О моделировании интерактивного итеративного процесса поиска документальной информации в документальных базах данных

© Максимов Н.В. Свириденко С.В.

Российский Государственный Гуманитарный Университет
svsvirid@mail.ru

Аннотация

В докладе рассматривается подход к повышению эффективности поиска документальной информации в документальных базах данных на основе моделирования и оптимизации интерфейса пользователя.

Существует гипотеза, что по первым действиям пользователя при работе с системой поиска можно предвосхитить реакцию системы и с некоторой долей вероятности определить дальнейшие действия пользователя по работе с системой. Подтверждение гипотезы позволяет оптимизировать интерфейс пользователя в системах поиска документальной информации таким образом, чтобы в каждый момент времени перед пользователем предъявлялось множество объектов и средств, необходимых и достаточных для эффективного продолжения процесса поиска на данный момент для данного конкретного пользователя с его конкретной поисковой потребностью.

В рамках проверки гипотезы предлагается построить модель интерактивного итеративного процесса поиска документальной информации в документальной базе данных, определить средства представления модели и ее анализа.

Предполагается, что исследование динамики модели процесса поиска документальной информации с варьированием уровня атомарности модели дает возможность просчитать вероятность достижимости того или иного состояния процесса поиска.

1 Концептуальная модель

1.1 Этапы поиска

В упрощенном виде процесс поиска можно разбить на 4 последовательных этапа:

1. формулирование пользователем поисковой потребности;
2. перевод поисковой потребности с естественного языка на язык системы;
3. собственно поиск (срабатывание внутрисистемных механизмов поиска);
4. просмотр и оценка пользователем результатов поиска;

Каждому этапу сопоставимо определенное множество объектов. Множество объектов, задействованных в системе, заранее определено и конечно. Каждому объекту, соответственно, можно сопоставить набор атрибутов и возможных действий с ними.

К примеру, для второго этапа можно выделить такие технологические объекты, как термин, поле документа базы данных, оператор, средства маскирования, рубрикатор, словник, тезаурус и т.д. Объекты, соответствующие этапам процесса поиска, описаны в Таблице 1.

Таблица 1 Объекты процесса поиска

	Объект	Атрибут	Действия над объектом
1			
2	Логическое выражение	- средства формирования (строка редактирования, конструктор запроса, QBE), - количество терминов в запросе	очистить, редактировать
	Термин	- частота встречаемости в БД	добавить, убавить,
	Поле		изменить
	Логический оператор	- вид оператора - порядок группировки слов (скобки)	
	Контекстный оператор	- вид оператора	
	Средства маскирования	- вид (маскирование вручную, автомаскирование)	
	Словарь		
	Тезаурус		
	Рубрикатор		
3	Множество выданных документов	- количество документов в выдаче	сортировать
4	Подмножество документов		сортировать, эвристический поиск, вывод в файл, статистика
	Документ	состояние (просмотрен, релевантен, не отмечен),	аналог, вывод в файл/на печать, перейти к следующему, литеральный поиск
	Средства обратной связи		

На первом этапе процесса поиска используются абстрактные объекты, такие как поисковая потребность и стратегия поиска, практически не поддающиеся формализации. В контексте системы

поиска эти объекты косвенно влияют на процесс поиска.

Целесообразно построить схему взаимодействия информационных объектов поиска (рис.1), что поможет определить возможные их сочетания.

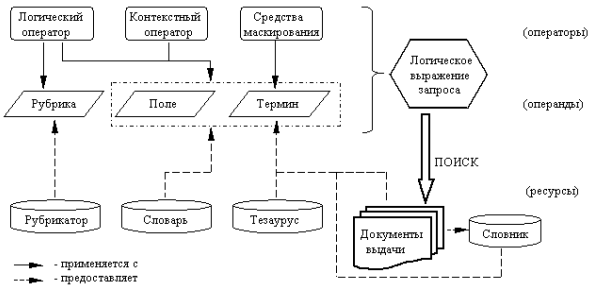


Рис. 1 Схема взаимодействия информационных объектов процесса поиска

1.2 Факторы, влияющие на ход процесса поиска

На выходе каждого этапа формируется четкий результат. Для каждого такого результата можно определить ряд признаков, по которым можно судить о модели поведения пользователя и, соответственно, предлагать средства (возможные действия над объектами этого же или следующего этапа), наиболее эффективные в данной точке процесса для этого пользователя. Под моделью поведения пользователя подразумевается набор особенностей поведения пользователя в системе поиска, таких как выбор стратегии поиска, частота использования определенных функций системы и др., относящий его к определенному типу пользователей. перечисленных и статистически обоснованных в [1].

Например, открыв базу данных по истории и философии, пользователь в строке редактирования поискового предписания набирает «история философия». Очевидно, что документов, содержащих данные термины, в базе данных по истории и философии будет намного больше, чем в состоянии просмотреть пользователь. Точность выдачи после третьего этапа минимальна. Существует возможность на основании количества терминов, частоты их встречаемости в базе данных и операторов, их соединяющих уже в конце второго этапа определить, что выдача будет непомерно велика, и предложить пользователю уточнить запрос, предоставив для выбора множество рубрик базы данных, статьи тезауруса или словник.

Таким же образом по факту применения или неприменения средств обратной связи пользователем на четвертом этапе можно определить намерение пользователя впоследствии проанализировать результат и продолжить поиск, что дает основания для отображения системой соответствующих средств.

Предположительные наборы признаков для каждого этапа процесса поиска, способствующие адаптации интерфейса системы в соответствии с моделью поведения пользователя, представлены в Таблице 2.

Таблица2. Свойства выходов этапа, влияющие на формирование системой модели поведения пользователя

№ этапа	Факторы
1	
2	Метод построения запроса (строка редактирования, конструктор запроса, QBE) Количество терминов в запросе Частота встречаемости этих терминов в БД Количество и имена используемых в запросе полей Факт пользования рубрикатором Факт использования логических операторов, Факт использования контекстных операторов, Факт использования средств маскирования
3	Количество выданных документов
4	Количество просмотренных документов и их порядок Факт обратной связи Соотношение релевантные/выданные Факт наличия новых выданных и новых релевантных в данном запросе Использование средств статистики

2 Инструментальные средства

Модель будет располагать средствами анализа, присущими формализму, с помощью которого будет реализована.

На данный момент для управления динамикой взаимодействия при итеративном тематическом или проблемном поиске представляется перспективным обращение к сетевым моделям. Исследования, предпринятые в этой области в 90-х гг. прошлого века [3,4,5] доказали перспективность такого направления научного поиска.

Одним из наиболее распространенных современных формализмов для моделирования и анализа распределенных динамических систем являются сети Петри. В настоящее время ведутся существует множество расширений сетей Петри, в частности для решения задач моделирования многоуровневых и параметризованных динамических систем [2]. Расширенные сети Петри обладают достаточной выразительной возможностью для проектирования реальных сложных систем. Для анализа такие расширенные сети путем структурных преобразований могут быть сведены к обыкновенным сетям Петри, после чего используются традиционные средства анализа: проверка свойств достижимости, живости переходов, наличия дедлоков и др..

В ходе машинной имитации предлагается осуществить серии опытов (прогонов модели) для

получения упорядоченных наборов данных – значений выходных переменных моделей. Каждая переменная по своему смыслу и допустимой области изменения должна быть адекватной некоторому понятию предметной области. После специальной, в том числе и статистической обработки полученных при имитации данных могут быть найдены экспериментально обоснованные интерпретированные оценки и прогнозы функционирования системы.

3. Модель итеративного поиска - варианты формализации модели

В итоге исходные данные для построения модели таковы:

O (objects) – множество информационных объектов процесса поиска, порождающих информационное пространство поисковой системы.

C (criterions) – множество критериев (условий, факторов), влияющих на формирование системой модели поведения пользователя.

P (phases) – множество контрольных точек (ситуаций, позиций, этапов) процесса поиска, в которых можно проверять истинность тех или иных критериев.

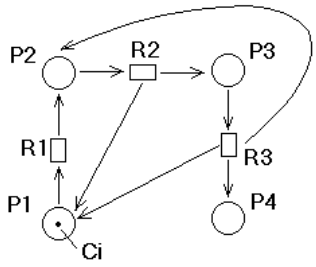
Также в процессе моделирования могут понадобиться следующие множества:

D (doing) – выполнение задач системой,

U (user actions) – действия пользователя, обусловленные множеством активных информационных объектов.

3.1 Первая интерпретация – процесс поиска

В самом общем виде сеть Петри, представляющая итеративный процесс поиска будет выглядеть следующим образом.



Состояния представляют собой выходные состояния каждого этапа, фишки – выполнение условий (критериев), перечисленных в таблице 2.

Состояния процесса поиска:

P1 – потребность ясна, запрос не сформулирован,

P2 – закончена формулировка вопроса,

P3 – выданы документы по запросу,

P4 – документы оценены.

Переходы:

R1 – процесс формулирования запроса ,

R2 – процесс поиска,

R3 – процесс оценки выдачи, где

R - множество функций-переходов $Pi+1 = R (Pi, \bar{C}i)$, отвечающих за переход к следующему этапу-состоянию процесса поиска в зависимости от выполнения условия $\bar{C}i$ на данном этапе

Метку целесообразно представить как вектор значений атрибутов, структура которого зависит от позиции, в которой расположена метка. Атрибуты в данном случае будут определять количество и свойства используемых на данном этапе объектов процесса поиска (табл.1).

Таким образом, представленная выше сеть является сетью высокого уровня, а точнее классической предикатной сетью. Переход считается возбужденным, если все его входные места содержат метки и их значения удовлетворяют определенному условию – «селектору перехода». После срабатывания устанавливаются новые значения вновь сгенерированных меток.

Как известно, предикатная сеть есть просто компактная запись сети Петри и может быть «развернута» в эквивалентную сеть Петри. Соответственно, все операции анализа сети можно проводить для сетей Петри.

Такая модель дает возможность анализировать адекватность критериев Ci и вероятность оказаться в той или иной контрольной точке в зависимости от критериев.

3.2 Другие интерпретации

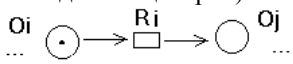
Модель динамической информационной структуры[3]

Совокупность одновременно активных в данный момент информационных элементов из допустимого множества всех информационных элементов названа сценарием $Si \subset O$. Тогда U (user) – действия пользователя, обусловленные Si. При наступлении какого-либо системного события или сразу после запуска приложения подмножество активных информационных элементов $Si \subset O$, названное сценарием, вызывает определенную реакцию пользователя, выражающуюся в инициализации одного из равновероятно разрешенных (для данного подмножества состояний Si) переходов. Каждый переход, в свою очередь, посредством отображения R определяет свое подмножество активных (видимых) элементов, определенное как сценарий.

R (reactions) – реакция системы - множество функций-переходов $S_{i+1} = R (U(S_i), P_i, C_i)$, отвечающих за формирование системой подмножества активных информационных объектов в зависимости от действий пользователя и значений критериев в каждой контрольной точке $Pi \subset P$. Графически учет этапа поиска возможен разбиением сети сценариев на подсети.

В рамках данной интерпретации информационные элементы приложения есть

множество позиций; R – множество всех возможных переходов; активному (видимому) элементу соответствует помеченная позиция; сценарию (множеству видимых элементов) – некоторая разметка M; каждая вершина сети может содержать не более одной фишки (один элемент не может быть одновременно дважды активизирован в рамках одного сценария).



Из свойств сетей Петри для ограниченных сетей, в приложении к задачам моделирования систем, взаимодействующих с пользователем, безопасность и активность являются наиболее важными свойствами. Другим важным свойством моделируемой структуры является ее активность, т.е. отсутствие тупиковых состояний или тупиковых разметок. Для реального процесса функционирования системы это означает, что должна существовать возможность либо продолжения поиска, либо возврата к некому предыдущему состоянию. Т.к. исходно количество фишек в любой позиции не может быть более одной, то данная модель безопасна в течение всего времени функционирования приложения. Разрешимость задачи достижимости переходов для ограниченной сети Петри также следует из введенного ограничения на количество фишек в каждой позиции и конечности дерева достижимости.

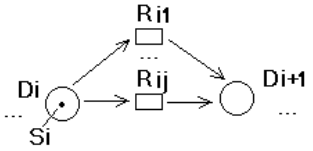
Представление графа достижимых маркировок такой сети цепью Маркова, отражающей вероятности переходов между маркировками графа, дает возможность определить оптимальные (имеющие максимально вероятную достижимость) маркировки.

Итак, сценарий есть множество активизированных в некоторый момент времени разнотипных элементов, используемых в одном контексте. Наиболее часто достижимые сценарии будут считаться оптимальными, на основе такого анализа и будет построен интерфейс системы поиска.

Модель диалога [4]

Над исходными объектами системы возможно производить некоторые действия, набор которых определен. Действия над объектом осуществляются субъектом воздействия (пользователем или системой).

Проблемы пользователя разбиваются на шаги-задачи для машины. Задачи в данной интерпретации представляют множество позиций, способы решения задач – множество переходов. Фишками сети будут наборы информационных объектов, требуемых для решения той или иной задачи. Например, задача «найти термин для добавления в запрос» может решаться четырьмя способами (используются такие объекты как тезаурус, словарь, словник или документ), а задача расширить/сузить запрос – обширным множеством вариантов решений с использованием многообразия наборов информационных объектов. Фрагмент сети задач:



Проблемы пользователя можно сопоставить с этапами процесса поиска, которые представляют собой повод для разбиения сети задач на подсети. Для подсетей должны быть определены согласующиеся между собой входные и выходные объекты.

Например, в данной предметной области возможны следующие пользовательские проблемы:

1 – построить запрос (O_1 – поисковая потребность и поисковая стратегия, O_2 – поисковое предписание),

2 – осуществить поиск (O_2 – поисковое предписание, O_3 – множество документов),

3 – оценить выдачу – (O_3 – множество документов, O_4 – множество документов, O_1 – измененная поисковая стратегия и потребность).

Для данной модели предполагается иерархичность сети (в каждой позиции простой сетью может решаться задача формирования необходимого набора объектов).

Построение такой функциональной сети диалога позволяет проследить корректность предполагаемой диалоговой системы и выявить её свойства.

Анализ сетевых моделей позволяет делать выводы о вероятности срабатывания того или иного перехода и проектировать интерфейс систем поиска документальной информации в соответствии с нуждами пользователя, основываясь на полученных типовых линиях решения задач поиска документальной информации.

Литература

[1] Забегаева Н.Н., Максимов Н.В. Информационный поиск и модели поведения пользователей // НТИ. Сер. 1.– № 11, 2001.

[2] Ломазова И.А. Объектно-ориентированные сети Петри: формальная семантика и анализ, Системная информатика, вып.8, "Наука" РАН, 2002.

[3] Смирнова Е.И. Моделирование ограниченными сетями Петри динамических информационных структур, дисс., Новгород, НГУ, 1998г.

[4] Туева Н.С. Моделирование диалоговых систем с помощью сетей Петри, ВЦРАН, Москва, 1991г.

[5] R.B. Coats, I. Vlaeminke. Man-computer interfaces, пер. под ред. В.Ф. Шаньгина, М.:Мир, 1990.

About modeling interactive iterative search process in documental data base

Maksimov N.V., Sviridenko S.V.

This paper regards various approaches to formalization and modeling of search process in documental data bases for user interface optimization.

Finding probability of using certain information objects on a certain step of searching documental information will provide an opportunity to increase the efficiency of user interfaces. It makes possible projecting a dynamic interface with providing elements of interface content exactly as necessary for user in order to adapt a computer system to this user's way to handle tasks.