УДК 004.8

ИССЛЕДОВАНИЕ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА

А.Б. Касекеева
Докторант 2 курса ЕНУ имени Л.Н.Гумилева, г.Астана, Казахстан

ЕНУ имени Л.Н.Гумилева, г.Астана, Казахстан.
E-mail: aibike_7474@mail.ru

А.Б. Касекеева
Окончила АГУ имени Абая, физико-математический факультет, докторант 2 курса ЕНУ имени Л.Н.Гумилева, г.Астана, Казахстан.

Аннотация. Приводится краткий обзор современных методов поиска. В обзор включены классические основы современных методов поиска. Рассматриваются модели булева поиска, векторная модель, вероятностная модель информационного поиска.

Ключевые слова: информационно-поисковая системы, информационно-поисковой язык, релевантность поиска, булева модель поиска, векторная модель поиска, вероятностная модель поиска.

Проблема поиска и сбора информации является одной из важнейших задач информационных систем. Компьютерные методы поиска информации являются активно развивающейся, актуальной в научно-практическом аспекте темой современных публикаций. Развитие компьютерных технологий влечет за собой значительное увеличение объема информации, представленной в электронном виде, влияние этого процесса на современные информационные технологии, в том числе на поиск информации [1].

Главная цель информационно-поисковой системы (ИПС) – наилучшим образом удовлетворить потребности пользователей в необходимой информации. Для реализации этой глобальной цели необходимо проделать ряд подготовительных операций: проанализировать информационный массив и представить его в форме, удобной для хранения и обработки. Не менее важной частью поискового аппарата ИПС является модель поиска информации. Она описывает способ и критерии сравнения запросов и документов, а также форму представления результатов этого сравнения [2].

Любая модель поиска тесно связана с информационно-поисковым языком. Информационно-поисковый язык (ИПЯ) – это специальный язык для формирования запросов к ИПС. Необходимость создания ИПЯ вызвана трудностями интерпретации естественного языка в компьютерной системе. Однако синтаксис информационно-поисковых языков обычно довольно прост и внешне они часто похожи на естественные. Перед использованием запросов на ИПЯ проводятся лексическая (например, удаление из запроса терминов, присутствующих в стоп-словаре), морфологическая (нормализация терминов запроса 1), реже синтаксическая и семантическая обработки (в основном в экспериментальных системах).

Информационная потребность пользователя. В определенный момент времени у

пользователя возникает информационная потребность в документах по той или иной нужной ему тематике, которая обычно достаточно узка.

В начальный момент эта потребность часто не может быть точно выражена словами. При этом любая поисковая система требует достаточно четких запросов. Следовательно, пользователь должен представить свою информационную потребность в виде некоторого выражения, которое может быть воспринято поисковой системой.

Выделяют четыре этапа представления информационной потребности:

- реальная информационная потребность – это неосознанная истинная информационная потребность пользователя (потребность в некоторой новой информации при решении стоящей перед пользователем задачи);
- осознанная информационная потребность – появляется после осознания пользователем стоящей перед ним проблемы (осознанная потребность отличается от реальной, более того, пользователь может понимать имеющуюся проблему неправильно);
- выраженная информационная потребность – результат описания осознанной информационной потребности с помощью естественного языка;
- формализованная информационная потребность – это результат представления выраженной потребности средствами формального поискового языка ИПС.

Информационная потребность, сформулированная на информационно-поисковом языке, называется *запросом*. Запрос поступает в ИПС, и система возвращает некоторый ответ. Степень соответствия ответов поисковой системы запросу пользователя, а значит, и его информационной потребности, называется *релевантностью*.

Релевантность – это фундаментальное понятие теории информационного поиска. Наиболее часто используются следующие виды релевантности:

Когнитивная релевантность, или *пертинентность* – характеризует степень соответствия информации из документа и реальной информационной потребности пользователя. Это истинная и наиболее трудноопределимая релевантность. Все остальные виды релевантности являются ее приближениями.

Тематическая (предметная) релевантность – характеризует степень близости предмета (тематики) информационной потребности и найденного документа. Под тематикой здесь понимается область интересов пользователя, в пределах которой существует его информационная потребность.

Ситуационная релевантность – определяет полезность информационного ресурса для задачи, решаемой пользователем, с точки зрения временных затрат, способа взаимодействия пользователя с системой и т. п.

Системная (алгоритмическая) релевантность – характеризует степень близости между формализованной информационной потребностью (запросом) и найденным поисковой системой документом.

Теоретически релевантность можно измерить некоторым числом в интервале от нуля (полностью непертинентный документ) до единицы (полностью пертинентный документ). На практике же релевантность в основном определяется при оценке просматриваемых документов – подходит ли данный документ пользователю или не подходит. Таким образом, релевантность обычно характеризуется бинарным значением.

Существуют также методы оценки релевантности по многозначной шкале. В этом случае коллектив экспертов просматривает доступные документы и присваивает каждому из них некоторое значение релевантности, которое определяется для каждого запроса. Такая методика используется главным образом при сравнении эффективности нескольких поисковых систем.

Один и тот же документ может быть релевантным согласно одному определению релевантности и нерелевантным согласно другому. Так, документ, релевантный алгоритми-

чески, может оказаться непертинентным. Например, если в неспециализированной ИПС искать информацию о компьютерных манипуляторах типа "мышь" и использовать для этого запрос, состоящий из слова "мышь", то велика вероятность получения ряда непертинентных документов из области зоологии и географии, описывающих мышей-грызунов. При этом найденные непертинентные документы формально будут соответствовать запросу, так как в их тексте содержится термин запроса – "мышь".

Качество поиска в информационно-поисковых систем обычно характеризуется двумя критериями – полнотой и точностью. Полнота поиска определяется общим количеством найденных документов, а точность – соотношением между найденными релевантными и нерелевантными документами. Из-за многозначности определения релевантности абсолютного критерия точности поиска не существует, и в результатах любого поиска всегда присутствуют как релевантные, так и нерелевантные документы.

Нерелевантные документы, которые сравниваются с релевантными, иногда называются шумом, по аналогии с теорией передачи информации К. Шеннона. Релевантные документы в таком случае можно называть сигналом, а эффективность поиска оценивать по соотношению "сигнал - шум".

Поиск в Интернет. Для поиска нужной информации в Интернете существует отдельный вид сетевых сервисов – *поисковые серверы*, или *поисковые машины*.

Среди поисковых серверов принято различать поисковые индексы и каталоги.

Серверы-индексы работают следующим образом: регулярно прочитывают содержание большинства веб-страниц сети ("индексируют" их), и помещают их полностью или частично в общую базу данных. Пользователи поискового сервера имеют возможность осуществлять поиск по этой базе данных, используя ключевые слова, относящиеся к интересующей их теме. Выдача результатов поиска обычно состоит из выдержек, рекомендуемых вниманию пользователя страниц и их адресов (URL), оформленных в виде гиперссылок. Работать с поисковыми серверами этого типа удобно в том случае, если имеется четкое представление о предмете поиска.

Серверы-каталоги представляют собой многоуровневую классификацию ссылок, построенную по принципу "от общего к частному". Иногда ссылки сопровождаются кратким описанием ресурса. Как правило, возможен поиск в названиях рубрик (категориях) и описаниях ресурсов по ключевым словам. Каталогами пользуются тогда, когда не вполне четко знают, что именно ищут. Переходя от самых общих категорий к более частным, можно определить, с каким именно ресурсом сети следует ознакомиться.

Современные поисковые серверы предлагают и полнотекстовый поиск, и поиск по категориям, сочетая в себе, достоинства индексирующего сервера и сервера-каталога.

Работа с поисковыми серверами состоит в следующем. В адресной строке браузера набираете его адрес, в строке запроса набираете на нужном языке ключевые слова или фразу, соответствующие ресурсу или ресурсам сети, которые вы хотите найти. Затем нажимаете мышью на кнопку "Поиск" и в рабочее окно браузера загружается первая страница с результатами поиска. Обычно поисковый сервер выдает результаты поиска небольшими порциями, например, по 10 на одну страницу выдачи. Под списком рекомендуемых ссылок будет находиться ссылка, предлагающая перейти к следующей "порции" результатов поиска.

Формирование запросов является ключевым моментом при работе с поисковым сервером. С первого раза удачно задать вопрос поисковому серверу получается не всегда. Если запрос короткий и в нем присутствуют только часто употребляемые слова, может быть найдено очень много документов. Наоборот, если запрос окажется слишком детализированным или в нем будут использованы очень редкие слова, вы увидите сообщение о том, что ресурсов, отвечающих вашему запросу, в базе сервера не найдено. Постепенное сужение или расширение фокуса поиска через увеличение или уменьшение списка ключевых

слов, замена неудачных поисковых терминов на более удачные помогут улучшить результаты поиска.

Ключевые слова, составляющие поисковый запрос, обычно просто разделяются пробелами. Необходимо помнить, что различные поисковые сервера по-разному интерпретируют это. Некоторые из них отбирают по такому запросу только документы, содержащие все ключевые слова, то есть воспринимают пробел в запросе как логическую связку "и". Некоторые интерпретируют пробел как логическое "или" и ищут документы, содержащие хотя бы одно из ключевых слов. При формировании поискового запроса большинство серверов позволяют в явном виде указать логические связки, объединяющие ключевые слова, и задать некоторые другие параметры поиска. Логические связки обычно обозначаются с помощью английских слов "AND", "OR", "NOT". На разных поисковых серверах при формировании расширенного поискового запроса используется разный синтаксис – так называемый язык запросов. С помощью языка запроса вы можете указать, какие слова обязательно должны встретиться в документе, каких быть не должно, какие желательны.

Наиболее распространенной моделью поиска является булева модель, позволяющая составлять логические выражения из набора терминов. Найденные документы определяются в результате описанных запросом логических операций над множеством поисковых образов документов. Пользователь получает только те документы (В базе данных ИПС термины обычно хранятся в так называемой нормальной форме. Например, для существительных - это именительный падеж единственного числа. Одновременная нормализация терминов запросов и документов позволяет существенно упростить процесс их сравнения при поиске), чьи наборы терминов точно совпадают с соответствующими комбинациями терминов запроса.

Поисковые образы запросов связывают термины с помощью булевых операторов ("И" – "AND", "ИЛИ" – "OR", "И НЕ" – "AND NOT"). Эти операции производятся над множествами документов, содержащих тот или иной термин, определенный запросом. Для обозначения объединения множеств ("ИЛИ" в запросе) применяется символ Y, пересечения множеств ("И" в запросе) – I, разности множеств ("И НЕ" в запросе) – \ .

Например, оператор "И", соединяющий два термина запроса, означает следующее. Из множества всех документов нужно сначала выбрать два подмножества. Одно из них содержит первый термин запроса, а другое – второй. Затем определяется общая часть (пересечение) этих подмножеств, то есть те документы, в состав которых одновременно входят и первый, и второй термины из запроса.

Рассмотрим, например, такой запрос:

(((Microsoft and Word) or (Microsoft and Excel)) and Macintosh) and not Windows.

В данном случае выражение на ИПЯ означает следующее: нужно найти все документы, которые одновременно содержат либо сочетание "Microsoft Word", либо сочетание "Microsoft Excel", а также содержат слово "Macintosh", но не содержат слово "Windows".

Этот запрос можно разбить на две части:

1. Microsoft and Word and Macintosh and not Windows
2. Microsoft and Excel and Macintosh and not Windows

Выполнение первого запроса происходит в два этапа. Сначала находятся все документы, содержащие термины "Microsoft", "Word" и "Macintosh". Затем из найденных документов отсеиваются те, которые содержат слово "Windows". Второй запрос выполняется аналогично. В конце производится объединение результатов работы первой и второй частей исходного запроса.

Часто пользователь строит свой запрос, не используя каких-либо логических операторов, и просто перечисляет ключевые слова. В таком случае обычно предполагается, что все термины соединены логической операцией "И".

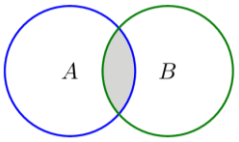
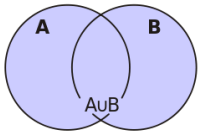
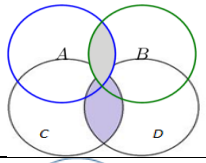
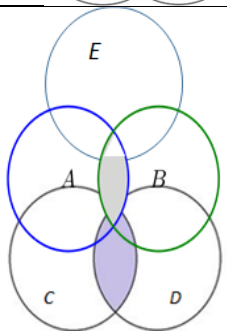
В некоторых поисковых системах вместо булевых операторов язык запросов позволяет использовать различные знаки. Так, знак "+" эквивалентен оператору "И", знак "-" – оператору "И-НЕ" и т. д.

В процессе поиска из исходного информационного массива выделяется часть, которая содержит найденные документы, соответствующие комбинациям терминов запроса. Какого-либо упорядочения (например, ранжирования по релевантности) не проводится: все выданные документы считаются одинаково важными.

Несколько типичных булевых стратегий поиска изображено в табл.1.

Таблица 1

Поиск с использованием булевых операторов

Формулировка запроса (термины)	Операции с множествами документов	Результат поиска
(a)	A	
(a AND b)	$A \cap B$	
(a OR b)	$A \cup B$	
(a AND b) OR (c AND d)	$(A \cap B) \cup (C \cap D)$	
((a AND b) OR (c AND d)) AND NOT e	$((A \cap B) \cup (C \cap D)) \backslash E$	

Здесь a, b, c, d, e – термины, из которых состоят запросы, а A, B, C, D, E - множества документов, содержащих эти термины (например, A - это множество документов, содержащих термин a, и т. д.).

ИПС, работающие с булевой моделью поиска, имеют ряд недостатков.

1.Обычные булевы запросы затрудняют варьирование глубины поиска с целью выдачи большего или меньшего количества документов в зависимости от требований пользователя. Для получения желаемого уровня эффективности необходимо найти правильную формулировку запроса: не слишком широкую и не слишком узкую. Оператор AND может привести к резкому сокращению числа найденных документов, а оператор OR, напротив,

может чрезмерно расширить запрос и выделить нужную информацию из информационного шума будет трудно.

2. При использовании булевой логики нельзя получить эффект от функций совпадения векторов, которые дают непрерывный спектр совпадений (полных, частичных или нулевых) между запросами поисковыми образами документов. Это обстоятельство приводит к жесткому требованию "все или ничего" на выходе.

3. Еще одним минусом является тот факт, что множество выданных документов не может быть представлено пользователю в ранжированном (Ранжирование – упорядочение результатов поиска по некоторому критерию соответствия их информационной потребности пользователя) виде, например, в порядке уменьшения сходства между документом и запросом. Документ либо полностью соответствует запросу, либо не соответствует совсем. Эта проблема может быть решена с помощью взвешенного булева поиска, при котором производится частичное ранжирование с использованием весов терминов - W_i. Результаты поиска располагаются в порядке уменьшения весов совпавших терминов [x_1, x_2].

Векторная модель является классическим представителем класса алгебраических моделей, реализована в 1968 г. Джерардом Солтоном в поисковой системе SMART (Salton's Magical Automatic Retriever of Text). Сокращенное обозначение TF*IDF – синоним наиболее распространенной современной векторной модели [26], основанной на математическом аппарате геометрии, в которой индексируемые текстовые ресурсы и запросы пользователей рассматриваются как векторы в пространстве слов, а релевантность – как расстояние между ними [12]. Векторные модели, в отличие от булевых, позволяют ранжировать результирующее множество документов запроса. В векторной модели каждому документу ставится в соответствие вектор $D_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$, где w_{ij} – вес j-го ключевого слова в i-документе, обычно вычисляемый по формуле нормированного представления TF*IDF $w_{ij} = a_{ij} log \frac{N}{dj}$, где a_{ij}- частота появления j-го ключевого слова в i-ом документе; d_j- количество документов, в которых встречается j-е количество слово; N – общее количество рассматриваемых документов. Аналогично для запроса Q вводится вектор $Q = \{q_1, q_2, \dots, q_n\}$, где q_j- 1, если j-е ключевое слово присутствует в запросе Q, иначе $q_j = 0$. Мера схожести документа D_i и запроса Q вычисляется как косинус угла между соответствующими векторами $r(D_i, Q) = (D_i, Q)/(\|D_i\| * \|Q\|)$, где (D_i, Q) – скалярное произведение, $\|D_i\|, \|Q\|$ – нормы векторов.

Вероятностная модель информационного поиска основана на теории вероятности и использует статистические показатели, характеризующие вероятность соответствия проиндексированных текстовых ресурсов запросу пользователя. Преимущество в том, что модель располагает документы в порядке убывания «вероятности оказаться релевантным». На практике эти модели не получили большого распространения. В рамках моделей вычисляется условная вероятность события, что документ соответствует данному запросу, то есть P(d|q) P (документ D релевантен | запрос Q) [4,5]. Для расчета используется формула Байеса и то, что вероятность $P(q)$ постоянна на протяжении всего поиска. Таким образом $P(d|q) = \alpha P(d)P(q|d), \alpha - const$. В качестве факторов, влияющих на безусловную релевантность документа $P(d)$, можно рассматривать его размер, источник, дату публикации. Вероятность запроса q при условии релевантности документа d зависит главным образом от веса ключевых слов запроса в документе d. Для ее расчета обычно принимают гипотезу независимости слов документа и запроса, что приводит к следующей формуле релевантностей: $R(d|q) = logP(d) + \sum_k logP(w_k d)$, $P(w_k|d)$ - вероятность появления k-го слова запроса в документе d.

В реальных поисковых системах, как правило, используется комбинация рассмотренных методов. При этом булев поиск используется для выделения из всего массива тех до-

кументов, которые содержат все слова запроса. Для определения релевантности документов и сортировки полученной выборки используются алгоритмы векторного и вероятностного поиска. Булева составляющая индексирования, сильно ускоряющая процесс поиска, – неотъемлемая часть поисковых систем, что говорит о необходимости создания и поддержки инвертированного индекса. Несмотря на описанные недостатки, булева модель поиска широко применяется в современных ИПС из-за простоты ее реализации.

Литература

[1]. Кнут Д. Искусство программирования. Т.3. Сортировка и поиск. – М.: Вильямс, 2000. – 844 с.

[2]. В.К. Иванов, К.В. Иванов. Введение в информационно-поисковые системы: метод. указания по изучению дисциплины "Мировые информ. ресурсы" для студентов спец. "Прикл. информатика (в экономике)" и "Информ. сист. и технол.". Ч. 2, Тверь: ТГТУ, 2005г

[3]. Вирт Н. Алгоритмы и структуры данных. - М.: Мир, 1989. – 360 с.

[4] Аграновский А.В., Арутюнян Р.Э., Хади Р.А. Современные аспекты проблемы поиска в текстовых базах данных // Телекоммуникации. 2003 – № 3.С. 25-30.

[5] Аграновский А.В., Арутюнян Р.Э. Способы индексации и поиска документов в интернет-порталах // Труды X Всероссийской научно-методической конференция «Телематика-2003». Санкт-Петербург. – 2003. – т.1. – С. 204-206.

RESEARCH METHODS OF INFORMATION RETRIEVAL

A. KASSEKEYEVA

Doctoral student of the 2nd course of the Eurasian National University named after L.N.Gumilyov, Astana, Kazakhstan

Abstract. A brief overview of current search methods is provided. The review includes the classic basics of modern search methods. The models of Boolean search, vector model, probabilistic model of information search are considered.

Keywords: information retrieval system, information retrieval language, search relevance, Boolean search model, vector search model, probabilistic search model.