УДК 004.89

Е.А. Дружинин, И.В. Шостак, А.А. Лысенко

Национальный аэрокосмический университет имени Н.Е. Жуковского «ХАИ», Харьков

ВЕРОЯТНОСТНО-ПОВЕДЕНЧЕСКАЯ МОДЕЛЬ ПОЛЬЗОВАТЕЛЯ УНИВЕРСИТЕТСКОЙ КАМПУСНОЙ СЕТИ

Рассмотрен базис вероятностных моделей выбора документов пользователями информационных систем. Приведены предположения о характере поведения пользователя информационной системы. Представлена интеграция модели задача-ориентированного выбора документов в рамках университетской кампусной сети. Рассмотрено применение вероятностно-поведенческой модели на примере класса пользователя – инженера конструкторского бюро.

Ключевые слова: *вероятностная модель выбора, модель задача-ориентированного выбора, университетские кампусные сети, анализ поведения пользователя, информационный поиск.*

Введение

История выбранных документов поисковой системы является источником информации о предпочтениях пользователя. Эта проанализированная информация используется при реализации и поддержке процессов, связанных с поиском данных, например: ранжирование результатов поиска [8], прогнозирование показателя выбора некоторого документа [6], прогнозирование степени удовлетворенности пользователя результатам поиска [9]. При анализе истории выбранных документов выделяют такую проблему, как построение модели – *модели выбора*, которая позволит определять релевантные пары запрос-документ, в случае когда на запрос была выдана огромная выборка результатов. Применение модели пользовательского выбора в поисковых системах в состоянии упростить прогнозирование поведения поиска документов пользователем в поисковой системе. Существует достаточное количество реализаций модели выбора, основанных на пользовательском просмотре и выборе документов информационного запроса:

- модель динамической байесовской сети [1];
- модель пользовательского просмотра [3];
- модель цепочки выбранных документов [4];
- модель релевантности [10].

Однако редко когда информационная потребность пользователя удовлетворяется единичным запросом к системе. Наиболее частым случаем является цепочка запросов и результирующая выборка документов на каждый из них, выполняющие по сути единственную задачу – поиска.

1. Анализ моделей выбора документов

1.1. Основной сценарий поведения пользователя при поиске информации

1. Произвести отправку информационного запроса поисковой системе.

2. Просмотреть результирующую выборку.

3. Выбрать некоторые документы для детального просмотра.

Результаты не удовлетворяют информационную потребность пользователя.

4. Построить уточненный информационный запрос.

5. Повторить действия, начиная с пункта 1.

Стоит отметить, что этот сценарий будет продолжаться до тех пор, пока пользователь не найдет желаемую информацию, либо же не завершит поиск. Другими словами, процесс поиска представляет собой множество пар запрос-выбор, которые составляют общую картину взаимодействия пользователя с поисковой системой.

Множество пар запрос-выбор для конкретных промежутков времени взаимодействия пользователя с системой представляют собой сессии — *сессии пользователя* [11]. Принято разделять сессии на две категории:

- сессии запроса (содержит информацию о конкретном информационном запросе);
- сессии поиска (охватывает все запросы и историю взаимодействия пользователя с результатами).

Принимая во внимание указанное обстоятельство, следует отметить, что реализованные модели выбора, приведенные выше, рассматривают только сессию единичного запроса и игнорируют значительный кластер информации сессии поиска. Таким образом, эти модели теряют точность в большинстве случаев, например: модель динамической байесовской сети предполагает, что пользователь всегда удовлетворен последним выбранным документом по некоторому запросу.

1.2. Основы вероятностных моделей выбора документов

Первичной проблемой моделирования выбора документов было предположение о позиции в результирующей выборке [12], утверждающее, что документ, находящийся на высоких позициях, явля-

ется наиболее привлекательным для выбора пользователями вне зависимости от того релевантный он, либо нет. После этого, было предложено понятие релевантности [6], которое позволяло документы, соответствующие запросу пользователя, перемещать на высокие позиции в результирующей выборке запроса. Рассматривая предложенное понятие, были предложены гипотезы поведения пользователей в поисковых системах [2], которые стало возможным формализовать при помощи теории вероятностей.

1.3. Гипотеза просмотра

Одной из гипотез поведения является *гипотеза просмотра документа*, которая предполагает, что пользователь осуществляет выбор документа только после просмотра его краткого описания:

$$P(C_i = 1 \mid E_i = 0) = 0\,; \qquad (1.1)$$

$$P(C_i = 1 \mid E_i = 1, q, \phi(i)) = a_{\phi(i)}\,, \qquad (1.2)$$

где обозначены вероятностные модели выбора: C_i – выбор документа, находящегося на позиции i результирующей выборки запроса; E_i – просмотр описания документа на позиции i; q – запрос, для которого была сформирована результирующая выборка; $\phi(i)$ – документ, находящийся на позиции i результирующей выборки; $a_{\phi(i)}$ – релевантность документа на позиции i.

Одним из расширений гипотезы просмотра является *модель пользовательского просмотра*, которая предполагает, что просмотр документа зависит не только от позиции в результирующей выборке, но и от предшествующего выбранного документа:

$$l_i = \max\left\{ j \in \{1,\ldots,i-1\} \mid C_j = 1 \right\}, \qquad (1.3)$$

$$P(E_i = 1 \mid C_{l_i} = 1, C_{l_{i+1}:i-1} = 0) = \beta_{l_i,i}\,; \qquad (1.4)$$

$$C_{i:j} = 0 \rightarrow C_i = C_{i+1} = \cdots = C_j = 0\,, \qquad (1.5)$$

где обозначены такие модели пользовательского просмотра: l_i – позиция предыдущего выбранного документа; $\beta_{l_i,i}$ – вероятность перехода с позиции l_i на позицию i (рис. 1.1).

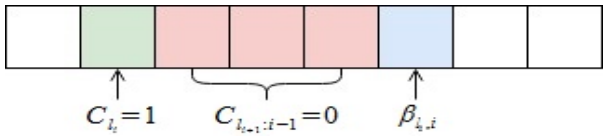

Рис. 1.1. Графическое представление $\beta_{l_i,i}$

Следующее расширение – *каскадная модель*, предполагающая, что пользователь постепенно просматривает каждый документ и выбор одного из них завершает процесс поиска информации. Указанное обстоятельство является ограничением данной модели, что способствовало её переработке, выразившейся в двух моделях - *модель цепочки выбранных документов* [4] и *модель динамической байесовской сети* [1].

Обе модели подчеркивают, что вероятность просмотра документа зависит от предшествующих выбранных документов и их релевантности. В модели динамической байесовской сети было введено понятие пертинентности, предписывающее, что пользователь не будет просматривать следующий документ, если его информационные нужды были удовлетворены текущим:

$$P(S_i = 1 \mid C_i = 1) = s_{\phi(i)}\,, \qquad (1.6)$$

$$P(E_{i+1} = 1 \mid S_i = 1) = 0\,, \qquad (1.7)$$

$$P(E_{i+1} = 1 \mid E_i = 1, S_i = 0) = \gamma\,, \qquad (1.8)$$

где обозначены такие модели динамической байесовской сети: S_i – пертинентность документа, находящимся на позиции I; γ – вероятность продолжения поиска информации.

Рассмотренные модели выбора различаются по предположениям и подходам относительно поведения пользователей, но ни в одной не анализируется сессия поиска.

1.4. Предположения о поведении пользователей поисковой системы

Принимая во внимание эволюцию как поисковых систем, так и моделей выбора (в частности, модель задача-ориентированного выбора), выделяют два предположения о поведении пользователя [7]:

1) пользователь стремится постепенно отражать свои информационные потребности и уточнять их путем просмотра результатов запросов;

2) пользователь стремится выбирать актуальные документы, которые не были включены в результирующую выборку предшествующих запросов.

Вышеуказанные предположения основываются на соответствующих им сценариях [7].

Сценарий поведения пользователя 1 предположения

1. Отправить информационный запрос поисковой системе.

В большинстве случаев результаты по первому запросу не удовлетворяют нужды пользователя.

2. Просмотреть результирующую выборку.

Анализируя результирующую выборку путем просмотра краткого описания документов без выбора одного из них, пользователь подчеркивает для себя как ему перефразировать, либо уточнить, запрос.

3. Построить уточненный информзапрос.

4. Повторить действия, начиная с пункта 1.

Сценарий поведения пользователя 2 предположения

1. Отправить информационный запрос поисковой системе.

2. Просмотреть результирующую выборку.

Выборка может содержать как документы, которые пользователь прежде не видел, так и те, для которых он прежде определял полезность.

3. Поиск документов, которые прежде выдавались на запрос.

Маловероятно то, что пользователь несколько раз будет просматривать документы, которые он видел в предшествующих запросах.

4. Просмотр фильтрованной выборки.

5. Выбор документов.

Пользователь идентифицирует полезность документа путем его выбора из результирующей выборки.

При рассмотрении сценариев поведения пользователя отражается тот факт, что формирование запроса и выбор документов в информационной системе имеет тенденцию к достижению нескольких итераций. Обработка и использование данных сессии поиска способствуют минимизации количества таких итераций. Таким образом, с целью повышения точности определения пертинентных документов поисковой системы и в то же время решения задачи прогнозирования поведения поиска была предложена модель задача-ориентированного выбора [7].

1.5. Модель задача-ориентированного выбора документов

Данная модель основана на двух предположениях поведения пользователя и ориентирована на работу непосредственно с сессией поиска. Структура модели состоит из двух слоев [7]: макро- и микромодели. Макромодель характеризует первое предположение о поведении, тогда как микро- – второе.

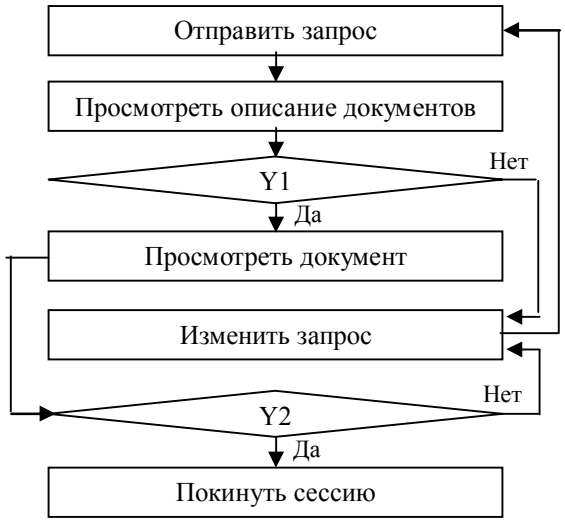

Рис. 1.2. Графическое представление сценария поведения пользователя 1-го предположения (Y1 – соответствует ли запрос нуждам пользователя; Y2 – завершить поиск)

В момент отправки запроса пользователем модель использует случайную величину для определения логического вывода условия У1 (соответствует ли запрос нуждам пользователей) на рис. 1.2 (У2 – завершить поиск). Стоит отметить, что значение величины У1 влияет на интерпретацию моделью поведения пользователя для текущего запроса. Значение этой величины зависит от того, просмотрит ли пользователь хотя бы один из документов результирующей выборки, либо нет. Процесс просмотра документов пользователем представлен в микро-модели; и свое графическое представление получил на рис. 1.3.

На вероятность просмотра документа влияют такие факторы как *релевантность* и *степень свежести* (времени, прошедшего с момента появления документа в сети до занесения его в индеексную базу). На рис. 1.3 эти факторы представимы в виде величин Y3 и Y4 соответственно. Значение свежести документа определяет, просмотрит ли его пользователь в следующий раз, либо пропустит.

Формально слои модели задача-ориентированного выбора документов представимы в виде формул (1.9) – (1.13) [7] (обозначения – в табл. 1.1).

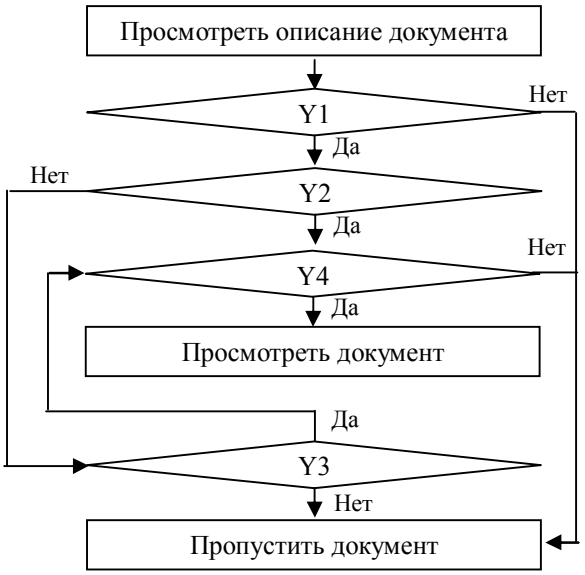

Рис. 1.3. Графическое представление сценария поведения пользователя 2-го предположения (Y1 – соответствует ли запрос нуждам пользователя; Y2 – просматривался ли ранее документ, Y3 – просмотреть ли повторно документ, Y4 – просмотреть ли документ)

Таблица 1.1

Обозначения модели задача-ориентированного выбора документов

Символ	*Назначение*
M_i	Соответствует ли i-й запрос нуждам пользователя
N_i	Продолжит ли пользователь поиск информации после успешного предшествующего запроса
$E_{i,j}$	Просмотр описания документа i-го запроса на позиции j
$H_{i,j}$	Предшествующий просмотр документа i-го запроса на позиции j
$F_{i,j}$	Свежесть документа i-го запроса на позиции j
$R_{i,j}$	Релевантность документа i-го запроса на позиции j
C_i	Выбор документа i-го запроса на позиции j
$S_{i,j}$	Пертинентность документа i-го запроса на позиции j

$$P(M_i = 1) = \alpha_1, \quad (1.9)$$

$$P(N_i = 1 \mid M_i = 1) = \alpha_2, \quad (1.10)$$

$$P(F_{i,j} = 1 \mid H_{i,j} = 1) = \alpha_3, \quad (1.11)$$

$$P(E_{i,j} = 1) = \beta_j, \quad (1.12)$$

$$P(R_{i,j} = 1) = r_d, \quad (1.13)$$

Следует отметить, что микромодель предполагает интеграцию с существующими моделями выбора, что обеспечивает масштабируемость модели задача-ориентированного выбора. Таким образом, если расширить модель моделью пользовательского просмотра, то (1.12) преобразуется в следующий вид:

$$P(E_{i,j} = 1 \mid C_{l_j} = 1, C_{l_{j+1}:j-1} = 0) = \beta_{l_j,j}. \quad (1.14)$$

Интегрируя динамическую байесовскую сеть в модель, вводится величина пертинентности документа, формально представимая в виде:

$$P(C_{i,j} = 1 \mid \begin{matrix} M_i = 1, E_{i,j} = 1, R_{i,j} = 1, \\ F_{i,j} = 1) = c_d, \end{matrix} \quad (1.15)$$

$$P(S_{i,j} = 1 \mid C_{i,j} = 1) = s_d. \quad (1.16)$$

2. Вероятностно-поведенческая модель пользователя университетской кампусной сети

Интегрируя модель задача-ориентированного выбора документов в рамки университетской кампусной сети, возникает возможность повышения точности определения пертинентных документов за счёт наличия специализированной информации о пользователе. Специализированную информацию составляют данные, зависящие от класса пользователя университетской кампусной сети. Классифицировать пользователей возможно по принципу выполняемых работ: студент, преподаватель, инженер конструкторского бюро и прочие. Рассматривая модель задача-ориентированного выбора документов, было установлено, что от значения случайной величины *Y1* будет зависеть вероятность просмотра документов результирующей выборки для запроса. Таким образом, используя данные класса пользователя можно повысить вероятность просмотра документов запроса пользователем.

Принимая во внимание вышеуказанное обстоятельство, модель задача-ориентированного выбора документов получает модификацию относительно двух слоёв. На рис. 2.1 представлен пример модификации макромодели.

Рис. 2.1. Графическое представление уточненного сценария поведения пользователя 1-го предположения

Данной модификации присуща абстрактность в плане целевого класса пользователя, т.е. различное число условий и их значений быть применены, учитывая специфику пользователя университетской кампусной сети. В табл.2.1 приведены обозначения случайных величин *Yx.x* для некоторых классов пользователей.

Таблица 2.1

Обозначения модифицированного слоя
модели выбора документов относительно классов пользователей

	Студент	*Преподаватель*	*Инженер конструкторского бюро*
Y1.1	Относится ли тематика документа к успешно обучаемой, либо обученной, дисциплине?	Относится ли тематика документа к направлению подготовки студентов?	Относится ли тематика документа к специальности инженера?
Y1.2	Относится ли тематика документа к дисциплине, связанной с *Y1.1*?	Относится ли тематика документа к направлению научной деятельности?	Относится ли тематика документа к решаемым задачам?
Y1.3	Соответствует ли запрос потребностям пользователя?		

Следует отметить, что вне зависимости от класса пользователя, последнее условие таблицы (в данном случае *У1.3)* всегда должно присутствовать в силу неопределенности потребностей конкретного пользователя и преследуемых им конечных целей в момента поиска информации.

Формализация слоев модели задача-ориентированного выбора примет вид табл. 2.

Таблица 2.2

Обозначения модифицированной модели задача-ориентированного выбора

	Студент	*Инженер конструкторского бюро*
$M_{T_{i,j}}$	Соответствует ли тематика документа i-го запроса на позиции j успешно обучаемой, либо обученной, дисциплине	Соответствует ли тематика документа i-го запроса на позиции j специальности инженера и решаемой им задаче
D	множество дисциплин	-
sp	-	Специализация
T	-	Множество решаемых задач
τ_d	$P(M_{T_{i,j}} = 1 \mid D)$ (2.1)	$P(M_{T_{i,j}} = 1 \mid sp, T)$ (2.2)

Принимая во внимание случайные величины Y1.1 и Y1.2, формулы вероятностей преобразуются к следующим формам:

$$P(M_i = 1 \mid \tau_d) = \alpha_1, \quad (2.3)$$

$$P(N_i = 1 \mid \alpha_1) = \alpha_2, \quad (2.4)$$

$$P(F_{i,j} = 1 \mid H_{i,j} = 1, l_{ind}) = \alpha_3, \quad (2.5)$$

где случайная величина τ_d дополняет M_i. Также при анализе документов на степень их свежести вводится величина l_{ind} – ограничение времени последнего обновления документа в системе. Таким образом, дополненные вероятности преобразуют собой составные:

$$P(E_{i,j} = 1 \left| \begin{array}{l} C_{l_j} = 1, C_{l_{j+1}:j-1} = 0, \alpha_1, \\ \alpha_3) = \beta_{l_j,j}, \end{array} \right. \quad (2.6)$$

$$P(C_{i,j} = 1 \left| \alpha_1, \alpha_3, \beta_{i,j}, r_d \right.) = c_d, \quad (2.7)$$

$$P(S_{i,j} = 1 \left| c_d \right.) = s_d. \quad (2.8)$$

Проведенная модификация способна повысить вероятность просмотра документов университетской кампусной сети, тематика которых соответствует нуждам пользователя.

3. Пример использования вероятностно-поведенческой модели инженера конструкторского бюро

Допустим, существует некоторое множество документов DC, выданных для пользовательского запроса. Элементы множества DC представляют собой следующую структуру:

$$dc = \{TP', t_{cur} - t_{ind} \le l_{ind}\}, \quad (3.1)$$

определяя множество тематик TP', экспертно установленных администратором УКС, и логический вывод относительно степени свежести документа. Логический вывод составляют величины:

t_{cur} – время отправки i-го запроса пользователем;

t_{ind} – время последнего индексирования (обновления) документа i-го запроса на позиции j в хранилище университетской кампусной сети;

l_{ind} – граничение времени последнего обновления документа i-го запроса на позиции j в университетской кампусной сети. Класс пользователя в данном случае является инженер конструкторского бюро, который ответственен за выполнение ряда задач, привязанных к нему. Формально сущность инженера представлена в следующем виде:

$$e \in E, e = \{sp\}, \quad (3.2)$$

$$sp = \{TP'\}, \quad (3.3)$$

$$t \in T, t = \{TP', e\}, \quad (3.4)$$

где E – множество инженеров конструкторского бюро; sp – специальность инженера e конструкторского бюро; T – множество задач конструкторского бюро, исполняющих их инженерам.

Тогда для множества DC возможно провести ранжирование документов относительно их тематике к специализации инженера и решаемых им задач, обеспечив (2.2):

$$T' = \{t \in T \mid t.e = e, t.TP' \in e.sp.TP'\}, \quad (3.5)$$

$$DC' = \{dc \in DC \mid dc.TP' \in e.sp.TP', dc.TP' \subset T'.TP'\}. \quad (3.6)$$

Получив ранжированное множество документов DC' необходимо также фильтровать документы относительно l_{ind} с целью выполнения условия 2-го предположения о поведении пользователя (в данном случае, инженера) в университетской кампусной сети (2.5).

$$DC'' = \{dc \in DC' \mid dc.(t_{cur} - t_{ind} \le l_{ind}) = true\}. \quad (3.7)$$

В конечном итоге, применяя методы определения параметров (2.3), (2.6), (2.7) модели, а также принимая во внимание неопределенность потребностей конкретного инженера, формируется конеч-

ная вероятностно-поведенческая модель инженера конструкторского бюро. Таким образом, появляется возможность проектирования программного агента, позволяющего ранжировать и предлагать пользователю (в данном случае, инженеру конструкторского бюро) университетской кампусной сети документы, позволяя ему не затрачивать время на поиск информации во время выполняемых им работ.

Заключение

Анализ моделей выбора документов информационных систем определил необходимость использования значительного кластера информации сессии поиска с целью проектирования эффективных моделей выбора. Группой ученых была предложена модель задача-ориентированного выбора документов, основанной на двух предположениях о поведении пользователей информационной системы, целью которой является повышение точности определения пертинентных документов поисковой системы и в то же время решение задачи прогнозирования поведения поиска. Модель обладает определенной степенью абстрактности и масштабируемости, что обеспечивает легкость её расширения.

Интегрируя модель задача-ориентированного выбора документов в рамки университетской кампусной сети была сформирована расширенная версия этой модели, позволяющая определить пертинентные документы для конкретного класса пользователя, работающих с университетской кампусной сетью. Помимо этого, была предложена идея о проектировании программного агента, позволяющего предлагать пользователю информацию с целью минимизации времени на поиск информации пользователем.

Список литературы

1. O. Chapelle. A dynamic bayesian network click model for web search ranking / O. Chapelle, Y. Zhang. In Proc. of the 18th Int. World Wide Web Conference, 2009, pages 1–10.

2. N. Craswell. An experimental comparison of click position-bias models / N. Craswell, O. Zoeter, M. Taylor, B. Ramsey. In Proceedings of the 1st ACM Int Conference on Web Search and Data Mining, 2008, pages 87–94.

3. G. Dupret. A user browsing model to predict search engine click data from past observations / G. Dupret, B. Piwowarski. In Proceedings of the 31st Annual ACM SIGIR Conference, 2008, pages 331–338.

4. F. Guo. Click chain model in web search / F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, C. Faloutsos. In Proceedings of the 18th International World Wide Web Conference, 2009, pages 11–20.

5. B. Hu. Characterize search intent diversity into click models / B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. In Proceedings of the 20th International World Wide Web Conference, 2011, pages 17–26.

6. M. Richardson. Predicting clicks: estimating the click-through rate for new ads / M. Richardson, E. Dominowska, R. Ragno. In Proceedings of the 16th International World Wide Web Conference, pages 521–530, 2007.

7. Y. Zhang. User-click Modeling for Understanding and Predicting Search-behavior / Y. Zhang, W. Chen, D. Wang, Q. Yang. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, Pages 1388-1396

8. C. Burges. Learning to rank using gradient descent / C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender. In Proceedings of the 22nd International Conference on Machine Learning, 2005, pages 89–96.

9. G. Dupret. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine / G. Dupret, C. Liao. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, 2010, pages 181–190.

10. R. Srikant. User browsing models: relevance versus examination / R. Srikant, S. Basu, N. Wang, D. Pregibon. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pages 223–232.

11. B. Piwowarski. Mining user web search activity with layered bayesian networks or how to capture a click in its context / B. Piwowarski, G. Dupret, R. Jones. In Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, 2009, pages 162–171.

12. L.A. Granka. Eye-tracking analysis of user behavior in WWW search / L.A. Granka, T. Joachims, G. Gay. In Proceedings of the 27th Annual ACM SIGIR Conference, 2004, pages 478–479.

Надійшла до редколегії 8.04.2017

Рецензент: д-р техн. наук, проф. С.В. Шабанов-Кушнаренко, Харківський національний університет радіоелектроніки, Харків.

ІМОВІРНОСНО-ПОВЕДІНКОВА МОДЕЛЬ КОРИСТУВАЧА УНІВЕРСИТЕТСЬКОЇ КАМПУСНОЇ МЕРЕЖІ

Є.А. Дружинін, І.В. Шостак, О.О. Лисенко

Розглянуто базис ймовірнісних моделей вибору документів інформаційних систем. Приведені припущення щодо характеру поведінки користувача інформаційної системи. Проведено інтеграцію моделі завдання-орієнтованого вибору документів у рамках університетської кампусової мережі. Розглянуто використання ймовірнісно-поведінкової моделі на прикладі класу користувача — інженера конструкторського бюро.

***Ключові слова**: ймовірна модель вибору, модель завдання-орієнтованого вибору, університетські кампусні мережі, аналіз поведінки користувача, інформаційний пошук.*

PROBABILISTIC AND BEHAVIORAL MODEL OF CAMPUS NETWORK'S USER

E.A. Druzhinin, I.V. Shostak, A.A. Lysenko

The basis of user-click models are considered. Assumptions of the user search-behavior are given. The task-centric click model integration to campus network are made. The usage of probabilistic and behavioral model, which rely on design-engineer class, are considered.

***Keywords**: user-click model, task-centric click model (TCM), campus networks, user search-behavior analysis, data browsing.*