Detecting the Eureka Effect in Complex Search

Hui Yang, Jiyun Luo, and Christopher Wing

Department of Computer Science, Georgetown University
37th and O Street NW, Washington DC, 20057, USA
huiyang@cs.georgetown.edu,jl1749,cpw26@georgetown.edu

Abstract. In search tasks that show a high complexity, users with zero or little background knowledge usually need to go through a learning curve to accomplish the tasks. In the context of patent prior art finding, we introduce a novel notion of Eureka effect in complex search tasks that leverages the sudden change of user's perceived relevance observable in the log data. *Eureka effect* refers to the common experience of sudden understanding a previously incomprehensible problem or concept. We employ non-parametric regression to model the learning curve that exists in learning-intensive search tasks and report our preliminary findings in observing the Eureka effect in patent prior art finding.

Keywords: Complex Search, Prior Art Retrieval, Learning Curve, Eureka Effect

1 Introduction

State-of-the-art Information Retrieval (IR) research is extremely valuable for a wide range of applications but they are subject to a limited number of search task types. Most search tasks attracting lots of current research efforts are one-shot query tasks. Although those search tasks account for a large portion of online Web search activities, a great deal of other complex search tasks remain understudied. These tasks lie along the spectrum of tasks that require plenty of professional expertise such as patent prior art finding [11] and e-discovery [5], to tasks that are complex but do not require much expertise such as travel search that [4]. Typically, these search tasks require multiple queries in a search session and involves rich user and system interactions.

One fundamental type of challenge in these complex search tasks is represented by the users' changing perception of document relevance. This problem can be understood by using the notion of learning curve, that is, the rate of a user's progress in gaining knowledge, experience or new skills. By modeling the learning curve, user's changing understanding of document relevance can be reflected in search algorithms and evaluation metrics. Commonly used learning curve formulas are observed from industrial production lines and are usually used to determine expected labor and materials costs. Most of them are linear formulations in the form of $Y_x = aX^b$, where Y is the cumulated average time required to produce X units, a is the time required to produce the first output, and b is the learning rate. Another popular formulation is the "S-shape" learning

Table 1. A prior art finding session.

query id	# returned docs	query	timestamp
1	109	(SAKAKI near1 YUZO).in	2012/06/07 13:27
2	855	428/827, 828, 829, 830.ccls	2012/06/07 13:27
3	195	428/836.2.ccls	2012/06/07 13:28
4	0	S1 and S2 and S3	2012/06/07 13:28
5	74829	CoCrPtRu((“Co.sub.”$2) same(Ru ruthenium))	2012/06/07 13:29
6	31	S2 and S3 and S5	2012/06/07 13:30
7	2	(“20040184176” — “20050181237”).PN	2012/06/07 13:45
8	402914	samung kikitsu.in.	2012/06/07 14:02
9	8	(S2 S3) and S8 and bernatz	2012/06/07 14:02
10	3456	lee.in and (Ku anisotropy)	2012/06/07 14:05
11	22	428/826-827.ccls and S10	2012/06/07 14:06
12	2	jp adj “2008090913”	2012/06/07 14:10
13	2	“20020012816”	2012/06/07 14:11
14	1	cn adj “1870145”	2012/06/07 14:12
15	2	“2006024791”.pn.	2012/06/07 14:15

curve using the Sigmoid functions [9, 6]. However, it is unclear whether existing learning curve formulas are suitable in the context of complex search.

This paper uses patent prior art search as a motivating example of complex search tasks constrained by time. As an illustration, we give an example taken from a query log from the U.S. Patent and Trademark Office (USPTO). We extracted and analyzed the prior art finding sessions that U.S Patent Examiners conducted for a patent application on “light controlling”. There are more than 15 distinct queries in the session. Most of these are structural queries, where Boolean operators (AND, OR) and proximity operators (within 2 words) are used to pose constraints on the query. Another common operator is browsing, where from a seed document, more documents are browsed from its references or from the document class it belongs to. The search lasted for around 2 hours. We noted that at the moment that the patent examiner came across the passage *“a control device for controlling hue of light emitted by a light source, device comprises: a body with a surface containing a visible representation of a plurality of selectable combinations of hue available for said light source”*, the time spent on examining a single document is suddenly decreased from 15 minutes per document to less than 1 minute per document. It is illustrated in Table 1 at query $S9$.

This sudden change of the reading time in general indicates a change of user's status of mind of understanding the related topics; which we call the “Eureka effect”. “Eureka!” is the word shouted out by Archimedes, the Greek mathematician, when he suddenly discovered how to calculate the volume of an irregular object and leap out of a public bath. Here we use “Eureka effect” to refer to the common experience of suddenly understanding a previously incomprehensible problem or concept.

On the other hand, this example suggests that if we are able to recognize the sudden drop of reading time per document, we could obtain a novel learning curve formulation specifically designed for complex search tasks, which will allow search engines to create better search algorithms and better evaluation mechanisms.

Based on these observations, we propose the following definition of Eureka effect in complex search tasks:

– *Eureka effect is the phenomenon that in complex search process where we detect a sharp increase of users' understanding of the domain and the related documents.*

There are two main issues involved in this definition, namely the computation of the gap between a document's user received relevance (URR) and user perceived relevance (UPR), and the modeling of the learning curve to detect the Eureka effect. We show that detection of Eureka effect can be tackled by a non-parametric regression algorithm. The solution consists of two steps. First, automatically extracting relevance judgments from office action documents submitted by patent examiners. Second, fitting the difference between user perceived relevance and user received relevance to a non-parametric regression model, in which the model parameters define the learning curve and the Eureka effect. Particularly, we are particularly interested in situations where users start the search with zero or little background and study the Eureka effect for them.

2 Related Work

In complex search tasks, retrieval results usually have different reading difficulties and users also show various reading proficiencies. Borlund [1] pointed out that relevance is a dynamic concept that depends on a user's judgment at a certain point of time. Heilman et al. [7] and Kidwell et al. [8] provided two statistical approaches to estimate a passage's reading difficulty by utilizing lexical and grammatical features. Collins-Thompson et al. [3] provided a Language Modeling Approach to estimate reading difficulties. In their further work, Collins-Thompson et al. [2] pointed out that users' satisfaction are enhanced when they are shown with materials that match with their reading proficiency. Scholer et al. [10] conducted a user study on eighty-two users and discovered that the relevance of documents viewed early impacts the assessment of subsequent documents.They also observed that the more difficult the search topics are, the more significant the difference between the two user groups. In this work, we conduct user study with students who has little background in searching patent documents, which increases the difficulty of the task and fits well with our purpose – detecting the Euraka effect.

3 Method

Our proposed method include the following main steps. Automatic extraction of human relevance judgments: (1) extract subtopics (claims) from the patent documents, (2) extract passage-level relevance from Office Actions as the truth data (user received relevance), (3) extract the user perceived relevance from query logs. Then, we (4) fit the difference between URR and UPR into a local polynomial regression model, and (5) based on the model parameters, determine the existence of the Eureka Effect.

Table 2. Final rejection data statistics.

#docs txt	#docs XML	avg # total claims	avg # claims rej.	avg # prior art cited	avg docs / claim rej.
1.6M	3.0M	12.74	8.94	1.55	2.20

3.1 Automatic Extraction of Relevance Judgments

We propose an automatic approach to generate ground truth (URR) from the official action (OA) documents that are available on USPTO PAIR.[1] An OA is written by patent examiners and explains which prior art they used as evidence to *reject* various claims in the patent application. We extract this information from their descriptions and transfer them to the input format of our metric scripts. For a patent application, its corresponding office actions $O = \{O1, O2, O3, ...\}$, including non-final office actions, final office actions and the examiners' answers, are processed to extract a set of evidence, including reference documents, reference passages and reasoning paragraphs. The evidence is then used to extract the actual passages and documents from the references. Most OAs used in this process issue rejections to patent applications.

We confine our data collection to Final Rejection office action documents. The dataset is constituted by a series of official actions from the year 2012. All image information and cover sheet have been removed. We assessed and appended the relevance score of each prior art cited within the Final Rejection to the patent application. Each Final Rejection typically cites between 1 to 4 prior arts with an average number of 1.87 citations. On average, 1.29 documents are used to reject a given claim. Each Final Rejection had an average number of 12.19 claims rejected. More dataset statistics can be found in Table 2.

3.2 Detecting the Eureka Effect

The Eureka effect can be formulated as a function about the difference between UPR and URR. We propose to use non-parametric regression to model the learning curve. To reduce the model bias, local polynomial regression estimator is chosen for our task instead of the most commonly used kernel regression.

For a polynomial $P_x(u; a(x)) = a_0(x) + a_1(x)(u - x) + \frac{a_2(x)}{2!}(u - x)^2 + ... + \frac{a_p(x)}{p!}(u - x)^p$, its coefficients $a(x)$ can be estimated by minimizing the weighted sums of squares

$$\sum_{i=1}^{n} w_i(x)(Y_i - P_x(X_i))^2,$$

where $w_i(x) = \frac{K(X_i - x)}{h}$.

The local polynomial regression estimation can be solved by minimizing the least squared error and we get:

$$\hat{m}_n(x) = \sigma_{i=1}^{n} l_i(x) Y_i$$

where $l(x)^T = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x$, $e_1 = (1, 0, ..., 0)^T$, X_x is a vector representation of the coefficients, and W_x is a diagonal matrix whose (i, i) component

[1] http://portal.uspto.gov/pair/PublicPair.

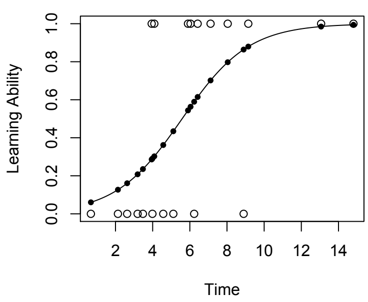

Fig. 1. User's learning ability.

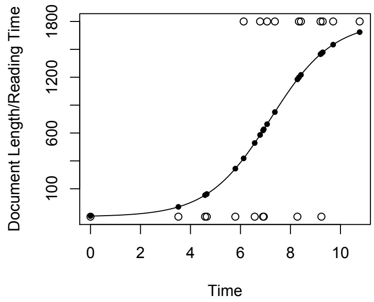

Fig. 2. Reading speed.

is $w_i(x)$. We then measure the gap between the adjacent learned coefficient $w_i(x)$. When a large gap is detected, we say an Eureka effect is found.

4 Preliminary Experimental Results

In this section, we report findings in our preliminary experiments. We conducted a user study to evaluate the relationship between UPR and URR. Twelve graduate students from various majors participated in the study. They are proficient with the use of computers, highly proficient in English, and have little knowledge about the topic described in the patent documents. This experiment setting makes sure that the user information needs are highly complex and the topics of search tasks are unfamiliar to the users. The dataset is the publicly available patent dataset[2] from the U.S. Patent and Trademark Office (USPTO). We automatically extract ground truth relevant documents from the official search reports published at PublicPAIR as described earlier.

One quantity to measure users' ability of making correct judgment is the difference between the relevance grade given by ground truth (URR) and by the users (UPR). We consider it is a measure of user's learning ability. Fig. 1 plots the curve for learning ability fitted by local polynomial regression (LPR). LPR suggests an Eureka effect happens in the middle of the S-shaped learning curve.

Fig. 2 plots the curve for average reading speed per document fitted by local polynomial regression. The plot suggests an S-shaped learning curve too. We can see that user's learning speed is low at the beginning for a relatively long time, and it accelerates steeply after the user spends more time learning and has accumulated enough background knowledge. As a user continues learning and the accumulated knowledge reaches a high plateau, the learning speed tapers off. An Eureka effect happens in the middle of the S-shaped learning curve.

5 Discussion and Conclusion

In search tasks that show a high complexity, users with zero or little background knowledge usually have the common experience of sudden understanding a pre-

[2] http://www.google.com/googlebooks/uspto-patents-applications-text.html.

viously incomprehensible problem or concept. In the context of patent prior art search, this paper introduces a novel notion of Eureka effect in complex search tasks that leverages the sudden change of user's perceived relevance observable in the search log data. An initial set of preliminary experiments are done using non-parametric regression to model the learning curve. The preliminary experimental results are encouraging – we are able to observe the S-shape learning curve in the search process. It suggests that in patent prior art search, at the beginning a user could not easily distinguish relevant documents from non-relevant ones since the terms used in patent documents are often very abstract, rare, and difficult. As the user learns more about the search topic from the retrieved documents, it is possible that he can suddenly understand quite a lot of related materials, which are not previously comprehensible, all at once.

As part of attempts to model the learning curve, our work focuses on detecting the Eureka effect in complex search. Learning curve is an important concept in learning-intensive search tasks, which will potentially enable search engines to improve on providing users with the right documents at the right time.

6 Acknowledgments

The research is supported by NSF grant CNS-1223825 and DARPA grant FA8750-14-2-0226. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Portions of this work were conducted under the umbrella of a larger project at the USPTO.

References

1. Borlund, P.: The concept of relevance in ir. Wiley Online Library 2003. Journal of the American Society for information Science and Technology 54(10), 913–925 (2003)
2. Collins-Thompson, K., Bennett, P.N., White, R.W., de la Chica, S., Sontag, D.: Personalizing web search results by reading level. In: CIKM '11
3. Collins-Thompson, K., Callan, J.: Predicting reading difficulty with statistical language models. Wiley Subscription Services, Inc., A Wiley Company 2005. Journal of the American Society for Information Science and Technology 56(13), 1448–1462 (2005)
4. Dean-Hall, A., Clarke, C.L.A., Hall, M., Kamps, J., Thomas, P., Voorhees, E.: Overview of the trec 2012 contextual suggestion track. In: TREC'12
5. Grossman, M.R., Cormack, G.V., Hedin, B., Oard, D.W.: Overview of the trec 2011 legal track. In: TREC '11
6. Hartz, S., Ben-Shahar, Y., Tyler, M.: Logistic growth curve analysis in associative learning data. Animal Cognition (2001)
7. Heilman, M., Collins-Thompson, K., Eskenazi, M.: An analysis of statistical models and features for reading difficulty prediction. In: EANL'08
8. Kidwell, P., Lebanon, G., Collins-Thompson, K.: Statistical estimation of word acquisition with application to readability prediction. In: EMNLP'09
9. Murre, J.M.: S-shaped learning curves. Psychonomic Bulletin & Review (2013)
10. Scholer, F., Kelly, D., Wu, W.C., Lee, H.S., Webber, W.: The effect of threshold priming and need for cognition on relevance calibration and assessment
11. Zhao, L., Callan, J.: How to make manual conjunctive normal form queries work in patents search. In: TREC (2011)