Workshop paper

Al-Maskari, A., Clough, P. and Sanderson, M. (2006) *Users' effectiveness and satisfaction for image retrieval.* In: Proceedings of the LWA 2006 Workshop. LWA 2006: Lernen - Wissensentdeckung - Adaptivität, October 9th-11th 2006, Hildesheim. , pp. 84-88.

Users’ Effectiveness and Satisfaction for Image Retrieval

Azzah Al-Maskari
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
Lip05aaa@shef.ac.uk

Paul Clough
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
p.d.clough@shef.ac.uk

Mark Sanderson
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
m.sandersonh@shef.ac.uk

Abstract

This paper presents results from an initial user study exploring the relationship between system effectiveness as quantified by traditional measures such as precision and recall, and users’ effectiveness and satisfaction of the results. The tasks involve finding images for recall-based tasks. It was concluded that no direct relationship between system effectiveness and users’ performance could be proven (as shown by previous research). People learn to adapt to a system regardless of its effectiveness. This study recommends that a combination of attributes (e.g. system effectiveness, user performance and satisfaction) is a more effective way to evaluate interactive retrieval systems. Results of this study also reveal that users are more concerned with accuracy than coverage of the search results.

1 Introduction

The performance of Information Retrieval (IR) systems is typically quantified using metrics derived from the number of relevant items found. Commonly used measures include Mean Average Precision (MAP), Precision at 10 documents retrieved (P@10), and bpref. Much of IR research has focused on improving these metrics, assuming that higher system effectiveness will help users to find more useful information. Some recent studies have shown that system performance. For example, Allan et al. (2005) have reported that a high increase in system effectiveness did not have detectable gains for the user. . Järvelin and Ingwersen (2004) assert that the real issue in IR systems design is not whether recall/precision goes up by a statistically significant percentage, but whether it helps the user solve the search task more effectively. Knowledge about what satisfies users is therefore crucial to improving retrieval systems. Factors such as prior search experience, search strategies and knowledge about the topic are also expected to influence the effectiveness of retrieval.

This paper examines the information seeking behaviour of users querying an interactive Arabic image retrieval system. The aim of this study is to compare users’ results with system performance and investigate the influence of factors such as users’ perception of the task/topic and their judgements of the search results. Users’ preference of coverage and accuracy of the results is also analysed.

We review past literature related to this topic in section 2; describe the system used in these experiments, the methodology and search tasks in section 3; present results of system and user evaluation in section 4 and provide discussion and conclusions in sections 5 and 6.

2 Related Research in User-Based Retrieval Evaluation

Recent studies have demonstrated that improvements in IR system effectiveness metrics do not translate into a direct benefit for end-users. A recent study by Turpin and Scholer (2006) attempted to address the relationship between the effectiveness of an IR system and how it matched up with user performance in a simple web search. Systems at various levels of MAP were used and assessment based on users performing a precision-search task measured by the length of time needed to find a single relevant document. Users also performed a recall-based task, measured by the total number of relevant documents users could identify in five minutes. There was no correlation between system performance measured with MAP and user performance on the precision task, and only a negligible improvement in performance on the recall task when MAP was increased.

A study by Hersh et al. (2000) showed that instance recall - where users try to identify different aspects of a question within a limited timeframe – did not improve with small increases in MAP of the underlying search system on the scale that is commonly reported in IR results. Allan et al. (2005) confirmed this result (using bpref), but also showed that for larger, specific increases in bpref, users did benefit on an instance recall task. Turpin and Hersh (2001) demonstrated a lack of improvement when users were engaged in a question answering task for a small number of questions.

The experiments of (Hersh et al., 2000), (Allan et al., 2005) and (Turpin and Hersh, 2001) have focused on recall-based tasks, whereas MAP is a precision-oriented measure. So, previous search tasks are different from what the employed effectiveness metrics are aiming to capture. The latest experiments of Turpin and Scholer (2006) were based on both recall and precision tasks and system effectiveness measured using MAP. In sum, from all these four studies, one can conclude that improvements in system effectiveness as measured using MAP, P@10, and bpref does not translate into a direct benefit to users. Therefore in this study, it was decided to use both recall-oriented and precision-oriented measures and compare them with the users' searching behaviour. In addition, in all the four previous experiments are based on text retrieval systems, whereas this experiment is based on image retrieval which we assumed could give different results but it did not. Furthermore, this study combines the results of both qualitative and quantitative analysis by taking into account system performance, users' performance, and users' perception of the tasks they performed (i.e., task difficulty, interestingness) and users' satisfaction of the search results.

3 Experiment Methodology

This study was conducted in conjunction with a submission to iCLEF 2006[1], and therefore restricted to the guidelines of iCLEF (e.g. methodology and number of topics). Previous studies had shown that a high increase in system effectiveness did not have a significant impact on the users' performance; only marginal gains had been reported. Therefore, for this experiment it was decided to test users' performance using just one system with acceptable retrieval effectiveness.

3.1 System Description

The system used on this experiment is based upon FLICKR[2]. Users query the system in Arabic which is translated into English, French, Spanish, German, Italian, and Dutch (with English used as an interlingua between Arabic and the other languages). Users are presented with results in which images are annotated in different languages, thereby increasing recall (different images are annotated with different languages). The user is able to edit the English translation of the Arabic query prior to search. More details of this system can be found in (Clough et al., 2006). The motivation for this study comes from wanting to experiment with Arabic users, and the availability of local resources to run the experiment. According to ABC news[3], there has been a rapid increase of Arabic users online and therefore we believe that many Arabic users would like to access FLICKR but don't have the necessary language skills to formulate multilingual queries.

3.2 Data Collection

Data collected for this experiment consisted of both qualitative (IR effectiveness metrics) and quantitative measures (pre-search questionnaire, task questionnaire, and exit questionnaire). Each user retrieved images for two types of tasks: 1) Classical ad-hoc task: "Find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside" and 2) Find five illustrations to the text "The story of saffron", the goal being to find five distinct instances of information described in a given narrative (saffron flower, saffron thread, picking the thread/flower, powder, dishes with saffron). The time allotted was 20 minutes per task

3.3 Users

Eleven Arabic students (postgraduate and undergraduate) with a median age 28 were recruited via email for this experiment. The work was conducted under the guidelines of Human Ethics Committee of Sheffield University. Most users reported having a great deal of experience with on-line searching (82%) and searching for images (45%).

4 Results

This section presents the results of system effectiveness, users' effectiveness, their perception and satisfaction of the results followed by a comparison between users' performance and system effectiveness.

4.1 System Effectiveness

A combination of binary[4] relevance and graded[5] relevance measures were used to evaluate system effectiveness. The system was assessed based on its retrieval results for the "European parliament" and "saffron" queries. The system was measured without query reformulation since more than half of the users did not reformulate the query during the search process. Table 1 illustrates that the system performs at similar levels of effectiveness for both tasks. Following is a brief description about each measure:

- Normalized P@100: precision over the first 100 images - normalised by the minimum of 100 or the number of retrieved images(Buckley and Voorhees, 2000).
- Q-measure: based on graded relevance and cumulative gain, designed for the task of finding many relevant items (Sakai, 2005). In this experiment Q-measure is computed until rank 10.
- bpref-10: Binary preference is the number of times nonrelevant images are retrieved/judged before relevant images (Buckley and Voorhees, 2004). In this experiment bpref-10 is computed until rank 50.
- R-precision is the precision after *R* images are retrieved where *R* is the number of relevant images for a given topic (Buckley and Voorhees, 2000).

Task	P @50	P@100	Q-	bpref-	10-
	norm	norm	measure	10	Precision
Parliament	0.48	0.46	0.27	0.48	0.58
Saffron	0.45	0.48	0.42	0.39	0.54

Table 1- System effectiveness (average)

4.2 Measuring Users' effectiveness

Tables 2 and 3 illustrate users' performance and satisfaction of both tasks. The evaluator assessed the images retrieved by users. Recall captures how well the subjects find different aspects of the topic. For the saffron

[1] http://nlp.uned.es/iCLEF/

[2] http://www.flickr.com/
FLICKR is a large-scale web-based image retrieval database. It is used to manage and share personal and commercial photographs and currently contains over five million freely accessible images.

[3] http://www.abc.net.au/news/newsitems/200604/s1624108.htm

[4] image is relevant or not relevant

[5] image is highly relevant, partially relevant, or not relevant

task: saffron flower, saffron thread, picking the thread/ flower, powder, dishes; the European parliament task: images of inside and out of different buildings from different European countries. Users are given one point for retrieving each true instance (unique images) and no credit for repeated instances (i.e. no credit for retrieving two saffron flowers. For the parliament task, one point was given for retrieving an inside image and another point for an outside image of the same building (unique images). Thus, users' recall for the parliament task was calculated as (number of unique images/ correct images retrieved) and for the saffron task as (number of unique images/ total required images), which is five in this case. Precision captures the proportion of correct relevant images to the total number of images retrieved. Precision for the parliament task was computed as (correct images/total images retrieved) and for the saffron task as (number of unique images/ total retrieved). Users' precision in the parliament task is better than the saffron task due to their perception of the topic according to the information obtained from the task questionnaire: familiarity with the topic, topic easiness and their interest in the topic. There is a moderate degree of correlation[6] between *familiarity* and user' precision (*p=0.7)* in the parliament task. Users' ability to achieve full recall and precision is fluctuates in both tasks (shown by Tables 2 and 3).

	Precision	U. Sat. Accuracy[*]	Recall	U. Sat. coverage[*]	Use ful[*]
user1	0.40	1.00	0.4	0.50	1
user2	0.80	0.50	0.8	0.50	0.5
user3	1.00	0.50	0.8	1.00	1
user4	0.60	1.00	0.6	1.00	1
user5	0.33	1.00	0.2	1.00	1
user6	0.80	1.00	0.8	1.00	0.5
user7	1.00	0.50	1	0.50	1
user8	1.00	0.50	1	0.50	0.5
user9	1.00	1.00	1	1.00	1
user10	0.40	0.50	0.4	0.50	0.5
user11	0.60	0.50	0.6	0.50	1
Average	**0.72**	**0.73**	**0.69**	**0.73**	**0.82**

Table 2 -Users' performance versus satisfaction-Parliament task

4.3 User' Prediction of the Search Results

Users' opinions and expectations of search tasks were extracted from the task questionnaire. Users rated the results in terms of their relevancy[7], satisfaction with usefulness of the results, satisfaction with the accuracy (efficiency) and coverage (completeness) of the results. According to the Tables 2 and 3, there is no significant correlation between users' estimation of results and their actual performance, except between usefulness of the results and users' recall (*p=0.02*). This suggests that users are not able to easily assess the success of their search. On average, users believed they had completed the task when they had actually achieved 69% recall. In general users were satisfied with the system despite the fact this did not reflect on their performance.

	Precision	U. Sat Accuracy[*]	Recall	U. Sat. Coverage[*]	Useful [*]
user1	0.50	1.00	0.50	1.00	1
user2	0.55	1.00	0.27	1.00	1
user3	0.73	1.00	0.88	0.50	1
user4	0.85	1.00	0.65	1.00	1
user5	0.92	1.00	0.67	1.00	1
user6	0.83	1.00	0.90	0.00	0.5
user7	1.00	0.50	0.67	0.50	0.5
user8	0.88	0.00	0.86	0.00	1
user9	1.00	1.00	0.75	1.00	1
user10	0.77	1.00	0.70	1.00	1
user11	0.90	1.00	0.78	1.00	1
Average	**0.81**	**0.86**	**0.69**	**0.73**	**0.91**

Table 3-Users' performance versus satisfaction saffron task

4.4 Comparison between System and User Effectiveness

The differences between system and user effectiveness were measured using an analysis of variance (ANOVA). According to the results in Table 4, there is a statistically significant correlation between users' performance, gauged by users' recall in addition to their satisfaction of the coverage of the results, and system effectiveness when measured by Q-measure. In general, there is no correlation between users' effectiveness and system as quantified by P@100 and 10-Precision, except for users' precision and the system P@100 in the parliament task. Although there is a strong correlation between users coincident of the usefulness of the results and system bpref-10 on both tasks it did not reflect on their performance or satisfaction of the accuracy and coverage of the results. The lack of correlation between users and the system as determined by precision-oriented metrics indicate that these metrics are not compatible with user satisfaction and performance.

Therefore, users' effectiveness is by and large inconsistent with the system effectiveness as measured solely by traditional IR metrics. This conclusion gives further credence to the findings of (Hersh et al., 2000), (Turpin and Hersh, 2001),(Allan et al., 2005), and (Turpin and Scholer, 2006) in that improvements in the metrics of systems (P@10, MAP, bpref-10) do not translate into a direct benefit for the users. The

[6] Measured using the Pearson correlation coefficient.

[7] If images directly address the core issue of the topic then *highly-relevant,* if they contain helpful information then *partially relevant*, otherwise *not relevant*.

[*] U=user; sat= satisfaction
1=very satisfied; 0.5=partially satisfied, 0=not satisfied

aforementioned experiment indicate that user satisfaction with the results provide a better picture of system' accuracy than classical measures.

Effectiveness measures	Parliament Task	Saffron Task
Users' recall vs. Q-measure	**p=5.97E-05**	**p=0.039**
Users' satisfaction with coverage vs. Q-measure	**p=0.011**	**p=0.018**
Users' precision vs. P@100	**p=0.005**	p=0.185
Users' precision vs. 10-Precision	p=0.09	p=0.262
Users' satisfaction with accuracy vs. 10-Precision	p=0.11	p=0.200
Users' precision vs. bpref-10	P=0.059	P=0.076
Users' recall vs. bpref-10	P=0.166	P=0.123
Users' satisfaction with accuracy vs. bpref-10	P=0.064	P=0.065
Users' satisfaction with coverage vs. bpref-10	P=0.210	P=0.065
usefulness of results vs. bpref-10	**P=0.028**	**P=0.013**

Table 4- System versus users' effectiveness

4.5 Accuracy vs. Coverage of Search Results

In the search task questionnaire, users identified their preference of accuracy or coverage according to the tasks. Accuracy was defined to the users as "relatedness of results to the search topic" and coverage as "coverage of results to all aspects of topic". Most users (68%) opted for accuracy over recall (32%). This implies that whether users are looking for few or many images, they are concerned with quality than quantity. Users seem to prefer having fewer highly relevant images than a larger proportion of relevant images. Hence systems with adequate precision may contain what the users are looking for.

5 Discussion

This study reports the relationship between IR system effectiveness and user effectiveness by using 11 subjects to search using an image retrieval system for recall-based tasks. Users are required to find as many relevant images of the European parliament and find for five different instances of saffron where users' recall was measured by the number of instances saved.

Results demonstrate that users were highly satisfied with the system's performance despite the system not being of high quality as measured using P@100, R-precision and the Q-measure. Results revealed a significant relationship between users' recall and the system' Q-measure in both tasks. Therefore, Q-measure, a recall-oriented measure, can be more useful when comparing system versus users' performance. Precision measures do not seem to correlate well with user performance as there is no significant relationship between users' precision when compared with the system P@100 for the saffron task. One possible explanation for the lack of correlation between the system and users in the saffron task is the users' familiarity with the search topic, affecting the quality of the search results. For both tasks, observations indicated that some users are just better than others at searching.

6 Conclusions

The conclusion of this experiment begins to answer the doubt expressed by Turpin and Scholer (2006) over whether a direct relationship between IR effectiveness measures and users satisfaction with search results exists. This experiment reinforces the findings of previous studies in that there does not appear to be a strong relationship between the performance of a system and the user. . It was found that users can find what they are looking for despite a fairly low level of system effectiveness. This indicates that results for experiments based on system measures are not comparable with experiments based on real users. The fact that system languages are query languages differ in this experiment, we are not generalizing our conclusion to all IR systems.

It is believed that different types of topics and tasks lead to different levels of quality in the search results. While this experiment is limited to two topics, assessment based on system performance only does not interpret system quality and additional analysis of a users' satisfaction with the results presents a more holistic view of search performance. We are planning to conduct further work to determine what really satisfies the user of an IR system. This includes a further study that look into measures of system performance such as speed, accuracy, coverage, presentation of the results, and language related aspects together with a larger number of topics and more diverse tasks. Additionally, more study to investigate what other measures correlate with users' performance besides Q-measure for both recall-based and precision-based tasks.

References

Allan, J., Carterette, B. & Lewis, J. (2005). "*When Will Information Retrieval Be "Good Enough"? User Effectiveness As a Function of Retrieval Accuracy. 2005. " In: Proc ACM SIGIR 433-440, Salvador, Brazil*

Buckley, C. & Voorhees, E. M. (2000). "*Evaluating Evaluation Measure Stability" In: Proc ACM SIGIR, 33 - 40 Athens, Greece*

Buckley, C. & Voorhees, E. M. (2004). "*Retrieval evaluation with incomplete information" In: Proc SIGIR, 25-32, Sheffield, United Kingdom*

Clough, P., Al-Maskari, A. & Darwish, K. (2006). (in Press). "*FLICKRArabic: multilingual access to photos" In: Proc, iCELF.*

Flickr. http://www.flickr.com/

Hersh, W., Turpin, A., Price, S. & Chan, B. (2000). "*Do Batch and User Evaluations Give the Same Results?" In: Proc SIGIR 17-24, Athens, Greece*

iCLEF. 2006. Interactive track for the cross-Language Evaluation Forum. http://nlp.uned.es/iCLEF/

Reuters and ABC Science Online. 2006. Search engine to target Arabic speakers. *ABC NEWSONLINE* http://www.abc.net.au/news/newsitems/200604/s1624108.htm

Sakai, T. (2005). "*The Reliability of Metrics Based on Graded Relevance " In: Proc Information Retrieval Technology: Second Asia Information Retrieval*

Symposium, AIRS 2005, Jeju Island, Korea, October 13-15, 2005. Proceedings Korea

Järvelin, K. & Ingwersen, P. (2004). Information seeking research needs extension towards tasks and *technology*. *Information Research*, 10(1)

Turpin, A. & Scholer, F. (2006). "*User Performance versus Precision Measures for Simple Search Tasks" In: Proc SIGIR, Seatle, Washington, USA*

Turpin, A. H. & Hersh, W. (2001). "*Why batch and user evaluations do not give the same results" In: Proc ACM SIGIR, 225 - 231 New Orleans, Louisiana, United States*