СПИСОК ЛИТЕРАТУРЫ

1. Казначеев В.П., Баевский Р.М., Берсенева А.П. Донозологическая диагностика в практике массовых обследований населения. М.: Медицина, 1992. 208 с.

2. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 1. М.: ФИЗМАТЛИТ, 2005. 144 с.

3. Евдокименков В.Н. Компьютерные технологии сбора, обработки и анализа данных медико-биологических исследований: Учебное пособие. М.: Изд-во МАИ, 2005. 436 с.

4. Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schutze Schutze. Introduction to Information Retrieval. Cambridge UP. Online edition (c), 2009. 544 c.

5. Нестерова О.А., Оленников Е.А. Проблема сбора и анализа данных для научных исследований в медицинских информационных системах / Искуственный интеллект: философия, методология, инновации: мат-лы III Всерос. конф. студентов, аспирантов и молодых ученых. М.: Связь-принт, 2009. С. 371-373.

6. Аветисян Р.Д., Аветисян Д.О. Теоретические основы информатики. М.: Вильямс, 2002. 168 с.

Игорь Николаевич ГЛУХИХ —
зав. кафедрой мат. методов, статистики
и информационных технологий в экономике
доктор технических наук, профессор
igluhih@utnm.ru

Михаил Владимирович ГУБИН —
ассистент кафедры мат. методов, статистики
и информационных технологий в экономике
Mikhail.gubin@gmail.com

Институт математики и компьютерных наук
Тюменского государственного университета

УДК 004.023: 004.031.42

ОПТИМИЗАЦИЯ ВЫДАЧИ КОНТЕНТА В ДИНАМИЧЕСКОМ ИНТЕРНЕТ-ПРОЕКТЕ НА ОСНОВЕ АНАЛИЗА ВЫБОРОВ НЕИДЕНТИФИЦИРОВАННЫХ ПОЛЬЗОВАТЕЛЕЙ

OPTIMIZATION OF CONTENT DELIVERY IN THE DYNAMIC INTERNET PROJECT ON THE BASIS OF SELECTION ANALYSIS OF UNIDENTIFIED USERS

АННОТАЦИЯ. В статье рассматриваются модели и алгоритмы, предназначенные для управления выдачей пользователям контента на основе учета действий множества неидентифицированных пользователей.

SUMMARY. The article describes models and algorithms designed to manage the content delivery based on the account of actions of a number of unidentified users.

КЛЮЧЕВЫЕ СЛОВА. Генерация контента, гипертекстовая статья, анализ данных, поиск информации.

KEY WORDS. Content generation, hypertext article, data analysis, search of information.

Одним из актуальных направлений развития крупных контентных проектов в интернете (информационных, новостных, образовательных и им подобных интернет-порталов) стало создание методов выдачи по поисковому запросу пользователя контента, отвечающего некоторым критериям оптимальности. Наиболее часто оптимизация поисковой выдачи обсуждается как задача ее персонализации. Решение этой задачи предполагает накопление знаний о пользователе путем сбора и анализа данных о его действиях за предыдущие периоды или же выявление предпочтений опросным путем [1].

В то же время опыт создания интернет-проектов [2], опыт анализа открытых данных аналитических систем (например, liveinternet.ru) показал, что большинство посетителей крупного контентного проекта (объемом десятки тысяч и более страниц) составляют посетители, которые пришли на портал впервые по запросу из поисковой системы. Доля таких посетителей может составлять более 70%. Повышение эффективности работы с этой категорией пользователей является одним из наиболее актуальных направлений повышения эффективности самого интернет-проекта.

В статье рассматриваются модели и алгоритмы, которые предназначены для оптимизации поисковой выдачи контента на основе анализа действий множества неидентифицированных пользователей. При этом формализуется представление контента, предлагаются критерии оптимизации с учетом целей владельца интернет-проекта.

Формализация представления контента. Пусть есть множество компонентов контента (КК):

$$X = \{x_1, x_2, x_3, \ldots\},$$

где по некоторым признакам можно выделить классы X1, X2, X3.... Каждый КК может быть представлен в множестве X набором различных форм представления:

$$x = < x_I, x_{S1}, x_{S2}, \ldots, x_{A1}, x_{A2}, \ldots >,$$

где x_I — основная форма представления; x_{S1} , x_{S2}, … — дополнительные и x_{A1}, x_{A2} , … — альтернативные формы представления КК. В качестве форм представления в зависимости от класса КК могут быть тексты, изображения, баннеры, гиперссылки, кнопки и др.

Результатом поисковой выдачи по запросу пользователя является гипертекстовая статья (ГС). Не уточняя форм представления, можно сказать, что множество компонентов гипертекстовой статьи H есть подмножество X:

$$H \subseteq X.$$

Сказанное пока относится лишь к составу, но ничего не говорит о структуре ГС, т.к. не учитывает, что в гипертекстовой статье, которая предъявляется пользователю, компоненты контента могут располагаться на различных, как правило, предопределенных местах. Задача синтеза ГС включает не только формирование $H \subseteq X$, но и расстановку x по «нужным» местам. Упорядоченное по некоторому признаку множество позиций для размещения КК в ГС обозначим M.

Пусть в результате обработки пользовательского запроса из множества X выделено упорядоченное по релевантности запросу подмножество X_z, причем

$|H| \le |X_z|$, где мощность $|H|$ ограничивается числом мест для размещения компонентов контента, т.е. $|H| = |M| = N$.

Задача синтеза ГС в ответ на запрос пользователя включает в себя этапы:

— выборка из X_z N элементов;

— размещение N элементов по N позициям с учетом $M \Leftrightarrow H$.

Выполнение обоих этапов порождает в итоге множество вариантов ГС = {ГС1, ГС2, … }, в результате чего задача синтеза становится задачей выбора, которая решается с помощью некоторой функции (оценки) ценности ГС. В случае крупного интернет-проекта мощность этого множества выбора будет большой, что наряду с трудностью формализации понятия ценности ГС делает задачу выбора весьма нетривиальной.

Традиционным является синтез ГС исходя из оценки полезности ГС для пользователя, что определяется семантической близостью контента ГС поисковому запросу пользователя и его известным предпочтениям и интересам, если в базах знаний интернет-проекта содержатся знания о данном пользователе. Мы, как уже говорилось ранее, будем задачу выбора рассматривать исходя из целей не пользователя, а владельца интернет-ресурса. Это же будет влиять и на формирование цели (критерия) выбора.

Введем функцию:

$$\rho: H \times M \to \{1,0\},$$

которая ставит каждой паре из $H \times M$ в соответствие 1, если i-й элемент H размещен на j-м месте из M и 0 — в противном случае.

Пусть $Z = \rho(H, M)$ — матрица, которая есть формальное представление ГС, где определено множество КК и их размещение. Пусть известна цель C выбора, тогда полезность ГС в смысле достижения цели обозначим как C(Z), и задача выбора лучшего варианта Z^* запишется так:

$$C(Z^*) \to \max, \qquad (1)$$

где $Z^* \in D=\{Z_k \mid k = 1, 2, 3,\ldots K\}$, D — множество вариантов, из которых осуществляется выбор, причем $\forall Z\ (Z \in D \to d(Z) \ge d_{min})$; d(Z) — некоторая функция допустимости, которая позволяет вычислить значение показателя соответствия Z требованиям допустимости. В простейшем случае область значений d(Z) есть {0,1}; d_{min} — минимальный порог допустимости.

Положим, что согласно запросу пользователя сформировано множество X_z, где на первом месте элемент $x \in X1$ — основной тематический КК, который по принимаемому критерию поиска определен как наиболее соответствующий запросу пользователя. В M выделим первый элемент, который есть позиция для размещения этого КК. Тогда элемент $z_{1,1}=1$ в Z есть представление главного КК, выдача которого и является ответом на запрос пользователя. Остальные элементы Z могут формироваться различными способами. Для генерации множества вариантов D используется преобразование R (далее для упрощения так же будем обозначать матрицу этого преобразования):

$$Z' = RZ, \qquad (2)$$

где элементы $r_{i,j} \in \{0, 1\}$.

Базовые условия допустимости преобразования при генерации вариантов выбора:

— в каждом столбце матрицы R только одна 1;

— в каждой строке матрицы R только одна 1;

— преобразование не является тождественным, т.е. $\exists i\ (i= j \rightarrow r_{i,j} \neq 1)$;

— $r_{1,1} = 1$ — условие сохранения инварианта — представления в ГС главного семантического КК.

Этим ограничениям отвечает операция перестановки столбов матрицы (начиная со второго), которая может быть использована для получения групп преобразований $\{R_k \mid k = 1, 2, 3,\ldots K\}$, каждое из которых порождает свой вариант выбора.

Базовый алгоритм А1 синтеза оптимальных ГС:

Шаг 1. Формирование упорядоченного множества X_z компонентов контента, релевантных запросу пользователя;

Шаг 2. Формализация исходного экземпляра ГС посредством Z;

Шаг 3. Генерация множества допустимых вариантов D по (2);

Шаг 4. Выбор из D согласно (1) и предъявление пользователю лучшей по составу и структуре размещения ГС.

Основная трудность появляется при попытке формализации цели С в условиях активности пользователя, отсутствия знаний о нем и возможных расхождениях целей владельца ресурса и пользователя. Формулировка целей владельца интернет-ресурса может быть разной. К наиболее интересным относится цель «добиться нужного действия», что практически может выражаться в том, что пользователь прочитал ту или иную информацию, воспользовался тем или иным сервисом, заполнил регистрационные данные и т.п. В применяемых здесь терминах эта цель формулируется как цель «Провести пользователя через траекторию ГС, последняя из которых содержит целевой компонент x* в одной из форм представления:

$$C \Rightarrow (ГС_{t1}, ГС_{t2}, \ldots, ГС^*),$$

где $ГС_{t1}$ — точка входа — первая ГС, которая выдается пользователю по его запросу; ГС* — целевая ГС, на которой выполняется «нужное действие».

Длина траекторий может быть разной и будет зависеть от многих факторов.

Важной промежуточной целью является синтез и выдача пользователю такой гипертекстовой статьи в точке входа, чтобы $ГС_{t1}$ не стала для него же и последней. Иными словами в точке входа нужна такая ГС, после которой пользователь не уйдет с сайта:

$$C \Rightarrow (ГС_{t1}, ГС_{t2}),$$

где $ГС_{t2}$ — любая гипертекстовая статья, которая может быть получена в результате активации (выбора) пользователем КК из $ГС_{t1}$.

Положим, что верна следующая гипотеза: в одной точке входа для L пользователей, т.е. за L случаев одного поискового запроса, можно определить такую $Z^* \in D$, что:

$$C\ (Z^*) \rightarrow \max,$$

где C (Z*) — вероятность такого выбора, что образуется последовательность $(Z^* = Z_{t1}, Z_{t2})$, Z_{t1}, Z_{t2} — матричные представления $ГС_{t1}$, $ГС_{t2}$ соответственно.

Согласно этой гипотезе, после L запросов элементы множества D упорядочиваются по числу $C\ (Z_k) = n_k/L_k$, где $L = \sum L_k$, $L_1 = L_2 = \ldots L_k$, n_k — число

случаев появления последовательностей ($Z_k = Z_{t1}$, Z_{t2}) из общего количества L_k выдачи варианта Z_k.

Тогда начиная с (L+1)-го пользователя можно предъявлять $Z^* = Z_1$ как вариант ГС, наиболее подходящий для достижения цели.

Базовый алгоритм А1 преобразуется в следующий

Алгоритм А2 синтеза оптимальной ГС по результатам анализа данных:

Шаг 1. Формирование упорядоченного множества X_z компонентов контента, релевантных запросу пользователя;

Шаг 2. Формализация исходного экземпляра ГС посредством Z;

Шаг 3. Генерация множества допустимых вариантов D по (2);

Шаг 4: Предъявления L пользователям вариантов $Z \in D$ и упорядочение элементов D по C(Z).

Шаг 5. Начиная с (L+1)-го пользователя, выбор и предъявление $Z^* \in D$ — первого элемента в упорядоченном множестве D.

Алгоритм А2 может быть дополнен другой полезной процедурой, которая позволяет проводить более «тонкую» настройку ГС за счет оптимизации внутреннего содержания по критерию достижения поставленной цели.

Процедура оптимизации набора компонентов гипертекстовой статьи (ОНК ГС).

В ее основе лежит следующая гипотеза: если в ответ на один запрос пользователям выдается ГС в виде множества размещенных определенным образом компонентов контента, то за L итераций (L случаев запроса) из исходного варианта $ГС_{t1}$ можно получить такую $ГС_L$, в состав которой входят компоненты с наибольшей значимостью для достижения цели.

Решение задачи предполагает последовательное уточнение множества H с оценкой полезности элементов из H в смысле поставленной цели C. Так как $x_1 \in H$ есть главный содержательный компонент, то последовательное уточнение должно касаться только остальных компонентов $x_2, x_3, \ldots, x_S$, причем полагаем, что

$$N - 1 < S \le |X_z|. \qquad (3)$$

Введем множество пар:

$$\mu^H = \{ <\mu_i, x_i> \mid i = 2, 3, \ldots; \mu_i \ge \mu_{i\ min}\},$$

где μ_i — весовой коэффициент, значение которого отражает значимость для достижения цели C компонента x_i.

Будем H трактовать как множество конкурентов за позиции из множества M, тогда можно задать правило: Если $\mu_i > \mu_j$, то позиция $m(x_i) \in M$ более предпочтительна позиции $m(x_j) \in M$. Учтем (3), при котором $|\mu^H| \le |M|$, тогда при некотором $\mu_i < \mu_{i\ min}$ $m(x_i) \notin M$, т.е. компонент контента x_i не входит в гипертекстовую статью. Нормализованное множество

$$\mu^{Hnorm} = \{ <\mu_i^{norm}, x_i>\},$$

где $\sum \mu_i^{norm} = 1$, $\mu_i^{norm} = \mu_i / \sum \mu_i$.

Алгоритм ОНК ГС основан на пересчете в течение L итераций значений μ_i^{norm}. На каждом t-m шаге (t-й итерации) генерируется ГС, в которой расположение КК зависит от их μ_i^{norm}. Пользователи, получая ГС и переходя по ее

гиперссылкам, выбирают, таким образом, тот или иной компонент. Каждый выбор подтверждает значимость одних компонентов для цели С и опровергает значимость других. Это подтверждение/опровержение фиксируется в вычислениях $\mu_i^{norm}(t+1)$.

Для реализации алгоритма предлагается использовать формулу:

$$\mu_i^{norm}(t+1) = \exp\{n_i(t)\gamma\} / \sum\exp\{n_i(t)\,\gamma\},$$

где $n_i(t)$ _ суммарное количество переходов на компонент x_i к t-й итерации включительно, при t = L выполняется $\sum n_i(t) = L$; γ — положительный коэффициент, значение которого влияет на скорость и величину расхождения значений весов подтверждаемых и неподтверждаемых компонентов.

Подбирая γ, с помощью процедуры ОНК ГС из некоторого варианта Z, полученного посредством (2), за несколько итераций можно получить новый Z', состав и размещение компонентов которого зависят от истории выборов пользователей и наиболее полно отвечают поставленной цели.

Отметим, что аналогичного результата можно добиться с помощью преобразований (2), т.е. всегда в D может найтись такой Z'=R'Z, что

$$C(Z'_{t2=L}) > C(Z_{t1=L}),$$

если $Z \neq Z^*$. Здесь $C(Z'_{t2=L})$ — найденное значение полезности за L случаев выдачи данной ГС. Однако с ростом $|X_z|$, |H| в еще большей степени растет и число вариантов в D, а обоснование $C(Z_k)$ потребует значительного L_k по всем k. Кроме того, формула оценки C(Z) отражает полезность варианта Z как неделимой единицы. В то же время процедура ОНК ГС позволяет оперировать содержимым и ценностью компонентов внутри вариантов. Обозначим преобразование Z в Z' с помощью процедуры ОНК ГС, как

$$\varphi: Z \rightarrow Z'.$$

Алгоритм А3 синтеза ГС с оптимизацией набора компонентов:

Шаг 1. Формирование упорядоченного множества X_z компонентов контента, релевантных запросу пользователя;

Шаг 2. Формализация исходного экземпляра ГС посредством Z;

Шаг 3. Генерация по (2) D_{less}, $|D_{less}| << |D|$, $D_{less}=\{Z_k \mid k=1, \ldots, R\}$, $R \geq 1$ — число, подбираемое экспериментально;

Шаг 4. За L итераций преобразование $\varphi: Z_k \rightarrow Z'_k$, $\forall Z_k \in D_{less}$;

Шаг 5. Модификация за дополнительные Q итераций множества D_{less} — исключение пар $Z'_i \approx Z'_j$ (где «≈» — символ нечеткого равенства) и упорядочение по C(Z').

Шаг 6. Начиная с некоторой (Q+1)-й итерации предъявление пользователям $Z'^* = \text{argmax}\, C(Z'_k)$ — первого элемента в упорядоченном D_{less}.

Заключение. Предложенные модели и алгоритмы позволяют решать практические задачи поисковой выдачи контента с учетом целей владельца интернет-проекта. Важным является то, что они ориентированы на работу с неидентифицированными пользователями, которые, возможно, впервые на данном интернет-ресурсе, и сведений о которых недостаточно для эффективного взаимодействия с ними.

Тот факт, что пользователи, чьи выборы обрабатываются, объединены одним поисковым запросом, позволяет вводить и учитывать свою классификацию посетителей интернет-ресурса. Дальнейшее развитие предложенного

подхода может предполагать формулировку новых целей владельца интернет-ресурса и введение новых мер эффективности компонентов контента; разработку алгоритмов синтеза оптимальных траекторий, приводящих к заданной точке выхода пользователя; уточнение классификаций, например, путем выявления устойчивых траекторий движения пользователей из одной точки входа с введением классов поведенческого таргетинга и др.

СПИСОК ЛИТЕРАТУРЫ

1. Гильманов А.С. Математическая модель и прикладные разработки технологий доставки контента в электронных образовательных системах: Автореф. дис. … канд. тех. наук. Тюмень, 2010. 21 с.

2. Разработка функциональных модулей программной платформы для реализации первой очереди сервисов виртуального офиса на портале электронных коммерческих сервисов. Закл. отчет по НИР / Рук. И.Н. Глухих № ГР 01200950860. Тюмень, 2010. 43 с.

3. Морозов В.П., Тихомиров В.П., Хрусталев Е.Ю. Гипертексты в экономике. Информационная технология моделирования. М.: Финансы и статистика, 1997.

4. Жилинскас А., Шалтянис В. Поиск оптимума: компьютер расширяет возможности. М.: Наука, 1989.

Валерий Алексеевич ШАПЦЕВ —
профессор кафедры информационных систем
Института математики и компьютерных наук
Тюменского государственного университета,
доктор технических наук
vashaptsev@ya.ru

УДК 519.72 : 004(075.8)

ИНФОРМАЦИЯ. ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ: АКТУАЛЬНАЯ ТОЧКА ЗРЕНИЯ

INFORMATION. INFORMATION TECHNOLOGY: UP-TO-DATE POINT OF VIEW

АННОТАЦИЯ. Анализируются понятие информация как потенциал отражения материального мира, и информационный процесс как совокупность этапов восприятия сигналов, их интерпретации и использования результата интерпретации. Обосновывается тезис: современные ИТ в основном работают с данными, а не с информацией. Данные — носитель информации. Выявлением информации из них занимается человек.

SUMMARY. The article offers the analysis of such notions as information (as a potential for material world reflection) and information process (as the totality of steps of signals perception), as well as their interpretation and application of interpretation results. The thesis is proved: modern IT systems basically work with data, but not with information. The data are regarded as information carrier. Detection of information from the data is carried out by a man.

КЛЮЧЕВЫЕ СЛОВА. Информация, сигналы, данные, информационный процесс, информационные технологии, информационные системы, системы работы с данными.

KEY WORDS. Information, signals, data, information process, information technologies, information systems, systems of data processing.