A Proposal for User-Focused Evaluation and Prediction of Information Seeking Process

Chirag Shah
School of Communication & Information (SC&I)
Rutgers University
4 Huntington St, New Brunswick, NJ 08901, USA
chirags@rutgers.edu

ABSTRACT

One of the ways IR systems help searchers is by predicting or assuming what could be useful for their information needs based on analyzing information objects (documents, queries) and finding other related objects that may be relevant. Such approaches often ignore the underlying search process of information seeking, thus forgoing opportunities for making process-based recommendations. To overcome this limitation, we are proposing a new approach that analyzes a searcher's current processes to forecast his likelihood of achieving a certain level of success in the future. Specifically, we propose a machine-learning based method to dynamically evaluate and predict search performance several time-steps ahead at each given time point of the search process during an exploratory search task. Our prediction method uses a collection of features extracted solely from the search process such as dwell time, query entropy and relevance judgment in order to evaluate whether it will lead to low or high performance in the future. Experiments that simulate the effects of switching search paths show a significant number of subpar search processes improving after the recommended switch. In effect, the work reported here provides a new framework for evaluating search processes and predicting search performance. Importantly, this approach is based on user processes, and independent of any IR system allowing for wider applicability that ranges from searching to recommendations.

Categories and Subject Descriptors

H.3: INFORMATION STORAGE AND RETRIEVAL **H.3.3: Information Search and Retrieval**: *Search process*; H.3: INFORMATION STORAGE AND RETRIEVAL **H.3.4: Systems and Software**: *Performance evaluation (efficiency and effectiveness)*

General Terms

Measurement, Performance, Experimentation

Keywords

Exploratory search, Evaluation, Performance prediction

Presented at EuroHCIR2013.

1 INTRODUCTION

IR evaluations are often concerned with explaining factors relating to user or system performance after the search and retrieval are conducted [20]. Most recommender systems, however, operate with an objective to suggest objects that could be useful to a user based on his/her or others' past actions [2][19]. We commenced our investigation by broadly asking how we could take valuable lessons from both IR evaluations and recommender systems to not only evaluate an ongoing search process, but also predict how well it will unfold and suggest a better path to the searcher if it is likely to underperform. The motivation behind this investigation was based on the following assumptions and realizations grounded in the literature.

1. The underlying rational processes involved in information search are reflected in the actions users make while searching. These actions include entering search queries, skimming the results, as well as selecting and collecting useful information [8][14][15].
2. A searcher's performance is a function of these actions performed during a search episode [7][22].

With these assumptions, we propose to quantify a search process using various user actions, and use it for user performance (henceforth, 'search performance' or 'performance') prediction as well as search process recommendations.

2 BACKGROUND

Past research on predictive models that relates to the approach we describe in this paper can be grouped into two main categories: (1) behavioral studies and (2) IR approaches. In both cases; however, the focus has been on end products instead of in the process required to produce them.

As far as the behavioral studies go, research has been conducted to explore users models that help anticipating specific aspects of the search process. One goal in this context has been the determination of whether a search process will be completed in a single or multiple sessions. For example, Agichtein *et al.* [3] investigated different patterns that can be identified in tasks that require multiple sessions. As a result, the authors devised an algorithm capable of predicting whether users will continue or abandon the task. Similar work is described in Diriye *et al.* [6], which focuses on predicting and understanding of why and when users abandon Web searches. To address this problem, the authors studied features such as queries and interactions with result pages. Based on this approach, the authors were able to determine reasons for search abandonment such as accidental causes (e.g. Web browser crashing), satisfaction levels, and query suggestions, among others.

There have been also attempts to understand past users' behaviors in order to predict future ones in similar conditions. For example, Adar *et al.* [1] visually explored behavioral aspects using large-scale datasets containing queries and other information objects produced by users. The authors were able to identify different behavioral patterns that seem to appear consistently in different datasets. While not directly related to performance prediction, this work focused on attributes of the search process instead of in final products derived from it.

Research like the ones described above often relies on historic data from large populations and the use of trend and seasonal components, which are used to model long-term direction and periodicity patterns of time-series [17]. For example, some have explored seasonal aspects in Web search (e.g. weekly, monthly, or annual behaviors) that provides useful information to predict and suggest queries [5].

From an IR perspective, Radinski *et al.* [18] explored models to predict users' behaviors in a population in order to improve results from IR systems. The authors also developed a learning algorithm capable of selecting an appropriate predictive model depending on the situation and time. As described by the authors, applications of this approach could go from click predictions to query-URL predictions. In contrast to this approach, our method presented in this paper considers both the population trends and an individual user behavior.

In a similar track, several works have been conducted on query performance prediction, focusing on developing techniques that help IR system to anticipate whether a query will be effective or not to provide results that satisfy users' needs [4][10][11]. For example, Gao *et al.* [10] found that features derived from search results and interactions features offer better prediction results than a prediction baseline defined in terms of query features. Results from this study have direct implications to individual users by aiding the auto evaluation process of IR systems.

In information search, users may be unaware of their individual performance when solving an information search task. For instance, Shah & Marchionini [23] showed how lack of awareness about different objects involved in searching (queries, visited pages, bookmarks) could result in mistaken perception about search performance during an exploratory search task. Even if an IR system is highly effective, users may run into multiple query formulation and evaluation of several pages before finding what they need. This process, which can be related to search strategies, implies effort and time that is usually underestimated by the users themselves. In this sense, instead of predicting end products (i.e., overall performance), the approach we introduce in this paper is oriented toward predictions at different times in order to increase the level of awareness of users about their own search process. Similar to weather forecast, this information could help users to be aware of possible trends based on past and current behavior.

For a more recent discussion on IR evaluations and their shortcomings, see [12]. To the best of our knowledge, search process performance prediction at different times from a user perspective has not been explored. Similar approaches can be found in weather and stock market studies. For example, using machine learning approaches such as Support Vector Machine (SVM), some models have been implemented to predict the trends of two different daily stock price indices using NASDAQ and Korean Stock prices [13][16]. In a similar fashion, our approach is oriented to forecast users' search performance *N*-steps ahead with the aim to aid their search process awareness and performance trends.

Unlike previous works in IR, we are not proposing to use time series analyses or seasonal components of historic data. Instead, we investigate predictive models based on machine learning (ML) techniques; namely: SVM, logistic regression, and Naïve Bayes which are trained over a set of features such as time, number of queries, and page dwell time. In contrast to most IR evaluations, our method focuses on user-processes. Also, unlike most recommender systems, our approach could output alternative strategies instead of similar/relevant products to help the searcher. In essence, the work reported here takes several lessons from tradition IR evaluations, recommender system designs, and weather/stock forecasting to come up with a new approach for evaluating and predicting search performance.

In the next section we provide a detailed description of our method, feature selection, and the measures we used in order to create ML-based predictive models.

3 METHOD

In order to analyze the search processes followed by different users/teams, we assume that the underlying dynamics of the search processes are expressed by a collection of activities that take place from the beginning to the end of the search processes.

The first part of our method is a feature extraction step in which we extract a wide array of features relating to webpages, queries and snippets saved from the search processes for each unit of time *t*. This step is performed in order to evaluate how well we could use those features to capture the underlying dynamics which would lead to recognizing whether a search process is going to lead to high or low performance in the future time steps at *t+n (n=1,2,….,N)*, where *N* is the furthest time step.

The decision to include or exclude a feature was based on literature (e.g., [7]) as well as our past experience [22] with representing and evaluating search objects and processes. Each feature is extracted for each user or team, *u,* up to time *t* from the search processes and they are explained in detail as follows.

- *Total coverage (u,t)*: The total number of distinct Webpages visited by a user (*u*) up to time *t*. This feature captures the Webpage based activity performed by a user and provides a measure to see how much distinct information has been found by the user up to this time.
- *Useful coverage (u,t)*: The total number of distinct webpages in which a user spent at least 30 seconds, up to time *t*. This measure evaluates out of the total pages he/she has visited how many of them were useful in finding relevant information leading to satisfaction with their context in completing the exploratory task [9][22][25].
- *Number of queries (u,t)*: Total number of unique queries executed by a user up to time *t*. This feature implicitly relates to how much effort and cognitive thinking a user has put in to this task.
- *Number of saved snippets (u,t)*: Total number of snippets saved by user *u* up to time *t*. This measures the amount of information that the user thought that might be relevant in the future to complete the task and needed to be remembered. In other words, this feature is an indication of explicit relevance judgments made by the user.
- *Length of Query (u,q,t)*: Length of each query(*q*) executed by a user *u* based on the character count of the query up to time *t*. This feature captures how the user imposed the

queries and how long they were at different times of the search process.

- *Number of tokens in each query (u,q,t)*: This is the count of tokens/words in each query(q) executed by user u up to time t. This query based measure takes into account how specific a user was in defining the query. By inspecting the datasets, we realized that queries with a less number of tokens tend to get general results. On the other hand, composed queries with multiple terms are related to more specific searchers. We also observed that typically the users started with general queries with few words at the beginning of the search process but then went into more detailed queries to find more specific information later. For all these reasons, we found it to be useful to capture the number of token used in a query.
- *Query entropy (u,q,t)*: This measures the information content in a given query (q), by finding the expected value of information contained in a query. We used the widely recognized notion of Shannon entropy [24] in Information Theory to calculate the information content of a query. We calculated the number of unique characters appearing in each of the queries, which represent the observed counts of the random variable. This was used as the input to Shannon entropy calculation and we used to the maximum-likelihood method to calculate the entropy. Query entropy feature has been used in the past to predict *goodness* of a query for making query expansion decision [21].

The method used to assess the search performance of a user is described below. We define a measure called *Efficiency (u,t),* for each user u up to time t *in order to* predict whether a given search process is going to yield in high/low performance in the future We first define *Effectiveness* of user u up to time t as the ratio of useful coverage and total coverage (both defined earlier). A similar measure was used in [7] and [22].

$$Effectiveness(u,t) = \frac{Useful\ coverage(u,t)}{Total\ coverage(u,t)} \qquad (1)$$

We then calculated *Efficiency* as defined in Equation 2.

$$Efficiency(u,t) = \frac{Effectiveness(u,t)}{NumberofQueries(u,t)} \qquad (2)$$

In other words, *Efficiency* is defined as the *Effectiveness* obtained per query, or how effective a query is in terms of achieving a certain level of useful coverage.

The performance for each user u at each time t was classified in to the two classes; high performance and low performance based on the following criteria:

$$Class = \begin{cases} high & ;if \quad Efficiency(u,t) \geq \overline{Efficiency(u,t)} \\ low & ;else \end{cases} \qquad (3)$$

Using various user studies data available to us, we constructed feature matrices which consist of all aforementioned features for each minute of time t for all the users in each dataset, and converted in to a long vector of features which we fed as the input to the classification models used.[1] The class labels were generated as high/low performance at minute $t+n$ based on the above mentioned criteria and threshold and used as the output class labels to be used in the n-step ahead prediction model. If a class label at n-step ahead was correctly predicted based on the features extracted up to time t from the classification model it was considered as correctly classified and if not as misclassified.

4 EXPERIMENTS

In order to evaluate whether users who are predicted to perform at low performance in the future based on the current search process, could benefit from this analysis to improve their search process, we conducted some simple simulation analysis.

We considered the individual user search processes as a collection of *search paths*, where each *search path* is defined as the search process from the time a user issued a query up to the time user issued another quite different query. This was found out using generalized Levenshtein (edit) distance, which is a commonly used distance metric for measuring the distance between two character sequences. If the Levenshtein (edit) distance between two subsequent queries were greater than 2 (assuming less than 2 was when there were changes in the queries due to simple spelling mistakes or refining of the query), we considered the search process from the former query to the next query as a single *search path*.

Following this method, we found the first *search path* of each user and based on the features extracted up to the end of the first *search path*, and based on the classification model learnt from that corresponding n-step ahead prediction we predicted whether the user is going to have low/high performance at the end of the session. If the user was going to have low performance, then out of the users who predicted to have high performance, we looked at which high performing user has the lowest Levenshtein (edit) distance between the queries issued by low performing user within the first search path and considered it as a pair of users, whom we are going to use in the simulation. Then, for each low performing user and high performing user that was matched, we switched the search process of low performing user at the end of the first *search path* with the high performing user's *search path* up to t=T minutes, where T is the total number of minutes for a session. Then we evaluated by switching the search process early during the overall process whether it would benefit each low performing user to improve their performance. We found that we were able to move most of the underperforming search processes to higher performance by early detection and switching, while keeping the higher performing processes unharmed.

These simulations provide verification that by realizing early during the search process whether a user is going to perform well or not, one could recommend better search processes/strategies for that user which would lead to uplifting the search performance of a previously destined to low performing user.

5 CONCLUSION

When it comes to prediction, information retrieval and filtering systems are primarily focused on objects while assessing what and if something could help the users. These approaches are often system-dependent even though the process of information seeking is usually user-specific. Personalization and recommendations are frequently exercised as methods to address user-specific IR and filtering, but still limited to comparing and recommending objects, not focusing on underlying IR processes that are carried out by the searchers. We presented a new approach to address these shortcomings. We began by asking

[1] In the interest of space and scope of work here, details of these experiments have been omitted, but will be available for discussion at the workshop.

whether we could model a user's search process based on the actions he/she is performing during an exploratory search task and forecast how well that process will do in the future. This was based on a realization that an information seeker's search goal/task can be mapped out as a series of actions, and that a sequence of actions or choices the searcher makes, and especially the search path he/she takes, affects how well he/she will do. Thus, in contrast to approaches that measure the *goodness* of search products (e.g., documents, queries) as a way to evaluate the overall search effectiveness, we measured the likelihood of an existing search process to produce *good* results.

Here we presented simulations to demonstrate what could happen if one can make process-based predictions, but one could develop an actual recommender system using the proposed method. Another potential application of such prediction-based method would be to use such approach in IR systems to provide the awareness to users how their future performance will be based on the current/past search process. The system could identify that a user will have low performance if, he continues this manner at an early stage of the process, and what could be done to provide suggestions to improve overall performance.

Given that the proposed technique is independent of any specific kind of system, and solely focused on user-based processes, it will presumably be easy to apply it to a variety of IR systems and situations irrespective of retrieval, ranking, or recommendation algorithms. Finally, while we have used datasets borrowed from previous user studies, one could easily apply the proposed method to Web logs, TREC data, and other forms of datasets with various user actions recorded over time.

6 ACKNOWLEDGEMENTS

The work reported here is supported by The Institute of Museum and Library Services (IMLS) Cyber Synergy project as well as IMLS grant # RE-04-12-0105-12. The author is also grateful to his PhD students Chathra Hendahewa and Roberto Gonzalez-Ibanez for their valuable contributions to this work.

7 REFERENCES

[1] Adar, E., Weld, D. S., Bershad, B. N., & Gribble, S. D. (2007). Why we search: visualizing and predicting user behavior. In *Proceedings of World Wide Web (WWW) Conference 2007.*

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749.

[3] Agichtein, E., White, R.W., Dumais, S.T., & Bennett. P.N. (2012). Search interrupted: Understanding and predicting search task continuation. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR) 2012.*

[4] Cronen-Townsend, S., Zhou, Y., & Croft, B. (2002). Predicting query performance. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR) 2002.*

[5] Dignum, S., Kruschwitz, U., Fasli, M., Yunhyong, K., Dawei, S., Beresi, U.C., & De Roeck, A. (2010). Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 2010*, vol.1, no., pp.425-430, Aug. 31 2010-Sept. 3 2010

[6] Diriye, A., White, R.W., Buscher, G., & Dumais, S.T. (2012). Leaving so soon? Understanding and predicting web search abandonment rationales. In *Proceedings of CIKM 2012.*

[7] González-Ibáñez, R., Shah, C., & White, R. W. (2012). Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems. In *Proceedings of the 75th Annual Meeting of the Association for Information Science and Technology (ASIS&T)*. Baltimore, MD, USA.

[8] Gwizdka, J. (2008). Cognitive load on web search tasks. *Workshop on Cognition and the Web, Information Processing, Comprehension, and Learning*. Granada, Spain. Available from http://eprints.rclis.org/14162/1/GwizdkaJ_WCW2008_short_paper_finalp.pdf

[9] Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM TOIS, 23*(2): 147–168.

[10] Gao, Q., White, R., Dumais, S.T., Wang, S., & Anderson, B. (2010). Predicting query performance using query, result and interaction features. In *Proceedings of RIAO 2010.*

[11] He, B., & Ounis, I. (2006). Query performance prediction, Information Systems, Volume 31, Issue 7, November 2006, Pages 585-594, ISSN 0306-4379, 10.1016/j.is.2005.11.003.

[12] Järvelin, K. (2012). IR research: systems, interaction, evaluation and theories. *ACM SIGIR Forum, 45*(2), 17. doi:10.1145/2093346.2093348

[13] Kyoung-jae, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, Volume 55, Issues 1–2, September 2003, Pages 307-319, ISSN 0925-2312, 10.1016/S0925-2312(03)00372-2.

[14] Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. *American Society for Information Science*, *47*(17). Available from http://dl.acm.org/citation.cfm?id=1920331.1920356

[15] Liu, J., Gwizdka, J., Liu, C., & Belkin, N. J. (2010). Predicting task difficulty for different task types. In *Proceedings of the Association for Information Science*, *47*(16). Available from http://dl.acm.org/citation.cfm?id=1920331.1920355

[16] Ming-Chi, L. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, Volume 36, Issue 8, October 2009, Pages 10896-10904, ISSN 0957-4174, 10.1016/j.eswa.2009.02.038

[17] Ord, J., Hyndman, R., Koehler, A., & Snyder, R. (2008). Forecasting with Exponential Smoothing (The State Space Approach). Springer, 2008.

[18] Radinski, K., Svore, K., Dumais, S. T., Teevan, J., Horvitz, E., & Bocharov, A. (2012). *Modeling and predicting behavioral dynamics* on the Web. In *Proceedings of WWW 2012.*

[19] Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM*, *40*(3), 56–58.

[20] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 138–146).

[21] Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 2-9). Sheffield, UK.

[22] Shah, C., & Gonzalez-Ibanez, R. (2011). Evaluating the Synergic Effect of Collaboration in Information Seeking. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)* (pp. 913–922). Beijing, China.

[23] Shah, C., & Marchionini, G. (2010). Awareness in Collaborative Information Seeking. *Journal of American Society of Information Science and Technology (JASIST)*, *61*(10), 1970–1986.

[24] Shannon, C. E. and Weaver, W. *Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press, 1963.

[25] White, R. W., & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. In *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*. Geneva, Switzerland.