

Вопрос 4. Поиск информации в информационных системах организации

1) Основные принципы информационного поиска. Проблема поиска документа возникает в любом хранилище данных. При создании систем хранения применяются две модели: иерархическая и гипертекстовая. Иерархическая модель хранения подразумевает многоуровневую рубрикацию системных ресурсов. Для определения пути к необходимому ресурсу используются описания, составленные при отправке документа на хранение. Гипертекстовая модель позволяет связывать документы ссылками, расположенными непосредственно в тексте документа.

При больших объемах информации, высокой скорости их обновления и разнородности запросов очевидны недостатки этих моделей. Многоуровневая рубрикация и простановка ссылок выполняется высококвалифицированными специалистами, поэтому объем обработанных ими документов становится ограниченным. Связанные документы ограничиваются определенной предметной областью, которая может разным образом трактоваться составителем и пользователем. При поиске документа целесообразно просматривать множество документов, содержащих лишь ссылки на другие ресурсы.

Этих недостатков лишены информационно-поисковые системы; будучи однажды созданными, они работают автономно. Принцип взаимодействия ИПС с пользователем заключается в том, что пользователь вводит в этой системе запрос, обрабатываемый системой, и получает список указателей на документы, удовлетворяющие запросу. Список может быть отсортирован по релевантности – степени соответствия документа запросу.

Основные принципы информационного поиска заключаются в том, что создается массив указателей на информационные ресурсы. Указатель (индекс) содержит некое свойство документа и ссылки на документы, обладающие этим свойством. Например, авторский указатель позволяет получить ссылки на работы определенного автора, предметный указатель – выбрать документы, затрагивающие определенные понятия (предметы). Процесс создания указателей называется индексированием, а термины, используемые для индексирования, называют терминами индексирования. В авторском указателе роль терминов индексирования выполняют фамилии авторов, работы которых хранятся в фонде. Совокупность используемых терминов индексирования называется словарем. Массив указателей, составленный после индексации информационных ресурсов, именуется индексной базой.

К индексной базе обращаются посредством запросов. Так, запрос пользователя должен быть переведен на язык индексирования. При поиске происходит сопоставление запроса с имеющимися данными и пользователю выдается список ссылок на подходящие ресурсы. Для повышения эффективности работы системы словарь и индекс должны быть упорядочены по системе, наиболее отвечающей задачам поиска в конкретной предметной области.

Первые информационно-поисковые системы были созданы в 1970–1980-х гг. и продолжают развиваться сегодня.

Любая информационно-поисковая система использует предметный указатель, позволяющий отыскивать документы, касающиеся некоего «предмета». Для составления предметного указателя анализируется содержание документа и определяется «предмет» или «предметы», о которых в документе идет речь. Названия этих предметов переводятся на информационно-поисковый язык (ИПЯ), в результате получают поисковый образ документа (ПОД). Проиндексировав (создав поисковые образы) все информационные ресурсы, получают индексную базу – основной массив данных ИПС.

Процесс поиска заключается в сопоставлении запроса пользователя с имеющимися данными, полученный запрос также переводится на информационно-поисковый язык. После сопоставления переведенного на ИПЯ запроса и поисковых образов документов

пользователь получает список ссылок на документы, соответствующие по мнению системы его запросу. Поиск происходит не по тексту документов, а по их поисковым образам, составленным на ИПЯ. Поэтому качество поисковой системы зависит в первую очередь от ее информационно-поискового языка. В состав информационно-поискового языка входят:

- 1) словарь индексационных терминов – множество терминов индексирования;
- 2) кодовый словарь – множество кодовых терминов;
- 3) словарь входов – множество входных терминов;
- 4) вспомогательные средства языка индексирования – используемые совместно с индексационными терминами для расширения или сужения определенных понятий;
- 5) правила использования языка индексирования.

Для повышения эффективности поиска словарь должен быть контролируемым, т.е. должен быть организован таким образом, чтобы полнота и точность поиска были оптимальными. Очевидно, что организация словаря зависит от многих факторов – предметной области, в которой будет функционировать ИПС, характера интересов пользователей, степени их подготовки и т.д.

Для улучшения результатов поиска необходимо определить степень специфичности терминов при индексации. Как правило, применяют два принципа – использование наиболее специфического термина, соответствующего объему и содержанию отражаемого понятия, и избыточное индексирование. В избыточном индексировании поисковый образ дополняется терминами, связанными с основным. Могут использоваться термины, связанные как с основным отношением обобщения или спецификации, так и ассоциативной связью. Дополнение поискового образа терминами с ассоциативной связью увеличивает полноту поиска, но неизбежно снижает его точность. К недостаткам избыточного индексирования относятся также увеличение объема поисковых образов. Для устранения этой проблемы во многих ИПС используется избыточное индексирование не документов, а запросов.

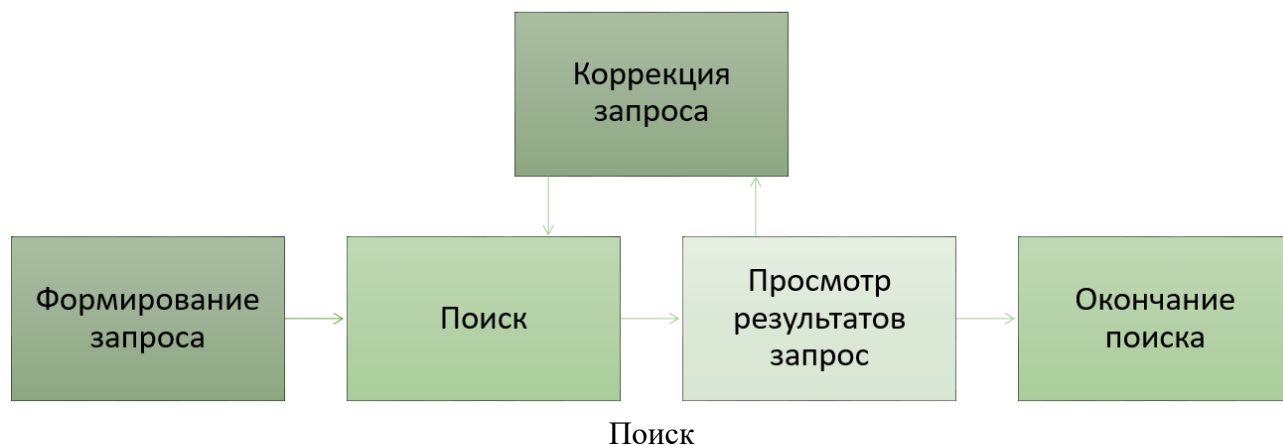
Предметное индексирование не исключает использование при создании поискового образа атрибутов документа. Это могут быть такие атрибуты, как данные об авторе, дата публикации, язык публикации и т.д.

Точность и полнота поиска зависят не только от характеристик самой ИПС, но и от того, как создается запрос. Идеальный запрос может быть составлен пользователем, в полном объеме знакомым с интересующей его предметной областью, а также с применяемой ИПС. Однако такому пользователю ИПС, очевидно, не требуется. Остальные пользователи вынуждены довольствоваться или низкой точностью поиска, или низкой полнотой.

Для повышения качества поиска существуют различные методы. Наиболее употребляемый из них – использование логических операторов И, ИЛИ, НЕ. Это довольно простой способ повысить релевантность выдаваемых документов. Недостатком считается плохая масштабируемость. Оператор И может сильно сузить поиск, а оператор ИЛИ – сильно расширить. Степень точности и полноты поиска зависит от того, насколько общие термины участвовали в формулировке запроса. Может быть неверным использование как наиболее общих терминов (возрастает уровень информационного шума), так и слишком специфичных терминов (снижается полнота поиска). Применение слишком специфичных терминов чревато еще и тем, что в словаре ИПС данного термина может не оказаться. В общем виде процедура поиска – процедура итеративная, т.е. за этапом выдачи результатов поиска следует коррекция запроса, поиск по этому запросу и т.д. Коррекция запроса происходит в зависимости от количества полученных документов и их релевантности и может выполняться как пользователем, так и самой информационно-поисковой системой.

В зависимости от соотношения полноты и точности найденных документов пользователь может сузить или расширить область поиска, перейдя к более общим или, наоборот, более специфичным терминам, а также использовав родственные понятия. В

случае поиска по нескольким терминам такая коррекция области поиска может происходить по одному из нескольких терминов, что позволяет изменять эту область достаточно плавно. Может оказаться полезным знание пользователя о наличии определенно релевантных документов. Не обнаружив их в списке найденных документов, область поиска следует расширить. Запрос корректируется системой информационного поиска на основании анализа документов, отмеченных пользователем как наиболее точно отвечающих его потребностям. В таком случае при следующем поиске система ищет те документы, где помимо заданных в первоначальном запросе содержатся термины, встречающиеся в документах, отмеченных пользователем. Улучшить результаты поиска можно различными способами, если функции для этого предоставляются интерфейсом информационно-поисковой системы.



В последнее время во многих ИПС появилась функция подсказки при вводе текста поискового запроса, учитывающая ранее введенные этим пользователем запросы по сходной тематике за некоторый период времени.

Основные тенденции (подходы) при организации поиска в ИПС:

1. Ориентация на создание ИПС, функционирующих в сети Интернет или корпоративных сетях. Это позволяет решить вопросы ограничений по размеру баз данных и индексов, а также скорости обновления информационных массивов и непосредственно выполнения поисковых запросов (на локальных компьютерах имеются ограничения как по объему внешней памяти, так и по производительности).

2. Использование в качестве хранилища информации готовой системы управления базами данных (СУБД), которая представляет собой совокупность программных и лингвистических средств общего или специального назначения, обеспечивающих управление созданием и использованием баз данных. Как правило, используются клиент-серверные СУБД, которые обычно располагаются на сервере, и все клиентские запросы на обработку данных обрабатываются централизованно.

3. Использование готовых систем поиска, как правило, бесплатных (например, Sphinx, Solr и др.), адаптированных для работы с правовой информацией.

4. Упрощение пользовательского интерфейса ИПС или создание интерфейса для различных категорий пользователей (обычный пользователь, юрист).

5. Организация и реализация механизма всевозможных "подсказок" непосредственно при вводе пользователем своих запросов.

6. Проведение анализа поисковых запросов пользователя и результатов поиска в целях использования механизмов управления релевантностью (соответствием поискового намерения, заложенного в запросе, результатам этого запроса, выданным системой поиска).

7. Использование классификаторов и словарей при организации поиска по реквизитам.

8. Анализ поискового запроса непосредственно после его ввода, выделение из него реквизитов, используемых (сформированных) в базе данных, сокращений различных слов и аббревиатур, их использование при поиске.

9. Использование в рамках поиска автоматического машинного переводчика запросов с одного языка на другой (например, с русского на белорусский) при хранении в базе данных разных документов на разных языках.

Указанный перечень направлений организации и развития поисковых возможностей ИПС не является исчерпывающим, а только отражает рассуждения автора публикации по данной проблеме.¹

2) Интерфейс системы. Важным фактором, во многом определяющим эффективность поиска, может быть вид представления информации в программе, т.е. ее интерфейс. По форме диалога, способу задания условия отбора и механизму поиска программные средства можно разделить на системы рубрикационного типа и структурно-логические системы.

Первые реализуются интерфейсом в виде иерархических последовательно раскрывающихся списков, через которые обеспечивается доступ к тематически связанным группам документов. Раскрывая очередную рубрику и перемещаясь таким образом по тематической иерархии, пользователь уточняет предметную область и увеличивает (усредненно) степень точности соответствия выдаваемых документов и информационной потребности. Предопределенность соотнесения документов с отдельными рубриками компенсируется логичностью естественно-научной классификационной схемы, заменяющей пользователю путеводитель.

Структурно-логические методы формирования запроса используются для работы с базами данных структурированной информации, когда каждый документ состоит из многих информационных полей, возможно, разного типа. Критерий отбора строится как логическая комбинация простых, сводящихся к проверке условия присутствия или отсутствия в документе слов (имен собственных или имен понятий, определяющих предмет поиска).

При составлении запроса к системе используют либо «меню-ориентированный» подход, либо командную строку. Первый позволяет ввести список терминов, как правило, разделяемых пробелом, и выбрать тип логической связи между ними. Логическая связь распространяется на все термины. Многие ИПС позволяют сохранять запросы пользователя – в большинстве систем это просто фраза на ИПЯ, которую можно расширить за счет добавления новых терминов и логических операторов. Но это только один способ использования сохраненных запросов, называемый расширением, или уточнением, запроса. Для выполнения этой операции традиционная ИПС хранит не запрос как таковой, а результат поиска – список идентификаторов документов, который объединяется или пересекается со списком, полученным при поиске документов по новым терминам.²

¹ Радиванович Н. Н. Статья: Модели поиска информации в информационно-правовых системах. Журнал «Информационное право» № 3, 2018.

² Гаврилов М. В. Информатика и информационные технологии. М., 2015.