УДК 681.3

ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ ДЕЯТЕЛЬНОСТИ ПОЛЬЗОВАТЕЛЕЙ С ВЫЯВЛЕНИЕМ ИНФОРМАЦИОННО-ЗНАЧИМЫХ ОБЪЕКТОВ

© 2021 Д. Л. Зайцев, А. Н. Зеленина

Воронежский институт высоких технологий (Воронеж, Россия)

Рассмотрена деятельность пользователя по поиску информационно-значимых объектов решения поставленной перед ним задачи. Для формализации этой деятельности использован аппарат теории множеств. Приведены математические модели формулировка запроса, анализа документов в выдаче и отбор информационных объектов для решения поставленной перед пользователем задачи. Кроме того, разработаны модели в аспекте интеллектуальной деятельности пользователя и когнитивных процессов. Приведены результаты экспериментальных исследований с такими поисковыми системами, как Google, Yandex, Rambler, Yahoo. Для оценки эффективности информационного поиска найденные документы разделены на пертинентные, релевантные и нерелевантные. Эффективность поиска определено отношением количества пертинентных и релевантных документов к количеству всех документов в выдаче. Информационные характеристики поисковых систем предлагается учитывать соответствующим коэффициентом.

Ключевые слова: информационная система, информационный поиск, поисковая деятельность, пертинентность, релевантность, модель пользователя, когнитивный процесс, интеллектуальная деятельность.

Введение. Развитие компьютерной техники и информационных технологий в значительной степени стимулировало создание и наполнение разнообразной информацией как общие, так и специализированные базы данных, обеспечивая управление ими. Кроме того, появилась возможность обработки не только текстовых материалов, но и различных изображений практически во всех областях деятельности человека. Однако, кроме того, огромные объемы данных делают практически невозможным непосредственную работу пользователя с ними, что, в свою очередь, стимулировало развитие соответствующих поисковых систем, основной целью которых является своевременное и полное обеспечение пользователя необходимыми ему данными.

Особенность поиска информационно-значимых объектов заключается в том, что если поиск текстовых документов осуществляется по нескольким десяткам символов, объединенных в группы – слова, для которых существуют специализированные словари, то поиск информационно-значимых объектов для решения конкретных задач значительно сложнее, поскольку создание запроса на их поиск требует немалых усилий для его формулировки. Поэтому актуальной проблемой, с которой сталкиваются пользователи, является обеспечение надежного, постоянного и полнофункционального доступа к актуальным данным в смысле поиска информационно-значимых объектов. Решение этой проблемы, по нашему мнению, следует начать с построения модели пользователя в системе человек-компьютер.

Общая постановка проблемы.

В системном аспекте пара «человек – компьютер» представляет собой сочетание двух подсистем: психофизических и функциональных особенностей человека и возможностей современной вычислительной техники в смысле поиска нужной информации в распределенных базах данных и в сети Internet. Первая из них, то есть человек, есть в основном непрограммирующий пользователь, но может быть как высококвалифицированным специалистом, хорошо знакомым с классом решаемых задач, методами их решения и подходами и принципами интерпретации полученных результатов, так и обычным пользователем, который, по крайней мере, умеет включить компьютер и выйти в сеть. Вторая – это современные высокопроизводительные компьютеры с

Зайцев Даниил Леонидович – Воронежский институт высоких технологий, аспирант.

Зеленина Анна Николаевна – Воронежский институт высоких технологий, канд. техн. наук, доцент, snakeans@gmail.com.

высокой скоростью обработки информации, огромными объемами памяти, объединенные в сети и построенные на их основе информационно-поисковые системы.

Проблема организации и обеспечения высокой функциональной эффективности информационного поиска в базах, хранилищах и пространствах данных заключается в том, что искомая информация сохраняется в различных формах ее кодирования, созданных в разное время и с разной целью; она сложно структурирована, для различных задач имеет различную информационную ценность и различными пользователями воспринимается по-разному. Зато при высокой надежности и стабильности аппаратного и программного обеспечения вся ответственность за результаты поиска возложена на человеческий фактор в смысле создания поискового запроса и отбора найденного материала. В этом плане объективно оценить эффективность поиска можно только на основании выданных документов.

Исторически, а в определенном смысле и политически (с целью защиты информации) различные источники информации (электронные библиотеки, общие и локальные базы, хранилища, пространства данных) имеют свои особенности по организации форм хранения, поиска, обнаружения, выдачи нужной информации, которые в основном заключаются в видах и тонкостях языков запросов и способов кодирования хранимой информации. Сегодня такой поиск осуществляют специальные поисковые системы, например, Google, Yandex, Rambler, Yahoo. Работа с одним или даже несколькими базами данных практически заключается в правильном формулировании запроса и именно здесь существующая поисковая система помогает найти необходимую информацию. Например, локальные базы данных даже больших предприятий довольно быстро дают информацию о изготовлены изделия, товарах, зарплаты работников и тому подобное. Однако, поиск данных в «чужих» базах данных может стать сложной проблемой. Здесь лучшим примером является поисковая система Google и аналогичные, которые выдают десятки тысяч документов, из которых пользователь выбирает лишь несколько, тратя огромное количество времени на поиск нужных среди предоставленных поисковой системой.

Задача поиска состоит в том, чтобы получить из одного или нескольких информативных источников системно интегрированные наборы с максимальным количеством релевантных и пертитентных информационнозначимых объектов, которые в совокупности обладают признаками полноты, целостности и непротиворечивости. Они фактически подаются в форме адекватной информационной модели проблемной области для ее анализа, обработки и использования в процессах поддержки принятия решений. Как правило, различные источники информации были созданы в разное время и по разным принципам и языках запросов, а главное, по разным профессиональным признакам и онтологиям.

Фактически в задачах поиска документов по информационно-значимыми объектами используют поисковые системы, основная роль в которых принадлежит именно пользователю-составление правильного запроса и отбора релевантных и анализа пертинентных документов, полученных в выдаче.

В общем виде процедура поиска является итеративной процедурой, то есть за этапом выдачи результатов поиска преимущественно осуществляется коррекция запроса и проводится новый поиск уже по исправленному запросу и т. д. Схематично такая процедура показана на рисунке 1.

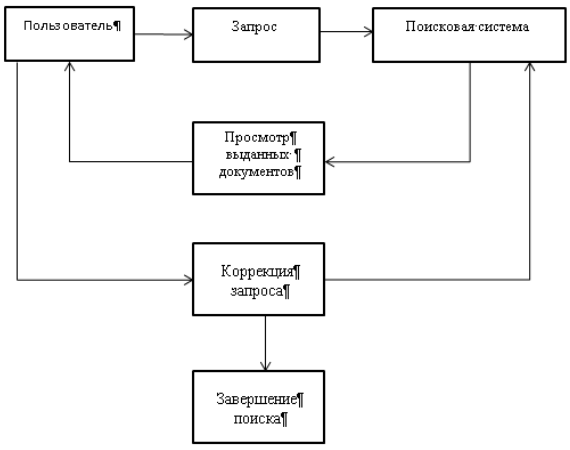

Рисунок 1. Общая схема процедуры поиска

В зависимости от соотношения полноты и точности найденных документов пользователь может сузить или расширить область поиска, перейдя к более общим или, наоборот, более специфических терминов, а также используя родственные понятия. В случае поиска по нескольким терминам такая корректировка области поиска может происходить по одному или по нескольким терминам, что позволяет изменять эту область довольно плавно. Если их

нет в списке выданных документов, область поиска должна быть расширена. Кроме того, оказывается чрезвычайно полезной априорная информация, сохраненная в памяти пользователя как о заведомо релевантных документах с нужными информационно-значимыми объектами, так и о методах решения аналогичных задач.

Процесс поиска требует не только корректной постановки задачи поиска «что надо найти», но прежде всего корректной постановки вопроса «где и как нужно искать». В конце концов возникает логичный вопрос: «найдено ли то, что надо». На основании анализа существующих подходов к оцениванию результатов информационного поиска можно сделать следующие выводы.

1. В теоретическом плане оценивают результат на основании математических моделей информационного поиска. Для этого используют преимущественно теоретико-множественный аппарат (реже вероятностный) и рассматривают отношения множеств релевантных и нерелевантных документов в выдаче их поисковой системой.
2. На практике используют критерии точности и полноты, реже включают и долю нерелевантных документов в выдаче.
3. Отсутствие интегрального критерия оценки результатов информационного поиска.
4. Отсутствие математической модели пользователя, поскольку именно он, в смысле поставленной задачи, определяет качество и соответствие полученных в выдаче документов.

Целью настоящего исследования является моделирование и оценка результатов информационного поиска в смысле математической модели интеллектуальной деятельности пользователя и принятия им решения в отношении документов, полученных в выдаче.

Такой подход должен учитывать полученный результат поиска как выборку документов в выдаче и должен осуществляться исключительно на основании выдачи первого запроса, после чего поиск можно уточнять дополнительными изменениями в запросе, например, дополнениями к ключевым словам, сроков, использованием отдельных фрагментов текста и т. д.

Основное содержание.

Процесс деятельности пользователя как составляющей системы «человек-компьютер» можно формально представить в виде когнитивной модели интеллектуальной работы с материалами в выдаче. Формализация поисковой деятельности человека в определенной степени касается моделирования трудового процесса [8]. Структура интеллектуальной деятельности человека является многоуровневой, адаптивной и существенно зависит от психофизиологии организма, а потому плохо поддается формализации. На общем уровне деятельность пользователя можно представить в виде последовательности процессов информационного поиска, восприятия информации, анализа информации и принятия решения. В этом аспекте деятельность пользователя является информационно-аналитической деятельностью, направленной на решение справочных и информационно аналитических задач.

Довольно часто перед пользователем стоит задача Z, решение которой требует конкретных данных, которые отсутствуют в ней. В таких случаях пользователь обращается к поисковым системам с соответствующим запросом для выявления и получения документов (материалов), которые содержат данные, необходимые для решения его задачи. Очевидно, организация такого поиска, как и выбор нужных информационно-значимых объектов, полученных в документах выдачи и принятия решения, а точнее, формулировка решения, фактически может быть отнесена к интеллектуальной деятельности пользователя. Структуру такой деятельности можно представить схемой, изображенной на рисунке 2, которая содержит следующие три составляющие: формулирование запроса, анализ выданных материалов и принятия решения.

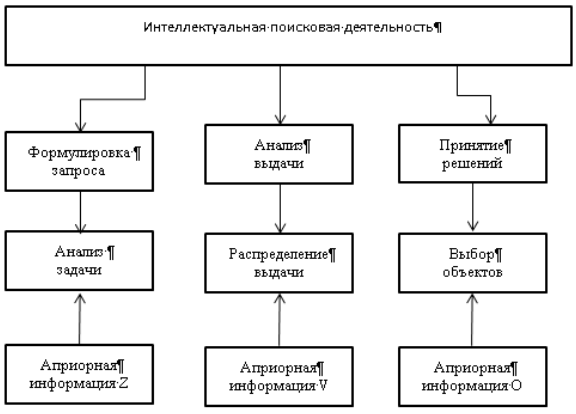

Рисунок 2. Схема интеллектуальной поисковой деятельности пользователя

Формулировка запроса, анализ выдачи и принятия решения об отобранных ин-

формационно значимых объектах для решения задачи фактически являются когнитивными процессами, в значительной степени связанными с априорными данными, хранимыми в памяти пользователя. Априорные информации A (Z), A (V), A (O) отличатся между собой, поскольку соответствуют различным процессам. Эти процессы можно формально представить следующим образом. Пусть для решения задачи Z={z_i : i=1,2,…q}, которая декомпозируется на z_q подзадач нужно использовать m≥q информационно-значимых объектов (констант, формул, конкретных величин, моделей, методов и т.д.) Пользователь представляет себе способ решения, однако ему также нужны объекты, находящиеся в соответствующих документах в библиотеках и базах данных. Названия таких документов зачастую не соответствуют названиям нужных объектов, а поисковые системы преимущественно выдают именно названия документов. Поэтому перед пользователем стоит задача: так сформулировать запрос, чтобы получить в выдаче как можно больше документов с информационными объектами. Итак, если в выдаче присутствуют релевантные Р, пертинентные П и нерелевантные Н объекты, то запрос должен обеспечить условие (Р + П) >> Н.

В процессе формирования запроса пользователь использует априорную информацию A(Z) в виде сохранённой в его памяти эталонной модели E(D_z), где D_z - документы с информационными объектами по этой задачи. В результате когнитивного анализа ξ он формирует множество ключевых слов K_z , т.е существует выражение:

$$\xi: E(D_z) \bigcap A(Z) \rightarrow Kz$$

Результатом реализации запроса является выдача V (D) - выборка произвольных документов, содержащих релевантные Р, пертинентные П и нерелевантные Н объекты. Пользователь, анализируя выданные документы, использует также априорную информацию A (V), которую ассоциирует с объектами в выдаче и разделяет их на классы Р, П, Н, то есть возникает отображение:

$$\chi : V(D) \bigcap A(V) \rightarrow \Omega,\ \Omega = \{\Omega_P,\ \Omega_П,\ \Omega_Н \mid \Omega_P \bigcup \Omega_П \bigcup \Omega_Н = \Omega,\ \Omega_P \bigcup \Omega_П \bigcup \Omega_Н = \varnothing\}.$$

Используя априорную информацию A (O) относительно необходимых информационных объектов, пользователь формирует их концептуальную модель в виде информационно-значимых объектов $w_q \in (\Omega_P \bigcup \Omega_П)$,, т.е выбор информационно-значимых объектов для решения задачи можно подать отображением

$$ύ:(\Omega_P \bigcap \Omega_П) \bigcup A(O) \rightarrow R_Z(\omega_q)$$

Деятельность пользователя по поиску нужных материалов информационно-значимых объектов начинается с анализа поставленной перед ним задачи. Очевидно, решение такой задачи прежде всего требует данных о способах решения аналогичных задач, выяснения, какие данные, критерии, условия, требования были для них использованы и дали полученные решения. Кроме того, пользователь, опираясь на собственные знания и опыт, заключает в своём воображении пути ее решения и, главное, пытается выявить, какой информации ему не хватает и как ее компактно сформулировать, чтобы увеличить вероятность получения релевантной информации для доступных источников. Собственно результат такой умственной деятельности используется им для построения запросов для различных баз данных. Представим множество документов ω, выданных пользователю в результате запроса $\xi_m \in Z$, где Z множество сделанных из m запросов, некоторым множеством Ω, т.е $\omega \in \Omega$. Среди документов ω содержатся, возможно, нужные пользователю. Разобьём это множество на несколько непересекающихся классов k = 1,2,…K г де K – множество классов, тогда:

$$\Omega = \{\Omega_k : \bigcup_{k=1}^{K} = \Omega, \bigcap_{k=1}^{K} = \Omega_k = \emptyset, k = \overline{1,K}\}$$

Определим классы следующим способом: k_1 – пертинентные, k_2 – ревалентные, k_3 – неревалентные. Тогда любой из выданных объектов ω_j^k будет принадлежать своему подмножеству – классу, т.е $\omega_j^{k_i} \in \Omega_{k_i}$ где $j=\overline{1,J_K}$, J_K- количество объектов в классе Ω_k. Объекты ω_j^k представлены ключевыми словами и определенными требованиями в запросе.

Фактически каждый объект в этом классе документов можно описать некоторым подмножеством $A_k^* \subseteq A_k$ признак ключевых слов

$$A_k = \left\{ a_k^i : \bigcup_i a_k^i = A_k, i = 1,2,..n \right\},$$

где A_k – множество всех ключевых слов в этом классе, а в некоторых ситуациях выданный объект может быть отнесен к своему классу и при отсутствии у него некоторых из них, то есть по набору признаков, которые являются булеан $B(A_k)$ – множества A_k. Признаки и их комбинации, которые входят в этот набор, определяют решение об отнесении выданного объекта к его классу распознавания и обеспечивают своей информативностью эффективность поиска.

При небольшом количестве четко выраженных информативных признаков определяется класс выданного объекта практически мгновенно, поскольку такой объект воспринимается целостно и совпадает со своим образом в воображении. Однако в большинстве случаев, даже когда принадлежность объекта данному классу очевидна, могут возникать определенные сомнения относительно его релевантности или пертинентности. Например, неизвестный автор, время и место публикации, авторитет издания и тому подобное.

Пользователь как динамическая система.

Деятельность пользователя в таких случаях требует значительного умственного напряжения, усиленной работы памяти и зрительного анализатора, причем при минимальной подвижности, монотонного режима работы. В результате прогрессирует психическое напряжение, снижается концентрация внимания, растет усталость, а затем снижается качество работы вследствие ошибок, повторного обработки материалов, увеличивается продолжительность поиска.

Поэтому в организации поисковой деятельности пользователя основной составляющей просмотра выданных поисковой системой материалов является когнитивный процесс.

Обозначим через $\bar{g}(V)$ вектор названий документов в выдаче, соответствующий ключевым словам в запросе, а через $\bar{g}(Z)$ вектор ключевых слов в запросе.

Тогда построить когнитивную модель поисковой деятельности пользователя можно, рассматривая пользователя как сложную динамическую систему S, в смысле теоретико-множественного подхода, приведенного в [10] и представленную отношением на множествах: X – объектов в выдаче и Y - отобранных пользователем.

Система S является функциональной, если каждому элементу множества Y однозначно отвечает единственный элемент множества X, то есть отношение S является функцией

$$S: X \to Y$$

Множество X полученных документов является областью ее определения

$$D(S)=\{x:(\exists y)((x,y) \in S)\}=X$$

а множество принятых решений Y является областью значений этой системы

$$R(S)=\{y:(\exists x)((x,y) \in S)\}=Y$$

В результате предварительного просмотра выданных материалов функцию S будет положительно определена только для некоторых документов множества X, то есть релевантных, а при повторных просмотрах к ним могут быть привлечены и пертинентные. Очевидно, функция будет отрицательно определена для всех остальных - не релевантных, отвергнутых документов.

Если система S является динамической системой, то для нее существует произвольное множество C такая, что функция R реализует некоторое отображение $R:(C \times X) \to Y$, для которого существует условие

$$(x,y) \in S \Leftrightarrow (\exists c)[R(c,x)=y]$$

Если это условие выполняется, тогда множество C является множеством состояний системы – уровнями функционального состояния пользователя. Собственно уровни функционального состояния пользователя определяются изменениями его нормального рабочего состояния, которые могут быть обусловлены снижением психомоторных функций и концентрации внимания, дискомфортом рабочей среды и внешними воздействиями (шумом, звуком, отвлечением и т. д.)

Поскольку процесс анализа и отбора выданных документов реализуется во времени, то, в принципе, в работе пользователя можно выделить следующие два момента:

- во-первых, функциональное состояние C пользователя на момент времени t

соответствует некоторому уровню C_t, и относительно документа X_t он принимает соответствующее решение Y_t, которое можно представить таким семейством отображений

$$\bar{p}=\{p_t: C_t \times X_t \rightarrow Y_t \ \& \ t \in \mathrm{T}\};$$

- во-вторых, смысл, содержание, форма, принадлежность документа X_t могут изменить уровень его функционального состояния C_t а также значительно увеличить срок принятия решения Y_t, что соответствует такому отображению

$$\bar{\varphi} = \{\varphi_{tt'}: C_t \times X_t \rightarrow C_{t'} \& \ t, t' \in \mathrm{T} \ \& \ t < t'\}$$

Итак, формально деятельность пользователя в течение некоторого времени T можно представить моделью, которая содержит последние два отражения $\bar{p}$ и $\bar{\varphi}$.

Когнитивная модель обработки полученных документов пользователем. Пусть V – множество выданных поисковой системой документов по сделанным пользователем запросу Z. Пользователь в результате просмотра этих документов разделяет их на релевантные V_r, неревалентные V_n и пертинентные V_p. В результате полученные по выдаче документы разделены на три класса, то есть множество выданных документов $V=V_r \cup V_n \cup V_p$ состоит из трёх подмножеств.

На «вход» пользователя с использованной информационно-поисковой системы поступает выдача V – множество выданных документов, согласно сделанного им запроса Z. В результате побочного просмотра часть L□V документов сразу можно отбросить как релевантную для этой задачи. Освобождение от заведомо ненужных документов обусловлено существующей в памяти пользователя эталонной модели относительно решения задач такого типа, содержит априорные знания о тех или других документах, которые должны помочь решить поставленную задачу. Таким образом, пользователь будет работать с множеством документов X=V\L. К этой множества относятся все документы множества $X=\{x_i : x_i = x(t_i), i = 1,2,\dots,m \ \& m = |X|\}$, причем связь со временем указывает на последовательность работы с документами, чем подчеркивается рассмотрение только в смысле теории динамических систем, а m - реальная мощность множества. Множество X фактически является информационной моделью M(X(x(t_i)) на «входы» пользователя. Отметим, что в этом исследовании представление документа как x_i означает конкретный документ в множестве X, а представление документа как $x(t_i)$ означает работу с документом в момент времени t_i, т. е возможное повторное обращение к уже просмотренному документу, и определяет реальную мощность m множества X. Иначе говоря, информационная модель допускает повторный просмотр документов, то есть повторение документа в множестве X, ведь количество элементов в модели определяется количеством моментов времени работы с документом. Эталонная модель содержит названия документов, сроки, ключевые слова, на основании которых разработан запрос. Она в информативном плане содержит больше данных и знаний, чем их содержит запрос, а потому значительно шире, чем полученная информационная модель. Эталонная модель E(G(g(t_i))) содержит названия известных пользователю документов, содержание многих документов, сроков, способов формулировка ключевых слов, возможные пути и способы решения поставленной задачи.

Имея на «входе» информационную модель M(X(x(t_i)) и используя свою эталонную модель E(G(g(t_i))), пользователь для выработки решения по каждому объекту выдачи строит в своем воображении когнитивную модель для выбора соответствующего решения из множества альтернативных решений

$Y=\{ \quad y_j : y(t_j), j = 1,2,\dots,r \ \& r = |Y| \ \& \ r \leq m\}$.

Множество альтернативных решений содержит следующее:

- документ является релевантным, соответствует сделанному запросу;
- документ является пертинентним, хотя и не соответствует ключевым словам, сформулированным в запросе;
- его беглый просмотр указывает на то, что он содержит необходимые для решения поставленной задачи данные;
- документ является не релевантным, поскольку его просмотр не дал ничего нового, что может быть использовано для решения задачи;
- документ уже с первого взгляда отвергается как таковой, не соответствует решаемой задаче и не стоит того, чтобы его подробнее обрабатывать;
- необходимо модифицировать и повторить поисковый запрос;
- изменить поисковую систему.

Когнитивная модель при таких альтернативных решениях соответствует пересечению моделей M(X(x(t_i)) и E(G(g(t_i))), т. е. имеем K(M(X(x(t_i)))⋂ E(G(g(t_i))).

Эти три модели фактически отражают интеллектуальную деятельность пользователя в процессе просмотра, анализа и отбора документов, полученных в выдаче по конкретному запросу.

С формальной точки зрения такую деятельность пользователя можно подать системой отображений:

$$\begin{cases} \alpha: Z(A_k) \times \mathrm{B} \times \Pi \times \mathrm{T} \to \mathrm{V} \\ \eta: V \backslash \Pi \times \mathrm{T} \to \mathrm{R} \cup \mathrm{P} = \mathrm{X} \\ \mu: X \to \mathrm{M}(\mathrm{X}(\mathrm{x}(t_i))); \\ k: \mathrm{M}(\mathrm{X}(\mathrm{x}(t_i)) \times \mathrm{E}(\mathrm{G}(\mathrm{g}(t_i))) \times \mathrm{C}(\delta_q, t_q) \times \mathrm{T} \to \mathrm{K}(\mathrm{M}(\mathrm{X}(\mathrm{x}(t_i))) \cap \mathrm{E}(\mathrm{G}(\mathrm{g}(t_i))) \\ p: K \times \mathrm{Y} \times \mathrm{C} \times \mathrm{T} \to \mathrm{R} \end{cases} \quad (1)$$

где α – отображение выявления поисковой системой в библиотеках, базах данных и других источниках документов по сделанному пользователем запросу; η – отображение визуализации названий и фрагментов аннотаций документов, найденных поисковой системой, которые в результате поверхностного просмотра уже на этом этапе могут быть приняты как релевантные или нерелевантные. Последние могут быть отнесены к поисковому «шуму» P, поскольку вероятно они попали в выдачу из-за несовершенства «понимания» запроса поисковой системой. Такие документы исключают из множества X вследствии разницы $V \backslash \Pi$; $\mu -$ отображение восприятия полученной выдачи пользователем, то есть формирование информационной модели на «входе» пользователя; k – отображение построения когнитивной модели интеллектуальной деятельности пользователя по обработке полученных в выдаче документов путем взаимодействия информационной модели M(X(x(t_i)) и эталонной модели E(G(g(t_i))), которая относительно решённой задачи сформирована в его памяти. Эта модель создается в результате подробного и аналитического просмотра и отбора документов, полученных в выдаче; p - отображение выбора и принятия в отношении каждого элемента решения R.

Итак, поисковую деятельность пользователя, то есть его взаимодействие с поисковой системой, результатом которой является создание выборки релевантных документов, можно представить моделью - системой отражений. Результат выдачи может удовлетворить пользователя, но может поставить перед ним новые задачи: изменение или уточнение ключевых слов, расширение запроса, изменение поисковой системы и других. Как правило, пользователь осуществляет несколько поисковых итераций, часто с привлечением различных информационных источников. В последнем случае поиск проводят на конкретных базах данных, специализированных электронных библиотеках и на соответствующих сайтах.

Соответствие выдачи запроса. Самым сложным моментом в оценке эффективности любого информационного поиска является установление соответствия между найденным и выданным документами и документами, а точнее, поисковыми признаками документов, представленных в заявке. Дело в том, что решение о соответствии (то есть являются релевантными выданные документы или нет) является весьма субъективным. Кроме того, если можно точно ответить, документ является релевантным, или нерелевантным, то четко указать, документ пертинентним, нельзя, поскольку он может быть пертинентным в разной степени. Из содержания понятий релевантности и пертинентности следует, что оценка эффективности поиска имеет принципиальные две составляющие. Напомним, что понятие релевантности означает соответствие информационного поиска сделанного пользователем запроса, а пертинентнисть – соответствие информационной потребности пользователя.

Первая из них – это оценка, а точнее, понимание поисковой системой составленного пользователем запроса. В этом аспекте информационно-поисковая система отбирает те документы, признаки которых указаны в запросе. Очевидно, что в таком случае семантический анализ выявленных документов в базе или хранилище данных, в файлах или библиотеках не производится, а только сопоставляются признаки выявленных документов, и при условии полного

или частичного совпадения документы подаются в выдаче.

Вторая составляющая – это оценка документов в выдаче, полученных пользователем, в результате информационного поиска. Здесь пользователь разделяет документы на три группы: релевантные (Р), пертинентные (П) и нерелевантные (Н). Документы в выдаче, как правило, сортируются информационно-поисковой системой по определенным критериям: по дате (собственная дата документа или последняя дата обращения к нему), по рейтингу пользования (сколько раз этот документ фигурировал в запросах различных пользователей в целом или за определенный период). Возможны и другие критерии, например, по объему или дате последнего обращения и др. Получив выдачу, то есть перечень найденных документов, пользователь последовательно или выборочно знакомится с документами, отбирая релевантные и пертинентные и отвергая нерелевантные. Последовательность релевантных, пертинентных и нерелевантных документов в каждой конкретной выдаче практически всегда случайна. Проверить этот факт можно экспериментально следующим образом.

Организация экспериментального исследования. В поиске необходимой информации для проведения научных исследований кроме отбора пертинентных документов фиксировались релевантные и нерелевантные. Содержание экспериментального исследования следующее.

Отбор ключевых слов. Для этого были сформулированы следующие ключевые слова, а точнее, словосочетания: информационный поиск; модели информационного поиска; информационно-поисковая система; эффективность информационно-поисковых систем; оценки эффективности информационного поиска.

Уточнение понятий. *Пертинентные документы* - это документы, которые по содержанию максимально отвечают потребности пользователя, хотя их названия и аннотации не содержат указанных в запросе ключевых слов и имеют все реквизиты для ссылки на них, то есть документы, являющиеся электронными копиями бумажных: монографий, статей в научных журналах и сборниках трудов, тезисах и трудах научных форумов и статьи, представленные в энциклопедиях и справочниках.

Релевантные документы – это те, которые по содержанию и ключевым словам полностью соответствуют потребности пользователя и имеют реквизиты своих бумажных оригиналов. Ими могут быть и фрагменты различных материалов, но тогда для ссылки на них надо использовать адрес их электронной почты. В некоторых случаях для выявления этого документа необходимо провести дополнительно еще и отдельный специальный поиск, причем результат не гарантируется. Все другие документы признаются нерелевантными.

Содержание эксперимента. Для экспериментальных исследований использованы информационно-поисковые системы Google, Яndex, Rambler, Yahoo, которые по ключевым словам выдают веб-страницы найденных документов. Настройки поиска обеспечила оптимальный вариант выдачи результата – 10 электронных документов на каждой странице. На основе предыдущих результатов поиска и по собственному опыту известно, что нужная информация по этому вопросу находится преимущественно на первых пяти страницах. Поэтому для экспериментов выбрано ограничение 5 полных страниц, то есть объем выдачи за каждым ключевым словом составлял 50 документов.

Для каждой страницы в результате пересмотра в каждом из десяти приведенных документов присваивались индексы Р, П, Н.

Задача пользователя заключается в том, чтобы среди этого множества выбрать именно те, которые соответствуют или способствуют решению его задачи. Очевидно, при любом поиске просмотр полученных документов будет аналогичным. Поскольку предоставленая выборка является конечной, можем оценить эффективность поисковой системы отношением благоприятных событий ко всем возможным, то есть отношением, например, количества релевантных документов в количестве всех предоставленных документов, полученных по этому запросу. Если документы классифицировать как в этом примере, то можно получить три частоты появления документов каждого класса.

$$_{\text{Р}} = \frac{1}{N}\sum_{i=1}^{n} P_i, \;\; _{\text{П}} = \frac{1}{N}\sum_{k=1}^{n} \text{П}_i, \;\; _{H} = \\ = \frac{1}{N}\sum_{j=1}^{m} H_i \; (1)$$

где N – количество документов, выданных конкретной поисковой системой, а величины l,m,n указывают на количество документов, и Н, П, Р соответственно.

На практике, как правило, информационный поиск осуществляется по разным запросам в зависимости от поставленных задач. В свою очередь, задачи могут касаться различных предметных областей, объема их онтологий, специфики конкретных объектов, требующих решения. С другой стороны, можно быть уверенным в том, что информационные источники имеют всю необходимую информацию по любой области знаний и деятельности человека. Поэтому количество предоставляемых пользователям документов является разным. Обычно поиск в источниках информации осуществляется поисковой системой, которая работает по определенному алгоритму и определенным формальным критериям соответствия, а потому можно предположить, что результаты различных поисков в одном и том же источнике информации будут статистически однородны, то есть иметь определенные статистические закономерности, которые могут отразиться по крайней мере на соотношении частот рассмотренных выше классов.

Результаты экспериментального исследования. Последовательность документов в выдаче можно изобразить графически в виде диаграммы, приведенной на рисунке 3.

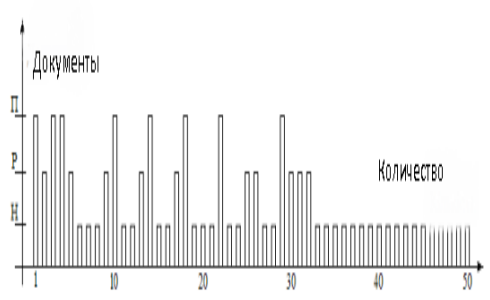

Рисунок 3. Распределение документов в выдаче: П – пертинентные, Р – релевантные, Н – нерелевантные

Как количественная оценка использована относительная частота появления того или иного вида документов. Для каждой поисковой системы результат определяем по соотношениям (1): релевантных $f_{\text{Р}}$, пертинентных $f_{\text{П}}$, нерелевантныхf_{H}.

Очевидно, что все эти значения в значительной степени зависят от объема документов в информационной системе (базе, хранилище данных, папках с файлами, библиотеке), возможностей информационно-поисковой системы, формы запроса, а также от информационной потребности пользователя – насколько глубоко он понимает задачу, для решения которой он осуществляет этот поиск.

Оценка эффективности информационного поиска. Сформированные практически соответствующими поисковыми языками, присущими тому или иному информационному фонду, запросы должны иметь довольно ограниченное количество поисковых признаков - ключевых слов, определенного типа расширений и объяснений или ограничений. Алгоритмы информационно-поисковых систем, используя эти данные в процессе сканирования-поиска существующего каталога, преимущественно используют как данные: имя, название и аннотацию документов, хотя возможно и сканирование самого документа. Поскольку ключевые слова в зависимости от контекста могут иметь несколько значений в выдаче могут попадать абсолютно нерелевантные документы.

В общем оценка эффективности основывается, как было сказано выше, на оценках точности и полноты. Попытка использовать дополнительные показатели поиска требует учета не только объема самого информационного фонда, но и объема релевантных и нерелевантных относительно данного запроса документов. Получить такие данные практически невозможно, так как для одной задачи документы могут быть релевантными, а для второй уже нет. С другой стороны, если знать все релевантные документы в фонде, то можно осуществить поиск только для них, и тогда в выдаче будут только релевантные документы, а это осуществить практически невозможно, по крайней мере по двум причинам: никто не сможет из многотысячного информационного фонда отбирать релевантные для данной задачи отдельного пользователя документы; а также присутствие конфиденциальной информации и отсутствие информации о самом фонде, за исключением лишь общих его характеристик. Поэтому наиболее правомерно оценивание эффективности поиска по его результатам, то есть на основе документов, которые являются в выдаче.

При наличии трех типов документов оценить эффективность информационного поиска можно так. Очевидно, что пертинентные документы имеют наибольшую

ценность для пользователя, поскольку могут содержать нужные новые, неизвестные или забытые информационные объекты.

Для своей задачи пользователь, как правило, использует только релевантные и отобранные пертинентные документы, то есть эффективной выдачей считается сумма $E_{\text{пои}}$=П+Р, а потому в пределах объема выдачи эффективность поиска можно определять как отношение.

$$E_{\text{пои}}=\frac{\text{П+Р}}{\text{П+Р+Н}} \qquad (2)$$

Учитывая особенности формы запроса, которая тесно связана с конкретной информационной системой, то есть с ее информационным фондом и его системой индексирования, необходимо ввести некоторый корректирующий множитель – коэффициент пропорциональности β, в результате получим

$$E_{\text{пои}}=\beta\frac{\text{П+Р}}{\text{П+Р+Н}} \qquad (3)$$

Оценка эффективности поиска в виде (3) характеризует осуществлен ли информационный поиск в конкретной системе, для конкретного ключевого слова и по результатам полученной выдачи, ограниченной N=Р+П+Н документами. Иначе говоря, показатель (3) характеризует предоставления преимущества источнику (системе поиска) относительно решения задачи пользователя.

Определить показатель β можно только на основании полученных в выдаче данных двумя и более поисковыми системами. Для этого необходимо:

1. Четко сформировать множество ключевых слов

2. Определить информационно-поисковые системы, которыми будет осуществляться поиск.

3. Найти для каждой поисковой системы среднее значение $E_{\text{пои}}$ по всем ключевым словам.

4. Вычислить сумму значений $E_{\text{пои}}$ всех использованных поисковых систем.

5. Разделить усредненные показатели эффективности для каждой системы на эту сумму.

Значение этого показателя для пяти поисковых систем приведены в таблице.

Очевидным является факт: чем больше объем информационного фонда, тем больше релевантных документов будет найдено. Однако, здесь надо иметь в виду и популярность или развитость этой тематики, поскольку именно ее популярность и востребованность определяют объем документов в фонде. Поэтому значение показателя b со временем меняется. Выражения (2) или (3) дают объективную оценку эффективности информационного поиска, но только при условии, что в выдаче будут присутствовать также и нерелевантные документы - по крайней мере хотя бы один.

Таблица 1

Оценка эффективности информационно-поисковых систем по количеству релевантных и пертинентных документов

№	Ключевые слова	Google	Yandex	Rambler	Yahoo
1	Оценка эффективности информационного поиска	0,42	0,46	0,22	0,34
2	Информационный поиск	0,50	0,32	0,36	0,36
3	Модель информационного поиска	0,34	0,44	0,22	0,46
4	Информационно-поисковая система	0,46	0,44	0,10	0,26
5	Эффективность информационно-поисковых систем	0,36	0,52	0,34	0,54
6	Усредненный показатель эффективности	0,41	0,42	0,24	0,39
7	Показатель отношения к системе β	0,212	0,218		0,202

Выражения (2) и (3) дают объективную оценку эффективности информационного поиска, но в первых (Левых) вариантах, только при условии, что в выдаче будут присутствовать и нерелевантные документы - по крайней мере, хотя бы один. Невыполнение этого условия означает деление на ноль. То есть, будет неправильный результат оценки. Такая ситуация может возникнуть тогда, когда количество релевантных документов в информационном фонде превышает объем выдачи.

В этом случае оценивается эффективность за выдачей для одного или нескольких запросов. если такой показатель использовать для каждого из нескольких запросов, но таких, которые касаются конкретной темы, можно оценить качество и самого запроса, точно установить, какой из

запросов или какие ключевые слова являются наиболее эффективными и уже по ним модифицировать следующие запросы.

Выводы и перспективы дальнейших научных исследований. Эффективность информационного поиска в разных системах хранения информации в смысле построения интегрального показателя практически нельзя определить, поскольку кроме двух показателей - полноты и точности - все остальные требуют знания количества релевантных и нерелевантных документов в этом информационном фонде по этой задаче. Получить такие данные для больших по объему фондов невозможно, поскольку: во-первых, осуществить такой подсчет означает пересмотр каждого документа, во-вторых, в больших базах данных переход от нежелательных документов к релевантным практически по любому запросу является нечетким и размытым, в-третьих – для разных задач понятие релевантности документов различается. Самым простым способом построения оценки эффективности поиска является использование логического подхода, который заключается в представлении отношением – количества нужных заказанных документов к количеству документов в выдаче, которые не соответствуют требованиям пользователя. На эффективность поиска влияет не только наличие в информационном фонде необходимых документов, но и правильность построения самого запроса согласно требованиям поисковой системы. Приведенный пример оценки эффективности информационного поиска демонстрирует правомерность использования найденных и выданных документов на пертинентные, релевантные и нерелевантные. В результате такого разделения оценку эффективности можно представить как усредненную, или суммарную, по результатам проведения информационного поиска в одном или в нескольких информационных фондах и на разных поисковых системах по одному набору ключевых слов.

Разработанный подход к построению оценки информационного поиска имеет практическое значение, поскольку полученные количественные значения локальных оценок дают основания для оптимизации набора ключевых слов и определения наиболее подходящих информационных фондов и поисковых систем.

ЛИТЕРАТУРА

1. Агеев М. Официальные метрики РОМИП 2010/ М. Агеев, И. Кураленок, Некрестьянов // Российский семинар по оценке методов информационного поиска: Труды РОМИП, 2010. (Казань, 15 октября 2010 г.) Казань, 2010. С. 172-187.
2. Целых А. Н. Оценка эффективности информационного поиска / А. Н. Целых, Э. М. Котов // Известия ТРТУ. Тематический выпуск «Управление в математических системах». – Таганрог: Изд-во ТРТУ. – 2006. – № 10 (65). – С. 43–45.
3. Яхина Е. П. Методы оценки информационных систем / Е. П. Яхина // В мире научных открытий. – 2010. – № 3 (09). – Ч. 1. – С. 63-66.
4. Попов С. В. Оценка функциональной эффективности систем текстового поиска на примере поиска патентных документов / С. В. Попов // Патентная информация сегодня. – 2010. – № 1. – С. 22-25.
5. Козлов Д. Д. Информационно-поисковые системы в Internet: текущее состояние и пути развития [Електронний ресурс] / Д. Д. Козлов // Информационно-поисковые системы в Internet: текущее состояние и пути развития. Технологический обзор. – М. 2000. – [28 с.]. – Режим доступу: http://lvk.cs.msu.su/~ddk/ir_and_ia_ review.pdf.
6. Тявкин И. В. Математическая модель информационного поиска и оценка єффективности поисковой системы / И. В. Тявкин, В. М. Тютюнник // Вестник ТГТУ. – 2008. – Т. 14. – № 3. – С. 478-481.
7. Козлов М. В. Метод оценки эффективности функционирования современных информационно-поисковых систем Интернета [Електронний ресурс] // М. В. Козлов, В. А. Яцко. – Режим доступу: – http://www.dialog-21.ru/dialog2006/materials/html/Kozlov.htm.
8. Кирхар Н. В. Модели деятельности пользователя компьютеризованной системы / Н. В. Кирхар, Д. В. Ходаков // Вестник ХНТУ № 4(27), 2007. – С. 370–378.
9. Багаев Д. В. Разработка системной модели технического объекта [Електронний ресурс]. – Режим доступу: http://systech.miem.edu.ru/ 2.doc.

10. Месарович М. Общая теория систем: математические основы / М. Месарович, Я. Такахара; под ред. С. В. Емельянова. – М.: Мир, 1978. – 312 с.

FORMAL PRESENTATION OF USERS 'ACTIVITIES WITH IDENTIFICATION OF INFORMATION-SIGNIFICANT OBJECTS

© 2021 D. L. Zaitsev and A. N. Zelenina

Voronezh Institute of High Technologies (Voronezh, Russia)

The activity of the user on the search for information-significant objects for solving the task assigned to him is considered. To formalize this activity, the apparatus of set theory was used. Mathematical models of the formulation of the request, the analysis of documents in the issue and the selection of information objects for solving the problem posed to the user are given. In addition, models have been developed in the aspect of the user's intellectual activity and cognitive processes. The results of experimental research with such search engines as Google, Yandex, Rambler, Yahoo are presented. To assess the effectiveness of information retrieval, the found documents are divided into pertinent, relevant and irrelevant. Search efficiency is determined by the ratio of the number of pertinent and relevant documents to the number of all documents in the search results. It is proposed to take into account the information characteristics of search engines with an appropriate coefficient.

Keywords: information system, information search, search activity, pertinence, relevance, user model, cognitive process, intellectual activity.