# МОДЕЛИ ИНФОРМАЦИОННОГО ПОИСКА

*Шарапов Р.В.,Шарапова Е.В., Саратовцева О.А.*

E-mail: info@vanta.ru

*Муромский институт Владимирского государственного университета, г. Муром*

**Аннотация**. *В работе дается понятие информационного поиска, рассматриваются основные модели информационного поиска, используемые в поисковых системах. Особое внимание уделено булевским, векторным, вероятностным моделям и сетям вывода.*

## Models of information search

*Sharapov R., Sharapova E., Saratovtceva O.*

**Abstract**. *In work the concept of information search is given, the basic models of information search used in search systems are considered. The special attention is given Boolean, vector, likelihood models and networks of a conclusion.*

В связи с бурным ростом различных источников информации все большее значение приобретает проблема выделения нужных данных из огромного числа документов. Часто вопрос стоит уже не в том, чтобы найти книгу или статью, где упоминается интересующая человека информация, а в выборе из сотен и тысяч подобных источников наиболее подходящий и наиболее полно освящающих нужный вопрос документов. Решить эту проблему позволяют информационно-поисковые системы.

Информационный поиск - направление исследований, изучающее вопросы поиска документов, обработки результатов поиска, а также целый ряд смежных вопросов: моделирования, классификации, кластеризации и фильтрации документов, проектирования архитектур поисковых систем и пользовательских интерфейсов, языки запросов, и т.д.

Документ - это содержательно законченная единица информации, представленная на каком-либо естественном языке, которая идентифицируется уникальным образом [1]. Документ - это порция информации, которой оперируют информационно-поисковые системы.

Информационно-поисковая система - это комплекс программных средств, обеспечивающих избирательный отбор по заданным признакам документов, хранимых в электронном (оцифрованном) представлении.

Одним из ключевых понятий, характеризующим выбор того или иного метода анализа текстовой информации, а также реализацию конкретного варианта поиска, является модель поиска [4, 5].

Модель поиска - это сочетание следующих составляющих [1]:

1. способ представления документов;
2. способ представления поисковых запросов;
3. вид критерия релевантности документов.

Вариации этих составляющих определяют большое число всевозможных реализаций систем текстового поиска. Рассмотрим некоторые из них, наиболее популярные в настоящее время.

Простейшие модели поиска - это модели, в которых документ представляется в виде набора ассоциированных с ним внешних атрибутов. К простейшим моделям поиска относится модель дескрипторного поиска и модель, основанная на Дублинском ядре.

В простейших системах дескрипторного поиска представление документа описывается совокупностью слов или словосочетаний лексики предметной области, которые характеризуют содержание документа. Эти слова и словосочетания называются дескрипторами. Индексирование документа в таких системах реализуется назначением для него совокупности дескрипторов. При этом дескрипторы могут приписываться документу:

1. на основе его содержания;
2. на основе его названия.

Эти два процесса называются соответственно индексированием по содержанию и индексированием по заголовкам документов [1]. В некоторых дескрипторных системах индексирование документов осуществляется вручную экспертами в предметной области системы, в других оно выполняется автоматически. Представление документа в дескрипторных системах называется поисковым образом документа. Дескрипторные системы можно отнести к классу систем, ориентированных на библиографический поиск или поиск “по каталогу”.

Дублинское ядро (Dublin Core) [1,4] — это набор элементов метаданных, смысл которых зафиксирован в спецификации определяющего его стандарта. В терминах значений этих элементов можно описывать содержание различного рода текстовых документов. Первоначальная версия Дублинского ядра была предложена

в 1995 году на состоявшемся в Дублине (США) симпозиуме, организованном Online Computer Library Center (OCLC) и National Center for Supercomuting Applications (NCSA) для описания информационных ресурсов библиотечных систем. В модели поиска, основанной на Дублинском ядре, представлением k-го документа является множество пар Dk = { (Nik, Vik) } , где:

Nik — имя i-го элемента метаданных Дублинского ядра в описании содержания k-го документа;

Vik — значение этого элемента метаданных.

Представлением запроса также является множество пар некоторых элементов Дублинского ядра и их значений Q = { (Nj, Vj) } , где:

Nj — имя j-го элемента метаданных Дублинского ядра в описании пользовательского запроса;

Vj — значение этого элемента метаданных.

Критерий релевантности k – го документа выглядит следующим образом:

$$Q \subseteq D_k.$$

Модели, основанные на классификаторах. Это одна из разновидностей простейших моделей поиска. Документ в данной модели представляется в виде совокупности ассоциированных с ним атрибутов. Атрибутами являются идентификаторы классов, к которым относится данный документ. Классы формируют иерархическую структуру классификатора. Запрос может быть представлен двумя способами:

1. Простой вариант - запросом является идентификатор какого-либо класса из заданного классификатора. Критерий релевантности документа запросу - класс документа совпадает с классом в представлении запроса или является его подклассом.
2. Сложный вариант — в запросе можно указать несколько классов классификатора. Критерий релевантности документа запросу - класс документа совпадает с каким-либо из указанных в запросе классов или является его подклассом.

Модели, основанные на классификаторах, близки к булевским моделям.

Булевские модели. В булевских моделях поиска пользователь может формулировать запрос в виде булевского выражения, используя для этого операторы И, ИЛИ, НЕТ. Термы запроса зависят от конкретного варианта модели поиска. В булевской модели, ориентированной на поиск “по тексту”, термами будут слова, соответственно, критерием релевантности будет условие вхождения некоторого слова или словосочетания в текст документа. В булевской модели, ориентированной на поиск по классификаторам, термами выражения будут идентификаторы классов классификатора. В булевской модели поиска с использованием Дублинского ядра термом будет значения элементов метаданных. Документ, имеющий совпадающие значения элементов метаданных со значениями, заданными в запросе, считается релевантным [4]. В общем случае критерием релевантности документа запросу в булевских моделях поиска является истинность булевского выражения, заданного в запросе. Одним из несомненных достоинств булевской модели поиска является простота ее реализации. Главными недостатками считаются:

1. отсутствие возможности ранжирования найденные документы по степени релевантности, поскольку отсутствуют критерии ее оценки.
2. сложность использования - далеко не каждый пользователь может свободно оперировать булевскими операторами при формулировке своих запросов.

Стоит отметить, что предпринимались попытки усложнения булевской модели поиска для обеспечения возможности ранжирования множества выдаваемых пользователю документов. А именно, предложено несколько вариантов так называемых расширенных булевских моделей [4]. В этих моделях вводятся специальные обобщения булевских операторов, позволяющие придать повышенный вес документам, в точности удовлетворяющих булевскому выражению запроса, и пониженный вес - всем остальным документам [1].

Векторные модели. В настоящее время векторные модели являются самыми распространенными и применяемыми на практике моделями поиска. Векторные модели, в отличие от булевых, без труда позволяют ранжировать результирующее множество документов запроса. Суть таких моделей сводится к представлению документов и запросов в виде векторов. Каждому терму ti в документе dj и запросе q сопоставляется некоторый неотрицательный вес wij (wi для запроса). Таким образом, каждый документ и запрос может быть представлен в виде k - мерного вектора [2]:

$$\vec{d_j} \stackrel{def}{=} (w_{1j}, w_{2j}, ..., w_{kj})$$

где k- общее количество различных термов во всех документах. Согласно векторной модели, близость документа di к запросу q оценивается как корреляция между векторами их описаний. Эта корреляция может быть вычислена, например, как скалярное произведение соответствующих векторов описаний [3]. Существуют различные подходы к выбору указанных весов. Одним из самых простых является использование нормализованной частоты данного терма в документе:

$$w_{ij} = \frac{n_{ij}}{N_j},$$

где nij — количество повторений данного терма в документе; Nj - общее количество всех термов в документе.

Более сложные варианты расчета весов учитывают частоту использования данного терма в других документах коллекции, т. е. учитывают дискриминационную силу терма [2]. Но эти варианты возможны только при наличии статистики использования термов в коллекции. Вариации всевозможных способов назначения весов термов и оценки меры близости векторов определяют широкий спектр различных модификаций данной модели поиска.

Вероятностные модели. Впервые идеи таких моделей были предложены в 1960 году. В их основе лежит принцип вероятностного ранжирования (Probabilistic Ranking Principle, PRP). Этот принцип заключается в следующем - наивысшая общая эффективность поиска достигается в случае, когда результирующие документы ранжируются по убыванию вероятности их релевантности запросу. Сначала для каждого для каждого документа оценивается вероятность того, что он релевантен запросу, а затем по этим оценкам выполняется ранжирование документов.

Существуют различные способы получения этих оценок, а также дополнительные предположения и гипотезы на основе априорных сведений относительно документов коллекции, которые и определяют конкретную реализацию вероятностной модели поиска. Например, эта оценка может быть вычислена, в соответствии с теоремой Байеса, по некоторой функции вероятностей вхождения термов данного документа в релевантные и нерелевантные документы. С помощью запроса определяется вероятность вхождения заданного терма в релевантные документы, а по полной коллекции документов определяется вероятность вхождения этого терма в нерелевантные документы [1]

Сети вывода. Так же, как и вероятностные модели, сети вывода основаны на принципе вероятностного ранжирования результирующих документов поиска [4]. Главное их отличие от вероятностных моделей заключается в том, что используется оценка не вероятности релевантности документа запросу, а вероятности того, что он удовлетворяет информационным потребностям пользователя.

В рамках данной модели процесс поиска документов описывается как процесс рассуждений в условиях неопределенности. В процессе такого рассуждения оценивается вероятность того, что информационные потребности пользователя, выраженные с помощью одного или нескольких запросов, удовлетворены.

Сеть вывода основана на Байесовской сети, которая включает узлы четырех видов. Узлами первого вида являются документы коллекции, изученные пользователем в процессе поиска. Узлами второго вида являются термы, которыми описывается содержание документов. Узлами третьего вида являются запросы, состоящие из термов, которыми описывается содержание документов. Узел четвертого типа в сети только один, и он соответствует информационным потребностям пользователя, которые не известны поисковой системе. Все узлы первого и второго вида формируются заранее для заданной коллекции. Узлы третьего вида и их связи с узлами термов, описывающих документы, и узлом информационных потребностей формируются для каждого конкретного запроса.

После того, как сеть построена, осуществляется оценка документов коллекции. Это реализуется распространением по сети оценки вероятности узла конкретного документа. Результатом распространения является вычисление вероятности узла информационных потребностей. При этом оценка для каждого документа строится независимо от оценок других документов, с учетом матриц описывающих связи между узлами документов и узлами термов, узлами термов и узлами запросов. Процесс оценки повторяется для каждого документа, затем они ранжируются на основе вычисленных оценок вероятности узла информационных потребностей [1].

В зависимости от постановки задачи поиска могут применяться любые из рассмотренных моделей. На наш взгляд для большинства задач поиска стоит использовать векторные модели, так как они позволяют не просто искать документы, но и осуществлять их ранжирование в зависимости с соответствием запросу пользователя.

**Литература**


1. Когаловский М. Р. Перспективные технологии информационных систем. - М.: ДМК Пресс; М.: Компания АйТи, 2003. – 288 с.
2. Некрестьянов И.С. Тематико - ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 / Санкт-Петербургский государственный университет - СПб., 2000. –
3. Дубинский А.Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. - 2001. - №4. - С. 77-83
4. Чугреев В.Л. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации: Диссертационная работа к.т.н.: 05.13.01 / Санкт - Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина) – СПб., 2003. – 156 с

5. Сэлтон Г. Автоматическая обработка, хранение и поиск информации: Пер. с англ. / Под ред. А.И. Китова. – М.: Советское радио, 1973. – 560 с.