Раздел IV. Новые информационные технологии

Э.М. Котов, А.Н. Целых

ОПИСАНИЕ ИНФОРМАЦИОННОЙ ПОТРЕБНОСТИ ДЛЯ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

Парадигма информационного поиска состоит в наличие потребности в информации и ряде информационных объектов, от которых зависит то, как может быть удовлетворена эта потребность. Рассмотрим общие модели, чтобы формализовать понятие информационной потребности.

Пользователь имеет потребность в определенной информации. Этой потребности могут быть присущи, как качественные, так и количественные аспекты. В контексте информационной коллекции, пользователь пробует сформулировать эту потребность в терминах данного информационного ресурса, переводя потребность в информации в потребность в информационных объектах (документах).

Информационная коллекция может быть фиксированной и в данном случае информационно-поисковая система может провести предварительное исследование коллекции для дальнейшего оптимального поиска. Суть этой предварительной обработки состоит в том, что информационно-поисковая система получает предварительные данные от содержания каждого документа. Это приводит к необходимости применения методики описания содержания документа.

Информационная коллекция может быть непрерывным потоком документов (например, телеконференция в Интернете). Для каждого следующего документа, информационно-поисковая система должна решить, будет ли это интересно для пользователя. Это может быть названо как парадигма информационной фильтрации. В данном случае не может быть никакой предварительной обработки всех документов. Содержание документа тогда является или основанным для некоторой системы описания содержания, или информационно-поисковая система должна пробовать строить такую систему с приращением, когда прибывают новые документы. Это актуально для Всемирной паутины, где присутствует чрезвычайно большое, динамическое и изменчивое собрание документов.

Различие между информационным поиском и информационным фильтрацией в том, что действительно интересует пользователя в течение некоторого промежутка времени, и том, что выражает информационную потребность в конкретный момент вовремя.

Основная цель информационно-поисковой системы обеспечить эффективный механизм извлечения n информационных объектов из информационной коллекции O. Предположим, что пользователь интересуется некоторыми из информационных объектов. Самый простой подход состоит в том, чтобы моделировать этот интерес, как некоторый запрос относительно всей информационной коллекции. Посредством данного запроса, возможно, провести разделение коллекции на релевантные и нерелевантные документы. Задача информационно-поисковой системы тогда может быть описана как создание запроса, который наилучшим образом удовлетворяет интересам пользователя. Эта модель определяется как *сравнительная модель* для информационной потребности пользователя.

Обычно не все информационные объекты имеют равную важность для пользователя. Информационную потребность пользователя, поэтому представим как функцию *N*:

$$N : O \to [0;1]. \quad (1)$$

Назовем *N(x)* потребностью в информационном объекте *x*. Эта функция может быть определена как субъективное значение релевантности ко всем документам. Это модель определяется как *взвешенная модель* для информационной потребности. Отметим, что сравнительная модель получена от взвешенной модели, сравнивая документы согласно их релевантности.

Возможно упрощение − выборочная информационная потребность, где каждый информационный объект или релевантен или нет:

$$N : O \to \{0;1\}. \quad (2)$$

Эта *дискретная модель* − традиционная основа для информационного поиска. Дискретная модель может быть охарактеризована как частный случай взвешенной модели. В дискретной модели, информационная потребность может интерпретироваться как функция, связанная с подмножеством всех документов, которыми интересуется пользователь.

Могут использоваться и более сложные модели описания информационной потребности. Например, отметим *возрастающую модель*. В этой модели, предполагается, что потребность в большем количестве документов возникает под влиянием того, что пользователь уже получил из коллекции. Это может быть отражено как функция:

$$I : P(O) \to (O \to [0;1]), \quad (3)$$

или, что эквивалентно, как функция:

$$I : P(O) \times (O \to [0;1]), \quad (4)$$

где *I(S,x)* интерпретируется как приращение в удовлетворении потребности пользователя, когда документ *x* предлагается пользователю после того, как установлена потребность в *S*, ранее предложенном пользователю. Функция *I* является *функцией приращения*. Отметим, что взвешенная модель может быть получена от возрастающей модели, определив $N = I(\emptyset)$. Таким образом, потребность в документе определена как удовлетворение потребности пользователя, полученное без предшествующего знания.

Возрастающая модель особенно полезна для динамических и распределенных архивов типа Всемирной паутины. Функция приращения учитывает вычисление в реальном времени, что является отличительным признаком от подходов, которые кластеризуют поисковый результат перед представлением кластера пользователю, но объединение в кластеры возможно только после получения всех документов.

Функция пользователя состоит в необходимости адекватным образом сообщить информационную потребность информационно-поисковой системе. С этой целью в информационно-поисковой системе реализован поисковый язык *Q*. Предположим, что пользователь сформулировал потребности в информации *N* как запрос *q*. Далее информационно-поисковая система будет искать соответствие информационным целям по запросу *q*, и приблизиться, насколько возможно к информации, которая необходима в соответствии с *N*, и сформулирована в запросе *q*.

Соответствующую функцию представим следующим образом:

$$M : Q \to (O \to [0;1]). \quad (5)$$

Для каждого *q*, *M[q]* − функция, которая назначает для каждого документа из коллекции его релевантность, оцененную информационно-поисковой системой на

основании запроса *q*. Чтобы разрешить возникающие проблемы при поиске соответствия, должен быть устранен неточный и некорректный перевод информационной потребности *N* в запрос *q*. С этой целью, мы предполагаем необходимость, что бы язык запроса *Q* приобрел семантическую интерпретацию через функцию Q_S:

$$Q_S : Q \to (O \to [0;1]). \quad (6)$$

Функцию Q_S называют *официальной интерпретацией* языка запроса *Q* в информационной базе *O*. $Q_S[q]$ называют *нормативной информационной потребностью*, связанной с запросом *q*. Нормативная информационная потребность является объективной релевантностью. Обычно эта функция не имеет никакого формального определения, и определяется вручную группой экспертов некоторой предметной области. В обычных СУБД язык запроса подразумевает существование запроса для каждой информационной потребности, которая может быть извлечена из базы данных. В информационно-поисковых системах язык запроса не имеет этой особенности.

Цель информационно-поисковой системы состоит в определении функция *M* такой, что *M[q]* является наилучшим приближением нормативной информационной потребности $Q_S[q]$ для всех запросов *q*.

Мера сходства может быть выражена как расстояние $R_O(M[q],Q_S[q])$ между функциями *M[q]* и $Q_S[q]$. Функция расстояния может быть определена несколькими способами. Например:

$$R_X(f,g) = (\sum_{x \in X} |f(x) - g(x)|^p)^{1/p}, \quad p \geq 1. \quad (7)$$

И справедливо следующее:

$$1. R(x,x) = 0$$

$$2. R(x,y) = R(y,x)$$

$$3. R(x,y) \leq R(x,z) + R(z,y)$$

Цель пользователя состоит в том, чтобы построить такой запрос *q*, который минимизирует $R(N, M[q])$. Этот процесс может быть понят из неравенства треугольника:

$$R(N, M[q]) \leq R(N, Q_S[q]) + R(Q_S[q], M[q]). \quad (8)$$

Качество информационно-поисковой системы определим как функцию расстояния:

$$R_{Q \times O}(M, Q_S). \quad (9)$$

Если *Q* представить рядом тестовых вопросов, тогда качество информационно-поисковой системы возможно оценить как:

$$\frac{1}{m} \sum_{q \in Q} R_O(M, Q_S), \quad (10)$$

где m – число тестовых вопросов.

Поиск соответствия информационной потребности *N* – довольно сложная задача. Информационно-поисковая система должна не только предложить документы в соответствии с запросом, но и необходимое требование к системе сводится к оценки документов, настолько они подходят, каким образом возможно провести ранжирование, в зависимости от информационной потребности *N*. Типичная поисковая система должна назначить каждому документу степень релевантности, и впоследствии сортировать документы согласно этим оценкам.

Информационно-поисковые системы прямо или косвенно базируются на моделях поискового процесса. Эти поисковые модели определяют, способы представления документов текста, информационной потребности и используемые методы для сравнения и оценки вероятности того, что документ будет релевантен. Оценки релевантности документов данному запросу являются основанием для ранжирования документов, которые признаны релевантными системой информационного поиска. Примеры простых моделей включают вероятностную или векторную пространственную модель. Предлагаются и используются так же и многие другие модели.

В процессе разработки более сложных моделей поиска проводилось множество экспериментов для определения эффективности предлагаемых подходов. В ходе ранних экспериментов наблюдалось, при применении различных моделей и алгоритмов поиска, удивительно низкое перекрытие в области релевантных документов, которые были найдены [1]. Недостаточность перекрытия между релевантными документами, найденными различными алгоритмами и привели к двум подходам в развитие систем информационного поиска и моделей поиска.

Один подход состоял в создании моделей поиска способных описать и скомбинировать различные критерии релевантности. Эти модели являются типичными вероятностными и мотивируются *принципом вероятностного ранжирования*, состоящем в том, что оптимальная эффективность поиска достигается путем ранжирования документов в порядке убывания относительно вероятности их релевантности и, что “вероятности оценены настолько точно, насколько возможно на основе исходных данных доступных системе” [2].

Другой подход состоял в том, чтобы проектировать системы, которые могут эффективно объединить результаты многократных поисков, основанных на различных моделях поиска. Эта комбинация может быть реализована в отдельной системной архитектуре или в распределенной, гетерогенной среде. Объединение многократных, гетерогенных поисков явилось основой для *метапоисковых* механизмов в Сети (например, MetaCrawler) и становится все более и более важным методом при поиске в базах данных мультимедиа.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *McGill, M.,Koll, M., and Noreault, T.* An evaluation of factors affecting document ranking by information retrieval systems. Final report for grant NSF-IST-78-10454 to the National Science Foundation, Syracuse University.
2. *Robertson, S.* The probability ranking principle in information retrieval. Journal of Documentation, 33:294–304.

В.В. Янушко, А.В. Далёкин С.Н. Еркин

ВИРТУАЛЬНЫЕ ОРГАНИЗАЦИИ. ДЕКОМПОЗИЦИЯ АРХИТЕКТУРЫ*

В условиях постиндустриальной экономики совместная работа над проектом с использованием различных способов взаимодействия как между людьми, так и между информационными системами становится конкурентным преимуществом. Поэтому видоизменяется само понятие «компания» (организация). Появляются «виртуальные организации», в которых границы между участниками, ресурсами и отдельными подразделениями, благодаря интенсивному информационному обмену, становятся нечеткими, «размытыми» [1,2].

* Работа выполнена при поддержке РФФИ (гранты № 05-08-18115, № 07-01-00511) и программ развития научного потенциала высшей школы 2006-2008 гг. (РНП.2.1.2.3193, РНП 2.1.2.2238).