Inf Retrieval J (2017) 20:25–52
DOI 10.1007/s10791-016-9289-z

Decoding multi-click search behavior based on marginal utility

Hai-Tao Yu[1] · Adam Jatowt[2] · Roi Blanco[3] · Hideo Joho[4] · Joemon M. Jose[5]

Received: 29 February 2016 / Accepted: 5 October 2016 / Published online: 22 October 2016

Abstract Query logs contain rich feedback information from users interacting with search engines. Therefore, various click models have been developed to interpret users' search behavior and to extract useful knowledge from query logs. However, most existing models are not designed to consider *novelty bias* in click behavior. The underlying hypothesis behind this paper is that given the previously clicked documents, a user tends to choose documents which provide novel relevant information to satisfy her information need, rather than redundant relevant information. Moreover, the prior click models have been mainly tested on frequently occurring queries, hence, leaving a large proportion of sparse queries uncovered. In this paper, we propose to predict users' click behavior from the perspective of *utility theory* (i.e., *utility* and *marginal utility*). In particular, *as a complement to the examination hypothesis*, we introduce a new hypothesis called *marginal utility hypothesis* to characterize the effect of novelty bias on users' click behavior by exploring the semantic

✉ Hai-Tao Yu
yuhaitao@slis.tsukuba.ac.jp

Adam Jatowt
adam@dl.kuis.kyoto-u.ac.jp

Roi Blanco
rblanco@udc.es

Hideo Joho
hideo@slis.tsukuba.ac.jp

Joemon M. Jose
joemon.jose@glasgow.ac.uk0

[1] Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan

[2] Department of Social Informatics Graduate School of Informatics, Kyoto University, Kyoto, Japan

[3] IRLab, Computer Science Department, University of A Coruña, A Coruña, Spain

[4] Research Center for Knowledge Communities Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan

[5] School of Computing Science, University of Glasgow, Glasgow, UK

divergence among documents in a result list. Moreover, to cope with sparse or unseen queries that have not been observed in the training set, we use a set of descriptive features to quantify the probability of a document being relevant and probability of a document providing marginally (novel) useful information. Finally, a series of experiments are conducted on a real-world data set to validate the effectiveness of the proposed methods. The experimental results verify the effectiveness of interpreting users' click behavior based on the marginal utility hypothesis, especially when query sessions contain sparse queries or unseen query-document pairs.

Keywords Click model · Query session · Novelty bias · Marginal utility

1 Introduction

Query logs constitute recorded user activities and interactions with search results pages. Many studies Silvestri (2010), Lucchese et al. (2013), Zhang and Nasraoui (2006), Baeza-Yates and Tiberi (2007) have demonstrated the high value of the feedback information contained in query logs. In this work, we focus on the class of models called *click models*, which simulate users' click behavior on search results. Click models have many applications. They can be used for predicting click behaviors, inferring documents' perceived relevance, or improving ranking functions. A successful click model allows to simulate real users' click behavior at a very low cost Radlinski et al. (2008). Moreover, we can predict the degree of a user's satisfaction by constructing a click model Guo et al. (2009), Liu et al. (2010), Wang et al. (2010) based on a large amount of real click data.

However, effective modeling of user search behavior and decoding of feedback information hidden in query logs are still an open research problem. One common issue is the so-called *position bias*, observed by eye-tracking experiments such as Granka et al. Granka et al. (2004). It refers to a phenomenon where users tend to click documents at higher positions of a ranked list, even if such documents provide less relevant information than the ones at lower positions.

The *examination hypothesis* Craswell et al. (2008), Richardson et al. (2007) has been typically applied to cope with the position bias. It assumes that a click depends on both the relevance and examination. If a surrogate of retrieved document is examined, then only its relevance determines the user click. The position bias is then reflected via the probability of user examining document at a given rank position, which merely depends on the rank. Follow-up studies have further extended the capabilities of the examination hypothesis. Specifically, cascade model Craswell et al. (2008) assumes that a user examines search results sequentially and stops examining once a relevant document is clicked. The probability of examination in the cascade model depends on the rank of the target document and on the relevance of all the documents ranked above. Another popular click model, dynamic Bayesian network model Chapelle and Zhang (2009), distinguishes between the perceived relevance (the relevance of the snippet presented by a search engine) and the actual relevance (the relevance of the landing page). The authors of dynamic Bayesian network model argue that a user can not examine the content of a document until she clicks on the corresponding url. Thus the decision to click on a document is determined based on its perceived relevance rather than based on its actual relevance.

Recent click models have built on more complex assumptions by incorporating additional knowledge or data about user behavior to better understand users' click behavior. For example, a number of studies Hu et al. (2011), Chuklin et al. (2013) tried to improve click models by exploring the *intent diversity*, which refers to the fact that users may submit the same query with different search intents. Take *Harry Potter* as an example of a common query. Users' search intention behind this query might be about a book, movie, or character. Hu et al. (2011) proposed the *intent hypothesis* to capture intent bias among different users to overcome the intent problem. Chuklin et al. (2013) studied how to combine the intent and layout information to improve a click model, where each query was assumed to have a number of categories or intents associated with it. Instead of exploring users' click behaviors based on individual queries, Zhang et al. (2011) proposed to characterize user behavior at a task level, which involved multiple queries and browsing actions that users performed to satisfy their search intents. Huang et al. (2012) explored how to use mouse movements and page scrolls to facilitate the detection of users' examination events. Their experimental results demonstrated that such additional data can improve the performance of click models in predicting user click behavior.

Another line of work to enhance click models is to use document features Wang et al. (2013), Zhu et al. (2010), Richardson et al. (2007), Zhang et al. (2014). A significant advantage of these feature-based models is that they can predict click behavior for unseen query-document pairs. Although feature-based models tend to involve additional computational costs to analyze documents' contents, such features can be extracted from indexing processes of search engines.

Despite the success achieved by the state-of-the-art models, they still have significant limitations. Before detailing these limitations, we first need to distinguish three key notions in click modeling: *relevance*, *diversity* and *novelty* by drawing lessons from the studies on search result diversification Clarke et al. (2008), Radlinski et al. (2009), Yu and Ren (2014). *Relevance* denotes how well a document meets the information need underlying a query. Note that, for simplicity, in many studies (e.g., Craswell et al. 2008; Dupret and Piwowarski 2008; Dupret and Liao 2010), the relevance between a query and a document is formulated as a constant parameter. *Diversity* indicates the necessity to resolve ambiguity. Since there is often a latent intent bias among different users issuing the same query, the corresponding click behaviors are expected to vary according to diverse intents Hu et al. (2011), Chuklin et al. (2013). Finally, *novelty* highlights the effect of redundancy on users' click behavior. Traditionally, search results are returned as a ranked list. When users sequentially examine the result list, the usefulness of an individual document is dependent on the previously ranked documents. The most extreme case is documents that are duplicates (this is actually a common phenomenon on the Web, e.g., a web page reprints the whole information or the main part of another web page).

Suppose two duplicate documents d_i and d_j are highly relevant to a query, and d_i is ranked higher than d_j. If a user has examined and clicked d_i, we assume that she would not click on d_j any more. In this paper, we call it *novelty bias*. It can be defined as follows (formalization of this bias will be given in Sect. 3).

> Given the previously clicked documents, a user tends to click on documents that provide marginally useful information to satisfy her information need, rather than on the ones that provide redundant information (even when this information is relevant).

However, most prior researches (e.g., models based on the examination hypothesis Chapelle and Zhang 2009; Dupret and Piwowarski 2008; Craswell et al. 2008 and ones based on the intent hypothesis Hu et al. 2011; Chuklin et al. 2013) suffer from the inability of correctly capturing the novelty bias. Specifically, Sect. 3.2 details the shortcomings of the examination hypothesis in modeling novelty bias. As a result, the models Chapelle and Zhang (2009), Dupret and Piwowarski (2008), Craswell et al. (2008) building on the examination hypothesis can degrade the performance of click modeling. The intent bias based model Hu et al. (2011) does not consider the effect of redundancy on a single user's click behavior within a query session (which refers to the search behaviors related to a single query, cf. Sect. 2), either.

Another limitation of existing work is that the previous click models have been tested and proven to work on frequently occurring queries. However, since query frequency is known to follow the power law distribution Baeza-Yates et al. (2007), Petersen et al. (2016), a large proportion of queries, namely, sparse queries, have not been considered in the evaluation.

The aforementioned limitations motivated us to develop a new model to predict users' click behavior. Our proposed method is designed to decode click behaviors from the perspective of utility theory. In particular, *utility* and *marginal utility* are introduced to denote whether a document is relevant or marginally relevant, respectively. We propose the *marginal utility hypothesis* to characterize the *novelty bias*.

Essentially, our proposed click model binds the probability of clicking on a document to not only the internal utility of this document (e.g., relevance) but also to its marginal utility. The latter represents the degree to which the document contains novel content with respect to the documents previously clicked by a user. We test our model on a real world query log containing multi-click query sessions and we prove its superiority over the state-of-the-art models in terms of simulating click behaviour and estimating relevance under the novelty bias.

The main contributions of this paper are as follows:

1. We propose the concept of novelty bias to motivate the need of considering semantic redundancy in click modeling, which has not been explored in depth by previous studies.
2. We propose the marginal utility hypothesis to characterize the effect of novelty bias on predicting users' click behavior. In particular, the probability that a document has a positive marginal utility value is quantified by exploring the semantic divergence among documents within a result list.
3. We demonstrate how to incorporate the marginal utility hypothesis into a click model both theoretically and empirically.
4. Based on a real-world data set consisting of both frequent queries and sparse queries with more than one click, a series of experiments are conducted to verify the benefit of the proposed method that considers novelty bias in click modeling.

In the next section, we survey the major click models developed for predicting users' search behavior. Section 3 details the proposed click model based on the concept of marginal utility. In Sect. 4, we describe the data set and the evaluation metrics as well as the baseline methods used in the experiments. Section 5 details the experimental results. In Sect. 6, some additional discussions are conducted to help understand the benefit of considering novelty bias as well as the advantages and limitations of this work. Finally, we conclude our work in Sect. 7.

2 Background

In this section, we give a brief survey of the typical models for interpreting users' search behavior. For a detailed review, please refer to the recent work by Chuklin et al. (2015). We begin by introducing some notations and background that will be used throughout this paper. Note that we focus here on the setting of organic search, thus, we omit the case of the ads-centric user behavior.

When a user submits a query to the search engine, she will get a ranked list of results, denoted by $L = \{d_1, \ldots, d_n\}$, where d represents a document, and $1, \ldots, n$ represent rank positions. The user then examines the result snippets, and clicks some or none of them. The following binary random variables are used respectively to represent the examination, relevance and click events when users interact with a search engine system,

- E_i: whether the user examines the i-th document,
- R_i: whether the i-th document is relevant,
- C_i: whether the user clicks on the i-th document.

Moreover, $P(E_i = 1)$ denotes the probability of examining the i-th document, $P(R_i = 1)$ denotes the probability of the document being relevant, and $P(C_i = 1)$ denotes the probability of clicking the i-th document. The parameter α_{q,d_i} is used to represent the relevance probability of the i-th document w.r.t. the query q, i.e.,

$$P(R_i = 1) = \alpha_{q,d_i} \tag{1}$$

Following the work by Chen et al. (2012), we use *query session* to refer to the search behaviors related to a single query, and we use *search session* to refer to the process that contains all the behaviors a user undertakes to perform a specific search task, which may span one or more query sessions. Given a query session or a search session in the query log, the events of examination and relevance are hidden, so only the click event is observable, since the clicked documents are recorded.

Before detailing the click models, an explanation of the typical assumptions (e.g., *examination hypothesis*, *cascade hypothesis* and *intent hypothesis*) is necessary, since they play a fundamental role within the state-of-the-art click models. The *examination hypothesis* Craswell et al. (2008), Richardson et al. (2007) refers to the assumption that a user clicks on a document *iff* this document has been examined and is relevant. It is formulated as

$$E_i = 1, R_i = 1 \Leftrightarrow C_i = 1 \tag{2}$$

Given a query q and a document d_i, Eq. 2 can be equivalently expressed as

$$P(C_i = 1 | E_i = 1, R_i = 1) = 1 \tag{3}$$

$$P(C_i = 1 | R_i = 0) = 0 \tag{4}$$

$$P(C_i = 1 | E_i = 0) = 0 \tag{5}$$

After summation over R_i and E_i, the probability of a click can be given as

$$\begin{aligned} &P(C_i = 1) \\ &\quad = \sum_{a\in\{0,1\}} \sum_{e\in\{0,1\}} P(C_i = 1|R_i = a, E_i = e)P(R_i = a)P(E_i = e) \\ &\quad = \underbrace{P(R_i = 1)}_{Component\text{-}1}\underbrace{P(E_i = 1)}_{Component\text{-}2} \end{aligned} \tag{6}$$

In particular, component-1 captures document relevance, which is independent of its position and of the above ranked documents. Component-2 reflects the position bias, i.e., the probability of examining a document merely depends on position. If a document is examined, then we have

$$P(C_i = 1|E_i = 1) = \alpha_{q,d_i} \tag{7}$$

The *cascade hypothesis* Craswell et al. (2008) assumes that: a user browses documents from top to bottom without skips, and the first document is always examined. All the following documents after the top-ranked one will not be examined if she stops examining. It is expressed as:

$$P(E_1 = 1) = 1 \tag{8}$$

$$P(E_{i+1} = 1|E_i = 0) = 0 \tag{9}$$

The *intent hypothesis* Hu et al. (2011) emphasizes that the submitted query partially reflects a user's search intent, thus, there is a gap between the query and a user's intent. It is formulated as

$$E_i = 1, N_i = 1 \Leftrightarrow C_i = 1 \tag{10}$$

$$P(N_i = 1|R_i = 0) = 0 \tag{11}$$

$$P(N_i = 1|R_i = 1) = \mu_s \tag{12}$$

where a latent event N_i is introduced to denote whether a relevant document is really needed by a user. μ_s represents the intent bias, namely, how well the query matches a user's intent.

There are also some other interesting assumptions for click modeling that are not explained here. The readers are encouraged to refer to Zhang et al. (2011), Chuklin et al. (2015) for detailed information.

Depending on whether click models utilize features (e.g., features extracted from a document), they can be classified into two categories: *feature-based click models* and *non-feature-based click model*s. In the following sections, we first introduce the typical non-feature-based click models, and then we show some representative feature-based click models.

2.1 Non-feature-based click models

For non-feature-based click models (e.g., Dupret and Piwowarski 2008; Chapelle and Zhang 2009; Dupret and Liao 2010), the relevance probability α_{q,d_i} is commonly initialized with a pre-defined value. This prior value is jointly adjusted by taking into account the observed events of examinations, skips, clicks, etc. We introduce the models separately based on the aforesaid examination hypothesis, cascade hypothesis and intent hypothesis. It should be noted that these hypotheses are not limited to non-feature-based click models as many feature-based click models are also build based on them.

2.1.1 Click models under examination hypothesis

The *cascade model* Craswell et al. (2008) builds on both the examination hypothesis and the cascade hypothesis. The click decision is made by taking into account previously examined documents,

$$P(C_i = 1) = P(R_i = 1)\prod_{j=1}^{i-1}(1 - P(R_j = 1)) \tag{13}$$

where $P(R_i = 1)$ depends on a document itself. This model suffers from the strong assumption that the user abandons the query session after the first click. Thus, it can not deal with query sessions characterized by multiple clicks.

To deal with query sessions that contain multiple clicks, the *dynamic Bayesian network* (DBN) model Chapelle and Zhang (2009) distinguishes perceived relevance and actual relevance. Moreover, the binary variable S_i is introduced to represent whether a user is satisfied at the i-th position or not,

$$P(E_{i+1} = 1 | E_i = 1, S_i = 1) = 0 \tag{14}$$

$$P(E_{i+1} = 1 | E_i = 1, S_i = 0) = \gamma \tag{15}$$

where γ is a pre-defined parameter. If the user is not satisfied by the current result, there is a probability γ that she examines the next document. The experimental results in Chapelle and Zhang (2009) showed that DBN outperforms other click models based on the cascade hypothesis. For inferring the model parameters, various methods Zhang et al. (2010), Zhu et al. (2010) have been proposed.

The *user browsing model* (UBM) follows the examination hypothesis, but it does not rely on the cascade hypothesis. The examination event E_i is determined by the current position i and by the distance from the current position to the rank of previous click $d = i - \max_{j<i}\{C_j = 1\}$ (i.e., Eq. 16). Particularly, UBM model is defined as

$$P(E_i = 1 | C_{1:i-1}) = \gamma_{id} \tag{16}$$

$$P(C_i = 1 | E_i = 0) = 0 \tag{17}$$

$$P(C_i = 1 | E_i = 1) = P(R_i = 1) = \alpha_{q,d_i} \tag{18}$$

where α_{q,d_i} is interpreted as the attractiveness of the document d_i w.r.t. the query q rather than as relevance. In fact, many follow-up click models Hu et al. (2011), Huang et al. (2012), Chuklin et al. (2013) directly extend the UBM model and the DBN model by adding new parameters or incorporating additional knowledge.

2.1.2 Click models under intent hypothesis

Under the intent hypothesis, Hu et al. (2011) investigated the effect of intent bias in decoding users' click behavior. By combining Eqs. 1, 10, 11 and 12, we can get

$$P(C_i = 1|E_i = 1) = \mu_s \cdot \alpha_{q,d_i} \tag{19}$$

Compared with Eq. 7 of the examination hypothesis, Eq. 19 of the intent hypothesis essentially uses the coefficient μ_s to discount the relevance.

Chuklin et al. (2013) explored the intent bias in a different way. Specifically, they use the intent distribution under a query, which is assumed to be known beforehand. Moreover, the presentation-type information of the result page is also incorporated.

2.1.3 Click models under non-sequential examination assumption

The eye-tracking study by Wang et al. (2015) showed that many query sessions (22.8 %) contain non-sequential (revisiting or duplicate) search behaviors. Furthermore, Wang et al. (2015) and Xu et al. (2012) investigated how to include non-sequential behaviors into click models and justified through the experimental results the benefit of incorporating non-sequential search behavior. The *session utility model* (SUM) Dupret and Liao (2010) maintains the click sequence in the chronological order. Under this model, each clicked document d_k provides some utility u_k (a positive real value) to the user. After clicking on a set of documents, the probability that a user stops a query session depends monotonely on the cumulative utility (i.e., the total utility that a user gathered from the clicked documents represented as a value between 0 and infinity). The underlying assumption is that users search and click on documents until their information needs are satisfied.

Although the effectiveness of the above-mentioned click models have been demonstrated through a series of experiments, a key shortcoming is that they can not handle unseen query-document pairs. Most of the models try to choose an appropriate prior to mitigate this problem, which needs enough evidence. Since query logs tend to be sparse and noisy, the performance of a non-feature-based click model is negatively impacted when dealing with sparse queries. We then next show how feature-based models can solve this problem and how they can model user search behavior.

2.2 Feature-based click models

As indicated in previous studies, the non-feature-based click models (e.g., Dupret and Piwowarski 2008; Chapelle and Zhang 2009; Dupret and Liao 2010) can demonstrate their effectiveness only when there are sufficient observations for given queries and documents. The capability of combining a large number of features is an advantage of feature-based click models, which enables them to deal with new and sparse queries and documents. Such a capability is highly needed, since it is almost impossible to precisely decode the complex search behaviors by only using a few factors.

Among the previous studies, a broad range of features have been explored. For example, some works have used the content information of queries and documents Richardson et al. (2007), the user-specific information (e.g., age and location) Wang et al. (2013), etc. A common way of linking model parameters and a set of features is to use the feature-based logistic model. For example, Richardson et al. (2007) proposed to predict the clickthrough rate (CTR) as: $P(C_i = 1) = \sigma(w^T f_{d_i})$, where $\sigma(x) = \frac{1}{1+exp(-x)}$, w is the weight vector and f_{d_i} denotes a feature vector. Their experimental results show that this model predicts CTR better than the random click model (cf. Sect. 3.1 in the book Chuklin et al. 2015).

Under the *bayesian sequential state* (BSS) model Wang et al. (2013), the dependencies among sequential search events (e.g., examination and click) are captured via a set of features. For instance, the conditional click probability is given as

$$P(C_i = 1|R_i, E_i, q) = \begin{cases} 0 & if \;\; E_i = 0 \\ \sigma({w^C_{R=0}}^T f^C_{d_i,q}) & if \;\; E_i = 1, R_i = 0 \\ \sigma({w^C_{R=1}}^T f^C_{d_i,q}) & if \;\; E_i = 1, R_i = 1 \end{cases} \qquad (20)$$

where $f^C_{d,q}$ represents the features corresponding to the click event, and $w^C_{R=0}$ as well as $w^C_{R=1}$ are weight vectors. The experimental results in Wang et al. (2013) show that the model achieves a poor perplexity value, but outperforms the DBN and UBM models in terms of relevance prediction.

To incorporate post-click behaviors (e.g., the dwelling time on a clicked document), Zhong et al. (2010) used a feature-based way to capture the satisfaction probability. In the *general click model* (GCM), Zhu et al. (2010) characterized the model parameters by using user-specific and URL-specific features. By deploying appropriate features and distributions over features, they showed that a number of click models (e.g., DBN) can be considered as special cases of this model.

2.3 Summary

In the above sections, we have shown how the main click models work to interpret users' click behavior. Different ways have been proposed to cope with the position bias, the intent bias, the non-sequential examination behavior, etc. Moreover, a number of feature-based click models have been designed to enhance the generalization capability. However, we have found that few works have attempted to investigate the effect of novelty bias on users' click behavior. Although the studies Zhang et al. (2011), Wang et al. (2013) have explored to certain extent the effect of redundancy, there are key differences between these works and ours. For example, Wang et al. (2013) have taken into account the case that a document may be skipped due to the content redundancy. The unigram and bigram segments of each document are used to compute the similarity of a pair of documents. Several features, such as the average value and variance, are then derived based on the similarity values to capture the redundancy effect. In contrast, we explore the effect of redundancy at a fine-grained granularity via a tensor-based approach. The dimensional semantic divergences are investigated in different ways to capture the novelty bias. Moreover, the rules defined for capturing the dependencies among users' click behaviors in this study are different from those in Wang et al. (2013). For example, the case of observing a click on an examined document that is not relevant is not allowed in our model, whereas it is possible in Wang et al. (2013). In another work, the *duplicate bias* Zhang et al. (2011) has been studied across multiple query sessions. The authors however have only focused on the case of duplicate documents. Hence, the effect of fine-grained granularity of redundancy has not been investigated. In this work, we investigate the effect of novelty bias on users' click behaviors by exploring the semantic divergence among documents co-occurring in a result list.

Moreover, we note that most of the prior click models (e.g., Hu et al. 2011; Wang et al. 2013; Dupret and Piwowarski 2008; Chapelle and Zhang 2009) have been tested and proven to work well on the frequent or "head" queries. This leaves a large fraction of sparse queries not investigated due to the fact that query frequency approximately follows the power law distribution Petersen et al. (2016). In this paper, the experiments are conducted on both sparse queries and on frequent queries.

3 Using marginal utility for click modeling

In this paper, we propose to interpret click behaviors from the perspective of marginal utility. In the following sections, we first give the definition of marginal utility in the context of information retrieval. Then we show how to decode users' search behavior based on marginal utility. Finally, we illustrate how to incorporate marginal utility into click models.

3.1 Utility and marginal utility

Analogous to the *utility theory* in economics Carl (2007), we use the concept of *utility* to refer to the satisfaction or contentment a user gets from a document in isolation from other documents. The factors for assessing a document's utility could include among others, topical relevance, authority of the site and the quality of display. On the other hand, we use *marginal utility* to refer to the satisfaction or contentment a user gets from a document by taking into account the previously browsed documents. Thus the utility of an individual document, say d_j, depends on the document itself, while its marginal utility is a context-dependent factor, which relies not only on the document itself but also the previous $j-1$ documents. Suppose two duplicate documents d_i and d_j are highly relevant to a query q, and d_i is ranked ahead of d_j. Although they have the same utility, the marginal utility of d_j is assumed to be zero, because d_j merely provides redundant information. Furthermore, we assume that: for a specific document d_i, its utility and marginal utility, say u_i and m_i, take values between 0 and infinity, i.e., $u_i \in [0, +\infty)$ and $m_i \in [0, +\infty)$. The utility of a document is always equal to or greater than its marginal utility, i.e., $u_i \geq m_i$. Namely, if a document contains redundant information given the previously browsed documents, its marginal utility would be smaller than its utility. If $u_i = 0$, then $m_i = 0$. On the contrary, if $m_i = 0$, u_i may not be zero. The aforesaid case of duplicate documents is an example.

In this paper, we do not explore how to determine the exact values of the utility and marginal utility of a document leaving it as future work. Instead, we focus on two probabilities. One is the probability that a document, say d_i, has a positive utility value. We assume that it is equivalent to the probability of being relevant, i.e., $P(R_i = 1)$. The feature-based expression of $P(R_i = 1)$ is given as

$$P(R_i = 1) = \sigma(w_u^T f_{q,d_i}^u) \tag{21}$$

where f_{q,d_i}^u represents the feature vector for capturing utility and w_u is the importance weight vector. To differentiate it from Eqs. 1, 21 is referred to *content-aware parameterization*, since the core composing features rely on the content information. The second probability is the probability that the document d_i provides marginally useful information, which is assumed to be equivalent to the probability of providing novel relevant information. It is expressed as

$$P(M_i = 1) = \sigma(\lambda_i w_u^T f_{q,d_i}^u + (1 - \lambda_i) w_m^T f_{q,d_i}^m) \tag{22}$$

where the binary variable M_i indicates the event whether d_i provides marginally useful information or not. The part $w_u^T f_{q,d_i}^u$ is used to capture the utility, and $w_m^T f_{q,d_i}^m$ is used to capture the semantic divergence between d_i and the previously clicked useful documents, where f^m_{q,d_i} is the feature vector and w_m is the weight vector. Finally, a linear combination of the two parts (λ_i is the coefficient) is used to capture the marginal utility. When $\lambda_i = 1$, Eq. 22 characterizes utility in the same way as Eq. 21. This will actually be the key point to understand the setting of Eq. 23 (explained in the next section). Section 3.3.1 discusses how to obtain the feature vectors f^u_{q,d_i} and f^m_{q,d_i}.

Using utility and marginal utility enables us to characterize the effect of redundancy on users' click behavior, which will be detailed in the following sections.

3.2 Marginal utility oriented click modeling

For interpreting users' click behavior, we propose the *marginal utility hypothesis*. Specifically, given a list of ranked results for a submitted query, we assume that a user sequentially examines this result list from the top to bottom. She would remember the previously examined and clicked documents when moving onto lower positions. A document must be examined before being clicked. If a document is not relevant or useful, the user will not click it. Given the previously clicked documents, the user will click a document if and only if this document has been examined and it provides marginally useful information.

Let $X = \{x_1, \ldots, x_m\}$ represent the sequence of observed clicks on the result list L, where x_j denotes the j-th click, and the value of x_j denotes the rank position of the document being clicked, thus $1 \le x_1 < x_2 < \ldots < x_m \le n$. The marginal utility hypothesis can be formally formulated as:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_{x_1} = 1 \tag{23}$$

$$\lambda_{x_1+1} = \lambda_{x_1+2} = \ldots = \lambda_n \tag{24}$$

$$P(C_i = 1 | E_i = 1, M_i = 1) = 1 \tag{25}$$

$$P(C_i = 1 | E_i = 0) = 0 \tag{26}$$

$$P(C_i = 1 | M_i = 0) = 0 \tag{27}$$

As explained in Sect. 3.1, Eq. 22 is characterizing utility in the same way as Eq. 21 when $\lambda_i = 1$. Based on this setting, Eq. 23 means that: when determining the usefulness of a document, marginal utility is equivalent to utility if there are no relevant documents clicked above. If some relevant documents were previously found, users consider the document's marginal utility rather than utility. In other words, within a query session, marginal utility is equivalent to utility if no relevant documents were found above. Equation 24 means that the same coefficient λ (set to 0.1 in experiments) is used for representing the marginal utility of any document below the first clicked document. To sum up, in the current setting λ is equal to 1 for every document ranked till the first clicked one (including the first clicked one), while it has the same fixed value (0.1 in the experiments) for each document ranked below. We leave the more complex setting with dynamically optimized λ values as future work.

Figure 1 illustrates the graphical models of the examination hypothesis and the marginal utility hypothesis, where the shaded nodes indicate observed variables, and the non-shaded nodes indicate hidden variables.

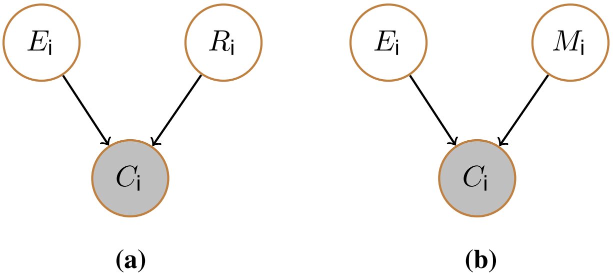

Fig. 1 The graphical models of the examination hypothesis and the marginal utility hypothesis

Table 1 Features for capturing utility and marginal utility

Features	Description
f^u_{q,d_i}	TF (term frequency), IDF (inverse document frequency), TF*IDF, BM25, Slash number in URL
f^m_{q,d_i}	Topic divergence, Text divergence

By viewing the variable R_i in the examination hypothesis as an indicator of utility (i.e., $u_i = 0$ or $u_i > 0$), it is straightforward to capture the distinction between the examination hypothesis and the marginal utility hypothesis. Namely, *from the rank position* 1 to x_1, *the marginal utility hypothesis is equivalent to the examination hypothesis*. When modeling users' search behavior on documents ranked below x_1 under the examination hypothesis, the usefulness of a document is considered independent of other documents. A non-clicked document, say $d_k(k > x_1)$, would be viewed as irrelevant or useless. On the other hand, *under the marginal utility hypothesis, the usefulness of a document is dependent on the previously clicked documents. A non-clicked document* $d_k(k > x_1)$ *can be relevant, yet, it may provide no marginally useful information*. Therefore, the marginal utility hypothesis can be essentially regarded as an improved version of the examination hypothesis.

3.3 Incorporating marginal utility hypothesis into click model

In this section, we theoretically show how to design a click model based on the marginal utility hypothesis. First, we explain the proposed ways to obtain the feature vectors for quantifying utility and marginal utility oriented probabilities. Then we detail how to incorporate the marginal utility hypothesis when implementing a click model.

3.3.1 Obtaining feature vectors f^u_{q,d_i} and f^m_{q,d_i}

The 2nd row in Table 1 shows the features composing f^u_{q,d_i}. For a particular document d_i, *TF*, *IDF*, $TF * IDF$ and *BM*25 are computed by using 3 data fields of d_i (i.e., the URL, title and body content) respectively, which generates 12 features. Following the studies on testing learning-to-rank algorithms (e.g., the benchmark collection LETOR[1]), the slash number in a URL is used as a feature indicating the depth or hierarchy of a webpage within a website. A total of 13 features are finally used.

[1] http://research.microsoft.com/en-us/um/beijing/projects/letor/.

For quantifying the semantic divergence between *a pair of documents*, we compute their *topic divergence* and *text divergence* separately between each field (i.e., the aforementioned 3 data fields). Thus, a total of 6 features is generated. Specifically, the topic divergence is computed based on the implicit subtopic distribution of the input objects. For example, given the body content information t_i and t_j of documents d_i and d_j, it is calculated as $\sqrt{\sum_k (P(z_k|t_i) - P(z_k|t_j))^2}$, where $P(z_k|t_j)$ is the subtopic probability computed using LDA model Blei et al. (2003). On the other hand, the text divergence is calculated as the cosine dissimilarity between the weighted term vectors, $1 - \frac{\mathbf{t}_i \cdot \mathbf{t}_j}{\|\mathbf{t}_i\| \|\mathbf{t}_j\|}$, where $\mathbf{t}_i$ and $\mathbf{t}_j$ represent the weighted term vectors of a single field w.r.t. d_i and d_j using $TF * IDF$ weighting scheme.

Assuming d_1, d_3, and d_5 are the clicked documents in the result list $\{d_1, \ldots, d_n\}$, we compute the semantic divergence between d_6 and all the three clicked documents when calculating the marginal utility of d_6 under the marginal utility hypothesis. Inspired by the tensor-based techniques for search result diversification Zhu et al. (2014) and relational learning Nickel (2013), we use a tensor-based method to capture the semantic divergence between the document d_i and the previously clicked documents in the result list. In particular, $T \in \Re^{\mathfrak{n} \times \mathfrak{n} \times \mathfrak{h}}$ is 3-way tensor that represents the pairwise divergence among n documents. The component T_i stands for the matrix of semantic divergence between d_i and other documents, T_{ij} denotes the divergence vector between d_i and d_j. $(T_{ij1}, \ldots, T_{ijh})$ represents the 6 divergence features (i.e., $h = 6$) computed based on d_i and d_j. Given the previously clicked documents, the feature vector f^m_{q,d_i} is extracted via a specific function, $f^m_{q,d_i} = D(T, X^{i-1})$, where X^{i-1} represents the sequence of observed clicks at the rank positions from 1 to $(i-1)$. Essentially, $D(T, X^{i-1})$ defines the semantic divergence between a single document and a set of documents. In this paper, $D(T, X^{i-1})$ is computed in the following three different ways:

$$D(T, X^{i-1}) = \left(\min_{x_j \in X^{i-1}} T_{ix_j1}, \ldots, \min_{x_j \in X^{i-1}} T_{ix_jh} \right) \tag{28}$$

$$D(T, X^{i-1}) = \left(\frac{1}{|X^{i-1}|} \sum_{x_j \in X^{i-1}} T_{ix_j1}, \ldots, \frac{1}{|X^{i-1}|} \sum_{x_j \in X^{i-1}} T_{ix_jh} \right) \tag{29}$$

$$D(T_i, X^{i-1}) = \left(\max_{x_j \in X^{i-1}} T_{ix_j1}, \ldots, \max_{x_j \in X^{i-1}} T_{ix_jh} \right) \tag{30}$$

Under Eq. 28, the target features rely on each dimensional minimum divergence between a document and the previously clicked documents. Under Eq. 29, they rely on the average divergence per dimension between a document and the previously clicked documents. Using Eq. 30, the features rely on the max divergence per dimension between a document and the previously clicked documents.

For a basic click model $\mathcal{M}$, the resulting model is denoted as MU-$\mathcal{M}$ when we extend $\mathcal{M}$ by incorporating the marginal utility hypothesis. Moreover, to investigate the effectiveness of the ways for characterizing marginal utility (i.e., Eqs. 28, 29 and 30), the suffixes *Min, Avg and Max* are used to distinguish the variants of MU-$\mathcal{M}$ that differ in obtaining the feature vector f^m_{q,d_i}, namely MU-$\mathcal{M}$-Min, MU-$\mathcal{M}$-Avg and MU-$\mathcal{M}$-Max.

3.3.2 Click model implementation

Given a click model and its parameters Θ, the goal is to find the optimal parameter setting Θ^* that optimizes the log-likelihood of the model given the training query sessions S:

$$\Theta^* = \arg\max_{\Theta} \sum_{s \in S} \log(\sum_{H} P(H, X^{(s)} | \Theta)) \tag{31}$$

where H is the vector of hidden variables, $X^{(s)}$ is the vector of observed clicks in a query session s. In fact, it is hard to optimize Eq. 31 due to the necessity of summing over all hidden variables. To cope with this problem, the Expectation Maximization algorithm is used to learn the optimal parameter setting as shown in Eq. 32,

$$\Theta^* = \arg\max_{\Theta} \sum_{s \in S} \sum_{H} P(H | C^{(s)}, \Theta) \log(P(H, X^{(s)} | \Theta)) \tag{32}$$

Specifically, in the E-step, the posterior distribution $P(H|C^{(s)}, \Theta^{(t)})$ is computed for hidden variables H under the current model $\Theta^{(t)}$. In the M-step, we derive the new parameter setting $\Theta^{(t+1)}$ by maximizing the expectation of the complete log-likelihood under $P(H|C^{(s)}, \Theta^{(t)})$.

For a click model $\mathcal{M}$ that follows the examination hypothesis, it is straightforward to switch from the examination hypothesis to the marginal utility hypothesis. The required step is to replace Eqs. 3, 4, 5 with Eqs. 23, 24, 25, 26, and 27. When the marginal utility hypothesis is adopted, the resulting click model is referred to as MU-$\mathcal{M}$. Let Θ be the parameters of the original model $\mathcal{M}$. Algorithm-1 illustrates how to infer the parameters of MU-$\mathcal{M}$, where two additional weight vectors w_u and w_m have to be estimated.

In Phase-C, given the estimated probabilities of being marginally relevant, we use the L-BFGS algorithm Liu and Nocedal (1989) to obtain the optimal weight vectors w_u and w_m.

Algorithm 1 A general framework for the marginal utility hypothesis based click models

```
Initialize $\Theta$ and weight vectors $w_u$ and $w_m$;
while the relative change of the right-hand side of Equation 32 is smaller than a threshold do
    Phase-A: calculate the posterior distribution $P(H|C^{(s)}, \Theta^{(t)})$ for the latent variables under the current setting $\Theta^{(t)}$, $w_u$ and $w_m$;
    Phase-B: calculate the new model parameters $\Theta^{(t+1)}$ by maximizing the expectation of $\log(P(H, X^{(s)}|\Theta))$ w.r.t. $P(H|C^{(s)}, \Theta^{(t)})$;
    Phase-C: update the weight vectors $w_u$ and $w_m$ based on $\mathbf{M}^{(t+1)}$;
end while
```

4 Experimental setup

In this section, we first describe the adopted data set. Its characteristics are specified from different dimensions (e..g, basic statistics, click distribution, query frequency, etc), which helps to better understand the experimental results. Then, we explain in detail the evaluation metrics and baseline methods.

4.1 Data set

To accelerate the research related to search logs, a number of publicly available data sets have been released, e.g., the AOL query log (2006)[2], the SogouQ data[3] and the query logs published as part of the Web Search Click Data (WSCD) workshop series in 2009[4], 2012[5], 2013[6], and 2014[7]. In regards to our study, a common limitation of these data sets is that they do not contain all documents (or URLs) displayed to a user. This makes it impossible to completely capture the semantic divergence among documents in a result list, since we have little information about the original content of documents. In view of this, the publicly available data sets are not adopted in this study.

We collected a 7-day (April 1–7, 2013) query log from a major web search engine, which contains 3, 802, 127 query sessions, 1, 473, 723 unique queries, 10, 847, 540 unique documents, and 1, 905, 104 clicks. The associated search activities include the anonymized user ID, requesting time, query string, the result list shown to a user, the corresponding clicks, etc. To train and evaluate the models, we have done the following pre-processing work on this query log: (1) Only query sessions having at least two clicks were considered. This step is motivated by prior studies Tyler and Teevan (2010), Teevan et al. (2007), Jiang et al. (2014), Lee et al. (2014) which demonstrated that a large percentage of queries are submitted with very simple search intents (e.g., finding or re-finding a specific homepage or locating a specific fact with known keywords). Search intents underlying such queries are usually satisfied through a single click. Among our adopted data set, there are 1,546,705 single-click query sessions (i.e., a ratio of 40.68 %). This high ratio is also confirmed to some extent in the previous studies Tyler and Teevan (2010), Teevan et al. (2007), Jiang et al. (2014), Lee et al. (2014). Since the effect of redundancy on click behaviors for this kind of queries is rather negligible, we decided to focus only on multi-click query sessions. The total number of multi-click query sessions is 129,043, which attains a ratio of 3.39 % of the entire data set. (2) Queries including characters outside a-zA-Z0-9 and the white spaces were filtered out. (3) Like Chapelle and Zhang (2009), we restrict ourselves to the top 10 search results of each query session. (4) Based on the URLs (both skipped and clicked) that were shown to search engine users, we downloaded the source files of these URLs for conducting content analysis such as extracting features to estimate document relevance probability. After these pre-processing steps, the final data set contains 10,778 unique query sessions, i.e., 8.35 % of the entire multi-click query sessions. It is summarized in Table 2.

Furthermore, Fig. 2 illustrates the number of clicks at each rank position. This figure again verifies the aforementioned *position bias*, i.e., users tend to click documents at higher positions.

Given the obtained data set, we perform 4-fold cross validation and report the average performance. The features are standardized using the Z-score normalization[8]. Particularly, in one round of cross-validation, the whole data set is split into the training set and the

[2] http://www.gregsadetsky.com/aol-data/.

[3] http://www.sogou.com/labs/dl/q-e.html.

[4] http://research.microsoft.com/en-us/um/people/nickcr/wscd09/.

[5] http://research.microsoft.com/en-us/um/people/nickcr/wscd2012/.

[6] http://research.microsoft.com/en-us/um/people/nickcr/wscd2013/datasets.htm

[7] http://research.microsoft.com/en-us/um/people/nickcr/wscd2014/.

[8] https://en.wikipedia.org/wiki/Standard_score.

Table 2 The summary of data set

#Unique query	#Unique clicked document /#Unique document	#Click/#Document	#Query session
9,995	27,400/105,756	28,891/118,203	10,778

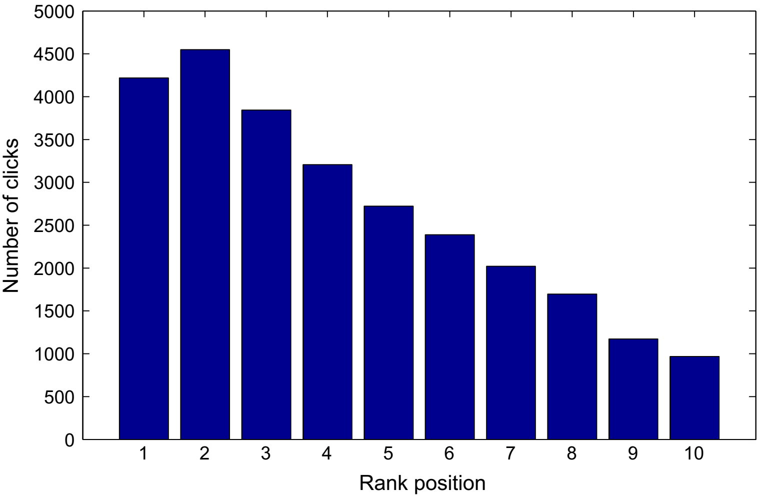

Fig. 2 The click distribution of data set

testing set at a ratio of 3:1. As for the testing set, it should be noted that *a subset of query sessions* are previously seen in the training set (i.e., each query-document pair within a query session has been observed in the training set). *The other complementary subset of query sessions* are unseen. In other words, each query session in such subset includes at least one query-document pair that has not been observed in the training set. When evaluating the performance based on the testing set, two ways of testing are conducted. One way is only using the subset of query sessions that have been observed in the training set (referred to as *Test-On-Seen*). The other way is using the entire testing set (referred to as *Test-On-Entire*). Using Test-On-Entire, we can only compare feature-based models, since the non-feature based models can not effectively deal with unseen query-document pairs. On the other hand, based on Test-On-Seen, we can perform a fair comparison among the non-feature based models and feature-based models. Table 3 shows the average number of testing query sessions when performing the 4-fold cross-validation.

Moreover, Table 4 illustrates the distribution of queries, documents and query sessions corresponding to query frequency.

Tables 3 and 4 clearly reflect the problem of query sparsity w.r.t. a query log. Specifically, Table 3 shows that a large portion of query sessions include unseen query-document pairs when we test a click model. From Table 4, we observe that a large number of queries are sparse queries. To investigate in depth the effectiveness and robustness of click models, these statistics suggest the necessity of conducting experiments on both frequent queries and sparse queries.

Table 3 The average number of testing query sessions when performing four-fold cross-validation

Testing way	Average number of testing query sessions
Test-On-Seen	58.75
Test-On-Entire	2,676.5

Table 4 The distribution of queries, documents and query sessions w.r.t. query frequency

#Query frequency	#Query	#Document	#Query session
1	9,503	103,949	9,503
2	391	8,799	782
3	61	2,005	183
≥ 4	40	3,450	310

4.2 Evaluation metrics

4.2.1 Log-likelihood

The metric of log-likelihood evaluates a click model's performance by looking at the likelihood of held-out test data. For each query session in the testing data $s \in S_{test}$, the likelihood of this session under a click model M is computed as $L(s) = P_M(C_1^s, ..., C_n^s)$, where $C_1^s, \ldots, C_n^s$ is the click events observed in s. Assuming the independence of query sessions, the log-likelihood of the test data is given as

$$LL(M) = \sum_{s \in S_{test}} \log P_M(C_1^s, \ldots, C_n^s) \tag{33}$$

Furthermore, let $|S_{test}|$ be the total number of query sessions, the average log-likelihood per query session is given as

$$Avg - LL(M) = \frac{LL(M)}{|S_{test}|} \tag{34}$$

The higher log-likelihood (or average log-likelihood) a model has, the better performance it achieves.

4.2.2 Perplexity

The perplexity metric used in the previous work Dupret and Piwowarski (2008) is defined as

$$p_k = 2^{-\frac{1}{N}\sum_{j=1}^{N}(\delta[C_k^j=1]\log_2^{q_k^j} + \delta[C_k^j=0]\log_2^{(1-q_k^j)})} \tag{35}$$

where C_k^j is a binary value that indicates an observed click on the k-th position of the j-th query session, and δ is the indicator function (i.e., [true]=1 and [false]=0). q_k^j is the predicted probability of a click on the k-th position of the j-th query session given the previously observed clicks. Perplexity measures how "surprised" the model is on observing the clicks in test data. The lower perplexity a model achieves, the better its performance is. In particular, perplexity of a naive model (predicting each click with a probability of 0.5) equals 2, and perplexity of a perfect model is 1. Across different rank positions, the average perplexity value is given as: $p_{avg} = \frac{1}{10}\sum_{k=1}^{10} p_k$. A standard way of comparing perplexity values is to compute the perplexity gain of a model A over a model B, i.e., $\frac{p_B - p_A}{p_B - 1}$. In the analysis, we use a mean average perplexity, which is a mean of average perplexity over query sessions, unless otherwise stated.

4.2.3 nDCG and MAP

In order to evaluate the quality in estimating relevance, the popular metrics for measuring ranking algorithms such as nDCG (Normalized Discounted Cumulative Gain) Burges et al. (2005) and MAP (Mean Average Precision) (see Manning et al. 2008 for the detailed definition) are used with different cutoff values 5 and 10. In view of the inherent sensitivity that a single metric has Radlinski and Craswell (2010), two metrics are adopted to ensure the reliability of the evaluation. Following Wang et al. (2013), the logged user clicks are regarded as binary relevance annotations. Particularly, if a documented is clicked, it is viewed as relevant, otherwise, it is non-relevant. If relevant documents are ranked at higher positions, the nDCG or MAP value will be large. Otherwise, the nDCG or MAP value will be small. The larger the nDCG or MAP value is, the more effective a click model is in estimating relevance.

4.3 Baseline methods

The UBM model and the DBN model are two widely compared click models that follow the examination hypothesis. The studies Hu et al. (2011), Chuklin et al. (2013) show that the UBM model outperforms the DBN model in terms of both perplexity and nDCG Järvelin and Kekäläinen (2002) (i.e., measuring the accuracy of relevance estimation). We thus decided to use the UBM model as the basis to study the effect of applying the marginal utility hypothesis. In the following, UBM refers to the original model proposed in Dupret and Piwowarski (2008), which is a non-feature based model.

There are two types of logistic regression models for click modeling, one is expressed as $P(C = 1|d_i, i) = \sigma(\alpha_{d_i} + \beta_i)$Craswell et al. (2008), i.e., a function of the document and the position. The other is built on features, e.g., Richardson et al. (2007) discussed in Sect. 2.2. In this paper, we use the feature based logistic regression model (denoted as *LR*) to show the effect of using features for capturing document relevance without considering other factors (e.g., rank position). Moreover, the content-aware model BSS introduced in Sect. 2.2 is also used as a baseline method (the code provided by the first author has been utilized). Due to the unavailability of the originally used data, we compare it based on our data set instead. In order to perform fair comparison, the features for estimating relevance are the same as in other methods compared in this paper. The features for estimating examination and click are generated in the same way as the original paper.

The proposed click models building on the marginal utility hypothesis are implemented by modifying the UBM model. The resulting click models are denoted as MU-UBM-Min, MU-UBM-Avg, and MU-UBM-Max, respectively.

5 Experimental results

In this section, we investigate the effectiveness of the marginal utility hypothesis oriented click models by comparing them with the examination hypothesis based UBM model and with the feature based models. In particular, we first use the testing set consisting of only seen query-document pairs (referred to *Test-On-Seen*), so as to perform a fair comparison between UBM and the proposed UBM variants that incorporates marginal utility. Then, we test the feature based models based on the entire testing set (*Test-On-Entire*) which contains both seen and unseen query-document pairs. Finally, we investigate the feature based models in terms of relevance estimation.

5.1 Results on seen query sessions

Table 5 shows the performance of four click models based on the Test-On-Seen set. UBM represents a non-feature based click model that relies on the examination hypothesis. MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max are three variants of UBM, which builds upon the marginal utility hypothesis.

From Table 5, we can observe that MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max outperform UBM in terms of Average log-likelihood. In terms of Avg-perplexity, MU-UBM-Min and MU-UBM-Max outperform UBM. This demonstrates the advantage of marginal utility based models over the baseline model on click modeling. Specifically, the advantage of MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max over UBM stems from the following key aspect: MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max use feature-based method to modeling the usefulness of a document. Moreover, the marginal utility hypothesis is adopted rather than the examination hypothesis, which helps to capture the effect of novelty bias.

We can also observe that the performance varies across the three implementations of marginal utility based models, although all outperform the baseline. MU-UBM-Min, which is based on the minimum divergence from previously clicked documents in the ranked list, performs best both in log-likelihood and average perplexity. This suggests that the way the marginal utility is calculated can affect the performance, and thus, a careful consideration is required for implementation.

To further investigate the behavior of the proposed models, we looked at the effect of query frequency on the performance. Here, the query frequency refers to the number of times that a testing query of a query session appears in the training data. The results are shown in Fig. 3.

Table 5 Performance comparison through the testing way: Test-On-Seen

Models	Average log-likelihood	Avg-perplexity
UBM	-10.66^*	3.69^*
MU-UBM-Min	$\boldsymbol{-6.69}$	$\boldsymbol{2.24}$
MU-UBM-Avg	-9.13^*	4.12^*
MU-UBM-Max	-8.17^*	3.06^*

For each metric, the best result is indicted in bold

* Indicates statistically significant difference to the best result under each metric (t test with $p<0.01$)

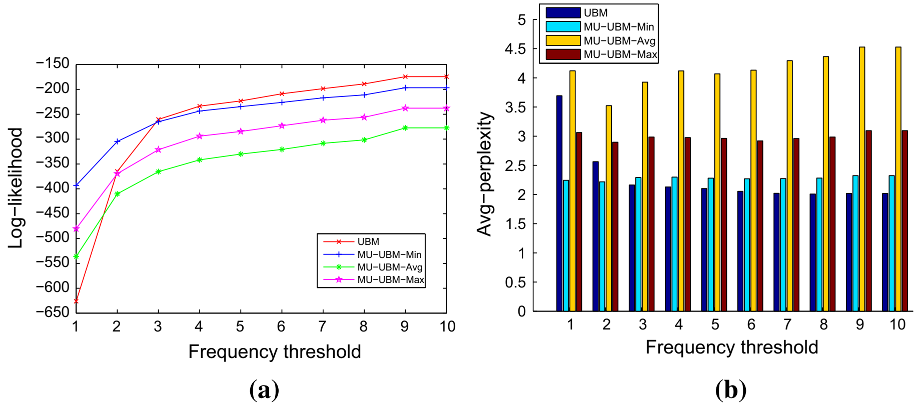

Fig. 3 Performance variation w.r.t. the frequency threshold of testing queries

We can observe that the effect of the query frequency is strong on UBM. In other words, the performance of UBM depends on the frequency of seen queries. In particular, UBM performs poorly on low frequency queries which constitute a large proportion of long tail distribution. On the other hand, MU-UBM-Min outperforms the baseline model at low frequency queries (i.e., Threshold = 1, 2). Average perplexity echoes a similar pattern, but the performance of marginal utility based models are more stable and robust against the change of query frequency than the baseline model.

In this section, we aim to show what would happen if we extend UBM by incorporating the marginal utility hypothesis. Since UBM is a non-feature based model, we perform more fair comparisons by comparing the proposed models to the state-of-the-art feature based models in the next section, where the *Test-On-Entire* set is used.

5.2 Results on entire query sessions

Next, we examine the performance of proposed models using the *Test-On-Entire* set, where a large number of unseen query-document pairs is included (cf., Table 3). Since the original UBM is not designed to handle unseen queries, we devised an extended UBM by using features to quantify the relevance probability. In particular, by replacing Eq. 18 with Eq. 21, we built the content-aware version of UBM (denoted as CA-UBM), which make it suitable to deal with unseen query-document pairs. Moreover, the feature based models LR and BSS described in Sect. 2.2 are also compared.

Table 6 shows the performance of the feature based click models, namely LR, BSS, CA-UBM, MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max based on the Test-On-Entire set.

From Table 6, we can observe that the proposed methods outperform LR, BSS and CA-UBM significantly. The perplexity gain over LR, BSS and CA-UBM for MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max are 36.0, 22.9 and 11.1 %, 31.0, 16.9 and 4.2 %, 34.0, 20.5 and 8.3 %, respectively. This provides a further evidence to support the advantage of marginal utility to capture novelty bias in click modeling. Since the position bias is prominent during users' search process, the lack of coping with position bias impacts the performance of LR to a large extent. For capturing the effect of redundancy, the content-oriented features of BSS depend on the pairwise similarity values of documents, such as

Table 6 Comparison based on the testing way: Test-On-Entire

Models	Average log-likelihood	Avg-perplexity
LR	−6.48*	2.00*
BSS	−5.75*	1.83*
CA-UBM	−5.09*	1.72*
MU-UBM-Min	***−4.64***	***1.64***
MU-UBM-Avg	−4.92*	1.69*
MU-UBM-Max	−4.78*	1.66*

For each metric, the best result is indicted in bold

*Indicates statistically significant difference to the best result under each metric (t test with $p < 0.01$)

the average value and variance (please refer to Table1 in Wang et al. (2013) for detailed information). The similarity of a pair of documents is calculated based on the unigram and bigram segments of each document. In contrast, we appeal to the dimensional semantic divergence by exploring the minimum, average and maximum combinations. The way of capturing redundancy is a reasonable reason for explaining the different results achieved by BSS and the proposed methods. Moreover, the original work Wang et al. (2013) also showed that UBM can achieve a better performance than BSS in terms of perplexity (Table 3 in Wang et al. 2013). Therefore, it is not surprising to observe that in Table 6 CA-UBM (i.e., content-ware version of UBM) outperforms BSS in terms of avg-perplexity.

Contrasting to CA-UBM, the exact reason for explaining why MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max perform better is the deployment of the marginal utility hypothesis. In addition, a comparison of the results in Tables 5 and 6 shows that the performance of MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max improved on the Test-On-Entire set.

Furthermore, we examine the effect of rank positions on the performance of proposed models. This is because users' clicks are known to vary across rank positions, as shown in Fig. 2. Therefore, we investigate the effectiveness of the feature based models in characterizing the click behaviors at specific rank positions. Figure 4 shows the average perplexity values of each model at different rank positions based on the testing way Test-On-Entire.

From Fig. 4, we can observe that the average perplexity value of LR is always around 2. Its performance can be approximately regarded as one of a naive model that predicts the click of each document with a probability 0.5. Conversely, it shows the necessity of considering position bias. Figure 4 again shows that CA-UBM, an extended version of UBM model, is more effective in modeling the click behaviors than BSS, even though BSS utilizes a more complex dependency framework to model click behaviors. Furthermore, both CA-UBM and BSS perform better at lower rank positions than higher positions. For MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max, Fig. 4 shows a noticeably better performance than the baseline models at the first rank position. For click behaviors at the 2nd and lower rank positions, these models show comparable performance. We will discuss the implications of these results in the Sect. 6.

5.3 Quality of relevance estimation

In this section, we investigate the performance of the aforementioned feature-based models in terms of relevance estimation, where the *Test-On-Entire* set is used. We rank candidate documents according to the estimated relevance given by each click model, and compare the ranked result against the *true* user clicks. The higher position a click model ranks a

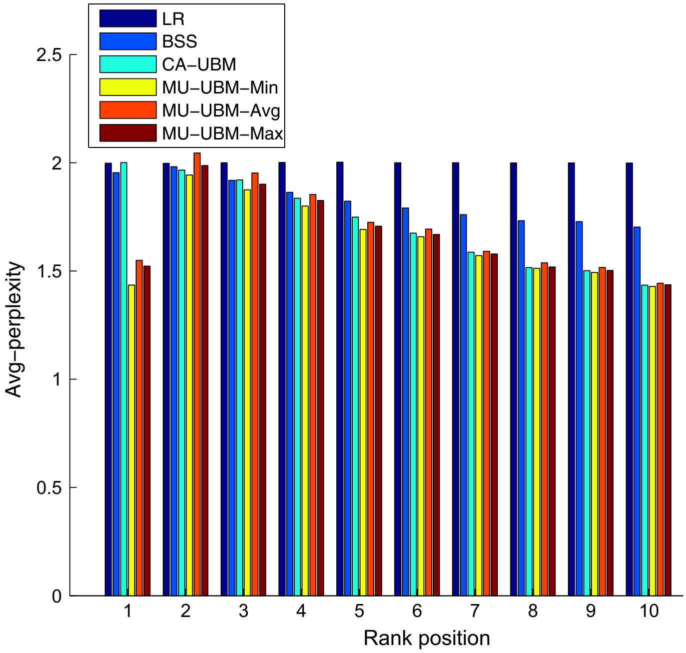

Fig. 4 Average perplexity values of each model at different rank positions

Table 7 The quality of estimating relevance based on the testing way: Test-On-Entire

Models	nDCG@5	nDCG@10	MAP@5	MAP@10
LR	0.455*	0.610*	0.178*	0.125*
BSS	0.462*	0.614*	0.181*	0.126*
CA-UBM	0.453*	0.608*	0.176*	0.125*
MU-UBM-Min	***0.665***	***0.769***	***0.298***	***0.179***
MU-UBM-Avg	0.650*	0.759*	0.289*	0.176*
MU-UBM-Max	0.657*	0.764*	0.293*	0.178*

For each metric, the best result is indicted in bold

*Indicates statistically significant difference to the best result under each metric (t test with $p < 0.01$)

clicked document, the better performance it achieves. Table 7 shows the results achieved by LR, BSS, CA-UBM, MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max, respectively.

By independently analyzing the results in Table 7, it is reasonable to say that the marginal utility oriented click models MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max significantly outperform the other baseline methods in estimating relevance in terms of both nDCG and MAP.

However, a joint analysis of the results in both Tables 6 and 7 helps to understand well the pros and cons of each click model. For example, since MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max builds upon UBM, the only difference between CA-UBM and a marginal utility oriented click model (MU-UBM-Min or MU-UBM-Avg or MU-UBM-

Max) is the modeling of relevance. The effectiveness in estimating relevance makes it easy to understand why MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max outperform CA-UBM in terms of average log-likelihood and avg-perplexity in Table 6. Another interesting observation is that: BSS outperforms CA-UBM in estimating relevance, but underperforms CA-UBM in terms of average log-likelihood and avg-perplexity. this is because computing average log-likelihood and avg-perplexity involves several factors jointly (e.g., the probability of examining, the probability of being relevant, or the distance to the last click, etc) rather than relevance itself. Therefore, a probable reason is that BSS fails to effectively modeling other factors, and, in result, it achieves a lower performance than CA-UBM in Table 6.

As an interesting exploration, Table 8 shows a subset of learned weights w.r.t. the features for estimating relevance, where CA-UBM, MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max are used as example models.

From Table 8, we observe that: for the same feature, different weights are learned by different models. Moreover, for the models build on marginal utility hypothesis, the weight vectors for any pair of models are not consistent. This implies different ways of estimating relevance. Another interesting observation is that the weight vector learned by CA-UBM is consistent with the weight vector learned by MU-UBM-Min (i.e., with the same positive and negative sign). Due to the fact that MU-UBM-Min jointly learns the weight vector for estimating relevance and the weight vector for estimating marginal utility, the relative values are different from CA-UBM. We leave the in-depth exploration of this consistence as well as the effectiveness of different feature combinations as a future work.

Furthermore, Table 9 shows a subset of learned weights by models building upon marginal utility hypothesis w.r.t. the semantic divergence features.

Table 8 Learned weights for estimating relevance

Model	TF*IDF			BM25		
	URL	Title	Body	URL	Title	Body
CA-UBM	−0.198	0.584	−0.050	0.417	−0.285	0.116
MU-UBM-Min	−0.791	0.180	−0.012	0.395	−0.339	0.068
MU-UBM-Avg	0.181	−0.055	−0.516	−0.552	0.014	0.384
MU-UBM-Max	−0.672	−0.730	0.058	0.390	0.389	0.414

TF*IDF and BM25 are the ways of computing feature scores. URL, Title and Body indicate the corresponding fields of a document for computing feature scores

Table 9 Weights w.r.t. the semantic divergence features

Model	Topic divergence			Text divergence		
	URL	Title	Body	URL	Title	Body
MU-UBM-Min	−0.313	1.338	−0.237	−0.348	0.411	1.381
MU-UBM-Avg	−0.278	−0.029	0.593	1.545	−0.029	0.162
MU-UBM-Max	0.855	−0.084	0.303	0.322	1.09	−0.601

Topic divergence and Text divergence denote the feature types. URL, Title and Body indicate the corresponding fields of a document for computing feature scores

From Table 9, we can observe that the weight vectors for any pair of models are not consistent. In other words, the way of balancing marginal utility features underlying each model differs from one another. For example, for the topic divergence based on the body field, feature values smaller than the mean value are preferred by MU-UBM-Min. On the other hand, for the text divergence based on the body field, feature values higher than the mean values are preferred.

6 Discussion

Our work was motivated by the intuition that users would prefer to click an URL which is likely to provide novel relevant information than redundant relevant information Tefko (1997) in multi-click sessions. A framework of utility and marginal utility hypothesis was employed to construct three new click models (MU-UBM-Min, MU-UBM-Avg and MU-UBM-Max) designed to capture novelty information from the contents of retrieved documents. We also extended UBM Model Dupret and Piwowarski (2008) to a feature-based model as part of our work. These models were evaluated by a total of 10,778 unique multi-click query sessions taken from a major Web search engine. This section discusses our major findings and implications on the development of effective click behavior models. We also discuss the limitations of our work.

6.1 Major findings and implications

The experimental results from the seen query-document set (Test-On-Seen) and mixed set (Test-On-Entire) demonstrated the advantage of the proposed models by outperforming baseline models Dupret and Piwowarski (2008) in terms of average log-likelihood, mean average-perplexity and relevance estimation (See Tables 5, 6, 7). This justifies our approach of considering novelty information in click modeling. It also demonstrates that the framework of marginal utility is an effective method to incorporate novelty properties into the modeling of click behavior.

Further analyses identified three aspects of strength in the proposed models. One was the robustness against the change of query frequency. Head queries are submitted very frequently, while the large number of tail queries are submitted much less frequently. Therefore, it is desirable for a model to accurately predict click behavior for those low frequency tail queries. Our analysis (Fig. 3) shows that the proposed model can perform well on those queries, while UBM needs enough evidences (i.e., the observed events of examinations, skips, clicks, etc.) to jointly adjust its prior parameter values. When the training data includes extremely low-frequency queries and documents, some parameter values, e.g., the relevance probability of a document w.r.t. a query, will be extremely small if no click is observed in the training data. Therefore, UBM shows poor performance on low-frequency queries.

The second strength was the robustness against the rank position. In particular, our result (Fig. 4) shows that the proposed marginal utility based models have a high prediction accuracy at the top ranked position. The difference between CA-UBM and MU-UBM can be explained as follows. The weight vector w_u for computing utility under the marginal utility based click models is different from the weight vector w_u of CA-UBM. The difference is that w_u of the marginal utility based click models was learnt jointly with the weight vector w_m for capturing marginal utility. Therefore, outperforming CA-UBM at the

1st rank position by a marginal utility based click model verifies the effectiveness of the jointly learnt importance weight vector w_u for computing utility. Meanwhile, outperforming CA-UBM at a rank position lower than 1 by a marginal utility based click model justifies the potential value of considering novelty bias via marginal utility.

The third strength was the effectiveness in estimating relevance. Different from the typical feature based model like BSS that relies on pairwise similarity values to detect redundancy, we explore the effect of redundancy at a fine-grained granularity via a tensor-based method. The dimensional semantic divergences are investigated in different ways to capture the novelty bias. Our analysis in Table 7 shows that the proposed models significantly outperform other baseline methods in terms of relevance estimation.

Finally, as for the comparison of three marginal utility based models, MU-UBM-Min outperformed MU-UBM-Avg and MU-UBM-Max. This indicates that relying on the dimensional minimum divergence seems to be the best choice when characterizing marginal utility. It also suggests that users are likely to click a novel but relatively similar relevant (i.e., low marginal utility value) document as subsequent search in multi-click sessions. This seems to echo the implications of Berry-Picking Model Bates (1989) where searchers are expected to learn little by little in the course of multiple queries and document reading. However, further studies are needed to clarify the relationship between the marginal utility measurements and user click behavior.

6.2 Limitations

The following practical issues have not been addressed well in this work. First, for preparing the data set, we downloaded the original documents (i.e., web pages) according to the displayed URLs in a result list in August and September, 2015. Some documents are however not available any more. The query sessions including unavailable documents have been then filtered out. In addition, we have used query sessions with more than one click, hence, the experiments were essentially conducted on a subset of the entire query sessions, which actually involves novelty bias. Second, for some documents on the Web, their content changes over time. Thus the utility and marginal utility would change correspondingly. This factor is not taken into account due to its inherent complexity. Third, for quantifying the probability (i.e., Eq. 21) that a document is relevant and the probability (i.e., Eq. 22) that a document provides marginally useful information, we appeal to the feature based method. Thus the quality of the adopted features is crucial to the final performance. To disentangle the underlying factors for representing both utility and marginal utility, it is worthy to investigate other alternative methods, e.g., using the technique of distributed representation Hinton et al. (1986), Paccanaro and Hinton (2001). Fourth, modern search engines sometimes provide federated search results. That is, some vertical results (such as image and video) from multiple specialized search engines tend to be incorporated Wang et al. (2013), Hong and Si (2013). The effect of the novelty bias on this kind of result pages is also not considered.

7 Conclusions and future work

This paper proposed a new approach to modeling click behavior based on marginal utility hypothesis. The proposed method was designed to capture novelty information among retrieved documents to predict users' click behavior. Experiments with over 10 K multi-click query sessions demonstrated that (1) the proposed method can handle unseen query-document pairs for click prediction; (2) the proposed method can outperform the state-of-the-art models; and (3) the performance of our model is robust against tail queries with low frequency of submission.

Overall, our work shows that it is possible and useful to capture novelty information of retrieved documents to model users' click behavior. Also, marginal utility was found to be an effective framework to implement novelty information for user click behavior modeling.

There are several future directions. One is to extend our work to predict click on advertisements in search results. Although we focused on the setting of organic search, in principle, there is no reason that the proposed method is not applicable to users' click behaviors on advertisements. Another is to deploy the learnt parameters for document ranking. Specifically, while in Sect. 5.3 we only evaluated the ranked documents from the perspective of relevance, the effect of novelty is not thoroughly explored. It would be interesting to use the estimated utility and marginal utility as signals for machine learning methods in search result diversification (e.g., Zhu et al. 2014; Xia et al. 2015). In addition, the studies Liu et al. (2015), Borisov et al. (2016) have demonstrated the potential values of using technique of deep-learning for click prediction. For example, Borisov et al. Borisov et al. (2016) explored the way of using distributed representations to represent the user's information need and search behaviors. Furthermore, the underlying dependencies among click behaviors and the information need are captured through the proposed models rather than a set of hand-crafted rules (e.g., Eq. 2). We also plan to investigate these approaches, as it is a good direction to further improve our model.

References

Baeza-Yates, R., Gionis, A., Junqueira, F., Murdock, V., Plachouras, V., & Silvestri, F. (2007). The impact of caching on search engines. In *Proceedings of the 30th SIGIR*, pp 183–190.

Baeza-Yates, R., & Tiberi, A. (2007). Extracting semantic relations from query logs. In *Proceedings of the 13th KDD*, pp. 76–85.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, *13*(5), 407–424.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borisov, A., Markov, I., de Rijke, M., & Serdyukov, P. (2016). A neural click model for web search. In *Proceedings of the 25th WWW*, pp. 531–541.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005) Learning to rank using gradient descent. In *Proceedings of the 22nd ICML*, pp. 89–96.

Carl, M. (2007). *Principles of Economics*. Auburn, Alabama: Ludwig von Mises Institute.

Chapelle, O., & Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th WWW*, pp. 1–10.

Chen, W., Wang, D., Zhang, Y., Chen, Z., Singla, A., & Yang, Q. (2012) A noise-aware click model for web search. In *Proceedings of the 5th WSDM*, pp. 313–322.

Chuklin, A., Markov, I., & de Rijke, M. (2015). *Click Models for Web Search*, (Vol. 7). Synthesis lectures on information concepts: Retrieval, and services.

Chuklin, A., Serdyukov, P., & de Rijke, M. (2013) Using intent information to model user behavior in diversified search. In *Proceedings of the 35th ECIR*, pp. 1–13.

Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008) Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st SIGIR*, pp. 659–666.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008) An experimental comparison of click position-bias models. In *Proceedings of the 1st WSDM*, pp. 87–94.

Dupret, G., & Liao, C. (2010). A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the 3rd WSDM*, pp. 181–190.

Dupret, G. E., & Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st SIGIR*, pp. 331–338.

Granka, L.A., Joachims, T., & Gay, G. (2004) Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th SIGIR*, pp. 478–479.

Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y., & Faloutsos, C. (2009). Click chain model in web search. In *Proceedings of the 18th WWW*, pp. 11–20.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1. chapter Distributed representations, pp. 77–109.

Hong, D., & Si, L. (2013). Search result diversification in resource selection for federated search. In *Proceedings of the 36th SIGIR*, pp. 613–622.

Hu, B., Zhang, Y., Chen, W., Wang, G., & Yang, Q. (2011). Characterizing search intent diversity into click models. In *Proceedings of the 20th WWW*, pp. 17–26.

Huang, J., White, R. W., Buscher, G., & Wang, K. (2012). Improving searcher models using mouse cursor activity. In *Proceedings of the 35th SIGIR*, pp. 195–204.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th SIGIR*, pp. 607–616.

Lee, C., Teevan, J., & de la Chica, S. (2014). Characterizing multi-click search behavior and the risks and opportunities of changing results during use. In *Proceedings of the 37th SIGIR*, pp. 515–524.

Liu, C., Guo, F., & Faloutsos, C. (2010). Bayesian browsing model: Exact inference of document relevance from petabyte-scale data. *ACM Transactions on Knowledge Discovery from Data*, *4*(4), 19:1–19:26.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Journal of Mathematical Programming*, *45*(3), 503–528.

Liu, Q., Yu, F., Wu, S., & Wang, L. (2015). A convolutional click prediction model. In *Proceedings of the 24th CIKM*, pp. 1743–1746.

Lucchese, C., Orlando, S., Perego, R., Silvestri, F., & Tolomei, G. (2013). Discovering tasks from search engine query logs. *ACM Transactions on Information Systems*, *31*(3), 14:1–14:43.

Manning, C . D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Nickel, M. (2013). *Tensor factorization for relational learning*. PhD thesis, Ludwig Maximilian University.

Paccanaro, A., & Hinton, G. E. (2001). Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, *13*(2), 232–244.

Petersen, C., Simonsen, J. G., & Lioma, C. (2016). Power law distributions in information retrieval. *ACM Transactions on Information Systems*, *34*(2), 8:1–8:37.

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, *43*, 46–52.

Radlinski, F., & Craswell, N. (2010) Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd SIGIR*, pp. 667–674.

Radlinski, F., Kurup, M., & Joachims, T. (2008) How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th CIKM*, pp. 43–52.

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th WWW*, pp. 521–530.

Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, *4*(1–2), 1–174.

Teevan, J., Adar, E., Jones, R., & Potts, M. A. S. (2007). Information re-retrieval: repeat queries in Yahoo's logs. In *Proceedings of the 30th SIGIR*, pp. 151–158.

Tefko, S. (1997). The stratified model of information retrieval interaction: extension and applications. In *Proceedings of the American Society for Information Science*, pp. 313–327.

Tyler, S. K., & Teevan, J. (2010). Large scale query log analysis of re-finding. In *Proceedings of the 3rd WSDM*, pp. 191–200.

Wang, C., Liu, Y., Wang, M., Zhou, K., Nie, J., & Ma, S. (2015). Incorporating non-sequential behavior into click models. In *Proceedings of the 38th SIGIR*, pp. 283–292.

Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. In *Proceedings of the 36th SIGIR*, pp. 503–512.

Wang, D., Chen, W., Wang, G., Zhang, Y., & Hu, B. (2010). Explore click models for search ranking. In *Proceedings of the 19th CIKM*, pp. 1417–1420.

Wang, H., Zhai, C., Dong, A., & Chang, Y. (2013). Content-aware click modeling. In *Proceedings of the 22nd WWW*, pp. 1365–1376.

Xia, L., Xu, J., Lan, Y., Guo, J., & Cheng, X. (2015). Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th SIGIR*.

Xu, D., Liu, Y., Zhang, M., Ma, S., & Ru, L. (2012). Incorporating revisiting behaviors into click models. In *Proceedings of the 5th WSDM*, pp. 303–312.

Yu, H., & Ren, F. (2014). Search result diversification via filling up multiple knapsacks. In *Proceedings of the 23rd CIKM*, pp. 609–618.

Zhang, Y., Chen, W., Wang, D., & Yang, Q. (2011). User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th KDD*, pp. 1388–1396.

Zhang, Y., Dai, H., Xu, C., Feng, J., Wang, T., Bian, J., Wang, B., & Liu, T. Y. (2014). Sequential click prediction for sponsored search with recurrent neural networks. In *Proceedings of the 28th AAAI*, pp. 1369–1375.

Zhang, Y., Wang, D., Wang, G., Chen, W., Zhang, Z., Hu, B., & Zhang, L. (2010) Learning click models via probit bayesian inference. In *Proceedings of the 19th CIKM*, pp. 439–448.

Zhang, Z., & Nasraoui, O. (2006). Mining search engine query logs for query recommendation. In *Proceedings of the 15th WWW*, pp. 1039–1040.

Zhong, F., Wang, D., Wang, G., Chen, W., Zhang, Y., Chen, Z., & Wang, H. (2010). Incorporating post-click behaviors into a click model. In *Proceedings of the 33rd SIGIR*, pp. 355–362.

Zhu, Y., Lan, Y., Guo, J., Cheng, X., & Niu, S. (2014). Learning for search result diversification. In *Proceedings of the 37th SIGIR*, pp. 293–302.

Zhu, Z.A., Chen, W., Minka, T., Zhu, C., & Chen, Z. (2010). A novel click model and its applications to online advertising. In *Proceedings of the 3rd WSDM*, pp. 321–330.