Rankings, 2009), Google's search engine accounts for 63.3% of the total searches on the internet, Yahoo's accounts for 17.3% and MSN Live accounts for 9.4%. These three search engines execute 90% of the total searches conducted on the internet. It is for this reason that they have been selected for comparison purposes in this study.

Experiment

To carry out the evaluation, a user-centered interpretative approach, based on the actual information-seeking behavior of real users (Smithson, 1994) was employed. Since this research is focused on the domain specific evaluation of the system, a subject matter expert in the domain of the game of cricket was used to evaluate the existing search engines. To evaluate the system, the expert was given a problem scenario for each of the query types; the expert then entered a query on the search engine in an undifferentiated, arbitrary search strategy based on his previous experience in search and retrieval of images. Once this is completed the expert attempted to discover the targeted image by selecting a subset of images to view from the set of all retrieved images. In this manner, the expert is, in fact, selecting these images based on his prediction of relevance (images relevant to the query) and evaluates the number of relevant images retried for the query.

Five queries for each query type are chosen. The queries consist of multi-word queries, related to the game of cricket, as shown in Table 1.

Query Types	**Queries**
Unique Images	MS Dhoni Vijay Bharadwaj Ricky Pointing Gary Sobers Abey Kuruvilla
Unique Images with refiners	Kapil Dev lifting World Cup Sreesanth + beamer + Pietersen Andy Flower + protest + black band Allan Donald + run out+ WC semifinal '99 Inzamam Ul Haq hitting a spectator + Canada
Non-Unique Images	Indian Cricket Players Surrey Cricket Team Ashes (Eng vs Aus) Cricket Players Huddle Rajastan Royals + IPL
Non-Unique Images with refiners	Victorious Indian Team + 20-20 WC SA chasing 438 Aus players with World Cup 2007 SL protesting against Aus + walking out of the ground Eng vs SA + WC stalled by rain + 1992

Table 1- Query Formulations for various Query types

The queries associated with “unique images” are all internationally known cricket players from different playing eras. The queries associated with “unique images with refiners” are related to a cricket player involved in a context such as winning a world cup. The queries associated with “non-unique images” are internationally well known cricket teams and finally the queries associated with “non-unique images with refiners” are internationally known cricket teams involved in a context similar to winning a world cup.

Each query is run on each of the three search engines. The first forty images retrieved in each search run are evaluated for relevance by the subject matter expert based on his knowledge. Relevance is determined in a binary manner. That is, the image is either deemed relevant or not relevant. In instances when the same image appears on different websites, these are evaluated as different images and each is evaluated for relevance. In instances where the same image appears in multiple

places on the same website, the first image is evaluated for relevance and the other images are considered not relevant. Additionally, if the image retrieved is not accessible due to technical difficulties in the site domain, the image is considered to be non-relevant. In order to obtain a stable performance measurement of image search engines, all the searches are performed within a short period of time (one hour) and the relevance of the images is decided by the subject matter expert.

RESULTS

Traditionally evaluation for information retrieval has been based on the effectiveness ratios of precision (proportion of retrieved documents that are relevant) and recall (proportion of the relevant documents that are retrieved) (Smith, 1998). Since the World Wide Web is growing constantly, obtaining an exact measure of recall requires knowledge of all relevant documents in the collection. Given the sheer volume of documents this is, for all practical purposes, impossible. Because of this, recall and any measures related to recall cannot be readily used for evaluation. This necessitates that the evaluation be based on the effectiveness ratios of precision. For purposes of this evaluation precision is defined as the number of relevant images retrieved to the total number of images retrieved. The search engines were evaluated based on the precision at a retrieval length R at R=10, 20, 30 and 40.

To check the adequacy of the factors thought, a factorial analysis was conducted and the results were analyzed using the analysis of variance (ANOVA) method. As previously discussed, the factors that were hypothesized to have a significant effect on the average precision of the retrieved results are Query Type, Search Engine and Retrieval Level. The independent variables for the factorial analysis and their levels are shown in Table 2.

Factor	Description	Level 1	Level 2	Level 3	Level 4
A	Query Type	Unique Images	Unique Images with Refiners	Non Unique Images	Non Unique Images with Refiners
B	Search Engine	Google	Yahoo	MSN Live	
C	Retrieval level R	10	20	30	40

Table 2- Independent Factors affecting quality of image retrieval and their levels

The response variable is the average precision at retrieval length R which is defined as the ratio of the relevant retrievals to the overall number of images retrieved.

Hypothesis

Null Hypothesis: H_0: There is no significant effect of Query Type, Search Engine or Retrieval level R on the precision of the retrieved results.

Alternate Hypothesis: H_1: There is significant effect of Query Type, Search Engine or Retrieval level R on the precision of the retrieved results.

Statistical Analysis

Data were collected for the 48 experimental trials of the 4 x 3 x 4 full factorial design that was run five times. Table 3 shows the ANOVA results obtained to check the accuracy of the model.

	ss	df	ms	f	p
average	386317.8873	1	386317.9		
A	106622.284	3	35540.76	55.70395	5.97836E-26
B	17902.46826	2	8951.234	14.0295	2.05665E-06
C	8543.665978	3	2847.889	4.463569	0.004684785
AB	3609.18176	6	601.5303	0.942794	0.465557359
AC	636.8045037	9	70.75606	0.110898	0.999405639
BC	41.88312333	6	6.980521	0.010941	0.999994074
ABC	925.57028	18	51.42057	0.080593	0.999999891
Error	122501.6533	192	638.0294		
Total	647101.3985	240			
model	138281.8579	47	2942.167	4.611334	2.40093E-14

Table 3- Analysis of Variance for various factors

At the 99 % confidence level, the ANOVA results show that there is a significant effect of the main effects, (A) Query Type, (B) Search Engine and (C) Retrieval Level R and there are no effects due to interactions between the main effects. The ANOVA results of the overall model (taking all the main effects and interactions into consideration) are also significant at the 99% confidence level.

The results clearly show that all the main effects are significant; hence we can further analyze the response variable.

Performance Evaluation

The performance of the search engines for various queries and retrieval levels is discussed in this section. The average precision of the retrieved images for Unique Images for different levels & search engines is tabulated in Table 4 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 2.

	Average Precision			
Search Engine	**R @ 10**	**R @ 20**	**R @ 30**	**R @ 40**
Google	0.76	0.69	0.61	0.57
Yahoo	0.68	0.59	0.56	0.52
Live	0.62	0.56	0.51	0.44

Table 4- Average Precision of retrieved images for Unique Images

Figure 2 clearly illustrates that Google has the best average precision at any cut-off point for unique images, followed by Yahoo and MSN Live respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased.

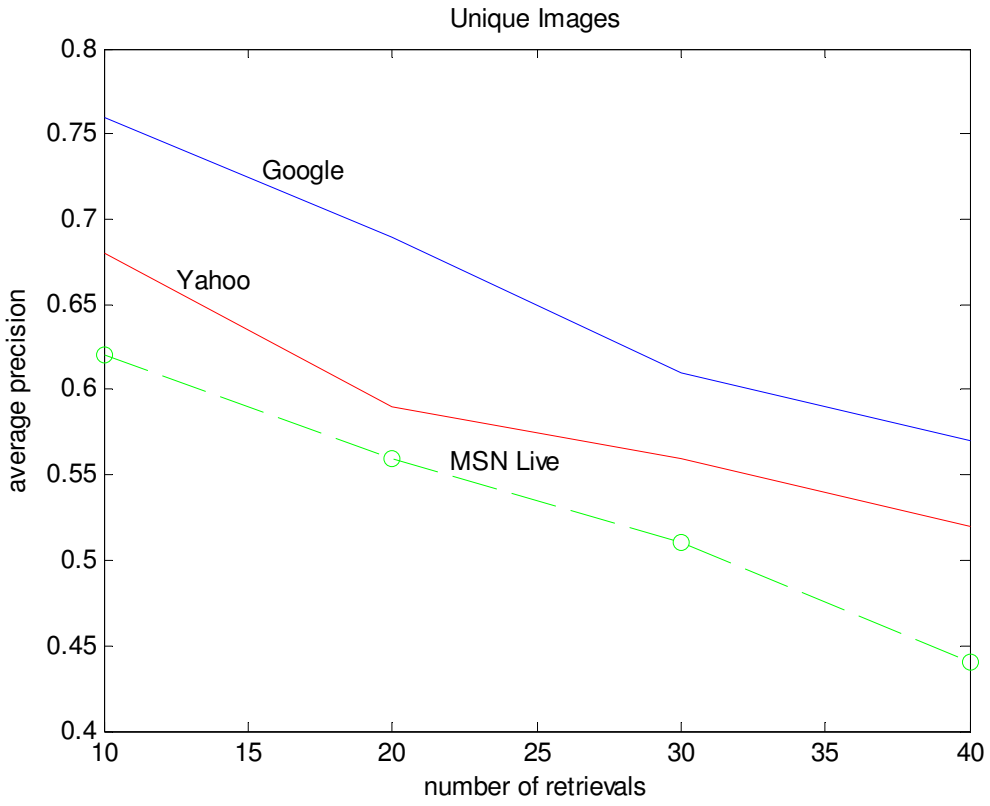

Figure 2- Average Precision of retrieved images for Unique Images

The average precision of the retrieved images for Unique Images with refiners for different levels and search engines is tabulated in Table 5 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 3.

	Average Precision			
Search Engine	**R @ 10**	**R @ 20**	**R @ 30**	**R @ 40**
Google	0.38	0.27	0.21	0.18
Yahoo	0.08	0.05	0.03	0.025
Live	0.2	0.14	0.09	0.07

Table 5- Average Precision of retrieved images for Unique Images with Refiners

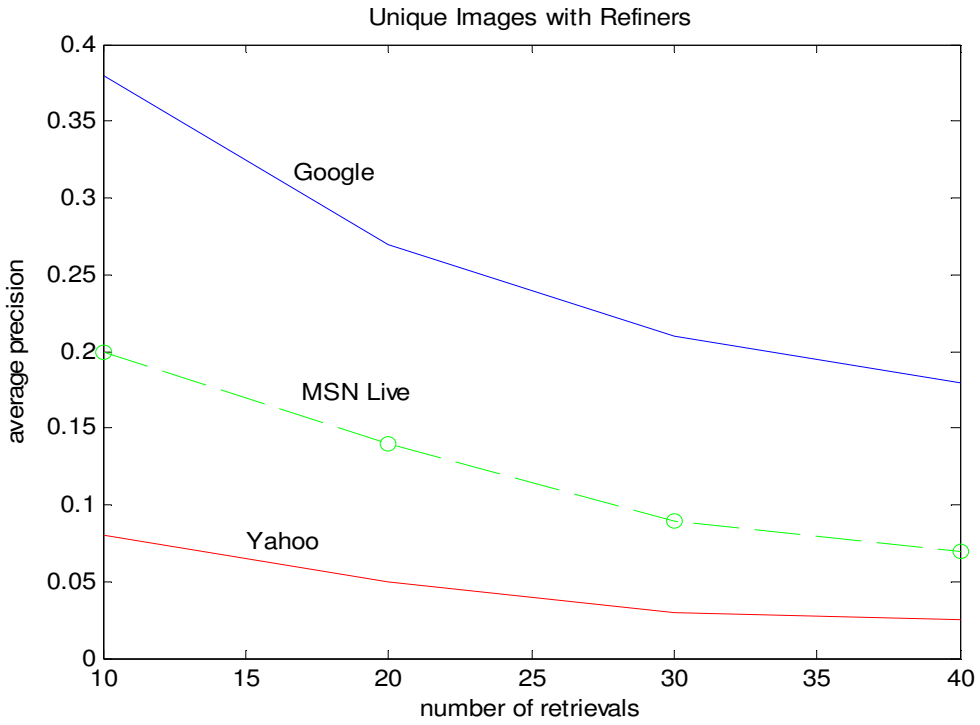

Figure 3- Average Precision of retrieved images for Unique Images with Refiners

For unique images with refiners, Google has the best average precision at any cut-off point, followed by MSN Live and Yahoo respectively. The precision has dropped off drastically as compared to the precision levels for unique images (without refiners).

The average precision of the retrieved images for Non-Unique Images for different levels & search engines is tabulated in Table 6 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 4.

	Average Precision			
Search Engine	**R @ 10**	**R @ 20**	**R @ 30**	**R @ 40**
Google	0.86	0.78	0.74	0.69
Yahoo	0.66	0.5	0.413	0.345
Live	0.72	0.66	0.6	0.525

Table 6- Average Precision of retrieved images for Non-Unique Images

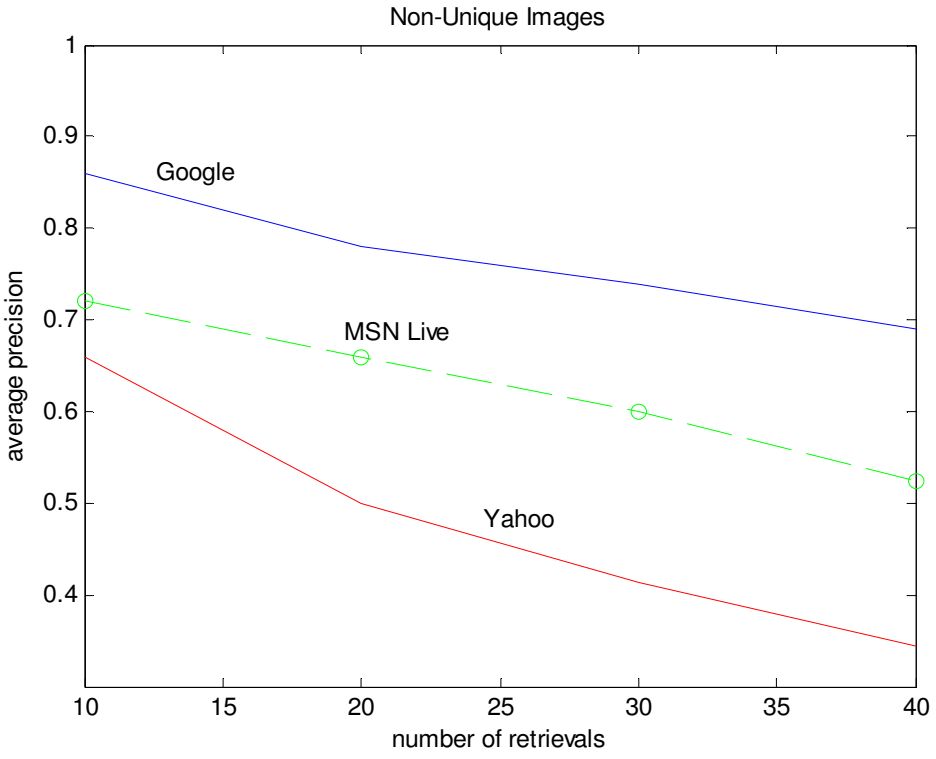

Figure 4- Average Precision of retrieved images for Non-Unique Images

Figure 4 illustrates that Google has the best average precision at any cut-off point for non-unique images, followed by MSN Live and Yahoo respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased. The average precision of the retrieved images for Non-Unique Images with refiners for different levels & search engines is tabulated in Table 7 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 5.

	Average Precision			
Search Engine	**R @ 10**	**R @ 20**	**R @ 30**	**R @ 40**
Google	0.42	0.41	0.39	0.355
Yahoo	0.24	0.17	0.113	0.085
Live	0.26	0.17	0.153	0.135

Table 7- Average Precision of retrieved images for Non-Unique Images with Refiners

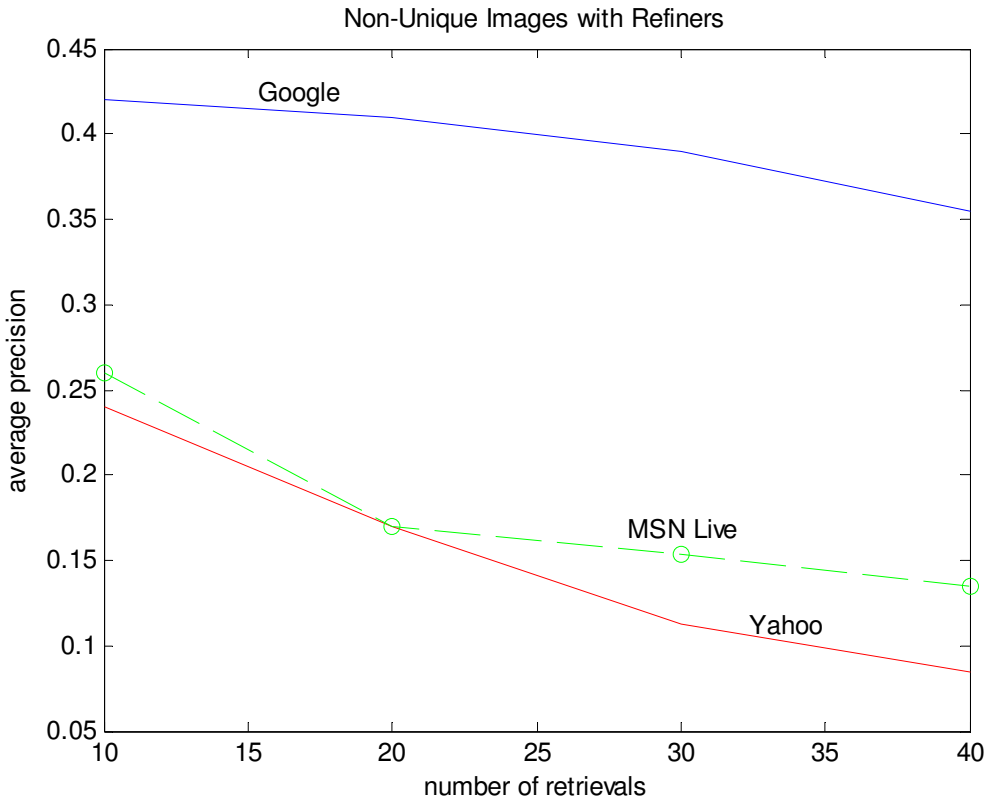

Figure 5- Average Precision of retrieved images for Non-Unique Images with Refiners

Tukey's Honest Significant Difference for Search Engines (Figure 6) clearly shows that Google Image Search Engine outperforms Yahoo and MSN Live.

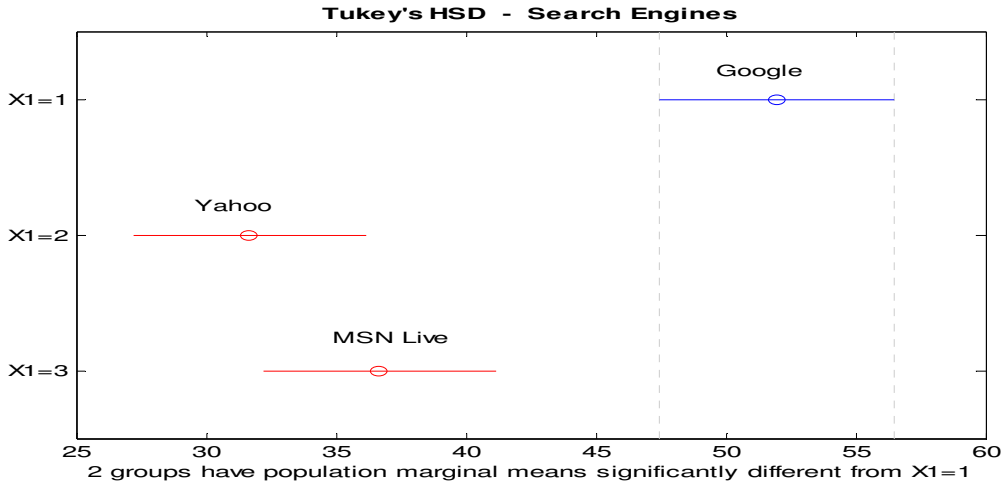

Figure 6- Tukey's Honest Significant Difference for Search Engines

This analysis also indicates that the performance of Yahoo and MSN Live does not differ statistically. Tukey's Honest Significant Difference for Query Types (Figure 7) clearly shows that the performance of search engines is better whenever there is no additional refiner.

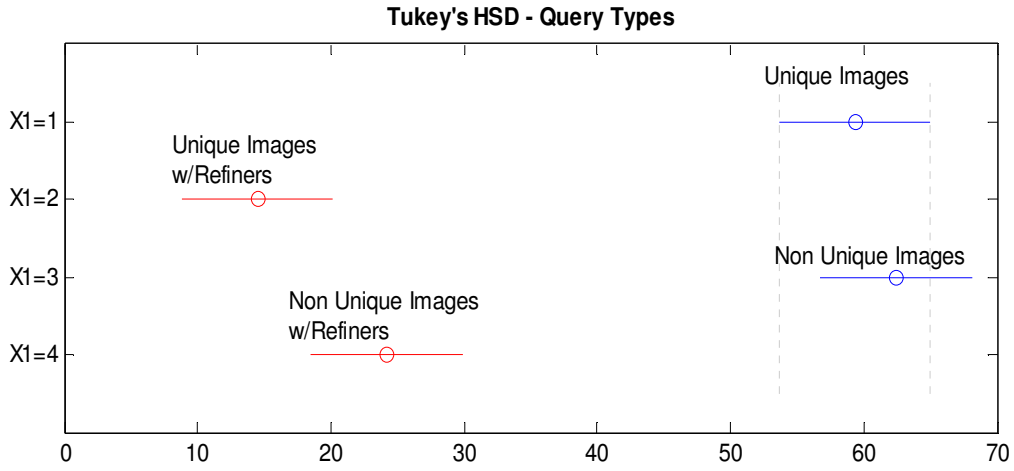

Figure 7- Tukey's Honest Significant Difference for Query Types

CONCLUSION

The results of the research suggest that overall, commercial search engines continue to have significant difficulties effectively executing image retrieval tasks. The Google search engine performs significantly better than Yahoo or MSN Live in any query type. The results also indicate that the precision of the search engines tended to drop with the increase in the number of retrievals. This performance reduction was noted across-the-board, that is irrespective of the search engines and the query types. The performance of the search engines also dropped dramatically when the queries had refiners (unique or non-unique).

One likely reason for the low performance in the image search engines is the continued reliance on manual annotation. The problem with manual annotation is that it is expensive, cumbersome, time consuming, may be imprecise due to variations between human analysts, and perhaps unreliable. Often, the text following the image is not be related to the image itself (or there may be no text at all) resulting in a mismatch between image and annotation. Thus, it is easy to see why image retrieval using this method has low precision and recall rates. It will be interesting to explore the integration of annotation-based image methods with content-based image retrieval methods in the future. Additional work developing principled methods of integrating human reasoning so as to further enhance image retrieval performance is also possible.

REFERENCES

1. Phani Kidambi, S. Narayanan, 2008, A human computer integrated approach for content based image retrieval, Recent Advances In Computer Engineering, Proceedings of the 12th WSEAS international conference on Computers, pp: 691-696, ISBN ~ ISSN:1790-5109 , 978-960-6766-85-5.

2. Baeza Yates, Ribeiro Neto, 1999, Modern Information Retrieval, ACM Press, ISBN-10: 020139829X, ISBN-13: 978-0201398298.

3. I.H. Witten, A. Moffat and T. Bell, 1999, Managing Gigabytes: Compressing and Indexing documents and images. Morgan Kaufmann Publishers, ISBN-10: 1558605703, ISBN-13: 978-1558605701.

4. I. E. Kuralenok, I. S. Nekrestyanov, 2002, Evaluation of Text Retrieval Systems, Programming and Computer Software, Vol. 28, No. 4, pp: 226-242.

5. Text Retrieval Conference (TREC), 1992, National Institute of Standards and Technology (NIST) and U.S. Department of Defense, http://trec.nist.gov/.

6. Hernon et al, 1990, Evaluation and Library Decision Making, Alex Publishing, ISBN: 0-89391-686-2, ISBN-13: 978-0-89391-686-2.

7. Meadow et al, 1999, Text Information Retrieval Systems, Library and Information Science series, Elsevier publications, ISBN: 9780124874053, ISSN: 1876-0562.

8. William Hersh, 1995, Information Retrieval – A Health Care perspective, Springer publications, ISBN-10: 0387944540, ISBN-13: 978-0387944548.

9. Lancaster et al, 1993, Information Retrieval Today, Information Resource Press, ISBN-10: 0878150641, ISBN-13: 978-0878150649.

10. John Smith, 1998, Image Retrieval Evaluation, IEEE Workshop on Content-based Access of Image and Video Libraries, Vol. 21, pp: 112-113.

11. Cooper W.S., 1968, Expected Search Length – A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. Journal of the American Society for Information Science, Vol. 19, pp: 30-41.

12. E. Cakir, H. Bahceci, Y. Bitirim, 2008, An Evaluation of Major Image Search Engines on Various Query Topics, The Third International Conference on Internet Monitoring and Protection, IEEE Computer Society, pp: 161-165.

13. A. Broder, 2002, A Taxonomy of web search, SIGIR Forum, Vol. 36, No. 2, pp: 3-10.

14. P.G.B. Enser, C. McGregor, 1993, Analysis of Visual Information Retrieval Queries, British Library Research, and Development Report 6104.

15. Nielsen Search Rankings, 2009, http://www.nielsen-online.com/pr/pr_090616.pdf.

16. Steve Smithson, 1994, Information Retrieval evaluation in practice: A case study approach, Information Processing and Management, Vol. 30, No. 2, pp. 205-221.