Understanding an Enriched Multidimensional User Relevance Model by Analyzing Query Logs

Understanding an Enriched Multidimensional User Relevance Model by Analyzing Query Logs

Jingfei Li
School of Computer Science and Technology, Tianjin University, Room B510, Building 55, Haihe Education Park, Tianjin, P.R. China. E-mail: jingfl@foxmail.com

Peng Zhang*
School of Computer Science and Technology, Tianjin University, Room B510, Building 55, Haihe Education Park, Tianjin, P.R. China. E-mail: pzhang@tju.edu.cn

Dawei Song*
School of Computer Science and Technology, Tianjin University, Room B510, Building 55, Haihe Education Park, Tianjin, P.R. China. E-mail:dawei.song@open.ac.uk

Computing and Communications Department, The Open University, Milton Keynes, United Kingdom. E-mail: dwsong@tju.edu.cn

Yue Wu
School of Computer Science and Technology, Tianjin University, Room B510, Building 55, Haihe Education Park, Tianjin, P.R. China. E-mail: yuewuscd@foxmail.com

Modeling multidimensional relevance in information retrieval (IR) has attracted much attention in recent years. However, most existing studies are conducted through relatively small-scale user studies, which may not reflect a real-world and natural search scenario. In this article, we propose to study the multidimensional user relevance model (MURM) on large scale query logs, which record users' various search behaviors (e.g., query reformulations, clicks and dwelling time, etc.) in natural search settings. We advance an existing MURM model (including five dimensions: topicality, novelty, reliability, understandability, and scope) by providing two additional dimensions, that is, interest and habit. The two new dimensions represent personalized relevance judgment on retrieved documents. Further, for each dimension in the enriched MURM model, a set of computable features are formulated. By conducting extensive document ranking experiments on Bing's query logs and TREC session Track data, we systematically investigated the impact of each dimension on retrieval performance and gained a series of insightful findings which may bring benefits for the design of future IR systems.

Introduction

There have been numerous attempts (Tombros, Ruthven, & Jose, 2005; Xu & Yin, 2008; Xu & Chen, 2006; Zhang, Zhang, Lease, & Gwizdka, 2014) to understand users' search behaviors when retrieving information with search engines, foe example, relevance judgment, satisfaction or dissatisfaction with search results. Understanding how users conduct relevance judgment and what factors influence users' satisfaction with the search results would help researchers design more effective retrieval models and better evaluation methodologies, aiming to further improve users' search experience.

Judging the relevance (or utility) of a retrieved document with respect to a user issued query (representing the user's current information need) is a central task for search engines. A large number of studies (Barry, 1998; Tombros et al., 2005; Xu & Yin, 2008; Xu & Chen, 2006; Zhang et al., 2014) have revealed that there exist a range of complex factors (e.g., topicality, novelty, reliability, understandability, and scope)

*Corresponding Authors.

Received March 21, 2016; revised December 20, 2016; accepted February 8, 2017

© 2017 ASIS&T • Published online 00 Month 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23868

affecting users' perception of relevance for the retrieved documents. However, the existing work is mainly based on small scale user studies, which may not reflect users' natural search scenarios, and the relevance judgments involved were made in a static way that cannot capture the dynamics of a user's information need, search interest, and habit.

To address aforementioned limitations, in this article, we propose to study and understand the multidimensional relevance through analyzing real query logs that record real world user interactions (e.g., query reformulations, clicks, and dwelling time, etc.) with the search engine, as an important supplement to the numerous existing work based on user studies. Specifically, we analyze how different factors affect the users' perception of relevance on retrieved documents within the framework of the Multidimensional User Relevance Model (MURM; Xu & Chen, 2006; Zhang et al., 2014). The existing MURM model includes five dimensions: topicality, novelty, reliability, understandability, and scope. However, it does not explicitly formulate users' search interests and habits, which are important for personalized relevance judgment (Bennett et al., 2012; Dou, Song, & Wen, 2007). For example, different users may prefer different web sites, genres of display, or even languages, and the same query may imply diversified intents for different users with different topics of interest. Therefore, we enrich the existing MURM by proposing two additional user-oriented dimensions, that is, habit and interest.

Furthermore, how to quantify different relevance dimensions in query logs is a key challenge for our study. To address this challenge and make it possible to conduct the log-based study, we formulate a series of computable features for each dimension of the enriched MURM. The features can be utilized to reflect the extent to which users focus on the corresponding dimension, and help us analyze the MURM quantitatively. Based on the proposed features, we conduct extensive document ranking experiments, from which various interesting phenomena have been observed. For example, we find that the dimensions of reliability, interest, novelty, and habit contribute to users' relevance judgment mostly. The topicality dimension, typically considered as a basic factor of relevance, has a relatively little contribution. This is an unexpected, but interesting observation that will be analyzed in a later section. We further analyze the contributions of different dimensions in-depth by observing ranking performance on different types of queries (with different lengths or different Click Entropy values (Click Entropy is a direct indication of query click variation, the less click entropy means the more focus of clicks on a query. Please refer to Dou et al. [2007] for a detailed computation method of click entropy—to be detailed in later sections). We observe that different dimensions dominate users' relevance judgment with respect to different type of queries. For shorter queries or queries with lower click entropy (which can be roughly considered as easier queries), people tend to focus on the reliability (or authority) of the retrieved documents when they decide which to click or to spend more time in reading. For difficult queries, on the other hand, people are more likely to take into account multiple factors, such as interest, novelty, habit, and so on. Another important finding is that "Interest" and "Habit" dimensions are important, confirming the necessity of the enrichment of the MURM by adding these two personalization-related dimensions.

In addition, within the presented framework, we conduct additional empirical experiments on a different application scenario, that is, TREC session search, where the session data are sparser than the real world query log data, search tasks (and the ground truth) are designed (and assessed) manually. In the session search scenario, "Interest," "Understandability," "Topicality," and "Scope" are the most important factors contributing to the relevance judgment, which shows that different dimensions may dominate in different search scenarios. This experiment can be a supplement to that on Bing's query logs.

In a nutshell, the main contributions of this article are summarized as follows:

1. We propose an enriched multidimensional relevance model, which consists of topicality, novelty, reliability, understandability, scope, habit, and interest;
2. We formulate multiple features for each dimension of the enriched MURM to quantify the extent to which the relevance of retrieved documents is correlated to the corresponding dimension;
3. We carry out large scale document ranking experiments on a real-world Web search log and the TREC session search data, from which we gain insightful findings that would provide guidance for the design of future IR algorithms and evaluation methodologies.

The remainder of this article is organized as follows. Section 2 reviews the related work. In Section 3, we present an enriched MURM and formulate the MURM-based features. Extensive empirical experiments and user studies are conducted in Section 4. Finally, Section 5 concludes the article and outlines future work.

Related Work

Our work is highly related to the existing work in relevance judgment. We briefly review this line of research and illustrate the motivation of our work.

Users usually submit queries into a Web search engine to find information in retrieved documents that is relevant to (thus satisfies) the users' information needs. Therefore, judging the relevance of retrieved documents is a central task for search engine users. Indeed, relevance has been regarded as one of the fundamental and central concepts in information retrieval (Saracevic, 1975; Tombros et al., 2005), which influences the design and evaluation of IR models.

Existing research in relevance judgment can be divided into three categories according to the main characteristics of relevance. First, relevance is multidimensional, that is, users tend to adopt multiple criteria or factors beyond topicality when performing their relevance judgments (Barry, 1998; Xu & Chen, 2006; Zhang et al., 2014). Second, relevance is a dynamic phenomenon in a sense that a user may judge the same document as relevant at a certain point of time but irrelevant at another point (Tiamiyu & Ajiferuke, 1988; Katzer & Snyder, 1990). Third, relevance is subjective, that is, different users may express different relevance perceptions (Janes & Mckinney, 1992; Regazzi, 1988). There has been a large body of literature on relevance. In this article, we focus on the MURM that is most related to ours.

Xu and Chen (2006) studied the multicriteria of users' relevance judgment (such as topicality, novelty, reliability, understandability, and scope) based on the semicontrolled user survey data, and drew a conclusion that the topicality and novelty are the essential relevance criteria. Xu and Chen's study focused on the psychological/subjective state of users when answering questionnaires, whereas ours focuses more on the interaction behaviors of users, for example, clicking and dwelling on webpages in a real-world search environment and obtain different findings. Zhang et al. (2014) stated that users' criteria in relevance judgment change in different tasks, some criteria may dominate in some domains (tasks) while being entirely dispensable in others. In this article, we observe similar findings that users would consider more factors for relevance judgment when they are facing difficult tasks (i.e., difficult queries), and consider simple factors when solving easier tasks.

In addition, Yilmaz, Verma, Craswell, Radlinski, and Bailey (2014) studied the relevance from the perspective of "document utility." They demonstrated that the amount of "effort" required for finding the relevant information in a document plays an important role in determining the utility of that document to a real user, which has been ignored by current evaluation mechanisms. Jiang, Ahmed, Shi, and White (2015) proposed to model search satisfaction (which was assessed on multi-point scale by human annotators) using features indicating search outcome, search effort and changes in both outcome and effort during a search session to predict subtle changes in user's satisfaction. In terms of relevance judgment, the "effort" is closely related to the "Understandability" dimension in MURM. Specifically, the better "understandability" of a document, the easier it is for users to read and the less "effort" is required to understand the relevant information in the document.

Recently, Saracevic (2016) published a new book titled *The Notion of Relevance in Information Science* which synthesized what researchers have learned about relevance in several decades of investigation on the notion of relevance in information science. In this book, the author discussed a number of fundamental questions about relevance from the human perspective. To some extent, our work in this article addresses one of them: "What affects relevance assessments?" Our work is different in the sense that we analyze relevance judgment and the corresponding features quantitatively, whereas the book is focused on qualitative analysis and does not cover how the retrieval systems deal with the relevance algorithmically.

Overall, to some extent, the study in this article can be viewed as a supplement to numerous existing user-study based research on relevance.

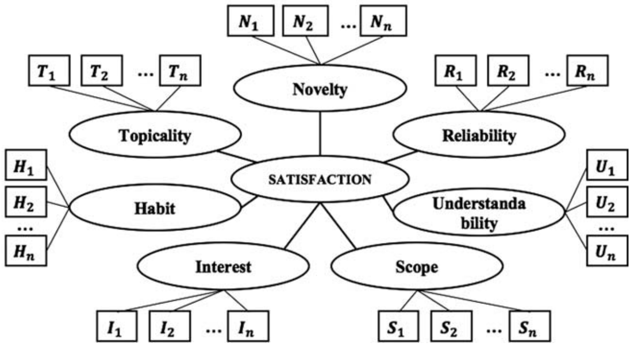

FIG. 1. The enriched multidimensional user relevance model with seven dimensions including topicality, novelty, reliability, understandability, scope, interest, and habit. Corresponding to a specific dimension, there are a list of computable features. This graph is similar to the graph illustrated in Zhang et al. (2014), with two additional dimensions "interest" and "habit" added.

An Enriched MURM

Users' relevance judgment is a complex decision making process (Xu & Chen, 2006; Zhang et al., 2014), which is influenced by multiple dimensions and factors (e.g., topicality, novelty, reliability, readability, and scope, etc.). In this section, we present an enriched MURM (see Figure 1, by adding two personalized dimensions, that is, interest and habit. Furthermore, we formalize a set of computable features for each dimension so that we can quantitatively analyze the MURM on large-scale query log data. In the formulation of features, the criteria of SAT-Click (Satisfied Click) are used, namely, a user dwells on a clicked document for more than 30 seconds or the click is the last click of a search session (Bennett et al., 2012; Jiang, Pei, & Li, 2013).

Topicality (T)

Xu and Chen (2006) defined the topicality in a subjective way as "the extent to which a retrieved document is perceived by the user to be related to her current topic of interest." We consider the "current topic of interest" as the user's current information need, which is different from the "interest" dimension discussed later in section Interest (I). Topicality is traditionally regarded as a fundamental factor for relevance judgment. In this article, for the convenience of quantitative analysis on query logs, we adopt an objective view and redefine the topicality in a computable way as the topical relevance of a document for a query.

In this article, we formalize three features to reflect the topical relevance of retrieved documents (see Table 1). Each feature reflects the extent to which a retrieved document (d) is topically relevant to the user issued query (q). These features have proved effective in ad-hoc retrieval.

Novelty (N)

The novelty of a document can be viewed as either a temporal concept in the absolute sense, or a psychological concept in the relative sense (Xu & Chen, 2006). The former refers to the retrieved document being published recently.

TABLE 1. Features for the topicality dimension.

No	Features and descriptions
T1	$tf \cdot idf$: $tf \cdot idf(q,d) = \sum_{t_i \in q \cap d} c(t_i,d) log \left(\frac{\lvert C\rvert}{df(t_i)}\right)$ where $c(t_i,d)$is the occurrence frequency of t_i in the document d, $\lvert C\rvert$ is the total number of documents in the corpus, $df(t_i)$ is the document frequency of term t_i (Liu et al., 2007) in the corpus.
T2	Query Likelihood (QL): QL measures the probability of topical relevance of a document given a query (Zhai, 2008).
T3	BM25: The Okapi BM25 is an effective probabilistic model (Manning, Raghavan, and Schütze, 2008). In this article, the parameters k_1, b and k_2 are set to 1.2, 0.75, and 7 respectively.

TABLE 2. Features for the novelty dimension.

No.	Features and descriptions
N1	MinKLN: $MinKLN(d) = min_{d_s \in SAT}\ KL(\theta_d, \theta_s)$, is based on the psychological aspect of novelty, where SAT is the set of documents that are previously SAT-Clicked by the user, and $KL(\theta_d, \theta_s)$is the Kullback-Leibler (KL) divergence between the language model of a retrieved document d and the language model of SAT-Clicked documents.
N2	ForgN: $ForgN(d) = \sum_{d_s \in SAT} \lambda^{T_{now} - T_{d_s}} KL(\theta_d, \theta_s)$, considers a Forgetting factor λ (we set $\lambda = 0.9$ in the experiments), where T_{d_s} is the last click time of d_s, T_{now} is the retrieval time. The unit of time is "day" in our experiments, and the feature assumes that the documents clicked a long time ago have a decaying influence on the novelty of a current retrieved document.
N3	WordN: $WordN(d) = \frac{1}{\lvert Set(d) \cap Set(SAT)\rvert}$, is based on the psychological view, but computes the value based on the number of repeated words in the retrieved document. $Set(d) = \{w \mid w \in d, tf_{w,d} > 2\}$ and $Set(SAT) = \cup_{d_s \in SAT} Set(d_s)$. Intuitively, if a retrieved document contains a large number of words that are previously viewed by the current user, the document tends to be not novel to the user.
N4	TempN: $TempN(d) = e^{T_{pro} - T_{now}}$, is based on the absolute temporal view, where T_{pro} and T_{now} are the production time and retrieval time of the document respectively. Note that, because of the lack of necessary data, we only formalize this feature here, but do not analyze it in the experiments.

The latter refers to the document containing information that is new to the user, relative to the users' background knowledge. Both views are important, but reflect different situations. For example, if a user issues a query to search for a news article, he or she will likely prefer the recently published news documents. On the other hand, if the user searches about a series of knowledge about a topic, the retrieved documents are expected to contain relevant and new knowledge to the user, which is not necessarily newly published. Note that, the documents containing old information that the user is already familiar with are not considered as novel documents, although they may have never been read by the user.

In this article, we propose four features that reflect absolute and relative novelty respectively, detailed in Table 2.

Reliability (R)

It is desirable that the retrieved documents or information are not only topically relevant but also reliable. When a user decides to click on and read a document, he or she will be likely to consider the reliability of the document. Reliability can be understood from two aspects: content reliability and source reliability. Xu and Chen (2006) defined reliability as "the degree to which the content of a retrieved document to be true, accurate, or believable." This definition is ideal but impractical in query log analysis, because it is difficult to quantify the "true, accurate, or believable" documents by content. Therefore, we define the reliability in an indirect way (i.e., from the source reliability point of view) by the popularity or satisfaction rate of retrieved documents. Intuitively, if a document is frequently clicked, and the clicks are mostly because of relevance but not position bias, the document is likely to be reliable (following the wisdom of population). Note that there may exist exceptions, for example, when frequently clicked documents are widely spread rumors (not reliable information), but this is beyond the scope of this article.

In Table 3, we present seven features that can reflect the reliability of retrieved documents based on the wisdom of population by mining the global click-through data.

Understandability (U)

Understandability, also known as readability, is a complex cognitive concept that measures the extent to which the content of a retrieved document is perceived by the user as easy to read and understand (Xu & Chen, 2006). Understandability can be influenced by various factors, such as the difficulty of the vocabulary, the complexity of sentences, the layout of webpages, and the reading level of users, etc. Understandability is a relative concept. For example, a document about medical science may be difficult to read for lay users, but less difficult for professional users. Moreover, understandability is also a dynamic concept. Intuitively, a medical document for a new student in medicine is difficult to read, but when the user becomes an expert in the field, the document will become more readable. Search engines should provide relevant results at the right level of reading difficulty.

To model the understandability in a relative and dynamic way, we present seven understandability features based on existing understandability or readability measures and user's click-through data (see Table 4). Although the layout of webpages also influences the understandability, in this article, we focus on the content-based understandability features.

Scope (S)

Xu and Chen (2006) defined the *scope* as the extent to which the topic or content covered by a retrieved document is appropriate to the user's need, that is, both the breadth and depth of the document are suitable. This relevance criterion requires search engines return results with an appropriate amount of information (broad enough but without unnecessary information) according to user's current information need.

TABLE 3. Features for the reliability dimension.

No.	Features and descriptions
R1	SATNum: *SATNum(d)* = *#SAT-Clicks*, is the number of SAT-Clicks on a retrieved document. The larger SATNum, the more reliable the document tends to be.
R2	SATRatio: $SATRatio(d) = \frac{SATNum(d)+\mu \cdot GlobalRatio}{ClickNum(d)+\mu}$, is the ratio of SATClicks on the document *d* clicked by all users, where *ClickNum(d)* is the total number of clicks on *d*, *GlobalRatio* is the ratio of SAT-Clicks in the whole query log, μ is the average number of clicks over all clicked web pages (as a smoothing parameter). The larger SATRatio indicates the better reliability of the document.
R3	SWNum: $SWNum(d) = \sum_{d_w \in W(d)} SATNum(d_w)$, is the number of SAT-Clicks on the source website *W(d)* where the document d comes from. It reflects the source reliability of a specific document. The larger *SWNum* reflects the better source reliability of the document.
R4	SWRatio: $\mathrm{SWRatio(d)} = \frac{SWNum(d)+\mu \cdot GlobalRatio}{CWNum(d)+\mu}$, is the ratio of SATClicks on the source website that the document *d* belongs to, where *SWNum(d)* is the number of SAT-Clicks on the same website; similarly, *CWNum(d)* is the number of all clicks on the same website, μ is the average number of clicks over the websites that have received at least one clicks (as a smoothing parameter). The larger SWRatio also corresponds to better source reliability of the document
R5	ClickRank4Website: it computes the reliability of the website which the retrieved document comes from (Zhu & Mishne, 2009). This feature is mined from the session data, and has been shown effective.
R6	PageRank: it measures the importance of a webpage by computing the number and quality of links to a page (Brin & Page, 1998).
R7	SpamPercentileScore: it indicates the percentage of documents in the corpus that are "spammier." The larger spamPercentileScore is, the better the quality of the webpage will be. Detailed description can be found in (http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/).

In Table 5, we formalize three features to reflect the information scope of a document with respect to an issued query. Note that, we focus on the breadth of information need, and the depth will be left as future work.

Interest (I)

The above dimensions of the existing MURM model have covered a range of factors influencing users' relevance judgment for traditional ad-hoc retrieval task, but do not sufficiently take into account the personalized search scenario. We propose to add the "interest" dimension, because the literature has shown that considering user's interests is important to improve search performance and better satisfy users' information need (Bennett et al., 2012; Dou et al., 2007; Li, Song, Zhang, Wen, & Dou, 2014; Vu, Song, Willis, Tran, & Li, 2014). In this article, the interest dimension refers to the user's search preference for specific topics, for example, IT, literature, politics. Formally, we define *Interest* as the extent to which the retrieved documents are preferred by the user according to his or her topics of interest. Search engines should provide results according to users' search interests, especially for ambiguous queries.

The "Interest" features are based on either the term space or topic space, built upon three temporal views (i.e., session, day, and long term), detailed in Table 6.

Habit (H)

Similar to the Interest dimension, habit also considers user preference, but differs in several perspectives. Interest focuses on the topical preference of users, whereas habit pays more attention to the behavioral preference. Intuitively, one user may be used to accessing some specific web sites for obtaining the desired information, whereas another may prefer other web sites for the same information. Essentially, we define the habit dimension as the extent to which the retrieved documents are preferred by a user according to their sources, genre, and language, and so on. Search engines are expected to satisfy users' habit as much as possible.

To model a user's habit, we propose three features measuring the probability that the retrieved document satisfies the user's habit (see Table 7 for details).

Experiments and Analysis

For an in-depth understanding the multidimensional relevance in IR with the enriched MURM model, a series of extensive learning-to-rank (Burges, 2010) based document ranking experiments are conducted on a real search scenario (i.e., with the query logs from the prominent Bing search engine). To conduct the experiments, we adopt the SAT-Click criteria as an indication of relevance and assign each retrieved document with a corresponding relevance degree. Specifically, we assign the relevance degree 0 to the unclicked documents, 1 to the clicked but not SAT-clicked documents, and 2 to the SAT-Clicked documents. We also conduct a user study to verify the usefulness of the SAT-Click criteria and thus the credibility of our findings.

Within the learning-to-rank experimental framework, we further conduct a supplementary document ranking experiment on TREC session tracks 2013 and 2014 to verify the generalizability of the proposed research method.

Data Statistics of the Query Log Data and Feature Extraction

The first set of experiments are conducted on a subset of query log data, which includes 1,166 randomly sampled users, collected from the Bing search engine for the period between July 1 and July 31, 2012. The detailed information of our query log dataset is shown in Table 8. The distributions of the number of queries over different Click Entropy intervals and query lengths (Len) are also shown in the table. "#Query" indicates the number of selected queries in our study. In the query log, each query is followed by an originally ranked document list returned by the search engine. In the table, we also report the original retrieval performance with respect to *NDCG@10*. It demonstrates a trend that

TABLE 4. Features for the understandability dimension.

No.	Features and descriptions
U1	EWRatio (Easy Word Ratio): $EWRatio(d)=\frac{\lvert EWSet(d)\rvert}{\lvert WordSet(d)\rvert}$, where $EWSet(d)=\{w\lvert w\in d, w\in DaleList\}$, $DaleList$ is the Dale-Chall Word List (Dale & Chall, 1948), a sight words list, and *WordSet(d)* is the word set of a document d.
U2	ReciDWN (Reciprocal Difficult Words Number): $ReciDWN(d)=\frac{1}{\lvert DWSet(d)\rvert}$, where $DWSet(d)=\{w\lvert w\in d,\ w\notin DaleList\}$. It reflects the absolute number of difficult words in a document.
U3	ReciAWL (Reciprocal Average Word Length): $ReciAWL=\frac{\lvert words\rvert}{\lvert Characters\rvert}$, assumes that the understandability of a document is inversely proportional to the average word length in the document. */words/* and */Characters/* are respectively the numbers of words and characters in a document.
U4	FRES (Flesch-Kincaid): $FRES=\ 206.835-1.015\times\frac{total\ \ words}{total\ \ sentences}-84.6\times\frac{total\ \ syllables}{total\ \ words}$, is a standard readability measure (Kincaid, Fishburne, Rogers, & Chissom, 1975). A higher FRES score indicates that the document is easier to read (more readable).
U5	ReciGFI (Reciprocal Gunning Fog Index): $ReciGF=\frac{1}{Gunning\ \ Fog\ \ Index}$, where Gunning Fog Index (GFI) is formalized as $GFI=0.4\times\left[\frac{words}{sentences}+100\times\frac{complex\ \ words}{words}\right]$. GFI measures the reading difficulty of text (Thomas, 2013).
U6	ReciCLI (Reciprocal Coleman-Liau Index): ReciCLI$=1/(0.0588\times\mathrm{L}-0.296\times\mathrm{S}-15.88)$, where L is the average number of letters per 100 words and S is the average number of sentences per 100 words (Coleman & Liau, 1975).
U7	ReciSMOG (Reciprocal SMOG Index): $eciSMOG=\ 1.0430\ \times\sqrt{number\ \ of\ \ polysyllables\times\frac{30}{number\ \ of\ \ sentences}}+3.1291$, where polysyllables are those words containing 3 or more syllables (Hedman, 2007).

TABLE 5. Features for the scope dimension.

No.	Features and descriptions
S1	JaccardIndexQD (Jaccard Index between the sets of query topics and document topics): $JaccardIndexQD=\frac{TopicSet(q)\cap TopicSet(d)}{TopicSet(q)\cup TopicSet(d)}$, where $TopicSet(d)=\{t\lvert p(t\lvert d)>\theta_1\}$, $TopicSet(q)=\cup_{w\in q}\{t\lvert p(t\lvert w)>\theta_2\}$ (we adopt this special method to infer the topic probability for short text fragment such as a query), $p(t\lvert d)$ is the inference probability of topic t given a document d, $p(t\lvert w)$ is the probability of topic t given a word w, $\theta_1\ and\ \theta_2$ are thresholds that are determined experimentally.
S2	CoverRatio (Covering Ratio of query terms in document): $CoverRatio=\frac{uwL(q)}{windows}$, where $uwL(q)$ is the number of fixed windows which contain more than one query terms in the document, *#windows* is the total number of windows in the document, and L is the window size (we set L=16 in this article). A larger CoverRatio value means that the document has a narrower scope and focuses on the query-related content.
S3	CoherenceQD (Coherence of Query meaning in Document content): $CoherenceQD=\sum_{w\in d}cos\left(v_w,v_q\right)/docLength$, where v_w is the word embedding vector for a word w, derived from Google word2Vec toolkit; v_q is the uniform-weighted sum of word vectors for all query terms. We also penalize longer documents. A larger CoherenceQD value indicates a better semantic coherence between a document and the query.

TABLE 6. Features for the interest dimension.

No.	Features and Descriptions
I1	SWI (Session Words-based Interest model): $SWI(d,s)=cosine(V_d,V_s)$, where V_d is $tf\cdot idf$ vector representation of the document d, V_s is the vector for the concatenation of all SAT-Clicked documents in the same session s.
I2	DWI (Day Words-based Interest model [Jiang et al., 2013]): $DWI(d,D)=cosine(V_d,V_D)$, where V_D is the $tf\cdot idf$ vector representation for the concatenation of all SAT-Clicked documents in the same day.
I3	LWI (Long term Words-based Interest model [Jiang et al., 2013]): $LWI(d,L)=cosine(V_d,V_L)$, where V_L are the $tf\cdot idf$ vector representation for the concatenation of all SAT-Clicked documents in the long term history.
I4	STI (Session Topic-based Interest Model [Vu et al., 2014]): $STI(d,s)=cosine\left(V_d^T,V_s^T\right)$, where V_d^T is the topic vector of the document d, each element in the vector corresponds to the probability that the document is relevant to a specific topic T_i (The topic space can be constructed from all SAT-Clicked documents in the global logs by a typical topic modeling approach such as the Latent Dirichlet Allocation [LDA], and we set the total number of topics as 200 when training the topic model), $V_s^T=\sum_{d_s\in SAT_s}V_{d_s}^T$.
I5	DTI (Day Topic-based Interest Model): $DTI(d,D)=cosine\left(V_d^T,V_D^T\right)$, where D is the set of SAT-Clicked documents in the same day.
I6	LTI (Long term Topic-Based Interest Model): $LTI(d,D)=cosine\left(V_d^T,V_L^T\right)$, where L is the set of SAT-Clicked documents in the long term history.

shorter queries (e.g., query length of 1 and 2) and queries with smaller click entropy (e.g., less than 2) tend to have a better initial retrieval performance. Moreover, the original retrieval performance can reflect the difficulty of queries. To some extent, the longer (shorter) queries or queries with a larger (smaller) click entropy can be seen as difficult (easy) queries.

For each query-document pair, we extract the features (as formulated in the previous section) corresponding to different dimensions in the enriched MURM. Because the

TABLE 7. Features for the habit dimension.

No.	Features and descriptions
H1	ProbW: $ProbW(d\|u)=\frac{Domain(d,u)+\mu\cdot ProbW(d\|G)}{SAT(u)+\mu}$, measures the probability that a user u prefers a source Website that a document d belongs to. *Domain(d, u)* is the number of times that the user has accessed (SAT-Clicked) the source website of d, *SAT(u)* is the total number of SAT-Clicks by the user, *ProbW(d/G)* is the background probability for the global query logs, μ is the average number of SAT-Clicks per user, as a smoothing parameter.
H2	ProbDL: $ProbDL(d\|u)=\frac{Level(d,u)+\mu\cdot ProbDL(d\|G)}{SAT(u)+\mu}$, measures the probability that a user u prefers a certain length level of the retrieved document. To do this, we devide the length of documents into several intervals, that is, [1,100], [100,200], [200,300], [400, 500], [500, ∞]. This feature assumes that different users may prefer different document lengths. *Level(d, u)* is the number of SAT-Clicks by the user u on documents with a specific document length level, μ is the smoothing parameter. *ProbDL(d/G)* is similar to *ProbW(d/G)* in H1. Note that, this feature can also be understood as a kind of tendency/habit for different users to select documents of different layouts (e.g., document length).
H3	ProbL: $ProbL(d\|u)=\frac{Language(d,u)+\mu\cdot ProbL(d\|G)}{SAT(u)+\mu}$, measures the probability that a user u prefers a specific language. Note that we do not analyze this feature in the experiments because the query log data used are in English. *ProbL(d/G)* is similar to *ProbW(d/G)*.

TABLE 8. Statistics of the query log data used in our experiments. Ori.Perform.

Items	Count	Len	#Query	Qri.Perform.	CE	#Query	Ori.Perform.
Users	1,166	1	53,051	0.9746	[0,2)	193,049	0.8739
Queries	540,258	2	51,438	0.9011	[2,4)	10,387	0.5532
Clicks	474,553	3	38,656	0.8035	[4,∞)	1,1124	0.4162
SAT-Clicks	359,902	4+	61,415	0.6936	-	-	-

Note. Represents the original retrieval performance with respect to NDCG@10.

absolute values of a feature for different queries might be incomparable (Liu, Xu, Qin, Xiong, & Li, 2007), we perform a query-based normalization for each feature according to Equation 1. Suppose that, for a query q, a search result list with N^q documents is retrieved. A feature for i^{th} document d_i corresponding to the query q is denoted as $f^q_{d_i}$, the normalized feature $normf^q_{d_i}$ is computed as follows:

$$normf^q_{d_i}=\frac{f^q_{d_i}-\min\left\{f^q_{d_k}\right\}}{\max\left\{f^q_{d_k}\right\}-\min\left\{f^q_{d_k}\right\}|},\quad k=1,\ldots,N^q \qquad (1)$$

To guarantee that each query has sufficient long-term historical interaction information when extracting the features for the dimensions of interest, novelty and habit, we only extract features for the queries that occur in the last 12 days (some noisy queries which contain only special characters are excluded). Note that, all stopwords in the data set are removed and stemming is performed with Porter stemmer.

Understanding the Enriched MURM With Document Ranking Experiments

This subsection reports the document ranking experiments to understand the enriched MURM from a novel perspective of real query log analysis (instead of the traditional user study with artificial task settings). To this end, we integrate features of different dimensions into a well-known learning to rank algorithm called LambdaMART (Burges, 2010), which is implemented in the open-source RankLib (https://sourceforge.net/p/lemur/wiki/RankLib/, with default settings) to re-rank the original results returned by the search engine. The performance of ranking models with respect to different dimensions for different categories of queries are reported and analyzed. In this way, we can gain insights about how different factors (dimensions) contribute to the relevance judgment in real search scenario.

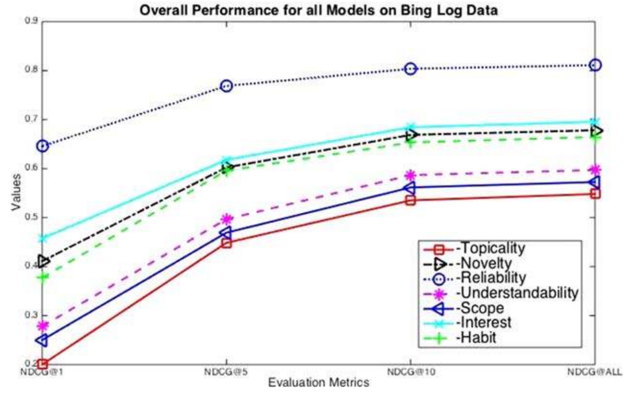

FIG. 2. Overall performance of ranking models considering different dimensions. [Color figure can be viewed at wileyonlinelibrary.com]

Overall performance for ranking models. Figure 2 shows the overall performance of different ranking models, which are trained with features for specific dimension. The training-target/evaluation metrics are nDCG@1, nDCG@5, nDCG@10 and nDCG@all, respectively. We can find that “Reliability,” “Interest,” “Novelty,” and “Habit” are four most effective ranking models, which shows that, on

DOI: 10.1002/asi

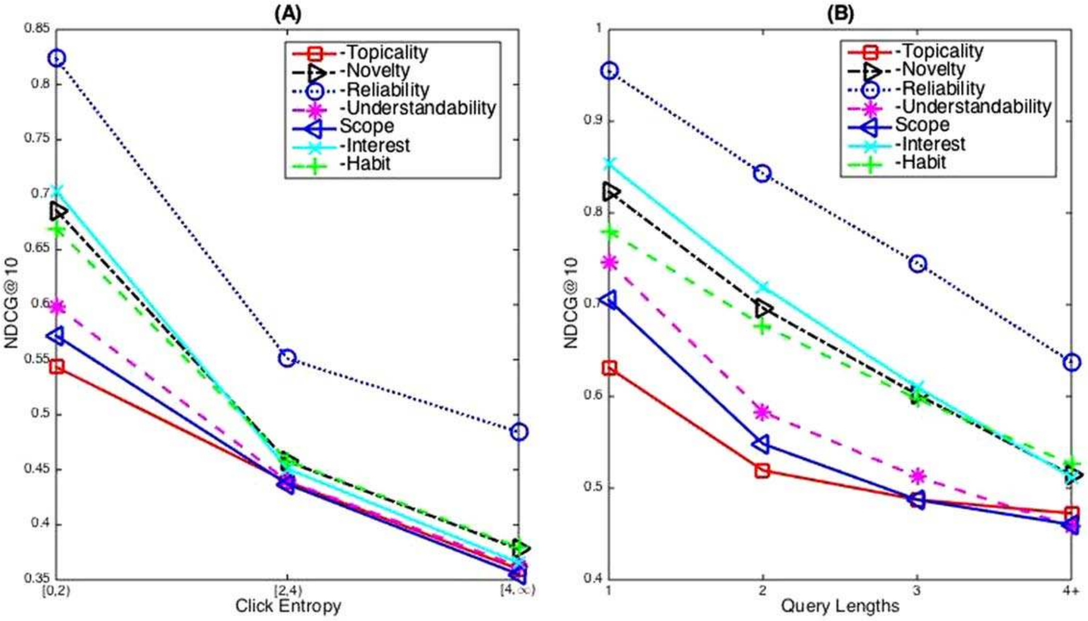

FIG. 3. Ranking performance of different models with respect to NDCG@10 for queries with different click entropy values and query lengths. (A) Is for queries with different click entropies. (B) Is for queries with different query lengths. [Color figure can be viewed at wileyonlinelibrary.com]

average, these four dimensions are mostly used for relevance judgment in the natural Web search setting.

Surprisingly, the ranking model trained with the "Topicality" dimension, which has long been seen an essential factor for relevance judgment, gains the lowest performance. Note that we do not regard the dimensions with lower ranking performance as unimportant, but consider these dimensions (e.g., Topicality) not sufficiently discriminative for retrieved documents. Traditionally, topicality is the basic factor for users' relevance judgment. However, in the process of judging relevance of the initial search results, the topical relevance scores of the top-ranked documents tend to be similar, given that modern search engines have already got a good ability to filter out the topically irrelevant results from the top ranked results, especially for easier queries. In the situation where the top-ranked documents have similar topical relevance scores, users may shift their focus to other relevance dimensions, such as reliability, interest, novelty, and habit. Similar interpretations apply to dimensions of "Understandability" and "Scope."

To gain a deeper understanding of the multidimensional relevance model, we conduct more experiments on different categories of queries in next subsection.

Performances over different categories of queries. In this subsection, we investigate how different dimensions contribute to the document ranking, in light of different types of queries. Figure 3A and B report the distributions of nDCG@10 results over different click entropy values and query lengths, respectively. Similar to the results reported in previous section, we can still find that the "Reliability," "Interest," "Novelty," and "Habit" are most effective ranking models. The enriched dimensions "Interest" and "Habit" are among the best performing, which shows that adding these two dimensions into the MURM is necessary. In addition, the ranking performance for each dimension decreases with the increase of click entropy values and query lengths. (The click entropy and query length can be regarded as important indicators of the query difficulty, because the original ranking performance drops with increasing click entropy values and query lengths, as shown in Table 8.) The ranking model corresponding to the reliability dimension significantly outperforms the other models on queries with less click entropy and shorter query length, demonstrating that the reliability of documents (with high authority) dominates users' relevance judgment criteria when they are searching with relatively easier queries, that is, shorter queries or the queries with less click entropy. For example, when a user issues a popular query "Facebook" (with click entropy in [0,1]), the user may mainly consider the reliability (or authority) of the retrieved webpages rather than the topicality or readability. However, the advantage of reliability over other dimensions becomes less obvious when users search with relative difficult queries, that is, the longer queries or queries with larger click entropy. For example, when a user is solving a problem on "Java" programming language (e.g., a query "new features of java," with click entropy in [4,∞]), he or she will synthesize a number of different relevance dimensions to judge whether a document is topically relevant, understandable, or novel. Therefore, it seems that users tend to consider less factors when searching with simpler queries, but will consider more factors (which interact with each other in a complex way) when searching with difficult queries.

Investigating the Correlation Between Dwell Time and Satisfaction

The work in this article is highly underpinned by the SAT-click assumption that dwelling on a specific webpage more than 30 seconds (SAT Click) would indicate users' perception of satisfaction. This is an assumption that has been widely used in the IR community, especially for query log analysis. Given it is infeasible to obtain the users' real perception of satisfaction by analyzing query logs, the dwell time, as an easy-to-access observable, was naturally utilized as an approximated reflection of satisfaction. However, one may argue that there is a lack of sufficient and strong evidences to support the assumption (Kim, Hassan, White, & Zitouni, 2014). Therefore, it is necessary to clarify whether or not this assumption is reasonable. To this end, in this subsection, we conduct a crowdsourcing user study to investigate how the dwell-time correlates to users' perception of satisfaction.

TABLE 9. Details of the crowdsourcing data.

Items	Number
Number of Users	19
Total Number of Search Tasks/Clicks	180/356
Number of Search Tasks/Clicks with Navigational Search Intent	50/93
Number of Search Tasks/Clicks with Informational Search Intent	92/187
Number of Search Tasks/Clicks with Transactional Search Intent	38/76
Pearson Coefficient between Dwell Time and Satisfaction for All Search Tasks	0.5847
Pearson Coefficient between Dwell Time and Satisfaction for Navigational Tasks	0.5412
Pearson Coefficient between Dwell Time and Satisfaction for Informational Tasks	0.5949
Pearson Coefficient between Dwell Time and Satisfaction for Transactional Tasks	0.6269

We recruited 19 participants (frequent users of search engines) in our study, which involves four PhD students, nine master's students and two graduate students in different subject areas, and four IT professionals. We asked them to record their natural search tasks in the period of July 4 to 6, 2016. In this way, the collected crowdsourcing data can naturally reflect the real search behaviors of users. Each search task contains seven fields, including the used search engine (e.g., Google, Bing, and Baidu, etc.), issued queries, search intent (i.e., navigational, informational, and transactional), screenshots of viewed snippets, clicked URLs, dwell time and corresponding satisfaction degrees (nonsatisfied:0, satisfied:1 and highly satisfied:2). For each recorded search task, the participant was paid a small amount of participation fee. The details of the collected crowdsourcing data are reported in Table 9. By observing the Pearson coefficient between the dwell time and satisfaction degree, we find that users' satisfaction correlates with the dwell time significantly. Figure 4 shows the distribution of average satisfaction degree over different intervals of dwell time for different search intents. We can find that there exists significant trend that users' satisfaction increases with the increase of dwell time. If the dwell time is very small (e.g., [0,15]), in general, there

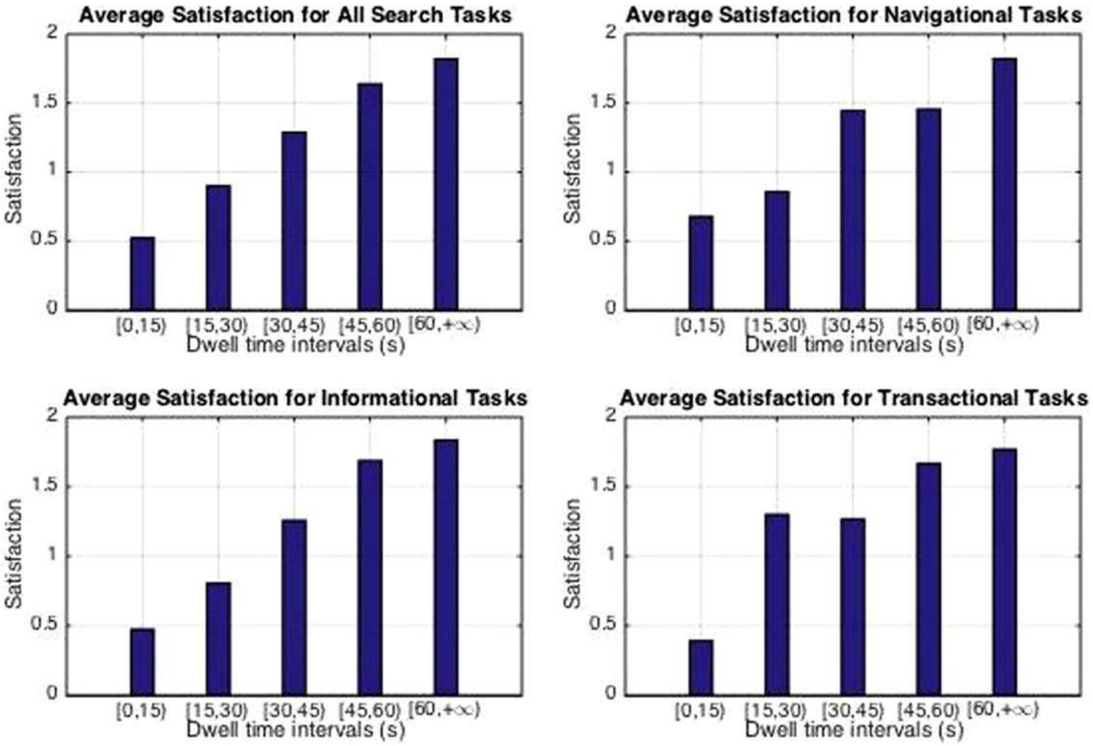

FIG. 4. Average satisfaction distribution over dwell-time intervals. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

is a low possibility that users are satisfied with the result, because they may find the result irrelevant and stop viewing it quickly. Note that, there also exist some special cases where users find the relevant information by only viewing the snippets or find the needed information soon after clicking the result. However, these cases account for a small proportion in all collected crowdsourcing data. Overall, our user study can verify that the use of the dwell-time based SAT-click criteria as an indicator of satisfaction is reasonable and the currently best possible solution to approximate user's ground truth in our query log analysis tasks.

A Further Study on TREC Session Track Task

In this article, we have proposed a novel idea to study the multidimensional relevance model for information retrieval with the query logs collected from a real search engine. This research methodology can be extended to a wider range of applications, for example, session search, recommendation system and e-commerce websites, etc. In this subsection, we will conduct a further study on session search, with the multidimensional relevance model to show the generalizability of the proposed method. Session search allows the search engine to retrieve documents with the short-term historical information within a search session. In the TREC session track data (http://trec.nist.gov/data/session.html), the human assessors have given relevance judgment, with a six-grade scale (-2,0,1,2,3,4), for the retrieved documents. Therefore, we can use the official relevance judgment information for evaluation. We conducted our experiments on the session tracks of TREC 2013 and 2014. There are 87 assessed session tasks in TREC 2013, and 100 in TREC 2014. Note that, we combined them and carried out 5-fold cross validation to gain an average evaluation results for all tasks. Clueweb12 Full corpus is used here, all words are stemmed with Porter's stemmer and stop words are removed.

According to the multidimensional relevance model formalized in the section "An Enriched MURM," we first extracted a series of features corresponding to different dimensions, and then conducted document re-ranking with the LambdaMART algorithm in order to investigate how different factors contribute to users' relevance judgment for session search. Note that, some features for reliability and habits based on clicks were not extracted, because the sessions are usually very short and sparse. The long-term "Interest" dimensional features were not extracted because there is no long-term historical information in the session data.

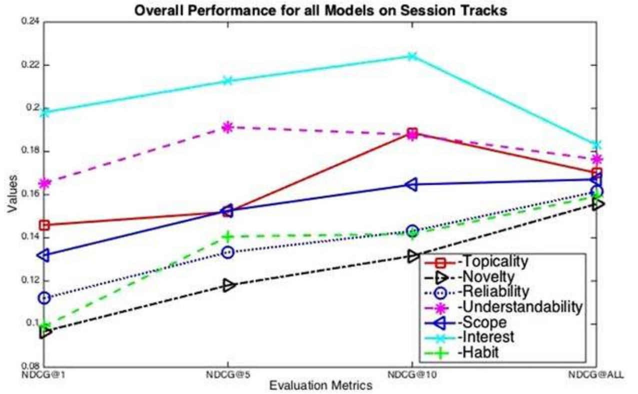

FIG. 5. Average performance for different models on session tracks. [Color figure can be viewed at wileyonlinelibrary.com]

In Figure 5, we report the average performance for all ranking models on session track data. From the results, we find that "Interest," "Understandability," "Topicality," and "Scope" are the most important factors that are considered in users' relevance judgment. Intuitively, for the session search task, we should filter out the nontopically relevant (or nonreadable) documents from a large set of candidates by considering the essential "Topicality" factor (or "Understandability"). Then, users will consider the "Interest" as the most important decision factor to determine the relevance of each returned document. Other dimensions seem less important, which shows that the official assessors do not consider them sufficiently.

The results on session search are somehow different from the findings revealed in Bing's query log. Specifically, in session tracks, "Topicality," "Understandability," and "Scope" are important factors for relevance judgment, whereas, in query log, they are less significant. This inconsistency may be resulted from the difference of data properties between two scenarios. Intuitively, the query logs are collected from the real-world search engine in a natural way, whereas the session tracks are designed and assessed manually. Through a detailed data analysis, we find that users assessed more relevant documents in session tracks than the relevant documents in Bing's query logs (estimated by SAT-Clicks) for each search task. This shows that the recruited assessors in session tracks are rather "tolerant" to all looking-like or topical relevant information, whereas the real search engine users may be more rigorous and only select the most "right" and useful information. The different findings also show that users may consider different relevance dimensions in different search scenarios. In addition, the "Interest" dimension is shown an important factor for both scenarios, which shows that enriching the multidimensional relevance model by adding the "Interest" dimension is beneficial for different search scenarios.

Conclusions and Future Work

In this article, we presented an enriched MURM consisting of seven dimensions that influence users' relevance judgment in different ways. For each dimension, we formalize a series of computable features for quantifying the dimension and allowing quantitative analysis of MURM. Extensive document ranking experiments have been conducted on a subset of Bing's query logs, which represent real Web search scenarios. The results reveal various meaningful phenomena, from which we obtain a series of important findings. Different dimensions demonstrate different degrees of contributions to users' relevance judgment for different types of queries. Specifically, "Reliability," "Interest," "Novelty," and "Habit" are the most significant dimensions that contribute to users' relevance judgment. On the other hand, "Topicality," traditionally regarded as the essential relevance decision factor, does not show a significant contribution to users' relevance judgment. The contributions of dimensions for different queries (with respect to click entropy, query length) are different. Specifically, the ranking performance for each dimension decreases with the increase of click entropy values and query lengths. (The click entropy and query length can be regarded as important indicators of the query difficulty.) For easier queries (e.g., navigational queries with shorter query length or smaller click entropy value), the "reliability" dimension dominates users' relevance judgment. For difficult queries (e.g., informational and transactional queries with longer query length or larger click entropy), the difference of ranking performance among different dimensions become smaller, although the "reliability" is still the most significant dimension. This phenomenon shows that users tend to consider more relevance factors to judge the relevance of documents when search tasks are complex and difficult. In addition, we also conduct an extensive user study, which verified the credibility of the widely used SAT-Click based evaluation strategy.

Furthermore, within the same experimental framework, we investigated how different factors contribute to users' relevance judgment for a different search scenario, that is, session search. The experimental results reveal that "Interest," "Understandability," "Topicality," and "Scope" are most important dimensions, which is different compared with the findings from Bing's query logs. This may be resulted from the difference of data properties, and also shows that users may consider different dimensions in different search scenarios. In addition, "Interest" is an important factor for both scenarios, which shows that enriching the multidimensional relevance model by adding the "Interest" dimension is necessary for different search scenarios.

Our experimental findings in this article can potentially bring beneficial inspirations to the design of the future IR algorithms. For example, we can extract and incorporate useful features that better correlate with relevance judgment process to improve the search quality and users' experiences. We can design different retrieval strategies focusing on different relevance dimensions for different types of queries and users. Moreover, we may develop more intelligent IR models, which dynamically adapt their relevance judgment strategies to different search tasks.

Furthermore, there exist different search scenarios, for example, Web search, intranet/enterprise search, medical search and entity search, in which users may exhibit different search behaviors and features. This article is focused on Web search (Bing search engine and TREC session track), and the findings may not necessarily hold for other search scenarios. However, the proposed methodology and framework for studying relevance is general. In the future, we will formalize novel relevance dimensions (and features) and study them in the context of more search scenarios

In addition, we can study the MURM from a cognitive perspective and adopt more cognitive signals (e.g., eye tracking, EEG and some other peripheral physiological signals) to further understand the process of users' multidimensional relevance judgment.

Acknowledgments

This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2014CB744604, 2013CB329304), the Chinese 863 Program (grant No. 2015AA015403), the Natural Science Foundation of China (grant No. U1636203, 61272265, 61402324), the Tianjin Research Program of Application Foundation and Advanced Technology (grant no. 15JCQNJC41700), and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie (grant agreement No. 721321).

References

Barry, C.L. (1998). Document representations and clues to document relevance. Journal of the American Society for Information Science, 49, 1293–1303.

Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the impact of short and long-term behavior on search personalization (pp. 185–194). In SIGIR. ACM.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks & Isdn Systems, 30, 107–117.

Burges, C.J. (2010). From ranknet to lambdarank to lambdamart: an overview. Learning, 11, 23–581.

Coleman, M., & Liau, T., L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60, 283–284.

Dale, E., & Chall, J.S. (1948). A formula for predicting readability: Instructions. Educational Research Bulletin, 27, 37–54.

Dou, Z., Song, R., & Wen, J.R. (2007). A large-scale evaluation and analysis of personalized search strategies (pp. 581–590). In *WWW*. ACM.

Hedman, A.S. (2007). Using the smog formula to revise a health-related document. American Journal of Health Education, 39, 61–64.

Janes, J.W., & Mckinney, R. (1992). Relevance judgments of actual users and secondary judges: A comparative study. Library Quarterly, 62, 150–168.

Jiang, D., Pei, J., & Li, H. (2013). Mining search and browse logs for web search. A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 4, 57.

Jiang, J., Ahmed, H.A., Shi, X., & White, R.W. (2015) Understanding and predicting graded search satisfaction. In *WSDM* (pp. 57–66). ACM.

Katzer, J., & Snyder, H. (1990). Toward a more realistic assessment of information retrieval performance. In Proc. ASIS (pp. 80–85).

Kim, Y., Hassan, A., White, R.W., & Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. In WSDM (pp. 193–202). ACM.

Kincaid, J.P., Fishburne, Jr, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (no. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Li, J., Song, D., Zhang, P., Wen, J.R., & Dou, Z. (2014). Personalizing web search results based on subspace projection. The 10th Asia

DOI: 10.1002/asi

Information Retrieval Societies Conference (AIRS 2014), LNCS 8870 pp. (160–171). Kuching, Malaysia.

Liu, T.Y., Xu, J., Qin, T., Xiong, W., & Li, H. (2007). Letor: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR* (pp. 3–10).

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, No. 1, p. 496). Cambridge: Cambridge University Press.

Regazzi, J.J. (1988). Performance measures for information retrieval systems-an experimental approach. Journal of the American Society for Information Science, 39, 235C251.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321–343.

Saracevic, T. (2016). The notion of relevance in information science: Everybody knows what relevance is. but, what is it really? Synthesis Lectures on Information Concepts Retrieval and Services, 8(3), i-109.

Thomas, D. (2013). Oxford guide to effective writing and speaking - how to communicate clearly. Oxford, UK: Oxford University Press.

Tiamiyu, M.A., & Ajiferuke, I.Y. (1988). A total relevance and document interaction effects model for the evaluation of information retrieval processes. IPM, 24, 391C404.

Tombros, A., Ruthven, I., & Jose, J.M. (2005). How users assess web pages for information seeking. Journal of the Association for Information Science and Technology, 56, 327–344.

Vu, T., Song, D., Willis, A., Tran, S.N., & Li, J. (2014). Improving search personalisation with dynamic group formation. The, International ACM SIGIR Conference (pp. 951–954). ACM.

Xu, Y., & Yin, H. (2008). Novelty and topicality in interactive information retrieval. Journal of the American Society for Information Science and Technology, 59, 201–215.

Xu, Y.C., & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? JASIST, 57, 961–973.

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. (2014). Relevance and effort: An analysis of document utility. In *CIKM* (pp. 91–100). ACM.

Zhai, C. (2008). Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, 1, 1–141.

Zhang, Y., Zhang, J., Lease, M., & Gwizdka, J. (2014). Multidimensional relevance modeling via psychometrics and crowdsourcing. In *SIGIR* (pp. 435–444). ACM.

Zhu, G., & Mishne, G. (2009). Mining rich session context to improve web search. In *SIGKDD* (pp. 1037–1046). ACM.

DOI: 10.1002/asi