СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р10 - 9051

Д.Д.Арнаудов, Н.И.Янев

ДОПОИСКОВОЕ ПРОГНОЗИРОВАНИЕ ЧИСЛА РЕЛЕВАНТНЫХ ДОКУМЕНТОВ В ИПС ОИЯИ

1975

P10 - 9051

Д.Д.Арнаудов, Н.И.Янев

ДОПОИСКОВОЕ ПРОГНОЗИРОВАНИЕ ЧИСЛА РЕЛЕВАНТНЫХ ДОКУМЕНТОВ В ИПС ОИЯИ

Основные характеристики, организационная структура и стратегия поиска информационно-поисковой системы ОИЯИ описаны в работах [3,4]. Под "информационным поиском" понимается область обработки данных, включающая совокупность ряда арифметических, логических операций, ввод-вывод, конечной целью которых является выявление по заданным признакам отбора всех данных /в виде документов, фактических справок/, содержащих требуемую информацию и представляющих собой ответы на вводимую.

Анализ работы поисковых систем [2] показывает, что множество выдаваемых ими документов определяется критерием смыслового соответствия. Для большинства ИПС этот критерий приблизительно одинаков. Наличие или отсутствие в документе какой-либо информации сверх того, о чем говорится в запросе, как правило, не влияет на решение вопроса о выдаче документа. Но если в документе отсутствует существенная часть информации, упоминаемой в запросе, то такой документ обычно не выдается. Не выдаются, как правило, и документы, в которых содержится информация более общая, чем та, о которой идет речь в запросе. Исходя из этого, сущность стратегии поиска в ИПС ОИЯИ заключается в поиске тех документов, для которых совокупность положительных дескрипторов запроса является подмножеством совокупности дескрипторов, входящих в поисковые образы документов. В связи с этим необходимо заметить, что ИПС ОИЯИ относится к т.н. включающим информационно-поисковым системам [2]. Эти системы обычно основаны на использовании ключевых слов /дескрипторов/, и их ос-

новная особенность заключается в том, что с ростом числа дескрипторов запроса ожидаемое число релевантных документов уменьшается.

ИПС ОИЯИ нацелена на обработку информационного массива большого объема /оперативное хранилище на 500000 документов/. Исследования показывают [2,4], что при таком большом объеме число релевантных документов становится значительным, что увеличивает время их выдачи и затрудняет пользователю выбор интересующей его информации. Для решения этого вопроса необходимо создать модуль системы, который на основе запроса "прогнозировал" бы число ожидаемых ответов. Если это число достаточно велико, то посредством обратной связи с потребителем может быть совершена переформулировка запроса. В таком случае ИПС выступает в качестве автоматического словаря, обеспечивая потребителей информацией словарного типа о дескрипторах, используемых в формулированных запросах. В нашем случае даются сведения о "более узких" терминах, соответствующих данному дескриптору, что приводит к уменьшению числа релевантных документов. Другой путь к разрешению данного вопроса - это увеличение дескрипторов конъюнкции, поскольку, как уже отметили, с ростом числа дескрипторов запроса ожидаемое число релевантных документов уменьшается.

С целью выяснения идеи построения модуля прогнозирования релевантных документов рассмотрим работу системы с точки зрения теоретико-множественной модели, описанной в [2].

Пусть D означает конечное множество документов /количество документов - K /, которые используются для обслуживания запросов Q . Содержание документов индексируется дескрипторами, выбираемыми из дескрипторного словаря N . Элементы множества Q являются любыми наборами из элементов множества N , связанными знаками конъюнкции, дизъюнкции и отрицания [5]. Пусть R означает некоторую функцию /ее называют еще поисковой/, определяемую в Q со значениями в 2^D /где 2^D означает множества всех подмножеств D /. Тогда каждому запросу $q \in Q$, $R(q)$ ставится в соответствие

определенный элемент из 2^D , который называют выдачей на запрос q /иначе говоря, документ, релевантный запросу q /.

Как уже отмечено в [1] , ИПС может быть формально определена как совокупность конечного множества документов D , дескрипторного языка N , поисковой функции R и языка запросов Q .

Сэлтон [2] отмечает, что ИПС частично упорядочена, если Q является частично-упорядоченным множеством, т.е. если на Q определено транзитивное и антисимметричное отношение $\geq$. Причем, если $q \in Q$ и $S \subset Q$, то q меньше $S (q < S)$, если $q \wedge S = q$. Такое соотношение введено и для элементов множества 2^D. Исходя из этого, ИПС является включающей, если функция R такова, что для $q < S$, для любых $\{q, S\} \in Q$ выполняется условие $R(q) \geq R(S)$.

Поскольку ИПС ОИЯИ является включающей информационно-поисковой системой, то для нее гарантируется нахождение всех документов по запросу S , если найдены документы по запросу q . Имеется в виду, что q тождественно S за исключением того факта, что данный дескриптор, включенный в S, исключен при формировании q . При этом по запросу q , однако, может быть найдено дополнительное число документов, не найденных раньше по S , так что $R(q)$ может быть больше чем $R(S)$.

Взаимосвязь между порядком, определенным в пространстве запросов Q и соответствующим порядком в пространстве документов 2^D, задается теоремой [6] .

Теорема 1. Во включающей ИПС образом уменьшающейся цепи в Q при отображении R является увеличивающаяся цепь в 2^D , а образом увеличивающейся цепи в Q - уменьшающаяся цепь в 2^D. Доказательство теоремы дано в [6] . Будем использовать эту теорему для решения задачи прогнозирования числа релевантных документов. Для этой цели введем еще два обозначения: $\bar{R}(q)$ и L , где $\bar{R}(q)$ означает ожидаемое число релевантных документов /напомним, что $R(q)$ - точное число релевантных документов, найденное на основе проведенного поиска в системе [4] /, а L - пороговое число ожидаемых документов. Его физический смысл следующий.

Если вследствие прогнозирования ожидается получить $\bar{R}(q)$ документов, причем $\bar{R}(q) > L$, то тогда необходима переформулировка запроса /т.е. либо увеличение числа дескрипторов конъюнкции, либо использование "более узких" дескрипторов/.

Итак, задача прогнозирования сводится к установлению верности следующих неравенств:

$$R(q) \leq L \qquad /1/$$

$$\bar{R}(q) > L . \qquad /2/$$

При установлении неравенства /1/ запрос считается "корректным" и по нему производится поиск согласно принятой в ИПС стратегии поиска [4]. Если, однако, установлено неравенство /2/, то следует, как мы уже отметили, переформулировка запроса.

В дальнейшем в качестве отдельного запроса q будем рассматривать только отдельные конъюнкции из положительных дескрипторов [4,5]. Если имеется дизъюнктивная форма запроса, то дизъюнкцию рассматриваем как сумму отдельных конъюнктивных групп, т.е. как состоящую из нескольких запросов.

Булевский оператор отрицания рассматриваем в смысле дополнимости в отношении множества A [4]. Следовательно, если в запросе имеются дескрипторы со знаком отрицания, то:

$$\bar{R}(q) = \bar{R}(a_1, a_2, \dots a_S, \bar{a}_{S+1}, \dots \bar{a}_t) = \qquad /3/$$

$$= \bar{R}(a_1 a_2 \dots a_S) - \bar{R}(a_1, a_2, \dots a_S, a_{S+1}, \dots a_t) .$$

Итак, пусть $q = (a_1, a_2, \dots a_S)$ обозначает запрос, требующий выдачи всех документов, в поисковых образах которых участвуют все дескрипторы запроса. По теореме 1 получаем, что

$$\bar{R}(a_1 a_2 \dots a_S) \leq \bar{R}(a_i) \quad \text{для } i = 1,2,\dots S , \qquad /4/$$

так как $(a_1 a_2 \dots a_S) > a_i$ для $i = 1,2,\dots S$.

Если найдем такое i^*, при котором $\bar{R}(a_{i^*}) \leq L$, то, как уже отметили, вопрос считается корректным.

Если, однако, $\bar{R}(a_{i^*}) = \min \bar{R}(a_i) > L$ для $i=1,\dots S$, вопрос о выполнении неравенств /1/ или /2/ остается открытым и нуждается в дальнейшем исследовании.

Заметим, что как первое приближенное решение задачи прогнозирования берется $R(a_{i^*})$, для чего используется $f(i^*)$ [7] - частота встречаемости дескрипторов в поисковых образах документов. При этом в таблицу сведены только те $f(i^*)$, для которых выполняются неравенства

$$f(i^*) > L .$$

Если при исследовании $f(i^*)$ $i^*=1,\dots S$ всех дескрипторов конкретного запроса q окажется, что среди них имеется хоть один, у которого

$$f(i^*) \leq L \qquad /5/$$

/т.е. его частота не указана в таблице/, то запрос, как уже заметили, считается корректным и подлежит дальнейшей поисковой обработке. При этом в качестве $f(i^*)$ неравенства /5/ используется минимальная частота дескрипторов запроса. Если, однако, условие /5/ не выполняется, то тогда переходим к дальнейшему исследованию вопроса, рассматривая все возможные пары дескрипторов запроса. Это основано на следующем суждении:

Если

$$\bar{R}(a_{p^*}, a_{q^*}) = \min_{i,j} \bar{R}(a_i, a_j)$$
$$i = 1,\dots S ; \quad j = 1,\dots S ; \quad j > i ,$$

то $\bar{R}(a_{p^*}, a_{q^*})$ можно использовать в качестве второго приближения задачи прогнозирования, так как

$$\bar{R}(a_{p^*}, a_{q^*}) \leq R(a_i, a_j) \leq \min\{\bar{R}(a_i), \bar{R}(a_j)\} = \bar{R}(a_{i^*}) .$$

Тогда, если $\bar{R}(a_{p^*}, a_{q^*}) \leq L$, то запрос считается корректным, а в противном случае можно искать $\bar{R}(a_i, a_j, a_k)$ и т.д. Однако это дальнейшее исследование, хотя и дает повышение достоверности прогноза, ведет к большим затратам машинной памяти и значительно уве-

личивает время ответа. Поэтому дальнейшие рассмотрения имеют место только тогда, когда совместное распределение любых двух и более дескрипторов не зависит /или слабо зависит/ от распределения других дескрипторов, т.е. если:

$$P(a_i, a_j / \bar{q}) = P(a_i, a_j). \qquad /6/$$

/Обозначение $P(a_i, a_j / \bar{q}$ / следует понимать как условную вероятность одновременного участия дескрипторов $a_i a_j$ в произвольно выбранном документе, если среди других дескрипторов, которыми индексирован документ, участвует $\bar{q}$. Символом $\bar{q}$ обозначена любая конъюнкция из дескрипторов, отличных от a_i, a_j).

Пусть $L^{**} = \min_{j>i} \bar{R}(a_i, a_j) > L$.

Если запрос $q = (a_1, a_2, \ldots a_S)$ состоит только из двух дескрипторов, то следует принять достоверным неравенство $\bar{R}(q) > L$. Если $S > 2$, то, имея в виду /6/, можно применить следующее выражение:

$$\bar{R}(a_1, a_2, \ldots a_S) = KP(a_1, a_2, \ldots a_S) = KP(a_1, a_2) \cdot P(a_3, a_4) \ldots P(a_{S-1}, a_S). \qquad /7/$$

Последнее решение принимается в зависимости от того, является ли величина $KP(a_1,a_2)P(a_3,a_4)\ldots P(a_{S-1},a_S)$ большей или меньшей L.

В случае нечетного числа дескрипторов конъюнкции последний участвует со своей вероятностью $P(a_S)$, т.е.

$$KP(a_1,a_2) \cdot P(a_3,a_4) \ldots P(a_{S-2},a_{S-1}) \cdot P(a_S).$$

В связи с вышеизложенным мы считаем, что для запросов, состоящих из трех-четырех дескрипторов, можно принять в качестве оценки прогнозирования исследования неравенств до второго ранга включительно /т.е. исследования пар дескрипторов/.

Для программной реализации модуля прогнозирования необходимо формировать две таблицы. В первой заносятся,

на основе определенной выборки, частоты "высокочастотных" дескрипторов, т.е. такие $f(i)$, для которых $f(i) > L$. Приблизительное число элементов этой таблицы можно вычислить на основе следующего неравенства:

$$L \geq \frac{S}{j(\ln N + \gamma)}.$$

где S - число всех дескрипторов, использованных для индексирования в поисковых образах документа, N - число различных дескрипторов, γ - константа Эйлера, /γ = 0,5772/. Дальнейшие подробности по использованию данной формулы описаны в [7].

Для ИПС из 100000 документов, каждый из которых индексирован в среднем десятью дескрипторами, при $N = 12000$ и $L = 10$, получаем $j \sim 1000$, т.е. число дескрипторов, превышающих пороговое число, не больше чем 1/12 общего числа различных дескрипторов.

Вторая таблица задает ассоциации дескрипторов /т.е. это матрица дескриптор-дескриптор/. Формально матрицу можно определить следующим образом. Рассмотрим каждый документ ИПС как N-мерный вектор-столбец, i-ая компонента которого равняется единице или нулю, в зависимости от того, содержит ли документ i-ый дескриптор, тогда связь дескриптор-документ задается как бинарная матрица размерностью $N \times K$ /где N - количество дескрипторов словаря, K - количество документов ИПС/.

Если обозначим эту матрицу через C, то симметричная матрица CC'/ C'- транспонированная матрица C / задает связь дескриптор-дескриптор. Элемент, который находится в i-ом ряду, j-ом столбце, является $\bar{R}(a_i, a_j)$.

Для решения задачи прогнозирования необходима не вся матрица CC', а только подматрица $\bar{C}\bar{C}'$, где $\bar{C}$ получается из C, вычеркивая все ряды, соответствующие дескрипторам a_i, для которых $f(i) \leq L$. Поскольку матрица $\bar{C}\bar{C}'$ симметрична, в памяти ЭВМ необходимо хранить только элементы над главной диагональю.

На основе рассмотренного материала проведем машинный эксперимент над реальным информационным массивом ИНИС со следующими параметрами: количество документов - 3236, среднее количество дескрипторов на

документ - 9, средняя длина дескрипторного списка - 9, пороговое число L=5. Результаты экспериментов сведены в таблицы. В *табл. 1* частично приведены некоторые цифры. Здесь собраны частоты дескрипторов больше критического числа L . Общее число элементов таблицы - 1100.

Таблица 1

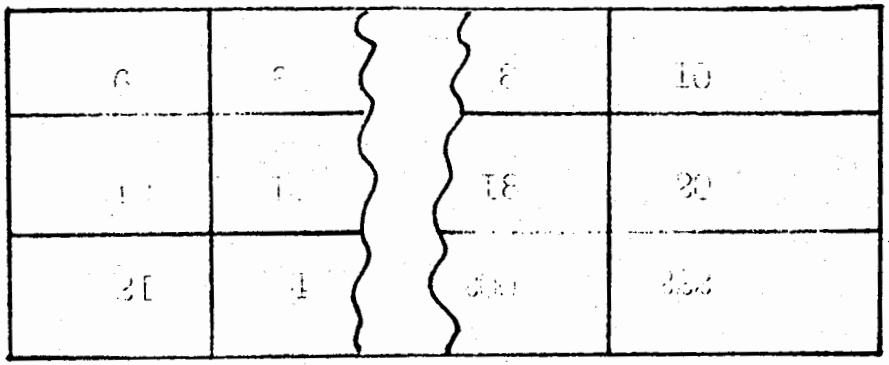
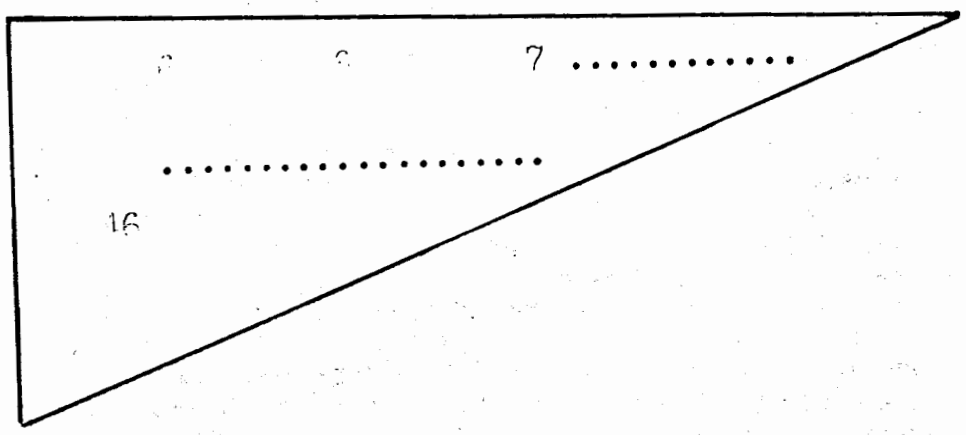


Таблица 2



В *табл. 2* собраны частоты пар дескрипторов больше критического числа L . Первоначальное число всех возможных пар - 60445О элементов. Из них с частотой, большей критического числа, имеется 1975 элементов, которые являются элементами *табл. 2.* Максимальная частота - 46.

Таблица 3

Вариант	Результат прогнозирования	Результат поиска
1	О,4	
2	1,9	2
3	О,7	
1	О,2	
2	О,6	О
3	О,5	
1	О,2	
2	1,9	2
3	О,7	
1	О,6	
2	О,6	О
3	О,4	
1	О,2	
2	1,7	2
3	О,5	

В *табл. 3* сведены результаты экспериментов, проведенных над запросами из трех дескрипторов. Вычисление ожидаемого количества документов проводилось по трем формулам. Первая формула, как уже отметили, представлена выражением /7/. Имеется, однако, еще две возможности вычисления вероятностей, когда S>2. Для выяснения этого вопроса отметим следующее. Пусть через $P(a_1, a_2, \ldots a_S)$ обозначим вероятность того, что в случайно выбранном документе встретятся дескрипторы $a_1, a_2, \ldots a_S$. Тогда математическое ожидание документов, релевантных запросу $(a_1, a_2, \ldots a_S)$, задается вы-

ражением $KP(a_1, a_2, \dots a_S)$, где K - число документов в информационном массиве ИПС. Тогда применяется /8/.

$$P(a_1, a_2, \dots a_K) = P(a_1) P(a_2/a_1) P(a_3/a_1 a_2) \dots P(a_S/a_1 a_2 \dots a_{S-1}) \quad /8/$$

Можно указать еще две возможности для вычисления вероятности прогнозирования документов.

$$P(a_1 a_2 \dots a_S) = P(a^*) P_1^{S-1} \quad /9/$$

$$P(a_1 a_2 \dots a_S) = P(a^*) P_2^{S-1} \quad /10/$$

В указанных выше формулах через $P(a^*)$ обозначена вероятность главного дескриптора запроса. В /2,6/ показывают, что в качестве главного дескриптора необходимо выбирать самый редкий дескриптор, т.е. тот, у которого минимальное $P(a_1)$. Через P_1 в выражении /9/ обозначена средняя условная вероятность какого-либо дескриптора в случайно выбранном документе, при условии встречаемости главного дескриптора в этом документе. Тогда

$$P_1 = \sum_{i=1}^{S} \delta_i P(a_i / a^*), \quad a_i \neq a^*. \quad /11/$$

Весовые коэффициенты δ_i пока можно выбрать только экспериментально, исходя из повышения точности прогноза. В нашем эксперименте $\delta_i = 1$ для всех i.

Через P_2 во второй формуле обозначена средняя условная вероятность появления дескриптора a_i при условии, что в этом же документе встретился a_j.

$$P_2 = \sum_{j=1}^{S} \sum_{i=1}^{S} \delta_{ij} P(a_i / a_j), \quad j > i \quad /12/$$

/δ_{ij} в нашем эксперименте равняется единице/. Вычисление условных вероятностей $P(a_i/a_j)$ производится по формуле:

$$P(a_i/a_j) = \frac{P(a_i/a_j)}{P(a_j)}; \quad P(a_j) > 0.$$

Вероятности $P(a, a_j)$, $P(a_j)$ находятся на основе табл. 1 и 2.

В *табл. 3* под номером варианта понимается соответственно использование формул /7/, /9/, /10/. Из приведенных данных видно, что формула /9/ дает лучшие результаты, и она положена в основу алгоритма прогнозирования релевантных документов в ИПС ОИЯИ.

Модуль допоискового прогнозирования необходимо вводить в ИПС только тогда, когда имеется значительный объем информационного массива /когда средняя длина дескрипторного списка превышает 100/. Тогда, выбирая подходящее L /например, L =40/ для коротких запросов /не больше трех-четырех дескрипторов/ можно ожидать удовлетворительных результатов.

Если запрос представляет дизъюнкцию конъюнкций, то прогнозируется число релевантных документов в случае, когда хоть бы для одной конъюнкции получается $\bar{R}(q) > L$. Здесь не учитывается случай, когда сумма конюнкций, для которых $\bar{R}(q) < L$, дает в сумме число, большее порогового.

Литература

1. *Д.Д.Арнаудов. Сообщение ОИЯИ, 10-7586, Дубна, 1973.*
2. *Сэлтон. Автоматическая обработка, хранение и поиск информации. Сов. радио, 1973.*
3. *Д.Д.Арнаудов. Сообщение ОИЯИ, Р10-8621, Дубна, 1975.*
4. *Д.Д.Арнаудов. Сообщение ОИЯИ, Р10-8622, Дубна, 1975.*
5. *Д.Д.Арнаудов, З.Коженкова. Сообщение ОИЯИ, 10-8799, Дубна, 1975.*
6. *W.A.Woods. A Mathematical Theoty of Retrieval Systems. Harvard University, Applied Mathematics 221, Student Research Report, Foll, 1964.*
7. *Д.Д.Арнаудов. Препринт ОИЯИ, 10-7953, Дубна, 1974.*

Рукопись поступила в издательский отдел
9 июля 1975 года.