The Water Filling Model and The Cube Test: Multi-Dimensional Evaluation for Professional Search

Jiyun Luo[1] Christopher Wing[1] Hui Yang[1] Marti A. Hearst[2]

[1]Department of Computer Science
Georgetown University
Washington, DC, USA
{jl1749, cpw26}@georgetown.edu
huiyang@cs.georgetown.edu

[2]School of Information
University of California, Berkeley
Berkeley, CA, USA
hearst@berkeley.edu

ABSTRACT

Professional search activities such as patent and legal search are often time sensitive and consist of rich information needs with multiple aspects or subtopics. This paper proposes a 3D *water filling* model to describe this search process, and derives a new evaluation metric, the *Cube Test*, to encompass the complex nature of professional search. The new metric is compared against state-of-the-art patent search evaluation metrics as well as Web search evaluation metrics over two distinct patent datasets. The experimental results show that the Cube Test metric effectively captures the characteristics and requirements of professional search.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*Systems and Software* [Performance evaluation]

Keywords

Evaluation, Patent Prior Art Retrieval, Professional Search

1. INTRODUCTION

Information Retrieval (IR) research has advanced to a stage where searching for complicated information needs is a common practice. Evaluation metrics need to reflect this complexity. Professional search activities such as medical record search [14], patent prior art search [19], and legal discovery [3] are examples of complicated information seeking tasks. Most IR research about professional search focuses on query formulation [19] and knowledge discovery [3]. In this paper, we focus on describing the information seeking process and evaluating professional search.

Professional search often involves information needs that consist of multiple *aspects* or *subtopics*. For example, a lawyer needs to discover various aspects of a lawsuit to find defensive materials. An ideal professional search evaluation metric should measure how well a search system allows the user to handle the tradeoffs between time taken to search and how well the returned documents cover the different aspects of the information need.

This paper uses patent prior art search as a motivating example of a complex information seeking task constrained by time. A patent application includes a series of independent and dependent *claims*, each of which defines some of the scope of the claimed invention. The dependent claims refer to their parent claims, which could be either independent or dependent claims. A patent examiner searches published literature, termed *prior art*, to determine whether the claimed invention is novel. In some cases, portions of the patent application document can be used as the input for prior art search rather than a user-issued query. In other cases, prior art search is done using complex Boolean queries whose results are ordered chronologically rather than by relevance.

In the terminology of this paper, each claim corresponds to a different aspect of the full information need. From a patent examiner's perspective, a prior art document is considered a higher quality search result if it can be used to reject a greater number of claims. An evaluation metric to compare ranking algorithms that automatically retrieve relevant prior art needs to consider: (a) the time spent to review the returned results, (b) the number of claims each returned document covers, and (c) how closely the text of the document matches the meaning of the claim in question.

The main metric used in the CLEF-IP 2009-2013 evaluations [9] is PRES [7]. PRES is a function of recall, the ranking of retrieved documents, and the maximum number of results to be checked by the user. However, PRES does not capture the embedded subtopics in patent documents and examiners' desire for wide subtopic/claim coverage in patent prior art search. It also fails to include time to evaluate the literature.

We propose that professional search can be understood by analogy to water filling a compartmentalized cube (see Figure 1). We also assume that searchers would like the multifaceted components of the cube filled as quickly as possible. Based on this model we propose a novel IR evaluation metric which we call the *Cube Test* (CT), which:

- Covers different aspects or subtopics,
- Allows for a single document to cover several subtopics,
- Is time-sensitive, and
- Expresses the tradeoff between time, quality of documents, and diverse coverage of subtopics.

We test the new metric using CLEF-IP datasets and patent applications from the U.S. Patent and Trademark Office

CIKM '13 San Francisco, CA
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2523648.

(USPTO). The results show that our metric effectively captures the characteristics of professional search.

2. RELATED WORK

2.1 Evaluating Patent Search

PRES [7] is the main metric for the CLEF-IP 2009-2013 evaluations [9]. PRES [7] was developed to bias towards recall and is defined as:

$$\text{PRES} = 1 - \frac{\frac{\sum p_j}{NRel} - \frac{NRel+1}{2}}{N_{max}} \tag{1}$$

where $NRel$ is the number of relevant documents, p_j is the rank position of the j^{th} relevant document, and N_{max} is the maximum number of retrieved documents to be checked by the user. PRES is based on normalized recall (R_{norm}), which measures effectiveness by ranking documents relative to the best and worst possible outcomes. The best outcome is defined by retrieval of all relevant documents at the top of the list, and the worst outcome is not retrieving them until the end of the list.

Since many patent searches consist of complex Boolean queries whose results are not relevance ranked, PRES defines a maximum number of documents to be checked by the user (N_{max}). If a relevant document is listed after position N_{max}, it will have zero recall. Although defining N_{max} is a rather ad-hoc procedure, PRES is among a minority of metrics that discuss stopping criteria in an evaluation metric for search tasks. Others include RBP [8] and expected global utility (EGU) [16].

The Cube Test presented in this paper defines a threshold value at which additional returns for the same concept provide no additional benefit for the user. We consider fulfilling the user's need, rather than stopping after an arbitrary number of documents.

2.2 Evaluating Subtopic Relevance

Researchers have looked into evaluating subtopic relevance. Zhai et al. [17] introduce subtopic recall, subtopic precision, and weighted subtopic precision, which evaluate a search system's ability to return documents that cover subtopics of a general topic. The TREC 6 interactive track [6] used *aspectual recall* and *aspectual precision* to measure how well a system allows a user to find documents which supply multiple instances for an aspect.

The α-nDCG metric proposed by Clarke et al. [2] expands upon nDCG by incorporating α, the probability that the assessor made an error when judging whether or not a document is relevant. α-nDCG is derived from a probabilistic model. It models the probability that a document d matches to any subtopic c_i in the information need Q as

$$P(R=1|d) = 1 - \prod_{i=1}^{m}(1 - P(c_i \in Q)P(c_i \in d)) \tag{2}$$

By making a few assumptions, the model simplifies to the gain function for the k^{th} document in a ranked list:

$$\alpha\text{Gain@}k = \sum_{i=1}^{m} rel(d_k, c_i)(1-\alpha)^{r_{c_i,k-1}} \tag{3}$$

where $rel(d_k, c_i) = 1$ if document d_k contains information for subtopic c_i, otherwise 0. $r_{i,k-1} = \sum_{j=1}^{k-1} rel(d_j, c_i)$ is the number of documents ranked up to position $k-1$ that contains information for subtopic c_i. Notice that when $\alpha = 0$, α-nDCG reduces to standard nDCG with the number of matching subtopics used as the graded relevance value.

α-nDCG has several limitations. Clarke et al. [2] self-acknowledges its limited relevance definition. The model regards the document as irrelevant if it contains a previously reported subtopic. Another limitation is that it assumes that concepts are independent of one another and equally probable to be relevant.

We argue that these simplifications as demonstrated in Eq. 3 are incorrect. The user may want to find several documents which all relate to the same subtopic, thereby providing a stronger degree of confidence. Our model takes this into account with a decreasing function. Second, after sufficient information has been collected to fulfill the subtopic need, the user does not benefit if further results only provide evidence for the same subtopic. Our model takes this into account by implementing a threshold value, after which further documents provide no gain. Third, we do not assume all subtopics to be independently important and thereby introduce another parameter (θ) in our model which measures the relative importance of subtopics with respect to one another.

Other recent metrics that consider subtopic relevance include I-rec [11], nERR-IA [12], and D-nDCG [12]. *I-rec@k* is the percentage of matched subtopics at k and is defined as: $I\text{-rec@}k = \frac{|\bigcup_{j=1}^{k} I(d_j)|}{|C_q|}$, where C_q denotes the complete set of subtopics for a query Q, d_j denotes a document at rank j, and $C(d_j)$ denotes the set of subtopics to which d_r is relevant at the cut-off rank k.

nERR-IA [12] assumes that, given a query q with several different subtopics c_i, the probability of each subtopic $P(c_i|q)$ can be estimated and $\sum_i P(i|q) = 1$. It also assumes that document relevance assessments $rel(d, c_i)$ are available for each subtopic. nERR-IA is defined as:

$$\text{nERR-IA} = \sum_i P(c_i|q) \frac{\sum_{j=1}^{k} P(j) P_{\text{dis\_satisfy}}(j-1)}{\sum_{j=1}^{k} P^*(j) P^*_{\text{dis\_satisfy}}(j-1)} \tag{4}$$

where $P(j)$ denotes the relevance probability of a document at rank j. The probability that a user doesn't find the relevant document from document rank 1 to rank j-1 is $P_{\text{dis\_satisfy}}(j-1) = \prod_{p=1}^{j-1}(1 - P(p))$. $P^*(j)$ is the relevance probability of a document at rank j in an ideal ranked list and $P^*_{\text{dis\_satisfy}}(j-1) = \prod_{p=1}^{j-1}(1 - P^*(p))$.

Similar to α-nDCG, D-nDCG [12] assumes that a document is relevant to a query if it is at least relevant to one subtopic. D-nDCG is calculated in the framework of nDCG, with the only change of introducing per subtopic relevance. Instead of a graded judgment for the whole query, D-nDCG uses a weighted combination of subtopic related relevance judgments as the global gain (GG):

$$\text{GG}(d) = \sum_i P(c_i|q) rel(d, c_i) \tag{5}$$

Both nERR-IA and D-nDCG employ $P(c_i|q)$, the probability of the i^{th} subtopic for query q. $P(c_i|q)$ indicates the importance of the i^{th} subtopic for the entire information need. The bigger $P(c_i|q)$ is, the more nERR-IA and D-nDCG favor systems that retrieve relevant documents for the i^{th} subtopic. Saika et al. [12] estimated $P(i|q)$ uniformly or non-uniformly. The former assumes uniform distribution of subtopics, which is equivalent to nDCG. The latter estimates the i^{th} subtopic probability as $2^{n-i+1}/\sum_{k=1}^{n} 2^k \approx 2^{-i}$ when n is reasonably large. Saika et al. reported that both approaches yield similar results.

2.3 Time-Based Evaluation

When evaluating an IR system, regardless of whether the user's focus is precision- or recall-oriented, the time for the user to actually acquire information needs to be considered. Intuitively, documents ranked higher in the returned list save the user time. Käki et al. [5] created a measure termed *immediate accuracy* that represents how often users found at least one relevant result by the nth result selection. Another metric *search speed* measures the relevant answers obtained per minute. This has been modeled through metrics such as nDCG [4], Rank Biased Precision (RBP) [8], and Expected Reciprocal Rank (ERR) [1].

The recent time-based-gain measure (TBG) [13] models a gain function that considers factors such as document length and duplicate documents. TBG models the time actually spent, rather than assuming a one-to-one relationship between document rank and time spent. The model considers whether a user spends more time reading a longer document, or if a summary is read prior to clicking and reading a document. The metric considers the time it takes the user to reach the position at which the document was ranked. Smucker et al. [13] calibrated their model based on a user study. According to their estimation, the time-based gain:

$$\text{TBG} = \sum_{k=1}^{\infty} g_k \exp\left(-\text{Time}(k)\frac{\ln 2}{\text{halflife}}\right) \tag{6}$$

where g_k is the gain of the k^{th} document in the ranked result list, $g_k = 0.4928$ if the k^{th} document is relevant, otherwise 0, and halflife=224 seconds. The expected time for a user to reach a document at rank k is: $\text{Time}(k) = \sum_{j=1}^{k-1} 4.4 + (0.018 l_j + 7.8) P(C = 1|R = r_j)$, where l_j is the length of the document at rank j, r_j is the binary relevance judgment associated with d_j. $P(C = 1|R = r_j)$ is the conditional probability that the user clicks the r^{th} document given document relevance, set to 0.65 if $r_j = 1$, otherwise the probability is set to 0.39.

We argue that TBG oversimplifies the parameters that define the time it takes the user to reach document rank k by keeping these variables constant with respect to time. In our model, we address the tradeoff between search speed and the subtopic relevance and leave for future work a calibration to more accurately model user behavior.

3. THE WATER FILLING MODEL AND THE TASK CUBE

We design a conceptual user utility model called the *water filling* model. We form an analogy between professional search and filling water into an empty container which we call a *task cube*. This model forms the basis of the *Cube Test* for evaluation.

Figure 1 shows the conceptual model of an empty *task cube*, which is intended to represent the user's entire information need. The task cube has unit length of 1. The segments of the bottom side of the cube represent subtopics, and the area of each segment represents the importance of the corresponding subtopic in view of the total information need. The area values are $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$, and θ_6 for subtopics c_1, c_2, c_3, c_4, c_5, and c_6, respectively. Each cuboid, or column, in the cube represents the information need of the corresponding subtopic.

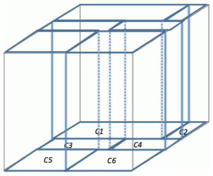

Figure 1: An empty task cube with 6 subtopics.

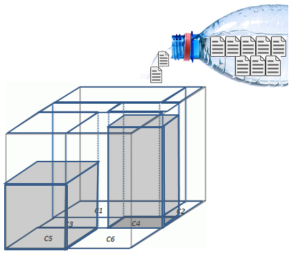

Figure 2: Filling "document water" into the task cube.

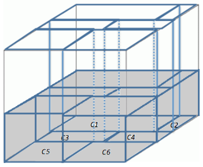

Figure 3: High scoring result.

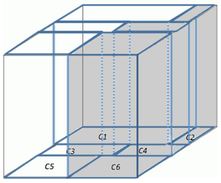

Figure 4: Low scoring result.

We imagine each retrieved document as water that flows into all relevant cuboids. In some cases, there is a one-to-one mapping between returned documents and subtopics; and in other cases, a document's contents can flow into several different chambers of the cube if it is relevant to several different corresponding subtopics. The volume of water added to each cuboid varies depending on its relevance to the subtopic. By filling the cube with "document water," each cuboid contains water filled to various heights ranging from empty to full. As users examine documents they accumulate more information for the total need, but gaining information for each subtopic changes at different rates. Figure 2 shows this imaginary "water filling" process.

The height of the cube constrains the maximum amount of relevant information that a cuboid can contain, mirroring the maximum amount of relevant information that the user cares about for a given subtopic. Finding additional documents that map to a cuboid which is already filled cannot contribute more volume to cuboid and does not provide any additional value to the overall information need of the user. Thus, once any cuboid of the task cube is filled, the corresponding subtopic need is considered satisfied.

The *water filling model* captures multiple dimensions in the general information seeking process. Recall-oriented IR problems prefer a set of documents that can contribute towards filling a greater number of cuboids, while precision-oriented IR problems prefer a few documents to produce high-volume cuboids quickly without requiring many cuboids to contribute. In professional search, we argue that having more cuboids partially filled is more appropriate than a smaller number of high-volume cuboids. We design the actual evaluation metric to favor a system that produces a filled task cube as in Figure 3 rather than a system that produces a filled task cube as in Figure 4.

The objective function of this water filling process is to fill up the entire cube with "document water" as quick as possible. This equates to the goal of information seeking: to find enough relevant information as soon as possible. This intuition is used as the main optimization objective for the Cube Test to evaluate search systems.

Our model does not assume linear document traversal as most IR evaluation metrics assume. Whether the returned documents are traversed linearly or non-linearly, our model takes into account time taken rather than traversal order. Our model does not require or assume that the returned documents are ranked, and thus can handle the output of unranked Boolean queries or a mix of ranked and unranked results. They can also come from results returned for multiple queries in a session.

4. THE CUBE TEST (CT)

Calculation of the *Cube Test* simulates the water filling model for information seeking. Given an information need Q, we can construct a task cube. Letting D be a set of documents returned by system S, we assume that the documents are ordered by the sequence that they are viewed by the user. This order does not necessarily match their ranking orders since the user can view them non-linearly. We calculate the gain for D to fulfill the task cube for Q.

Each document $d_j \in D$ examined contributes some gain of information equal to the volume of relevant "document water" that matches to subtopics in the task cube. It can be calculated as:

$$\text{Gain}(Q, d_j) = \sum_i area_i height_{i,j} KeepFilling_i \qquad (7)$$

Eq. 7 calculates the volume of the "document water" in the cuboids that represent the subtopics that d_j is relevant to. $KeepFilling_i$ is a function specifying whether more "document water" is needed for a subtopic, which depends on the current amount of "document water" in a subtopic cuboid.

Using mathematical representation, Eq. 7 becomes:

$$\text{Gain}(Q, d_j) = \sum_i \Gamma \theta_i rel(d_j, c_i) \mathcal{I} \left(\sum_{k=1}^{j-1} rel(d_k, c_i) < \text{MaxHeight} \right) \qquad (8)$$

where $rel()$, a score in [0,1], denotes the relevance between a document and a subtopic. θ_i represents the importance of subtopic c_i; $\sum_i \theta_i = 1$. $\mathcal{I}$ is the indicator function. Since the task cube is in unit length, $MaxHeight$ is set to 1. $MaxHeight$ provides a constraint such that once a cuboid is full of "document water", it cannot acquire further gain from any future document.

The discount factor Γ is the novelty discount factor that is related to previous gains for the same subtopic $c_i^{1,2,..,t-1}$. Γ can be any formula that models a discounting function. In our experiments, we set

$$\Gamma = \gamma^{nrel(c_i, j-1)}, \qquad (9)$$

where $nrel(c_i, j-1)$ is the number of relevant documents for subtopic c_i in the previously examined documents (d_1 to d_{j-1}). We study the effect of γ on the tradeoff between recall vs. precision in the experimental section.

Eq. 8 calculates the gain of document d_j as a sum of relevance between d_j and each subtopic. When d_j is under examination for subtopic c_i, if c_i has received enough evidence from previously examined documents (d_1 to d_{j-1}), d_j contributes nothing to providing evidence to c_i and therefore its contribution is zero. This is captured in the indicator function: if the accumulated gain for concept c_i is equal to or more than 1, i.e., the maximum amount of relevance this concept needs, even if document d_j is very similar to c_i, d_j's contribution is counted as 0 since it does not add more value in finding useful information for c_i.

For a list of documents D, the accumulated gain up to document d_j can be calculated as:

$$Gain(Q, D) = \sum_j Gain(Q, d_j) \qquad (10)$$

where $Gain(Q, d_j)$ is the per-document gain for document d_j. Note that we do not assume discounted gain according to document rank as in nDCG and many other metrics. This is because that we do not assume the result lists are ranked, which is common in professional search environment. It makes our metric general enough to handle both ranked and unranked search results.

The above models the gain that a user gets in the search process. However, this is only one aspect of the objective function. We also need to fill up the cube as quickly as possible. Thus, we model the overall utility and propose a new evaluation metric *Cube Test CT* as:

$$CT(Q, D) = Gain(Q, D)/Time(D) \qquad (11)$$

It is actually an average speed function that measures the volume (gain) change during a period of time. In more general cases for calculating the speed of how fast we can fill up the task cube at time t, we calculate the speed as the derivative of the gain over time:

$$CT(Q, D) = \frac{\partial Gain(Q, D)}{\partial Time(D)} \qquad (12)$$

which can be further written as:

$$CT(Q, D) = \frac{1}{|D|} \sum_t \frac{Gain(Q, D^t)}{Time(D^t)} \qquad (13)$$

where t is the moment examining the t^{th} document, D^t is the set of documents examined by the user from the beginning up to the t^{th} document, and Time(D^t) is the amount of time taken to examine the documents up to the t^{th} document.

5. EXPERIMENTS

We compare the Cube Test with a number of IR evaluation metrics, including Recall [18], MAP [18], PRES [7], TBG [13], nDCG [15], α-nDCG [2], I-rec [11], nERR-IA [12], and D-nDCG [12]. Among these metrics, Recall, MAP, PRES, TBG and I-rec use binary-relevance, while nDCG, α-nDCG, nERR-IA, D-nDCG and CT use graded relevance.

5.1 Datasets

In the experiments, we compare all the metrics on two patent datasets: USPTO and CLEF-IP 2012.

5.1.1 U.S. Patent Dataset

The USPTO[1] dataset is publicly available. It consists of three million patent applications and publications filed from 2001 to 2013 in XML format with images removed.

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

Table 1: Experimental Dataset Statistics.

Dataset	#docs	avg doc length	#unique terms	#topics	#runs	#subtopic per topic	#subtopic per rel. doc	avg rel. docs per subtopic	avg rel. docs per topic
CLEF-IP	3M	4K	30M	105	31	-	-	-	4.42
USPTO	3.3M	9.9K	26.7M	49	33	8.67	5.75	1.11	1.67

Some claims in a patent application refer to other claims. We treat the former as dependent claims, and treat other claims as independent claims. We assume that each claim is a subtopic query, and assign an importance weight of $\theta_{ind} = 1$ for all independent claims and $\theta_{dep} = 0.5$ for all dependent claims. Then we normalize the subtopic importance value by dividing it by the sum of all subtopics' importance values.

Using the Lemur search engine package,[2] we created 33 runs using various state-of-the-art retrieval algorithms, including tf-idf, language modeling, and Okapi BW25. We also applied several query expansion and refining strategies, including adding titles or important sentences into query keywords, filtering out terms with too high or too low IDF values, and filtering out verbs and adverbs. These runs were created for 49 patent prior art finding tasks. The input query is a complete patent application document. Patents cited by examiners in the office actions are considered relevant prior art. We generated ground truth automatically from the office actions publicly available at USPTO PAIR.[3]

5.1.2 CLEF-IP 2012

Another patent dataset is the European patents provided by CLEF-IP 2012 [9]. These are XML patent documents from the European Patent Office (EPO) prior to the year 2002 as well as over 400,000 documents published by the World Intellectual Property Organization (WIPO). The documents are multilingual, including English, German, and French. We evaluate the 31 official runs from 5 teams that were submitted to CLEF-IP 2012. The ground truth is provided by CLEF-IP.

CLEF-IP's ground truth does not provide graded relevance. However, it provides information about which and how many paragraphs in a relevant document are related to one subtopic. We generate the relevance grade by transforming the number of relevant paragraphs into a scale of [1∼4]. Let Rel be the relevant document set, $TSet$ the topic Set, $NRelPara(d, q)$ the number of paragraphs in document d relevant to topic q, then document d's relevance grade towards topic q is $\lceil \frac{NRelPara(d,q)}{\max_{(d_i \in Rel, q_j \in TSet)} NRelPara(d_i,q_j)} \times 4 \rceil$.

5.2 Discriminative Power

We experiment with several CT variations and differentiate them using subscripts as outlined in Table 2. We employ the measure of discriminative power [10], proposed by Sakai, to evaluate a metric's ability to distinguish IR systems.

In Table 3, we examine the discriminative power for metrics under comparison at the significance level 0.05, which indicates that the metric shows reasonably strong evidence to claim that two runs A and B are different. Our experiments show that when $(\theta_{ind}, \theta_{dep}) = (1, 0.5)$, CT metrics give the best discriminative power. Moreover, there is a large discriminative power drop from CT_{XXc} to $CT_{XX\bar{c}}$, which leads to the conclusion that CT_{XXc} is a better choice than $CT_{XX\bar{c}}$.

Table 2: CT Variations

l	Subtopic importance is generated by a decay function $\theta_k = \frac{1}{log_2(1+k)}$ and normalized the the range of [0∼1].
p	Subtopic importance is θ_{ind} for independent claims, and θ_{dep} for dependent claims. All subtopic importance values are normalized into the range of [0∼1].
t	Assume only the top 10 documents will be examined.
a	Assume documents will be examined from top down, and 150 words will be read from each document.
g	Calculate time using $Time(k) = \sum_{j=1}^{k-1} 4.4 + r_i \times (0.018 l_j + 7.8)$, where $r_i = 0.64$ if d_j is relevant, otherwise 0.39.
c	Assume the task cube has a top cover and subtopic's gain will not increase if the height reaches 1. $\bar{c}$ means that there are no such limit; the subtopic gain is discounted by $rel(d_j, c_i) = rel(d_j, c_i) \times e^{a(-t)+b}$, where t is the t^{th} document to be examined by the user. Smoothing parameters are set as $a = 1$ and $b = 1$.

Table 3: Discriminative power.

USPTO		CLEF-IP	
Metric	disc. power	Metric	disc. power
I-rec	51.5%	CT_{lac}	69.5%
Recall	42.4%	I-rec	69.2%
CT_{ltc}	28.5%	CT_{lgc}	67.1%
nDCG	27.7%	CT_{pgc}	67.1%
D-nDCG	27.1%	α-nDCG	63.9%
CT_{lac}	23.3%	nDCG	58.9%
CT_{pgc}	18.0%	D-nDCG	57.4%
CT_{lgc}	17.2%	Recall	55.9%
PRES	16.5%	PRES	55.7%
α-nDCG	4.9%	nERR-IA	55.5%
nERR-IA	2.5%	CT_{ltc}	51.6%
MAP	0.4%	TBG	47.1%
TBG	0.0%	MAP	45.8%

Furthermore, we find that in the CLEF-IP dataset, all CT metrics show high discriminative power. For the USPTO dataset, Recall and I-rec show the best discriminative power. CT metrics show good discrimination power. The relative low discriminative powers of the CT metrics may be due to the fact that the ground truth of the USPTO dataset are automatically generated from office actions, which unavoidably contains some errors. Nonetheless, the results suggest that the proposed CT metric works well in complex search tasks which demonstrate multiple subtopics.

5.3 Tradeoff between subtopic coverage and single relevant document

In this section we demonstrate CT's ability to be able to adjust its bias between recall-oriented tasks and precision-oriented tasks. We generate two artificial ideal runs. The first run is totally biased towards coverage. It arranges relevant documents to each subtopic in a round-robin fashion and sorts subtopics by their importance θ_i in descending order. The second run is totally biased towards precision, i.e.

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

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

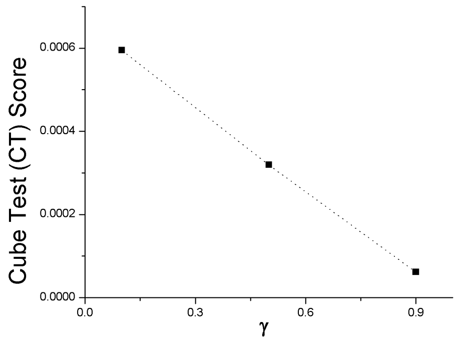

Figure 5: CT Score vs. γ for the coverage run.

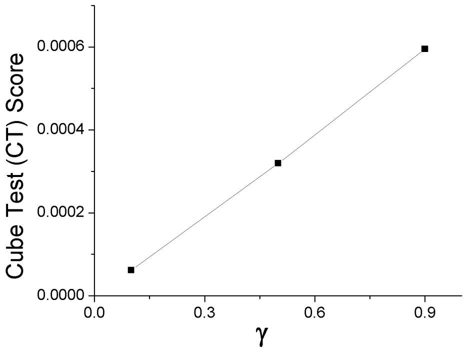

Figure 6: CT Score vs. γ for the single relevance run.

single document relevance. It puts all relevant documents by $rel(d, c_i)$ for a subtopic first, then for the next subtopic. The subtopics are sorted by their importance. We conduct this experiment using CT_{lgc}.

The coefficient γ in Eq. 9 can be used to adjust CT's degree of how much bias towards subtopic coverage. In Figure 5 and Figure 6, the CT metric evaluates the coverage run and the single relevance run with γ changes in the range of [0.1,0.9]. When γ is small, the CT metric has a big discount and rewards novelty heavily. It means that a relevant document will contribute little gain if it is not the first document that covers a subtopic. The CT metric is thus biased towards coverage and rewards more for runs that produce relevant documents that spread across different multiple subtopics. CT therefore shows higher values for the coverage run than the single relevance run. When γ is big, we observe the opposite effect. The CT metric is biased towards precision and rewards more for runs that produce highly relevant documents early. CT therefore shows higher values for the single relevance run than the coverage run. By introducing the tradeoff factor γ, the proposed CT metric is able to balance the tradeoff between recall and precision.

6. CONCLUSIONS

This paper presents a novel evaluation metric – the Cube Test (CT), based on a novel utility model – the water filling model. CT considers both subtopic relevance and the speed required to fulfill the overall information need. It well captures the features and requirements of complicated search tasks. These features include (1) multiple subtopics in a single document is allowed, which is close to reality; (2) the user's information need on certain subtopic can be fulfilled after collecting enough data; (3) different subtopics may have different importance with respect to each query; and (4) an IR system is considered performing better if it allow users to spend less time to gain more information. Experiments demonstrate that our new metric effectively distinguishes and rates IR systems.

Acknowledgments: Portions of this work were conducted to explore new concepts under the umbrella of a larger project at the US Patent and Trademark Office. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions.

7. REFERENCES

[1] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. CIKM '09.

[2] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. SIGIR '08.

[3] M. R. Grossman, G. V. Cormack, B. Hedin, and D. W. Oard. Overview of the trec 2011 legal track. In E. M. Voorhees and L. P. Buckland, editors, *TREC '11*.

[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[5] M. Käki and A. Aula. Controlling the complexity in comparing search user interfaces via user studies. *Information Processing and Management*, 44(1):82 – 91, 2008.

[6] E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: the trec-6 interactive track matrix experiment. SIGIR '98.

[7] W. Magdy and G. J. Jones. Pres: a score metric for evaluating recall-oriented information retrieval applications. SIGIR '10.

[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, Dec. 2008.

[9] F. Piroi, M. Lupu, A. Hanbury, A. P. Sexton, W. Magdy, and I. V. Filippov. Clef-ip 2012: Retrieval experiments in the intellectual property domain. In *CLEF-IP '12*.

[10] T. Sakai. Evaluating evaluation metrics based on the bootstrap. SIGIR '06.

[11] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. Simple evaluation metrics for diversified search results. In *EVIA '10*.

[12] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *SIGIR '11*.

[13] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. SIGIR '12.

[14] E. Voorhees and W. Hersh. Overview of the trec 2012 medical records track. In *TREC'12*.

[15] M. Weimer, A. Karatzoglou, Q. Le, A. Smola, et al. Cofirank-maximum margin matrix factorization for collaborative ranking. In *NIPS'07*.

[16] Y. Yang and A. Lad. Modeling expected utility of multi-session information distillation. ICTIR '09.

[17] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. SIGIR '03.

[18] P. Zhang and W. Su. Statistical inference on recall, precision and average precision under random selection. In *Fuzzy Systems and Knowledge Discovery (FSKD'12)*.

[19] L. Zhao and J. Callan. How to make manual conjunctive normal form queries work in patents search. In *TREC'11*.