ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL – 2017, Vol. 06, No. 4, 103-109

Topic entities for determination pertinent documents of university campus network

E. Druzhinin, I. Shostak, N. Slavinskaya, A. Lysenko

The National Aerospace University Kharkiv Aviation,
Sumska St, 124, 61000 Kharkiv, Ukraine, e-mail: alexander.lysenko.dev@gmail.com

Received October 10.2017: accepted November 24.2017

Abstract. This paper reviews user-click models and the typed query profile model in order to investigate principles of faster access to documents which are in a university campus network. There was synthesized university campus user model that is based on the task-centric click model and determines document pertinence conditions, which relies on a user class. A solution of documents search automation was suggested that assumes that users will enable to avoid wasting of time in scientific materials search.

Key words: Task-Centric Click Model, User Behavior, Micropost Search, University Campus Network.

INTRODUCTION

The university campus network (UCN) [2] is one of the ways of spreading scientific results of students' and researchers' scientific activities. There is an issue of faster access to documents that satisfy information needs of an individual user. It depends on necessities of making a chain of requests to a data source and analyzing documents for each request. So a researcher spends a lot of time to search materials, but there might be support in such kind of activity and they will spend that equivalent of time to achieve new scientific results.

A solution of the described issue involves two tasks:

1. synthesize a user-click model which identifies relevant pairs «query-document» in case of plenty of returned documents;
2. determine an entity that matches pertinent documents to a user profile.

There is enough amount of user-click model implementations based on user click behavior during the search of pertinent documents:

- the dynamic Bayesian network [2];
- the user browsing model [5];
- the click chain model [7];
- the pure relevance model [13].

These models have the following disadvantage: they describe that a user satisfies their information needs only by a single query to a search engine. But more often there could be noticed a situation when a user has to work with the sequence of queries and find their pertinent data. Using synthesized click-model's data will enable users to reduce the number of queries sequence.

Providing the estimation of document pertinence based on click-model's data relies on the entity that matches click-model with documents' content. The most perspective way of search of such kind of entity is investigation of social networks which provide efficient manner of distributing information [14].

DATA BROWSING PRINCIPLE IN SEARCH ENGINES

Data browsing involves the set of pairs «query-document». This set helps to analyze user interaction behavior in a search engine and it often calls a user session [9, 10, 12]. The session refers to one of following categories:

- the query session (holds details about certain informative query);
- the search session (covers all queries and user's interaction history).

Click-models described above considers only the query session and ignores the large number of search session's data. In the end, these models lose accuracy of determination of pertinent documents.

Click models basis

The first issue of click modeling was the position bias [6]. It is stated that documents at high positions are likely to attract user's attention. As an improvement of this idea there was introduced a concept of document relevance [11] which determinates relevant documents by a composite factor. The investigation of document relevance generates hypotheses of user-click behavior [3] that are based on the probability theory. The notations are presented in Table 1.

User-click behavior in search engine

One of the behavior hypotheses is an examination hypothesis, which assumes that a user clicks a document only after a short preview of its description.

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

$$P(C_i = 1 | E_i = 1, q, \varphi(i)) = a_{\phi(i)}\,. \tag{2}$$

The extension of the examination hypothesis is the user browsing model that assumes document click relies on its position and the last clicked position in the same query session as illustrated in Fig. 1. The user browsing model notations are presented in Table 2.

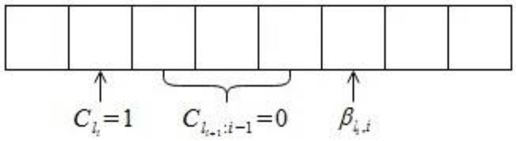

Fig. 1. The graphical view of $\beta_{l_i,i}$

$$l_i = \max\{\, j \in \{1,...,i-1\} \mid C_j = 1\}\,, \qquad (3)$$

$$P(E_i = 1 \mid C_{l_i} = 1, C_{l_{i+1}:i-1} = 0) = \beta_{l_i,i}\,, \qquad (4)$$

$$C_{i:j} = 0 \rightarrow C_i = C_{i+1} = ... = C_j = 0\,. \qquad (5)$$

The next extension is the cascade model that assumes that a user will examine documents from top to bottom without skipping any of them. Because of its main disadvantage it was reworked into the following models: the click chain model [7] and the dynamic Bayesian network model [2]. The cascade model notations are presented in Table 3.

They empathize that the examination probability depends on last clicked documents and their degrees of relevance.

The concept of pertinence was introduced in the dynamic Bayesian network model that prescribes that a user won't examine next document in case of their information needs were satisfied.

$$P(S_i = 1 \mid C_i = 1) = s_{\varphi(i)}\,, \qquad (6)$$

$$P(E_{i+1} = 1 \mid S_i = 1) = 0\,, \qquad (7)$$

$$P(E_{i+1} = 1 \mid E_i = 1, S_i = 0) = \gamma\,. \qquad (8)$$

Table 1 — Click-models notations

Symbol	*Purpose*
C_i	The document at the position i is clicked.
E_i	The description of the document at the position i is examined.
q	The retrieval documents query
$\varphi(i)$	The document at the position *i*.
$a_{\varphi(i)}$	The degree of relevance of the document at the position *i*.

Table 2 — User browsing model's notations

Symbol	*Purpose*
l_i	The last clicked position.
$\beta_{l_i,i}$	The transition probability from the position l_i to the position *i*.

Table 3 — Dynamic Bayesian network model’s notations

Symbol	*Purpose*
S_i	The degree of pertinence of the document at the position *i*.
γ	The continuation probability of documents browsing.

Considered click models differ by hypotheses and approaches of description of user-click behavior but none of them analyzes the search session.

More common user-click behavior assumptions were generated [15]:

1. a user tends to gradually express and enrich their information needs by documents examination;
2. a user tends to click on fresh documents that have not been seen before.

Since query generation and documents examination are likely to rich the large number of iterations, handling of search session data will reduce this number. To accurately determine pertinent documents and solve the click behavior manner the task-centric model was introduced [15].

Task-centric click model

This model is based on user-click behavior assumptions and relies on the search session. The task-centric click model has two layers [15]:

- the macro model that solves the first assumption;
- the micro model that solves the second assumption.

As illustrated in Fig. 2 there is a random variable M which determines whether a query matches user's information needs. The variable value reflects whether the user examines any document in this query. The examination of a document is described in the micro model and illustrated in Fig. 3. The probability that a document will be examined, relies on document relevance and the time passed from document indexing into data source.

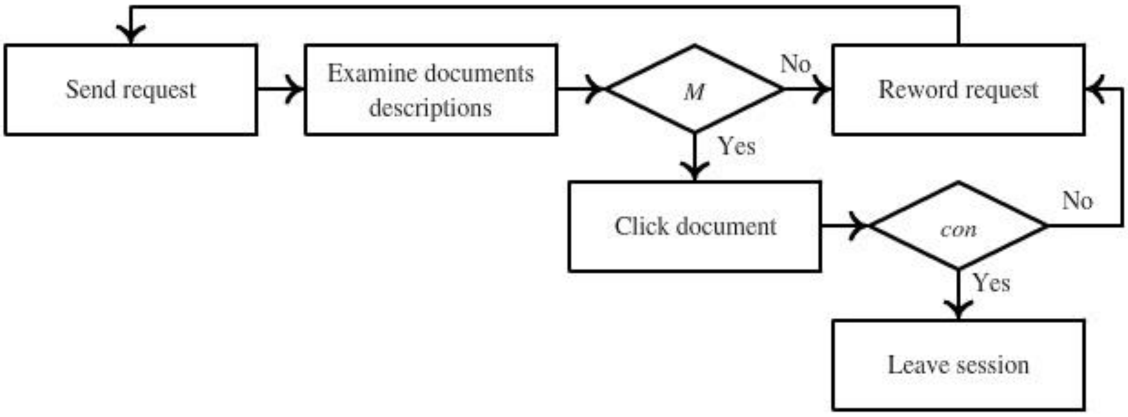

Fig. 2. The graphical view of the first assumption scenario: *con* – whether stop searching

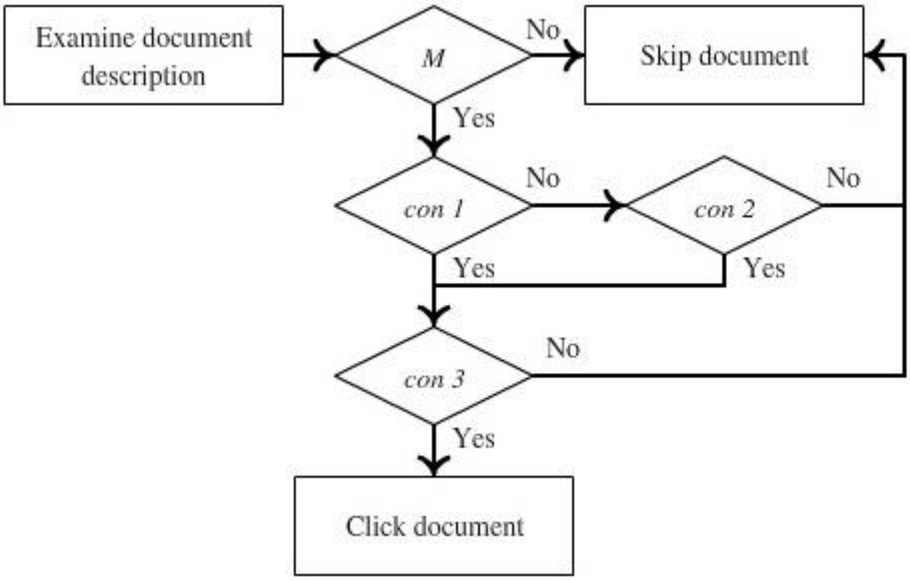

Fig. 3. The graphical view of the second assumption scenario: *con 1* – whether a document was examined before; *con 2* – whether examine a document description again; *con 3* – whether click a document

Layers of the task-centric click model might be formalized in the following manner [15]:

$$P(M_i = 1) = \alpha_1, \tag{9}$$

$$P(N_i = 1 | M_i = 1) = \alpha_2, \tag{10}$$

$$P(F_{i,j} = 1 | H_{i,j} = 1) = \alpha_3, \tag{11}$$

The task-centric click model notations are presented in Table 4.

$$P(E_{i,j} = 1) = \beta_j, \tag{12}$$

$$P(R_{i,j} = 1) = r_d. \tag{13}$$

The micro model is able to be extended by a set of existing or custom click models. For instance, if there is integration with the user browsing model, Eq. (12) will transform to:

$$P(E_{i,j} = 1 | C_{l_j} = 1, C_{l_{j+1}:j-1} = 0) = \beta_{l_j, j}. \tag{14}$$

There is the variable s_d of document pertinence into the task-centric click model, which was retrieved by integration with the dynamic Bayesian network model.

$$P(C_{i,j} = 1 | M_i = 1, E_{i,j} = 1, R_{i,j} = 1, F_{i,j} = 1) = c_d, \tag{15}$$

$$P(S_{i,j} = 1 | C_{i,j} = 1) = s_d. \tag{16}$$

Table 4 — Task-centric click model’s notations

Symbol	*Purpose*
M_i	Whether the *i*-th query matches user’s information needs.
N_i	Whether user examines remaining documents in the *i*-th query after her clicked and satisfied by the last document.
$E_{i,j}$	Examination of the document at the *j*-th position in the *i*-th query.
$H_{i,j}$	Previous examination of the document at the *j*-th position in the *i*-th query.
$F_{i,j}$	Freshness of the document at the *j*-th position in the *i*-th query.
$R_{i,j}$	Relevance of the document at the *j*-th position in the *i*-th query.
C_i	Whether click will be initiated on the document at the *j*-th position in the *i*-th query.
$S_{i,j}$	Pertinence of the document at the *j*-th position in the the *i*-th query.

UNIVERSITY CAMPUS USER MODEL

The determination accuracy of pertinent documents can be increased by the integration the task-centric click model and university campus user data, which relies on user's class. The classification of users might be provided by their activities: a student, a lecturer, an engineer. There was considered that the variable M defines probability of document examination. Thus it shows user data, which are presented in Table 5, are used.

The Fig. 2 was extended based on conditions from Table 5. It can be noticed that the variable M is still present because of user’s information needs uncertainty. The formalization of the Fig. 4 is described as follows. The notations are presented in Table 6.

$$P(M_i = 1 \mid \tau_d) = \alpha_1, \qquad (17)$$

$$P(N_i = 1 \mid \alpha_1) = \alpha_2, \qquad (18)$$

The variable l_{ind} contains the limit of time passed from the last document update event.

$$P(F_{i,j} = 1 \mid H_{i,j} = 1, l_{ind}) = \alpha_3, \qquad (19)$$

$$P(E_{i,j} = 1 \mid C_{l_j} = 1, C_{l_{j+1}:j-1} = 0, \alpha_1, \alpha_3) = \beta_{l_j, j}, \qquad (20)$$

$$P(C_{i,j} = 1 \mid \alpha_1, \alpha_3, \beta_{i,j}, r_d) = c_d, \qquad (21)$$

$$P(S_{i,j} = 1 \mid c_d) = s_d, \qquad (22)$$

Table 5 – Conditions of a document pertinence

	student	*lecturer*	*engineer*
M'	whether a document topic refers to a discipline that they are studying or have already studied.	whether a document topic refers to students’ course.	whether a document topic refers to a specialization.
M''	whether a document topic refers to a discipline that connects to disciplines from M'.	whether a document topic refers to scientific work.	whether a document topic refers to active tasks.
M	whether a query matches user’s information needs.		

Table 6 – University campus user model’s notations

	student	*engineer*
$M_{T_{i,j}}$	whether document’s topic at the *j*-th position in the *i*-th query matches M'	
D	a set of disciplines	-
sp	-	a specialization
T	-	a set of tasks
τ_d	$P(M_{T_{i,j}} = 1 \mid D)$ (23)	$P(M_{T_{i,j}} = 1 \mid sp, T)$ (24)

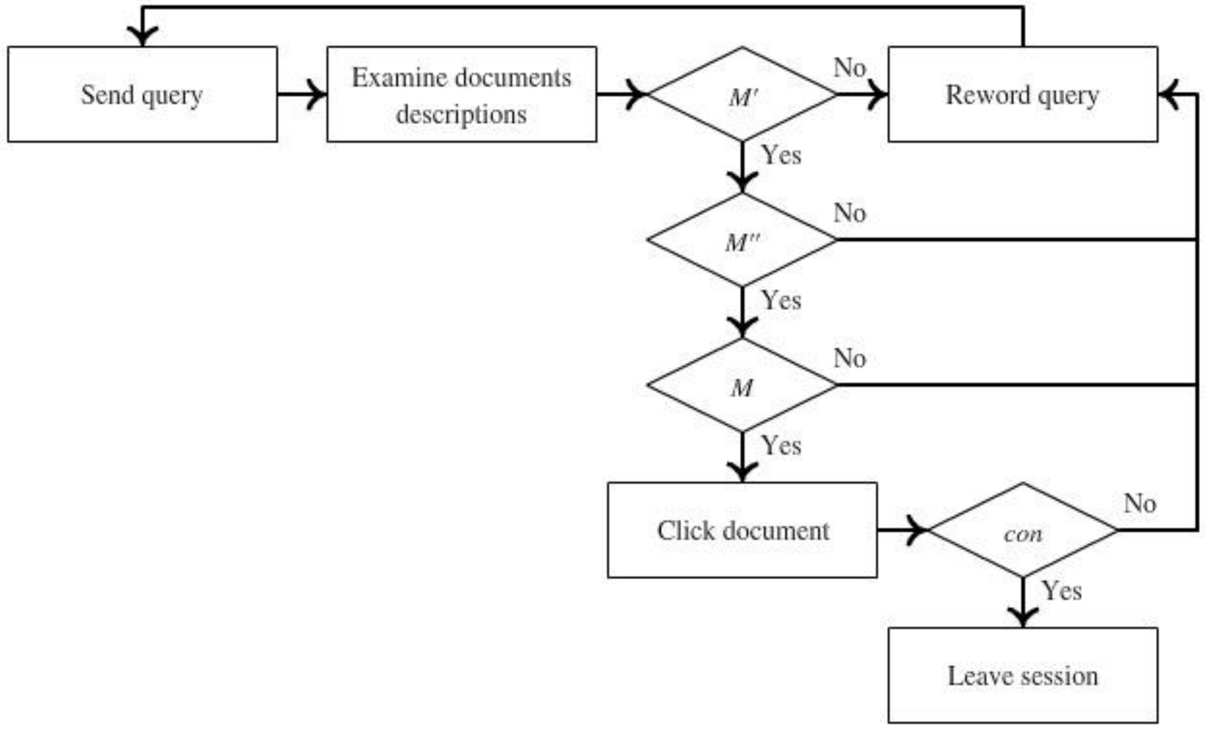

Fig. 4. The graphical view of the first enriched assumption scenario

DATA RETRIEVAL QUERY PROFILE

Social networks are one of the most useful data sources that allows to solve different tasks [14]:

- notifying about natural disasters;
- analyzing user's activity in the Web;
- reviewing trip routes;
- predicting results of political elections;
- supporting of decision making in the Web.

The main issue in case of using this type of data is determination of valuable piece of information that satisfies user's needs. Each network's resource (an image, a video or an audio, or a message) contains metadata that keeps information about resource's authors, counters and bound topics etc.

To make data determination easier and more efficient there was proposed the enrichment principle based on content semantics. In this way all data should contain key concepts (words, phrases or expressions which involve named entities, topics, reasons, damage-rates, locations, message type etc. The principle assumes enrichment of an original set of concepts by other topic-relevant concepts. The combined concepts sets are able to be used in data search and retrieve data with topics that were not specified in an original query but are relevant to search topics.

For instance, sending the following query "Ukrainian team in Canadian cup of synchronized swimming" user expects to retrieve information about the competition details and results of solo, duet and group programs. After enrichment of the original query, there will be able to examine translation records or a participants list, so it can be useful for the user. An example of the enrichment concept is presented in Table 7 and Fig. 5.

Table 7 – The enrichment original concepts

Topic: Ukrainian team in Canadian cup of synchronized swimming		
Concepts:	**- original:**	**- related:**
	Ukrainian team	Ukraine
	Canadian cup	Canada; the final cup of world series
	sync swimming	artistic swimming; sync; swimming

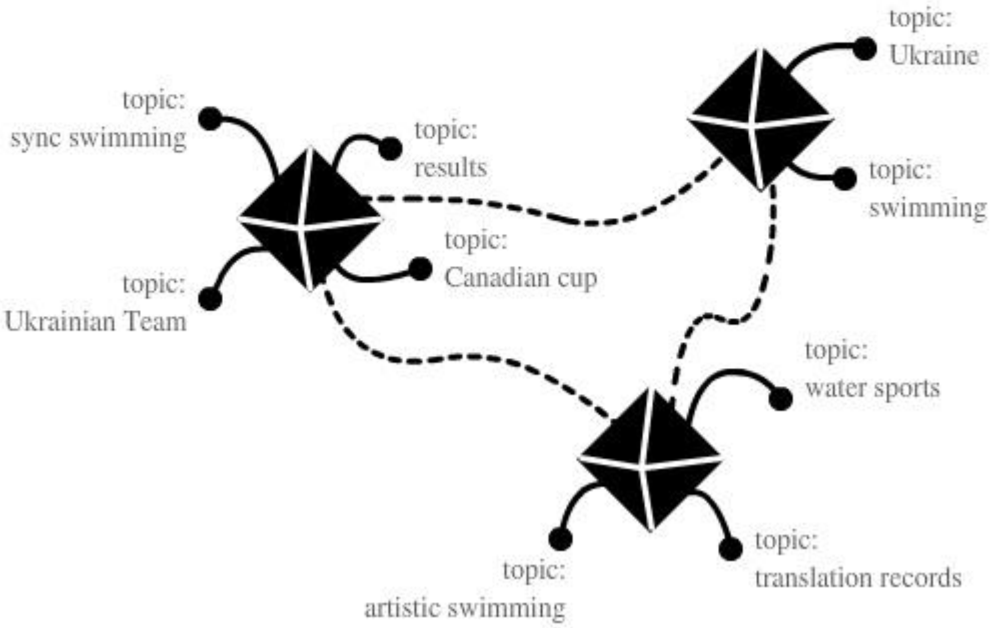

Fig. 5. The graphical view of enrichment principle by the example query “Ukrainian team in Canadian cup of synchronized swimming”

As an implementation of this principle the query profile model was introduced [14]. It encapsulates the determination of related concepts to the original set and assigns their weights. The concept’s weight affects in the estimation of the degree of data relevance.

The query profile is formalized in the following manner: the concepts set K of the original query Q is determined by a strategy s where the weight w is assigned to each concept.

$$QP_s(Q) = \{(k, w_s(k,Q)) | k \in K, K = s(Q)\} \quad (25)$$

There are following strategies of determination concepts [14]:

- by concepts semantics: synonyms, abbreviations, part of concepts (see Table 7);
- by the first found and relevant n-messages;
- by the specified time interval.

The determination strategy by the first n-messages M_n assumes that generation of the concepts set for each message m is based on its semantic. In the end, all sets combine into a result set S_{msg} :

$$QP_{S_{msg}}(M(n)) = \left\{ (k_{m_i}, w_{msg}(k_{m_i}, m_i)) | \, k_{m_i} \in \bigcup_{i=1}^{n} s_{msg}(m_i), m_i \in M(n), i = 1 \ldots n \right\}. \quad (26)$$

The query profile will enable to automate relevant data searching via strategies that described above.

CONCLUSION

1. Synthesized user-click model will reduce the number of queries sequence through data about user’s information needs that keep into a user class. Adaptation user-click model requires presence a mechanism of documents matching and estimation of their pertinence degree.

2. One of principles of data matching is the topic-based principle. It determines the original concepts set and enriches it through each concept semantic. The result set can be used in an automated query of documents retrieval. The principle relies on that user class data and documents content should contain concepts. They describe topics what they refer and are used in matching conditions of user's information needs satisfaction. In fact, it will make relevant data searching easier and efficient.

REFERENCES

1. **Burges C., Shaked T., Renshaw E., Lazier A., Deeds M., Hamilton N., Hullender G. 2005.** Learning to rank using gradient descent. Proceedings of the 22nd International Conference on Machine Learning. 89–96.
2. **Chapelle O., Zhang Y. 2009.** A dynamic bayesian network click model for web search ranking. Proceedings of the 18th International World Wide Web Conference. 1–10.
3. **Craswell N., Zoeter O., Taylor M., Ramsey B. 2008.** An experimental comparison of click position-bias models. Proceedings of the 1st ACM International Conference on Web Search and Data Mining. 87–94.
4. **Dupret G., Liao C. 2010.** A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. 181–190.
5. **Dupret G., Piwowarski B. 2008.** A user browsing model to predict search engine click data from past observations. Proceedings of the 31st Annual ACM SIGIR Conference. 331–338.
6. **Granka L. A., Joachims T., Gay G. 2004.** Eye-tracking analysis of user behavior in WWW search. Proceedings of the 27th Annual ACM SIGIR Conference. 478–479.
7. **Guo F., Liu C., Kannan A., Minka T., Taylor M., Wang Y., Faloutsos C. 2009.** Click chain model in web search. Proceedings of the 18th International World Wide Web Conference. 11–20.
8. **Hu B., Zhang Y., Chen W., Wang G., Yang Q. 2011.** Characterize search intent diversity into click models. Proceedings of the 20th International World Wide Web Conference. 17–26.
9. **Karpukhin A., Gritsiv D., Tkachenko A. 2014.** Mathematical simulation of infocommunication networks applying chaos theory. Econtechmod. An international quarterly journal. Vol. 3, No. 3, 33-42.
10. **Piwowarski B., Dupret G., Jones R. 2009.** Mining user web search activity with layered bayesian networks or how to capture a click in its context. Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. 162–171.
11. **Richardson M., Dominowska E., Ragno R. 2007.** Predicting clicks: estimating the click-through rate for new ads. Proceedings of the 16th International World Wide Web Conference. 521–530.
12. **Ryshkovets Yu., Zhezhnych P. 2013.** Information model of Web-gallery taking into account user’s interests. Econtechmod. An international quarterly journal. Vol. 2, No. 3, 59–63.
13. **Srikant R., Basu S., Wang N., Pregibon D. 2010.** User browsing models: relevance versus examination. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 223–232.
14. **Tao K. 2014.** Social Web Data Analytics: Relevance, Redundancy, Diversity. Delft University of Technology. Delft.
15. **Zhang Y., Chen W., Wang D., Yang Q. 2011.** User-click Modeling for Understanding and Predicting Search-behavior. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1388-1396.