УДК 004.82

ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ МОДЕЛЕЙ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ В СЕТИ

Крюкова Я.Э.[1], Гришунов С.С.[1], Рыбкин С.В.[1]

[1] Калужский филиал ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: fn1-kf@mail.ru

В статье был проведен обзор современных наиболее популярных моделей пользователей. В настоящее время исследование и создание таких модели являются одним из основных направлений для улучшения интернет поиска. Под моделью пользователя подразумевается математическая модель взаимодействия пользователя с поисковой системой или пользовательским сервисом. Были рассмотрены такие модели как: модель click chain model (CCM) в переводе означает модель цепочки кликов, модель dynamic Bayesian network (DBN) которая переводится как динамическая байесовская сеть, модель task-centric click model (TCM) в переводе модель клика, ориентированного на задачу и модель session click model (SCM) в переводе модель клика по сеансу. Были подробно рассмотрены исходные гипотезы, на которые опирались авторы при создании моделей. Были разобраны основные принципы взаимодействия пользователя с сетью в упрощенном виде с целью получения его математической модели. Для наглядного представления были приведены блок схемы моделей и байесовские сети. На основе описания поведения пользователя были построены математические модели, имеющие статистические зависимости. И в итоге были проанализированы основные преимущества моделей и их отличительные особенности.

Ключевые слова: модели поведения пользователей, модель CCM, модель TCM, модель DBN, модель SCM

RESEARCH OF MODERN MODELS OF USER BEHAVIOR IN THE NETWORK

Kryukova Y.E.[1], Grishunov S.S.[1], Rybkin S.V.[1]

[1] Bauman Moscow State Technical University (National Research University), Kaluga Branch, Kaluga, e-mail: fn1-kf@mail.ru

The article reviewed the current most popular user models. Currently, the study and creation of such models is one of the main areas for improving Internet search. A user model is a mathematical model of user interaction with a search engine or user service. The following models were considered: the click chain model (CCM) in translation means the click chain model, the dynamic Bayesian network (DBN) model, which is translated as a dynamic Bayesian network, the task-centric click model (TCM) model in translation, the click-based model task and session click model (SCM) in the translation of the session click model. The initial hypotheses on which the authors relied in creating models were considered in detail. Were dismantled the basic principles of user interaction with the network in a simplified form in order to obtain its mathematical model. For a visual representation, a block diagram of models and Bayesian networks were given. Based on the description of user behavior, mathematical models with statistical dependencies were constructed. And as a result, the main advantages of the models and their distinctive features were analyzed.

Keywords: user behavior patterns, CCM model, TCM model, DBN model, SCM model

Введение. Журналы кликов в поисковой системе являются бесценным ресурсом, который может предоставить богатый источник данных о пользовательских предпочтениях в их результатах поиска. Анализ журналов переходов по ссылкам может использоваться во многих связанных с поиском приложениях, таких как ранжирование веб-поиска, прогнозирование популярности сайта или прогнозирование удовлетворенности пользователей. При анализе журналов кликов основной вопрос заключается в том, как построить модель кликов, чтобы вывести воспринимаемую пользователем релевантность для каждой пары запрос-документ на основе огромного

количества данных поисковых кликов. Используя модель кликов, коммерческая поисковая система может лучше понять поведение пользователей при поиске и предоставить улучшенные пользовательские сервисы.

Модель CCM. В работе [1, 2] представлена модель click chain model (CCM), в переводе означает модель цепочки кликов. CCM основа на циклическом процессе взаимодействия пользователя с результатами поисковой системы.

При создании CCM авторы отталкивались от следующих предположений: пользователи являются однородными: их информационные потребности одинаковы при одинаковом запросе; вероятность клика определяется исключительно вероятностью просмотра и релевантностью документа в данной позиции; просмотр документов проводится в строго последовательном порядке без перерывов.

Модель в виде блок-схемы представлена на рисунке 1.

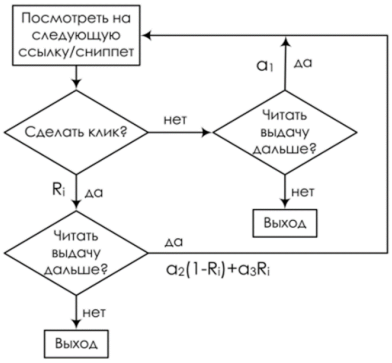

Рис. 1 Модель CCM. Блок схема

Модель в виде байесовской сети представлена на рисунке 2.

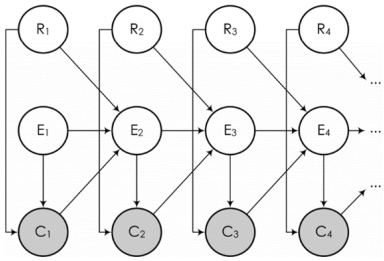

Рис. 2 Модель CCM в виде байесовской сети

Пользователь начинает проверку результатов поиска из документа с самым высоким рейтингом [3]. В каждой позиции i пользователь может выбрать или пропустить документ d_i в соответствии с предполагаемой релевантностью. В обоих случаях, пользователь может продолжить изучение или отказаться от текущего запроса. Вероятность клика на текущий документ составляет R_i. Вероятность пропуска документа d_i составляет α_1. Если же был сделан клик, то продолжение просмотра определяется релевантностью просмотренного документа и параметрами α_2, α_3.

$$p(C_i = 1 | E_i = 0) = 0 \quad (1)$$

$$p(C_1 = 1 | E_i = 1, R_i) = R_i \quad (2)$$

$$p(E_{i+1} = 1 | E_i = 0) = 0 \quad (3)$$

$$p(E_{i+1} = 1 | E_i = 1, C_i = 0) = \alpha_1 \quad (4)$$

$$p(E_{i+1} = 1 | E_i = 1, C_i = 1, R_i) = \alpha_2(1 - R_i) + \alpha_3 R_i \quad (5)$$

где C_i – клик на этой позиции

E_i – просмотр описания ссылки на документ, показанный на позиции i

Преимуществами данной системы является то, что она является масштабируемой и инкрементной, что идеально подходит для аналитических задач, которые используют журналы кликов, которые постоянно растут.

Модель DBN. Авторами работы [4] была предложена модель dynamic Bayesian network (DBN), которая переводится как динамическая байесовская сеть.

При создании CCM авторы основывались на гипотезах, предложенных ранее: клик происходит тогда и только тогда, когда пользователь изучил URL-адрес и посчитал его релевантным; пользователь делает линейный переход по результатам и решает, нажимать ли, основываясь на воспринимаемой релевантности документа; пользователь решает проверить следующий URL-адрес, если не удовлетворен выбранным URL-адресом (на основе фактической релевантности). Отличительной особенностью данной модели являются предположения: щелчок не обязательно означает, что пользователь удовлетворен выбранным документом (авторы пытаются различить воспринимаемую релевантность и фактическую релевантность); нет ограничений по количеству кликов, которые пользователь может сделать во время поиска. Ее модель в виде байесовской сети представлена на рисунке 3.

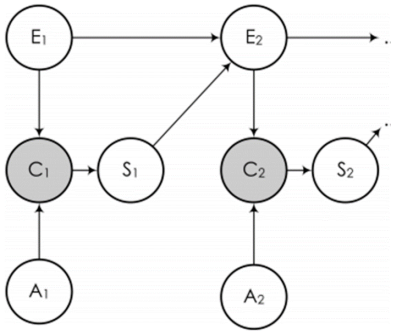

Рис. 3 Модель DBN. Байесовская сеть

Пользователь кликает на документ только в том случае, если пользователь посмотрел описание документа и был им привлечен. Вероятность того, что документ окажется привлекательным для пользователя составляет a_i. После того, как пользователь щелкнет и посетит URL-адрес, существует определенная вероятность s_i того, что он будет удовлетворен этим URL-адресом. С другой стороны, если он не перейдет по ссылке, он не будет удовлетворен. Как только пользователь удовлетворен посещенным им URL-адресом, он прекращает поиск. Если пользователь не удовлетворен текущим результатом, существует вероятность γ, что пользователь просмотрит описание следующего документа. С другой стороны, существует вероятность 1 - γ, что пользователь откажется от своего поиска. Другими словами, γ измеряет настойчивость пользователя. Если пользователь не исследовал позицию i, он не будет проверять последующие позиции.

$$E_i = 1, A_i = 1 \Leftrightarrow C_i = 1 \tag{6}$$

$$p(A_i = 1) = a_i \tag{7}$$

$$p(S_i = 1 | C_i = 1) = s_i \tag{8}$$

$$C_i = 0 \Rightarrow S_i = 0 \tag{9}$$

$$S_i = 1 \Rightarrow E_{i+1} = 0 \tag{10}$$

$$p(E_{i+1} = 1 | E_i = 1, S_i = 0) = \gamma \tag{11}$$

$$E_i = 0 \Rightarrow E_{i+1} = 0 \tag{12}$$

В отличие от других, DBN имеет две переменные a_i и s_i, связанные с релевантностью документа. Первая измеряет вероятность клика на основе описания документа и связана с воспринимаемой релевантностью. Вторая - вероятность того, что пользователь удовлетворен, если он нажал на ссылку.

Таким образом преимуществом модели является разделение воспринимаемой и фактической релевантности документа. В некоторых случаях она различается, несмотря на существование сильной корреляции между ними.

Модель TCM. В работе [5] представлена модель task-centric click model (TCM), в переводе модель клика, ориентированного на задачу. Традиционно пользовательские сеансы получают из последовательной последовательности действий пользователя по поиску и просмотру в течение фиксированного интервала времени. Эти сеансы могут быть разделены на две категории: сеанс запроса (рисунок 4) и сеанс поиска (рисунок 5),

где первый относится к действиям по просмотру информации в отдельном запросе, в то время как последний охватывает все запросы и просмотры, которые пользователь выполняет для удовлетворения своих информационных потребностей.

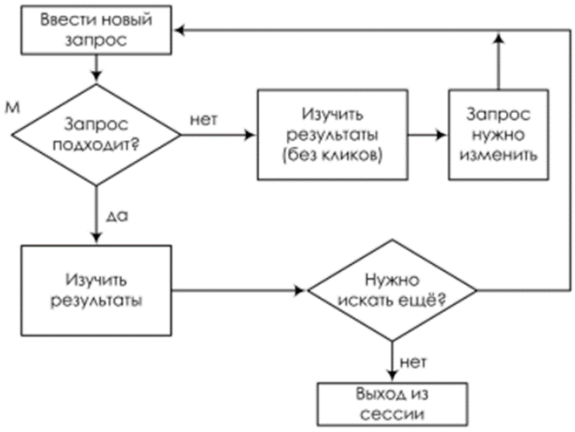

Рис. 4 Модель TCM Блок-схема сеанса запроса

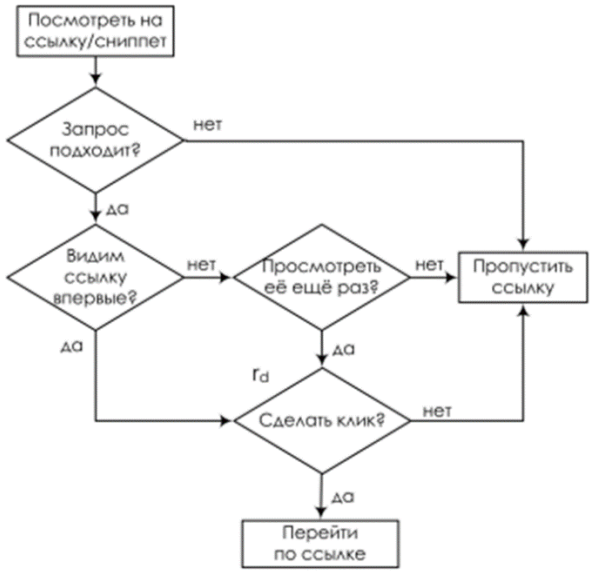

Рис 5 Модель TCM. Блок-схема сеанса поиска

В начале было выдвинуто два предположения: если запрос не соответствует намерению пользователя, он не будет выполнять щелчки, а будет создавать новый запрос; когда документ был исследован ранее, вероятность того, что пользователь нажмет на него, будет ниже, если повторный запрос выдаст его.

TCM использует две двоичные случайные величины, M и N, для характеристики. M представляет, является ли запрос хорошим или, другими словами, соответствует ли он намерению пользователя, а N представляет, хочет ли пользователь продолжить свой поиск, если его предыдущий запрос является хорошим. Стоит отметить, что значение M будет влиять на то, как модель интерпретирует поведение пользователя для текущего запроса. Вероятность клика определяется релевантностью документа, обозначенной как r.

С другой стороны, если документ был проверен ранее, пользователь решит, является ли документ все еще «свежим», который характеризуется как случайная переменная F. Для не свежего документа его поведение при щелчке больше не определяется исключительно его актуальностью, поэтому нужно охарактеризовать свежесть, чтобы интерпретировать поведение кликов пользователей. Другими словами, пользователь будет стремиться пропустить не свежий документ, так как он мог быть исследован ранее.

Задача содержит m запросов (сессий), причем каждый сеанс запросов содержит n упорядоченных документов. Коэффициент j-й обозначает позицию документа в i-й сессии запроса. α1 представляет вероятность Mi = 1, извлеченную из контекстной информации всей задачи. α2 и α3 являются параметрами, которые имеют аналогичные значения [6].

$$p(M_i = 1) = \alpha_1 \tag{13}$$

$$p(N_i = 1 | M_i = 1) = \alpha_2 \tag{14}$$

$$p(F_{i,j} = 1 | H_{i,j} = 1) = \alpha_3 \tag{15}$$

$$p(E_{i,j} = 1) = \beta_{i,j} \tag{16}$$

$$p(R_{i,j} = 1) = r_d \tag{17}$$

$$M_i = 0 \Rightarrow N_i = 1 \tag{18}$$

$$H_{i,j} = 0 \Rightarrow F_{i,j} = 1 \tag{19}$$

$$H_{i,j} = 0 \Leftrightarrow H_{i',j'} = 0, E_{i',j'} = 0 \tag{20}$$

$$C_{i,j} = 1 \Leftrightarrow M_i = 1, E_{i,j} = 1, R_{i,j} = 1, F_{i,j} = 1 \tag{21}$$

где β – обозначает вероятность того, что пользователь просмотрит описание.

Таким образом главным отличием и преимуществом данной модели является то, что модель рассматривает поведение пользователя на уровне сессии. Что в свою очередь ведет к повышению точности моделирования.

Модель SCM. Авторами работы [7] была предложена session click model (SCM), в переводе модель клика по сеансу. За основу была взята модель DBN.

Авторы работы основывались на следующих предположениях: пользователь может сформулировать свой запрос неточно и, при этом, удовлетвориться нерелевантным документом или проигнорировать релевантный; пользователь перед кликом на документ всегда просматривает описание ссылки; пользователь просматривает выдачу последовательно при этом может сделать один, несколько или не одного клика по

документам и покидает сессию, как только находит интересующий его документ; вероятность того что пользователь будет кликать на ранее просмотренный документ меньше.

Используя вышеперечисленные предположения, авторы разработали модель структура графа, зависит только от количества запросов в сессии, а не от конкретной сессии. На рисунке 6 приведена предлагаемая модель клика.

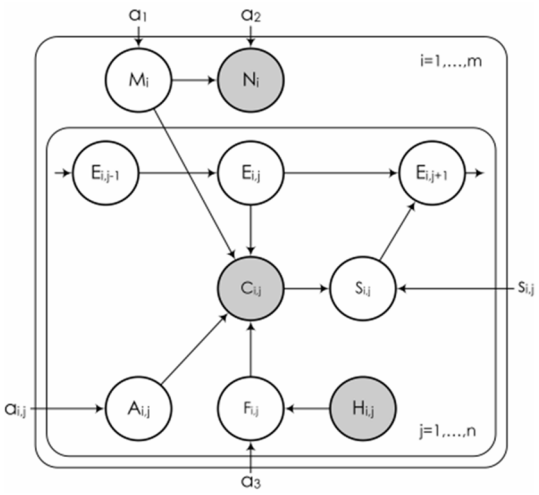

Рис. 6 Модель SCM. Байесовская сеть

Набор условных вероятностей, завершающий формальное определение модели.

Параметр $a_{i,j}$ определяет привлекательность документа, а точнее, соответствующего ему элемента на странице выдачи. Этот параметр влияет на вероятность клика. Параметр $s_{i,j}$ можно интерпретировать как пертинетность документа: если запрос соответствовал потребностям пользователя, и пользователь закончил просмотр выдачи кликом по этому документу, значит, документ оказался ему полезным [8].

$$E_{i,j}=1 \Rightarrow E_{i,j-1}=1, S_{i,j-1}=0 \tag{22}$$

$$p(S_{i,j}=1 \mid C_{i,j}=1)=s_{i,j} \tag{23}$$

$$p(S_{i,j}=1 \mid C_{i,j}=0)=0 \tag{24}$$

$$C_{i,j}=1 \Leftrightarrow M_i=1, E_{i,j}=1, S_{i,j}=1, F_{i,j}=1 \tag{25}$$

В качестве оценки релевантности документа, как и в DBN, используется:

$$\begin{aligned} r_{i,j} &= p(S_{i,j}=1 \mid E_{i,j}=1, M_i=1, F_{i,j}=1) = \\ &= p(S_{i,j}=1 \mid C_{i,j}=1)\, p(C_{i,j}=1 \mid E_{i,j}=1, M_i=1, F_{i,j}=1) = a_{i,j} s_{i,j} \end{aligned} \tag{26}$$

Заключение. Многообразие моделей поведения пользователя говорит о распространенности и важности данной проблемы. Используя данные математические модели для улучшения пользовательских сервисов и релевантности выводимой информации необходимо выбирать между высокой точностью модели и высокой производительность. Таким образом, извлечение релевантной информации из журналов пользователей является сложной, но ценной задачей для ранжирования в веб-поиске.

Список литературы

1. Guo F., Liu C., Kannan A., Minka T., Taylor M., Wan Y., Faloutsos C. Click chain model in web search // Proceedings of the 18th International Conference on World Wide Web. 2010. P. 11-20
2. Микуленя И.В., Белов Ю.С. Коллаборативная фильтрация // В сборнике: В МИРЕ НАУКИ И ИННОВАЦИЙ. сборник статей международной научно-практической конференции: в 5 частях. 2017. С. 72-74.
3. Клюквин Р.В., Белов Ю.С. Использование косинуса схожести в рекомендательных системах на основе коллаборативной фильтрации // Электронный журнал: наука, техника и образование. № 2 (12). 2017. С. 131-136. [Электронный ресурс]. – Режим доступа: http://nto-journal.ru/catalog/informacionnye-texnologii/340/ (дата обращения 27.01.2019)
4. Chapelle O., Zhang Y. A dynamic Bayesian network click model for web search ranking // Proceedings of the 18th International Conference on World Wide Web. 2010. P. 1–10.
5. Zhang Y., Chen W., Wang D., Yang Q. User-click modeling for understanding and predicting search-behavior // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, 2011. P. 1388–1396.
6. Artzi, Y., Pantel, P., Gamon, M. Predicting responses to microblog posts // Proceedings of the 12 th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT ’12, Stroudsburg, PA, USA. 2012. P. 602–606
7. Николенко С. И., Фишков А. А. SCM: новая вероятностная модель поведения пользователей интернет-поиска // Тр. СПИИРАН. № 20. 2012. С. 72–100.
8. Николенко С. И., Фишков А. А. Обзор моделей поведения пользователей для задачи ранжирования результатов поиска // Тр. СПИИРАН. № 22. 2012. С. 139–175.