The Economics in Interactive Information Retrieval

Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
Leif.Azzopardi@glasgow.ac.uk

ABSTRACT

Searching is inherently an interactive process usually requiring numerous iterations of querying and assessing in order to find the desired amount of relevant information. Essentially, the search process can be viewed as a combination of inputs (queries and assessments) which are used to "produce" output (relevance). Under this view, it is possible to adapt microeconomic theory to analyze and understand the dynamics of Interactive Information Retrieval. In this paper, we treat the search process as an economics problem and conduct extensive simulations on TREC test collections analyzing various combinations of inputs in the "production" of relevance. The analysis reveals that the total Cumulative Gain obtained during the course of a search session is functionally related to querying and assessing. Furthermore, this relationship can be characterized mathematically by the Cobbs-Douglas production function. Subsequent analysis using cost models, that are grounded using cognitive load as the cost, reveals which search strategies minimize the cost of interaction for a given level of output. This paper demonstrates how economics can be applied to formally model the search process. This development establishes the theoretical foundations of Interactive Information Retrieval, providing numerous directions for empirical experimentation that are motivated directly from theory.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval:Search Process; H.3.4 [**Information Storage and Retrieval**]: Systems and Software:Performance Evaluation

General Terms

Theory, Experimentation, Economics, Human Factors

Keywords

Retrieval Strategies, Production Theory, Prosumer Theory, Consumer Theory, Simulation, Evaluation

SIGIR'11, July 24–28, 2011, Beijing, China.

1. INTRODUCTION

Interaction in the search process is usually required in order to find the desired amount of relevant information, especially in the context of topic retrieval. Often the user will need to pose a number of queries and examine numerous documents before their underlying information need is satisfied [6, 18]. Given that searching for information requires user effort (and thus a cost), it is interesting to consider what kinds of *search strategies* a user could or should employ to efficiently undertake their search task. Broadly speaking, we can think of search strategies as different ways to interact with an Information Retrieval (IR) system. So, for example, to obtain the desired amount of relevant information in a cost efficient manner, what search strategy should a user employ? Should they:

1. pose a handful of queries and assess deeply into the ranks,
2. pose numerous queries and assess only the top ranks,
3. or, invoke some other combination of interactions?

In this paper, we aim to examine such strategies by applying microeconomic theory to Interactive Information Retrieval (IIR). We argue that the process of interaction between a user and a system can be modeled as a series of inputs (queries, assessments, etc) that "produce" an output (utility/gain from finding relevant items). Under this view, we can adapt techniques from microeconomics, in particular, production theory to analyze the interaction in the search process using formal methods. By framing the search process as an economics problem, it is possible to ask questions such as, what search strategy (i.e. *combination of inputs*) will minimize user effort (i.e. *cost*) for a given level of utility/gain (i.e. *output*) when using a particular retrieval system (i.e. *technology*)?

Being able to answer such questions is important for IIR because while numerous behavioural and observational studies have been conducted, there is a lack of any formal theory to explain why such behaviours and observations are witnessed [5]. For example, in practice users often issues short queries [22], but longer queries have been shown to be more effective [14]. Kestalous *et al* [16] tried to justify this strategy empirically, and showed that a series of extremely short queries can be quite effective for finding one highly relevant document. But searching for a number of relevant documents often requires numerous queries to be posed during the search session [15]. And generally, users will usually only examine the first page or so of the result list [22, 17]. However, users of Boolean systems will often examine up

to 200 documents [17, 11]. The variation in search strategy is believed to be, in part, due to user adaption. Smith and Kantor [19] showed that users can adapt to degraded systems by modifying their search strategy. In their experiments, users increased the number of queries they issued to compensate for a poor system. While, this enabled the users to find relevant material, it did come at a greater cost. While these are interesting observations, is it possible to explain why users interact and behave in such a manner? To understand why, and perhaps show that these observed behaviors are optimal or justified in some way, we need to be able to formally model the search process. By applying microeconomic theory to IIR it may be possible to: (i) develop such formal models, (ii) provide arguments for particular courses of interaction, and (iii) suggest alternative ways for users to interact with systems, such that they minimize their effort/cost. To this end, we describe how microeconomic theory can be applied to model the search process. Then, we explore its application by performing a large scale simulation that evaluates an array of search strategies on various retrieval systems to determine which strategies are feasible and which are cost efficient.

2. THE ECONOMICS IN IIR

Economics provides a series of tools and techniques for analyzing social phenomena [23], and can be applied to IR in a number of different ways. Varian in his SIGIR 1999 keynote address "Economics and Search" presented three suggestions on how economics could be useful in IR [24]: (1) to examine the economic value of information using consumer theory, "where a consumer is making a choice to maximize expected utility or minimize expected cost", (2) to obtain better estimates of the probability of relevance, and (3) to apply Stigler's theory on Optimal Search Behavior to IR. Despite these promising suggestions, little research has been undertaken investigating the use of economics and economic theory within IR. However, in line with (2), Wang and Zhu [25] used mean-variance analysis from economics to develop Modern Portfolio Theory to obtain better estimates of document relevance. While, the work in [2] employed methods from economics to conduct an analysis of query length showing that the law of diminishing returns applies to querying. These past works provide the inspiration and motivation for this research. In particular, we follow Varian's first suggestion on applying consumer theory in IR. However, here, we shall apply production theory, instead (for reasons which we shall explain later). Thus, in the remainder of this section, we shall describe how the theory of production can be adapted to model the search process.

2.1 Modeling the Search Process

Interactive Information Retrieval is a non-trivial process consisting of a multitude of factors, interactions and variables, from user context to system configurations [10]. Trying to incorporate all of these complexities would result in a rather unwieldy model. Since one of the goals of this paper is to inform on the cost efficiency of different search strategies, then we shall concentrate on the main interactions between a user and a system.

The model that we shall be defining is based upon production theory [23]. In production theory, a firm produces an output (such as goods or services), and to do so requires *inputs* to the process (usually termed, capital and labour) [23]. The firm will utilize some form of *technology* to then produce the *output* given the inputs. The process of production is similar to the search process, which we model as follows: the output of the search process is the utility or gain obtained from the relevant documents found, and the inputs to the process we have chosen consist of: (i) the number of queries, (ii) the length of queries, and (iii) the depth of assessment per query. Each of these inputs will directly influence the number of relevant documents found during the search process. For example, query length has been shown to directly relate to performance [2], while the number of documents that are assessed provides an upper bound on the total number of relevant documents which could be found. Given this abstraction of the search process, we can define a search strategy as a combination of inputs $(\mathbf{Q}, \mathbf{D})$ for a given query length L, which a user could employ, where:

Q the number of queries that the user will issue,

D how many documents the user will assess per query.

So the particular combinations of inputs describe potential user search strategies. Then the technology engaged by the user to produce/find relevant documents is, of course, a particular retrieval system. Through the course of interacting with the system the output of the search process, unlike production theory, is not a good or service, but a certain amount of utility. Here we consider the total Cumulative Gain acquired over the search session as a measure of utility/output. This abstraction reduces the search process down to the core variables which directly influence how much utility a user receives through the course of interaction with the system.

Now, depending on the particular retrieval system employed, different **technological constraints** will be imposed upon the search process such that only certain combinations of inputs will produce a given or specified amount of gain. In economic terminology, the set of all combinations of inputs and outputs that are technologically feasible can be referred to as the **production set**. However, for the purposes of analysis what is of interest is the boundary case given this production set which is defined by the maximum possible output for a given level of input. The function describing this boundary case is referred to as the **production function**. Applied to search, where we consider the two inputs Q and D for a given L, we can devise a search production function $f(Q, D)$ which will quantify the maximum amount of Cumulative Gain that could be obtained if the user issued Q queries, and assessed D documents per query for a given L using a particular retrieval system[1].

Figure 1 provides an example of the production set for BM25 on the Associated Press Collection for the different input combinations. The upper right hand region enclosed by the black dotted line denotes the set of combinations of Q and D which could be issued to obtain a particular level of gain (i.e. this is the production set). The boundary case denoted by the dotted line is referred to as an isoquant in microeconomics, and denotes the minimum amount of the inputs required to produce the particular level of gain. Note the isoquant represents the most efficient usage given the inputs. Using the isoquant it is possible to estimate the search production function.

[1]While, we could devise a function using all three inputs (Q, D and L), but this would be add a lot more complexity to the model. In order to facilitate explanation and to concentrate on introducing the

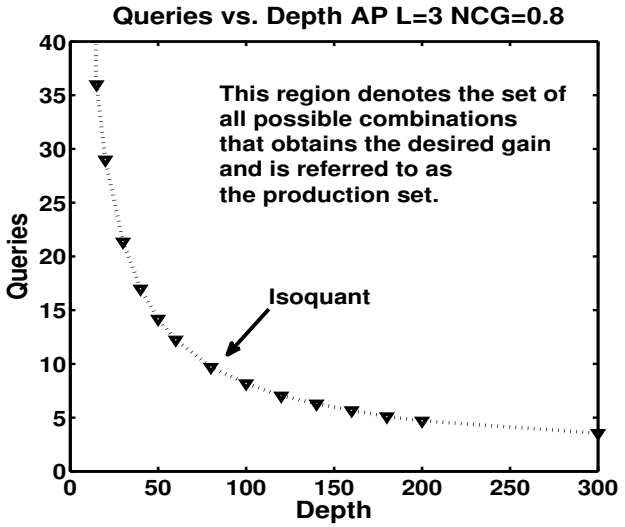

Figure 1: Example: Queries vs Depth on Aquaint collection for BM25. The isoquant denotes the minimum amount of the inputs to produce the specified level of gain.

2.2 Model Limitations and Caveats

Firstly, in terms of the analogy with production theory, it should be noted that the search process is not exactly like the production process. This is because relevance is not really produced, so to speak, it is found within documents. However, the relevant documents found provide the user with some utility or gain. In our formulation the gain is considered the output of the search process. In mapping the search process as an economics problems we also considered using consumer theory as suggested by Varian [24]. In consumer theory, a consumer receives utility from the bundles of the goods that they consume [23]. However, this analogy was less intuitive because searchers do not buy goods or services in the search process. Instead, they exert effort like labour in a production process when they query and assess. While neither production theory nor consumer theory exactly fits the search process[2]: the techniques used in both consumer and production theory are similar i.e. they derive a utility or production function that characterizes the consumer or production process, and then examine the rates of change, maximize utility/profit, minimize expense/cost, etc, see [23] for more details). So either way we shall be applying similar techniques.

Secondly, in terms of IIR, our abstraction of the search process makes a number of assumptions about possible interactions. In reality, users are likely to vary the depth of assessment, the length of queries, and the number of queries that they pose depending on how (un)successful their queries are at returning relevant results given the retrieval system. While, we assume the search strategies denoted by (Q, D) are fixed for a given L, i.e. the user will issue Q queries, each of length L, and assess D documents per query, this helps constrain and reduce the possibilities to a manageable size so that we can perform the analysis. Rather than thinking that these are fixed, if we consider that these variables reflect how a user would search on average, i.e. if a user on average examined D per query, and issued on average Q queries with an average length of L, then this model provides a reasonable approximation of usage. Nonetheless, this abstraction still provides a sufficiently rich representation of the search process which can still provide interesting insights and explanations.

economic concepts to IR, we shall leave such formulations to future work.

[2]Actually, the process appears to be an example of "prosumer theory", where the producer and the consumer are one and the same.

2.3 Research Objectives and Questions

Given this view of the search process, our main objective is to estimate or describe the search production function for interactive topic retrieval mathematically; and in doing so provide a formal model for IIR. During the course of this research we shall also consider the following research questions:

- What combination of inputs are required to achieve a particular level of utility?
- What is the trade-off between querying and assessing? Or, what is the rate of change between querying and assessing? and,
- Given a cost function, which search strategy or strategies minimize the cost of searching?

3. EXPERIMENTAL METHODOLOGY

For the purposes of this study, three TREC test collections were used: the AP 88-89 collection with TREC 1, 2 and 3 Topics (AP), the LA Times collection (LA) with TREC 6, 7, and 8 Topics, and the Aquaint collection (AQ) with TREC 2005 Robust Topics (See Table 1). Each test collection was indexed using the Lemur toolkit[3], where the documents were preprocessed using Porter Stemming and a standard stop list. Since we are interested in interactive ad-hoc querying and retrieval, where the goal is to retrieve a number of relevant documents, we have selected only those topics that have at least 50 relevant documents in Aquaint and AP, and at least 40 relevant documents in LA. We used these cut offs to ensure that there were enough relevant documents to produce sensible values when we examined the various levels of gain. Also, in terms of examining interaction, if we only had a few of relevant documents per topic, then it is likely that only one query would be needed, which would not be particularly interesting.

[3]http://www.lemurproject.org

Collections	AP	LA	Aqauint
Docs	164,597	131,896	1,033,461
Topic Set	TREC 123	TREC 678	Robust 05
No. of Topics Used	87	43	26
Avg. Query Len.	3.3	2.5	2.6
	Mean Average Precision		
BM25	0.2966	0.2145	0.2021
LM2K	0.2967	0.2143	0.2043
BM25AND	0.2038	0.1094	0.1331
TFIDF	0.1867	0.0683	0.0803
TF	0.1435	0.0501	0.0598
BOOL	0.1202	0.0683	0.0460

Table 1: TREC Collection and Topic Statistics for Associated Press (AP), LA Times (LA) and Aquaint (AQ), along with the Mean Average Precision for the retrieval models used in this study.

To explore the influence of different retrieval systems on search behavior we employed six different retrieval systems: two probabilistic systems, BM25 and a Language Model with Dirichlet Prior Smoothing (LM2K). The modified Okapi BM25 function was used with $b = 0.75$, while the Dirichlet Prior was set to 2000 for LM2K. Two vector space systems were also employed one with TF.IDF weightings and other TF weightings. These were included to contrast the probabilistic models as TFIDF and TF usually perform poorly in comparison. We also used two Boolean systems: one which was configured to be Boolean with an implicit AND, sorted by date order (referred to as BOOL), and another which was Boolean with implicit AND, ranked by BM25 (referred to as BM25AND). We used the implicit AND, because according to [17], over 90% of searches under taken using Boolean based models are AND queries, while other operators are rarely used. As previously mentioned, for our experiments we used session based Cumulative Gain (CG) as a measure of the utility/output. However, during the analysis, we also used normalized session based Cumulative Gain (NCG) so that we could aggregate the results across topics.

3.1 Simulated Interaction

Simulation in Information Retrieval has recently attracted a lot of attention, especially in Interactive IR [4]. Simulation enables researchers to conduct carefully designed and controlled experiments to elicit precise answers to research questions and obtain novel insights into the retrieval process [3, 20, 21, 27, 26]. In these studies, the simulations were designed to replicate and mimic the different aspects of the retrieval process as realistically as possible. In this paper, we also employ simulation as part of the experimental methodology but to explore an array of *possible* search strategies that *could* be employed. For example, it is unlikely that a web user would, on average, examine hundreds of documents per query, but it is of interest to see whether this strategy is better or worse than other strategies. In the following paragraphs we shall detail and justify the simulated querying and interaction that was employed to generate the data used in the analysis.

Querying: To provide the queries that will be issued during simulated search sessions, we needed to generate a number of queries per topic. In this paper, we adopt the approach taken in [12], where controlled queries are created, as opposed to probabilistically generating random queries as suggested in [3]. The reason is that we wish to generate high quality queries, as opposed to queries of varying quality. The query generation process was as follows: (1) given a document or set of documents d: construct a weighted term vector $w(t, d)$, (2) rank $w(t, d)$ from highest to lowest, and (3) select the top k terms to be the query q. For our experiments, the weighted term vector is simply the number of times a term appears in d. This is referred to as the popular sampling strategy, which for generating queries in English was shown to produce queries akin to real queries [3]. For our experiments, we generated one query per relevant document and an additional query given all the relevant documents[4]. By generating queries in the manner we should obtain high quality queries that provide the "best case" scenario. This should enable us to obtain a reasonable approximation of the isoquant i.e. boundary case. However, it is interesting to note that the quality of queries produced by the query generation method, far from being complete unrealistic, is in line with the performance of the TREC title queries. For example, when using BM25 the TREC short queries for the AP, LA, and AQ collections resulted in a MAP of 0.297, 0.215, and 0.202, respectively. While, for generated queries of length 3, the performance, in terms of MAP was 0.30, 0.1950 and 0.266, respectively. Also, on the whole the queries generated appeared quite sensible: and reflected the querying behavior observed in the study by Keskustalo *et al* [16], referred to as S3. S3 was the most common strategy they observed, and was where users would issue multiple queries of length three: pivoting on two key words, and then varying the third query term. See Table 2 for examples of some generated queries.

AP Topic 52	LA Topic 313	AQ Topic 303
TREC Title Query		
south african sanction	magnet levit maglev	hubble telescop achiev
Generated Queries		
group south govern	rail line studi	astronom galaxi univers
south govern state	rail line metro	galaxi astronom light
south africa sanction	line rail vallei	galaxi matter star
south africa apartheid	train french speed	telescop studi galaxi
south africa human	train tgv engin	telescop studi space
south jackson africa	train public counti	hubbl star dust

Table 2: TREC Title queries for topics 52, 313 and 303, along with the generated queries of length 3.

Interacting To build up the sequence of interactions we used a greedy best-first approach to select the subset of queries required to obtain the desired level of Cumulative Gain utility. While this might not always achieve the optimal subset of queries it should provide a good approximation for the analysis. Thus, we assumed that the user will issue the best possible query out of all the generated queries first, then issued the next best, and so on until they have found the desired amount of relevant material. The best or next best query is determined by selecting the query which provided the largest increase to the total Cumulative Gain at the given depth D. If the desired level of total Cumulative Gain was not been reached, then the process is repeated until the desired level of total Cumulative Gain has been reached, or all queries were posed. For the total Cumulative Gain during the search session the query had to retrieve relevant documents which had not been seen at previous query iterations. Once the desired level of utility had been reached, the number of queries required Q was recorded for the given assessment depth D and query length L. The number of queries was a free parameter which was determined through the simulated interaction - and averaged over all topics, so from here on when we refer to Q as the number of queries, strictly speaking we mean the average number of queries, and similarly with (normalized) Cumulative Gain, we mean the average (normalized) Cumulative Gain over all topics.

3.2 Experimental Setup

For each topic, we generated a series of queries of length $L = 3$ which is a typical length for user queries [1, 17]. The assessment depths D considered were $D = \{5, 10, 15, 20, 25, 30, 40, 50, 60, 80, 100, 120, 140, 160, 180, 200, 300, 400, 500, 700, 1000\}$ documents per query. We selected this subset to cover the top ranks, and first few pages of search results (where it is typical for search results to be dispatched in groups of 10 to 25) [17]. Also, we consider significantly deeper depths (or multiple pages of search results) up to

[4]The query generation software was written using the Lemur 4.10 Toolkit and the code is available on request.

the typical depth of assessment used at TREC (i.e. 1000). While, most studies report that users only examine the top ranks or first page or two of results [22, 17], it is of interest to determine whether there are alternative strategies which are more cost effective. As stated above, given the depth D and length L, the number of queries Q required to obtain Normalized Cumulative Gain levels of $NCG = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ was determined according to the interaction algorithm employed. For each level of output NCG the corresponding inputs D and Q which were needed to obtain that output given length L were recorded and used in the analysis.

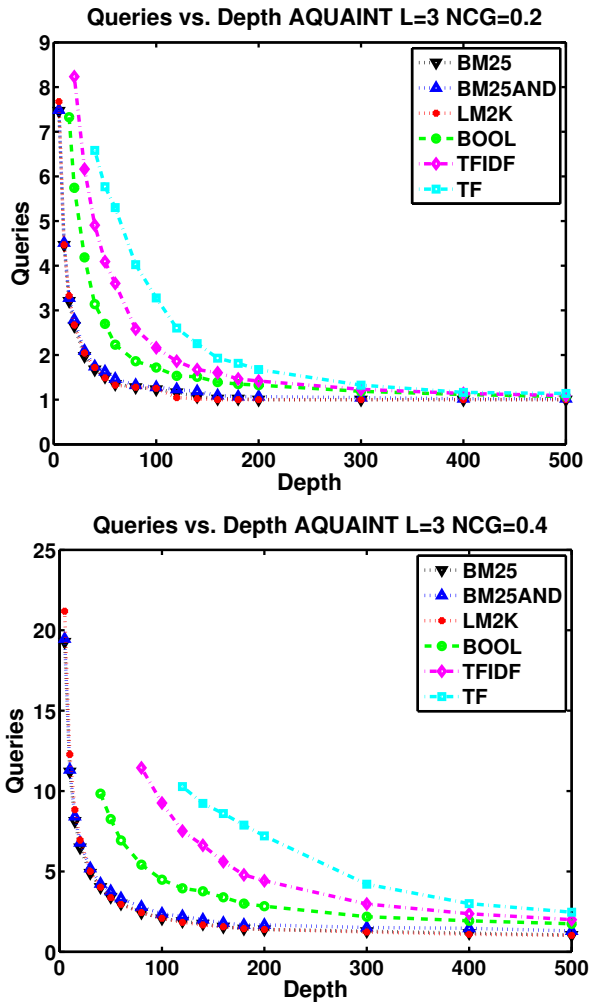

Figure 2: The trade off between the no. of queries and the depth of assessment per query across retrieval models on the Aquaint collection. Top Plot: NCG = 0.2, Bottom Plot: NCG=0.4.

4. ECONOMIC ANALYSIS

In this section, we shall focus mainly on presenting the analysis using the simulated interaction data from the Aquaint collection. However, the general findings, trend and patterns observed were also obtained on the two other collections (see Figure 4 for examples on these collections). We shall start the analysis by examining the production set and estimating the search production function. The two input variables which we shall pay most attention to is the number of queries issued Q and the number of documents assessed per query D (or assessment depth).

Production Set: Figure 2 shows the isoquants for each of the different retrieval models for two gain levels. Combinations where Q and D are greater than the boundary case will yield similar or greater utility. The top plot shows the combinations required to yield an NCG of 0.2, whereas the bottom plot shows the combinations required to yield an NCG of 0.4 for each retrieval model on the Aquaint collection (where $L = 3$). On inspection of these two plots, there are a number of interesting observations to be made:

- As the gain level is increased from 0.2 to 0.4, more queries are required and/or more assessments per query. Naturally, this is to be expected because more documents need to be examined in order to achieve a higher gain. For example, with BM25 it is possible to obtain NCG=0.2 with a combination of $Q = 2$ and $D = 25$, whereas to obtain NCG=0.4 while keeping one of the inputs fixed then either: (a) $Q = 2$ and D increases to 75, or (b) $D = 25$ and Q increases to 5.
- For particular retrieval models some combinations do not yield the desired level of gain. That is, given a particular depth there is no subset of queries which will obtained the specified gain. For example, in the bottom plot, the Boolean Model, TFIDF, and TF do not provide complete sets of feasible combinations over the range of depths examined i.e. their production sets contain fewer technically feasible input combinations than the other retrieval models. For example, using the TF model, a user would have to examine to a depth of at least 140, before they could find a viable search strategy where they could pose enough queries to obtain an NCG of 0.4.
- On the other hand, BM25, BM25AND, and LM2K provide a greater number of combinations ranging from $D = 5$ up to $D = 1000$. This means that BM25, LM2K and BM25AND provide more search strategies to the user, such that: if a user prefers to issue many queries or only a few queries, then provided they examine deeply enough, there are combinations which could yield the desired gain.
- Also, the plots graphically show the trade-off between querying and assessing: as depth D increases, we see that the number of queries required decreases. Whereas, if the number of queries is increased, then the depth can be decreased. We shall examine this phenomena in more detail later in our analysis.

With regards to standard IR evaluation, it is interesting to note the difference between retrieval models. In Table 1 we reported the standard measure of retrieval performance, mean average precision, for each of the different retrieval models. The results, not surprisingly, show that BM25 and LM2K deliver substantially greater retrieval performance than TFIDF, TF, BM25AND and BOOL. However, the graphs of the production set for these retrieval models are far more illuminating: they show the array of search strategies that are possible. Specifically, the plots show how much querying and/or assessing is required in order to obtain the same level of gain. These plots also show

			NCG=0.2			NCG=0.6			NCG =1.0		
Col.	**Model**	**L**	**K**	α	$\mathbf{r}^2$	**K**	α	$\mathbf{r}^2$	**K**	α	$\mathbf{r}^2$
AP	**BM25**	3	5.118	0.608	0.952	6.161	0.566	0.996	4.417	0.586	0.998
	LM2K	3	5.126	0.607	0.9501	5.921	0.563	0.994	3.918	0.565	0.999
	BM25AND	3	5.105	0.612	0.9523	6.062	0.585	0.99	4.365	0.6457	0.994
	BOOL	3	4.027	0.598	0.965	4.563	0.585	0.986	4.025	0.680	0.991
	TFIDF	3	3.928	0.559	0.981	3.693	0.523	0.997	2.127	0.512	0.999
	TF	3	3.704	0.571	0.968	2.998	0.501	0.992	2.246	0.479	0.998
AQ	**BM25**	3	4.819	0.624	0.943	5.394	0.576	0.995	3.733	0.606	0.999
	LM2K	3	4.741	0.618	0.947	5.033	0.562	0.998	3.223	0.586	0.999
	BM25AND	3	4.808	0.634	0.946	5.232	0.581	0.992	3.824	0.689	0.989
	BOOL	3	3.114	0.589	0.963	3.471	0.607	0.997	-	-	-
	TFIDF	3	2.466	0.569	0.984	1.694	0.503	0.996	-	-	-
	TF	3	1.989	0.556	0.981	1.445	0.500	0.979	-	-	-
LA	**BM25**	3	4.008	0.718	0.812	3.82	0.603	0.998	2.876	0.686	0.989
	LM2K	3	3.223	0.585	0.999	3.463	0.611	0.995	2.413	0.645	0.994
	BM25AND	3	4.021	0.747	0.794	3.501	0.68	0.978	2.543	0.799	0.969
	BOOL	3	2.805	0.871	0.977	2.443	0.663	0.986	2.805	0.871	0.978
	TFIDF	3	1.962	0.653	0.991	1.377	0.553	0.977	-	-	-
	TF	3	1.686	0.636	0.992	1.145	0.541	0.989	-	-	-

Table 3: Cobbs-Douglas Production Function Fitting Parameters (K,α) and the r^2 value for 3 gains levels across all models and collections. (-) indicates when no combinations were found for that gain level.

what kind of adaption is required to adjust to systems of varying performance. We know from the work of Smith and Cantor [19] that users can adapt to degraded systems, and in their study they observed users compensating for a degraded system by issuing more queries. If we examine the top plot in Figure 3, then we can see, for example, that for BM25 $Q = 2$ and $D = 25$ to obtain an NCG=0.2, but for TFIDF, if the user fixes D equal to 25, then they need to posed 4 additional queries, i.e. $Q = 6$ to obtain the same gain. These findings are consistent with the work in [19]. However, here we are enable to examine more precisely the differences between systems, and predict how much more interaction is required to compensate for degraded systems.

Search Production Function: While the plots are quite illustrative showing the relationship between querying and assessing, it is our objective to try and characterize this relationship mathematically. Given the shape of the plots, the data appears to be in the form of a Cobbs-Douglas production function (which is one of several types of production functions often used in microeconomics, see [23] for others). Thus, we hypothesize that the *search production function* would take the following form:

$$f(Q, D) = K.Q^{\alpha}.D^{(1-\alpha)} \quad (1)$$

where $f(Q, D)$ is the function that quantifies the total cumulative gain given the inputs Q and D (conditioned on query length L and retrieval system), K provides an indication of the efficiency of the technology (i.e. a greater K will result in more gain), and α is a mixing parameter determined by the technology used. If $\alpha = 0.4$, a 10% increase in querying would lead to approximately a 4% increase in output.

To determine whether the isoquants could be modeled by the Cobbs-Douglas search production function described in Equation 1, we tried to estimate the parameters K and α using the Curve Fitting tool provided in Matlab's statistics toolbox. Table 3 shows the fits for each model, for each collection, when the query length was three and for the NCG values 0.2, 0.6 and 1.0. The K and α values are shown along with the coefficient of determination (i.e. the r^2 value) for each fit. The closer the r^2 value is to 1 the better the fit is to the data, given the specified parameters. For the particular retrieval models and gain levels, which have incomplete isoquants across the range of D explored, then the estimate of $f(Q, D)$ will only hold where D is greater than or equal to the depth at which a combination is technically feasible. For example, when $NCG = 0.4$, TFIDF and TF in Figure 2 both have incomplete isoquants and so their production function is constrained.

For each model and collection, we can see that r^2 for most of the models is quite close to one, indicating that the Cobbs-Douglas function is quite a good fit to the data, and a reasonably good characterization of the gain produced through querying and assessing. This is an important finding because it means that instead of empirically estimating the rates of changes we can use differentiation to obtain the marginal product of querying and assessing, and the marginal rate of technical substitution (see below). If we consider the results for different retrieval models, we note that (1) TFIDF and TF tend to have the lowest values of K indicating these technologies are rather inefficient, while (2) BM25 and LM2K have the highest values indicating that they are more efficient at producing gain. While, this is consistent with the retrieval models retrieval effectiveness (i.e. MAP), the search production function quantifies the retrieval models performance under interaction (and across various possible ways in which the system could be used).

4.1 Marginals or Rates of Change

Of particular interest in microeconomics is the rates of change between the inputs and output (and are often referred to as marginals). Here we describe the *marginal product of querying* and the *marginal product of assessing*. These describe how much the output changes, when an additional query is submitted or an additional document is assessed (i.e. how much more Cumulative Gain do we obtain if we pose one more query, or assess one more document). Given these marginals, then it is possible to determine the *rate of technical substitution*. This would allow us to determine how much more assessing is required if one less query was posed, in the case where we substitute queries for assessments.

The **Marginal Product of Querying** is the increase in output given an increase in querying i.e. if we issue another querying how much more gain will we get for each additional query. Given the search production function defined in Equation 1 the marginal product of querying can be obtained by differentiating with respect to Q:

$$MP_Q = \frac{\delta CG}{\delta Q} = K.\alpha.Q^{(1-\alpha)}.D^{(1-\alpha)} \quad (2)$$

Similarly the **Marginal Product of Assessing** is the increase in output given an increase in assessing: and is obtained by differentiating with respect to D:

$$MP_D = \frac{\delta CG}{\delta D} = K.(1-\alpha).Q^{\alpha}.D^{-\alpha} \quad (3)$$

The marginal products of querying and assessing result in diminishing marginal returns; where the gain gets smaller if you hold one of the inputs constant while increasing the other. So, if querying is held constant, and the depth of assessment is increased then each additional document that is assessed will add less and less to the Cumulative Gain. This is consistent with what we would expect during search because documents are usually ranked in decreasing order of relevance [8]. One of the possible uses of the Marginal Product of Assessing would be to predict when a user would stop examining the ranked list. For example, if we assumed that the user would like to obtain at least an additional g of gain for every n documents that they assess, then we could determine at what depth D the rate of change equals g/n. Beyond that D the rate of change would be lower than the desired and so the user would stop at depth D. While this would be very interesting to determine and empirically test, we shall leave such a direction to future work, and focus on quantifying the relationship between querying and assessing.

The Technical Rate of Substitution: Instead of considering how much output changes by, the technical rate of substitution considers how much of one input we need to increase or decreases, if we decrease or increase the other input in order to hold output at the same level. So for example, how many more documents would we need to assess per query, if we issued one less query. The technical rate of substitution can be defined as:

$$TRS(Q,D) = \frac{\Delta D}{\Delta Q} = -\frac{MP_Q}{MP_D} \quad (4)$$

and it measures the rate at which querying can be substituted for assessing. While, $TRS(D,Q)$ measures the rate at which assessments can be substituted for querying. Note that the TRS is the slope of the line of the search production function. For convenience, we shall not report $-MP_Q/MP_D$, but MP_Q/MP_D, so that it can be intuitively interpreted as the number of additional documents that would have to be assessed per query, if one query was given up.

In Table 4, we report the technical rate of substitution of querying for assessing for BM25, TFIDF and Boolean at several different levels of gain and at particular depths[5]. If we take BM25 for example when the $NCG = 0.2$, then when $D = 5$, a user would have to assess, on average, 1.1 extra documents per query, if they give up one query. Whereas when $D = 30$, they would have to assess, on average, approximately 25 extra documents per query, if they forgo one query. Overall, the trend is that as depth increases, the TRS also increases, and this appears to be at an increasing rate. Essentially the technical rate of substitution enables us to quantify the trade-off between querying and assessing. Next we examine the cost of the different strategies to determine which strategy is the most cost-efficient.

[5]Note, we have not displayed the TRS value when the average number of queries is less than two . This is because the search process requires at least one query to be posed - i.e. it does not make any sense to pose less than one query.

4.2 Cost of Interaction

In order to determine what search strategies minimize the cost to the user, we have constructed a cost function to measure the effort required to obtain the desired level of output for the given inputs. The **user cost function** we shall employ is a linear combination of querying and assessing, and is defined as follows:

$$c(Q,D) = \beta.Q + Q.D \quad (5)$$

where the cost is composed of two parts: the total cost of querying, and the total cost of assessing i.e. the total number of documents examined[6]. Here, the total querying cost is proportional to the number of queries issued Q. And the relative cost of a document versus a query is dictated by the parameter β. If β is greater than one than it is relatively more expensive to pose a query than to assess a document, and vice versa. For these experiments, we determined β by drawing upon the user experiments conducted in [9], where the cognitive load of various interactions in the information seeking process were measured. In these experiments, a dual-task method was used which measured how long it takes for the participant to respond to a secondary task (in milliseconds). It was found, on average, that assessing documents placed a load of 2266 ms on the user, while posing queries was somewhat more taxing with a load of 2628 ms (values taken from Table 11 in [9]). Since we need a relative cost between querying and assessing then we assigned $\beta = 2628/2266 = 1.1598$. This estimate provides a reasonable indication of the relative costs based on the available data which provides some grounding for our analysis. Essentially, this user cost function estimates the relative cognitive effort of querying and assessing. Given that we are examining various search strategies under similar circumstances it should be adequate to make a reasonably fair comparison between strategies. However, in the future it would be interesting to explore alternative cost functions and different parameterizations.

Search Strategy Cost Efficiency: For each input combination, we calculated the cost to the user and we have reported the costs for various combinations in Table 4. The asterisk indicates if the combination was the minimum cost in the set. In Figure 3, for BM25, TFIDF and Boolean retrieval models, we have also plotted Q vs. D for each gain level in the top plots, whilst in the bottom plots we show the corresponding cost across the depths at each level. From inspecting the plots, we can see that as the gain increases so does the cost. For a given level of gain, we can also see that the most cost efficient strategies tend to be the ones where

[6]We acknowledge that this is a rather simple user cost function, but it does capture the main elements of interest and is similar to the cost function proposed in [13] for evaluating faceted browsing strategies. We shall leave the development of non-linear cost models, which cater for other factors like user frustration and fatigue when assessing [7], or the increasing costs of generating queries [18] to future work.

	NCG=0.2				NCG=0.4				NCG=0.6			
Model	**Q**	**D**	**Cost**	**TRS**	**Q**	**D**	**Cost**	**TRS**	**Q**	**D**	**Cost**	**TRS**
BM25	7.5	5	63	1.1	19.3	5	163	0.4	41.1	5	348	0.2
	4.5	10	60	3.7	11.2	10	151	1.2	24.7	10	333	0.5
	3.2	15	59*	7.8	8.1	15	150*	2.5	17.6	15	326*	1.1
	2.7	20	62	12.5	6.5	20	152	4.2	14	20	328	1.9
	2	30	66	25.2	4.9	30	164	8.3	10.2	30	343	3.8
TFIDF	8.2	20	193*	3.2	11.4	80	955	7	14.9	180	2731	12.1
	6.2	30	206	6.4	9.3	100	958	10.9	14.2	200	2882*	14.1
	4.9	40	213	10.8	7.5	120	928*	16.1	9.5	300	2887	31.4
	4.1	50	219	16.2	6.6	140	951	21.3	6.9	400	2787	57.7
	3.6	60	229	22	5.6	160	916	28.7	5.4	500	2705	92.8
BOOL	7.3	15	135*	2.9	9.8	40	428*	5.5	9.3	120	1154*	19.1
	5.7	20	135	5.0	8.3	50	442	8.2	8.2	140	1181	25.3
	4.2	30	140	10.3	6.9	60	440	11.7	7.3	160	1186	32.8
	3.1	40	137	18.3	5.4	80	452	19.9	6.8	180	1246	39.4
	2.7	50	144	26.6	4.5	100	464	30.1	6.3	200	1292	46.8

Table 4: Costs and Rate of Technical Substitutions: For each retrieval model, shown for the first five combinations of Q and D that result in the specified NCG, along with the cost $c(Q, D)$, TRS of Q for D (which is how many more documents would need to be assessed if one query was given up at the given depth). * indicate the depth length which obtains the overall minimum for a given level of gain.

less documents are assessed (i.e. assessing lots of documents imposes a high cost to the user). However, to get the most out of each query usually a number of documents need to be assessed, and of course, the strategy needs to be technically feasible. We can see from the top plots in Figure 3, that for TFIDF and BOOL, in particular, that not all combinations are possible. In these cases, it is the combination with the lowest depth that minimizes costs (usually around a depth of 100-200 documents). For BM25, which provides the user with a greater array of strategies to choose from, we see that the minimum cost is around a depth of 10 to 20 documents. If we inspect the costs in the Table 4, then we see that for BM25 the depth that, on average, provides the lowest cost to the user is examining 15 documents per query[7]. Recall, that this depth is approximately the number of documents a user typically examines [22, 17]. This suggests that state of the art systems which deliver reasonably good performance (like BM25 on these collections) may induce similar behavior in users (i.e. an emphasis on querying more, rather than assessing deeply). For TFIDF and BOOL, however, the strategies that minimized cost varied across the levels of gain. Essentially, as the desired gain increased, more queries and greater depths were required. Interestingly, for the Boolean retrieval model for higher gain levels, the user would have to examine up to 120 documents per query to reach an NCG of 0.6. According to the studies in [17] and [11], this statistic reflects the number of documents users typically examine when using Boolean systems. While for TFIDF, these results show that, a user may be able to adapt to a degraded system (i.e. BM25 vs TFIDF) but this is going to substantially increase the cost to the users. These findings are not definitive explanations for observed user behavior, but they do provide some credence to observed behaviors. Nonetheless, the application of economic theory to IIR process has thrown up a number of possibilities to hypothesize about and test search strategies: and this analysis shows that for a reasonably good retrieval system, like BM25 on these collections, then examining only the top 15 or documents, and posing as many queries as needed to obtain the desired level of gain is a viable and cost efficient strategy to employ.

What happens when β changes? On a hypothetical note it is interesting to consider what would happen if we varied the relative cost β. For example, let's assume that the search system employed provides query assistance functionality like query suggestion [15], which reduces the effort of posing queries (i.e. β decreases). In this case, queries become cheaper, and the preferred search strategy would tend to towards strategies where more queries are posed and less documents examined. While, if queries were more costly to pose (i.e. β increases), perhaps because there was no automatic spelling corrections or exact matching was enforced, then the preferred search strategy would would tend towards strategies where less queries were issued, but more documents examined. This is because assessing documents, under this condition, would be relatively cheaper than posing additional queries. This example shows how a formal model for IIR can be used to better understand and explain the dynamics within the search process.

5. DISCUSSION AND CONCLUSION

In this paper, we have shown how production theory from microeconomics can be applied to IIR by modeling the searching process as if it were a production process. Then, through the course of the analysis we have shown how the tools and techniques from microeconomics can used to describe the dynamics of querying and assessing on various test collections and retrieval models. For instance, we found that BM25 supported a greater variety of search strategies when compared to TFIDF and BOOL. A result that would not have otherwise been found through standard evaluations. We also found that it was possible to mathematically describe the search process through the Cobbs-Douglas production function. From this, we were able to derive the technical rate of substitution from the search production function in order to calculate the trade-off between querying and assessing. After we mapped the inputs to a cost function, we were able to show that for BM25 the search strategy that minimized the cost was when D was around 15. This suggests that a user would only need to examine the first page or so of re-

[7]This was observed for gain levels up to 0.8. To obtain $NCG = 1.0$ then at least 50 documents needed to be examined per query.

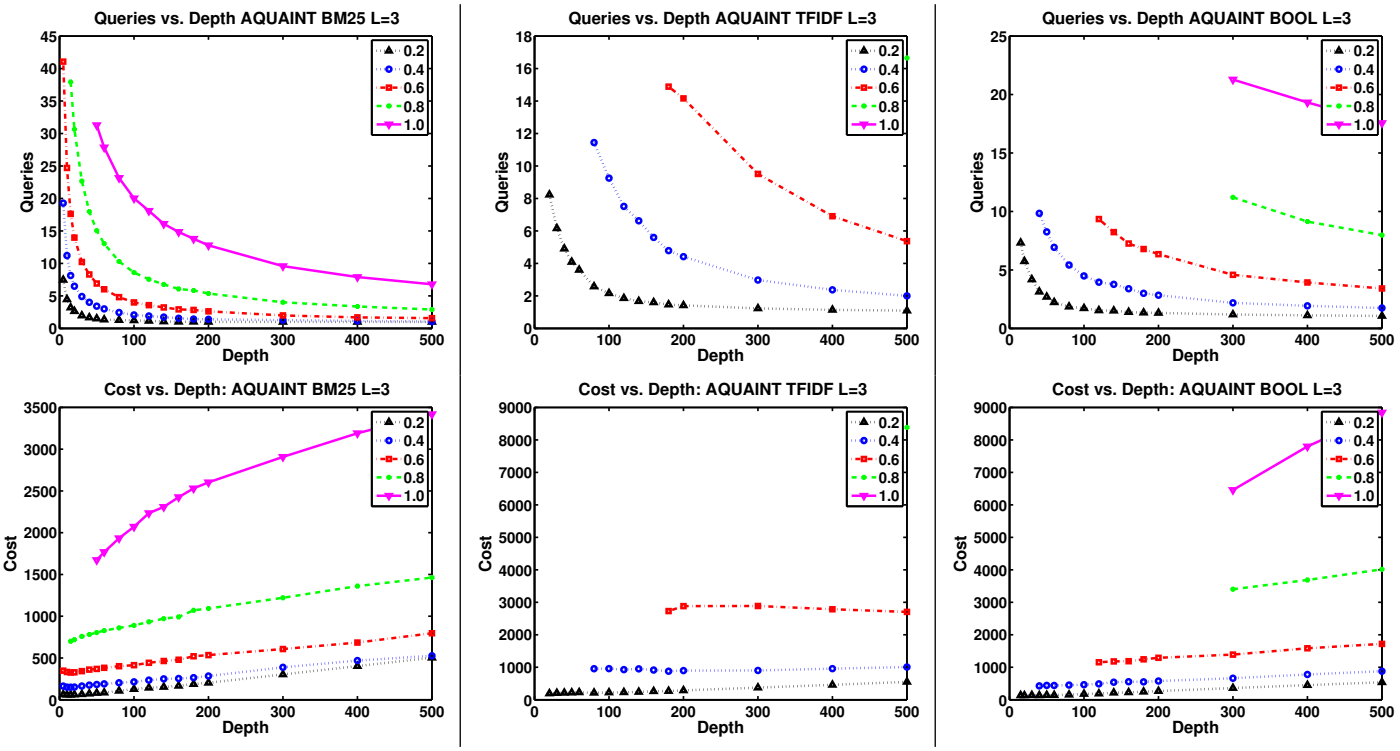

Figure 3: Top: Query vs. Depth, and Bottom: Cost vs. Depth, for the different levels of gain. From Left to Right: BM25, TFIDF and Boolean retrieval models.

sults per query, and continue to pose queries until they reach their desired level of gain to operate the retrieval system efficiently. On the other hand, for retrieval models like TFIDF and BOOL, our study suggests that users would needs to delve deeper into the rankings and issue substantially more queries in order to achieve the same level of gain. These findings are consistent with previous findings obtained from studying users [22, 17, 19] suggesting that there is an economic justification for such search strategies. In our final example, we showed how the theory can be used to generate hypotheses about how users would change their search strategy depending on whether the cost of a query increased or decreased. This work demonstrates that the formal models and methods from microeconomic theory are useful for describing, explaining and hypothesizing about IIR. However, further research is required to empirically explore and test the models and theory developed here, and to address the limitations of this work.

During the course of this research we have tried to point out the main limitations. In particular, we mentioned that the abstraction of the search process could be improved to include more variables and factors (such as query length, and interactions which are not fixed). However, given the initial model of the search process, we were still able to reveal a number of interesting findings regarding the economics of interaction. When we applied production theory to IIR we had to re-consider what the output was – and what was "produced" by the search process. However, this was overcome by considering the gain obtained from finding the relevant documents as the output of the process. Another limitation at the modeling level was that we employed a linear user cost function. However, there is a growing body of work examining the cost of interaction, which has begun to emerge over the past couple of years [20, 7, 9, 13]. As this area develops then the findings from such studies will enable the development of more accurate cost functions, so that search strategies can be evaluated more precisely.

To sum up, we have shown that microeconomic theory provides a number of valuable tools and techniques for understanding IIR. The mapping of production theory to the search process provided a novel way in which we can formally model IIR. Thus, this work provides the foundations on which to build formal models for describing, understanding and explaining the interactions between a user and system, i.e. we can explore the economics of IIR. Clearly, there are a number of ways in which we can develop this work further, from refining the initial models of the search process to applying the theory in practice. In future work we shall:

(i) develop production functions that incorporate other inputs such as query length,

(ii) refine the model of the search process to introduce other variations in interaction, for example, examining snippets in the results list, and then selectively picking documents to assess,

(iii) given the Marginal Product of Assessing predict when users will stop examining the ranked list,

(iv) develop non-linear user cost functions which incorporate factors like user fatigue and frustration, as well as the factor in the costs involved when trying to formulate the nth query on a particular topic, and

(v) investigate how the quality, length and order of the queries affects interaction.

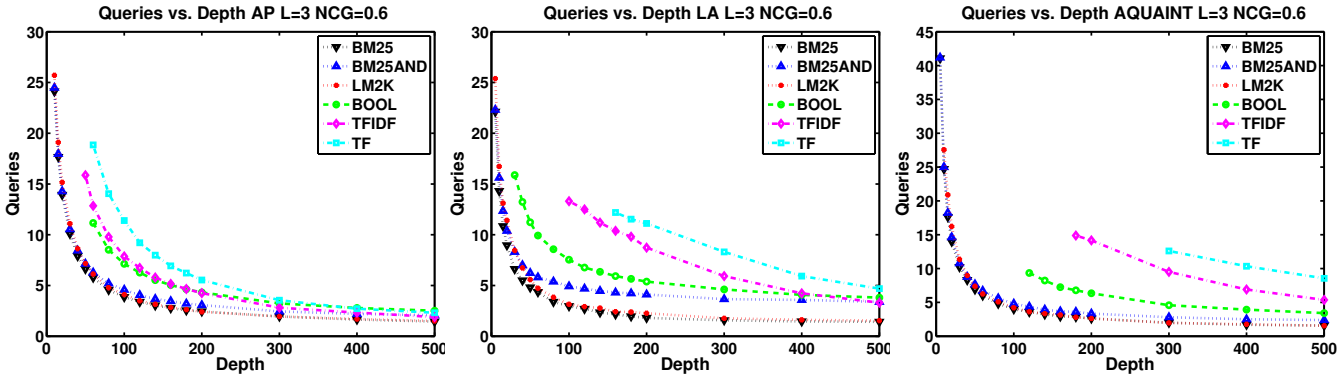

Figure 4: The trade off between No. of Queries and Assessment Depth across the different retrieval models on AP (Left), LA (Middle) and Aquaint (Right). Note how BM25 and LM2K require less input to obtain the desired gain. They also provide more possible combinations to obtain the desired gain compared to the other retrieval models.

Acknowledgments I would like to thank Richard Glassey and Guido Zuccon for the numerous discussions we had on this topic and for their useful comments and suggestions. Their feedback greatly improved the clarity of this work.

6. REFERENCES

[1] A. Arampatzis and J. Kamps. A study of query length. In *Proceedings of the 31st ACM SIGIR conference*, pg 811–812, 2008.

[2] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proceedings of the 32nd ACM SIGIR conference*, pg 556–563, 2009.

[3] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics In *SIGIR '07*, pg 455–462, 2007.

[4] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker. Sigir 2010 workshop report on the simulation of interaction. *SIGIR Forum*, 44:35–47, 2011.

[5] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42:47–54, 2008.

[6] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval: part i: background and theory; part ii: results of a design study. *Journal of Documentation*, 38(2) 61-71 and 38(3) 145-164, 1982.

[7] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proceeding of the 33rd ACM SIGIR conference*, pg 34–41. ACM, 2010.

[8] N. Fuhr. A probability ranking principle for interactive information retrieval. *Inf. Retr.*, 11:251–265, June 2008.

[9] J. Gwizdka. Distribution of cognitive load in web search. *J. Am Soc Inf Sci Tech*, 61:2167–2187, 2010.

[10] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag, 2005.

[11] H. Joho, L. Azzopardi, and W. Vanderbauwhede. A survey of patent users. In *3rd IIiX*, pg 13–24, 2010.

[12] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *6th JCDL*, pg 286–295, 2006.

[13] A. Kashyap, V. Hristidis, and M. Petropoulos. Facetor: cost-driven exploration of faceted query results. In *Proceedings of the 19th ACM CIKM conference*, pg 719–728. ACM, 2010.

[14] D. Kelly, V. D. Dollu, and X. Fu. The loquacious user In *Proc. of the 28th ACM SIGIR*, pg 457–464, 2005.

[15] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd ACM SIGIR conference* , pg 371–378. ACM, 2009.

[16] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based ir evaluation needs extension toward session, In *Proc. of the 5th AIR Symp.*, pg 63–74. 2009.

[17] K. Markey. Twenty-five years of end-user searching, part 1: Research findings. *J. Am. Soc. Inf. Sci. Technol.*, 58:1071–1081, June 2007.

[18] I. Ruthven. Interactive information retrieval. *Annual Review of Inf. Sci. and Technol*, 42:43–92, 2008.

[19] C. L. Smith and P. B. Kantor. User adaptation: good results from poor systems. In *Proceedings of the 31st ACM SIGIR conference*, pg 147–154, 2008.

[20] M. D. Smucker. Towards timed predictions of human performance for IIR evaluation. In *Proc. of the 3rd Workshop on HCIR*, 2009.

[21] M. D. Smucker and J. Allan. Find-similar: similarity browsing as a search tool. In *Proceedings of the 29th ACM SIGIR*, pg 461–468, 2006.

[22] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web *J. Am. Soc. Inf. Sci. Technol.*, 52:226–234, Feb. 2001.

[23] H. R. Varian. *Intermediate microeconomics : a modern approach*. W.W. Norton, New York, 1987.

[24] H. R. Varian. *Economics and search. SIGIR Forum*, 33:1-5, 1999.

[25] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd ACM SIGIR conference*, pg 115–122. ACM, 2009.

[26] R. W. White. Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Inf. Process. Manage.*, 42:1185–1202, Sept. 2006.

[27] R. W. White, I. Ruthven, J. M. Jose, and C. J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM TOIS*, 23(3):325–361, 2005.