Занятие №14

Индексирование предоставляет возможность классифицировать документы посредством метаданных и словарного индекса[1] текста, извлечённого из документа.

Индексирование (англ. «indexing») − это процесс выражения главного предмета или темы текста какого-либо документа в терминах информационно-поискового языка.

Применяется для облегчения поиска необходимого текста среди множества других. Индексация существует, главным образом, для поддержки развитых возможностей поиска документов, поскольку одно из главных условий быстрого и качественного поиска − это создание индекса документа.

Индексирование проводится как целого документа, так и его части. Для индексирования нередко используются заглавия текстов. При индексировании опускаются сопутствующие предметы или темы. Это служит причиной того, что при поиске ненайденными остаются тексты, для которых предмет или тема информационного запроса является не главной, а сопутствующей.

Различают два основных типа индексирования − классификационное и координатное.

При *классификационном (*предкоординатном, англ. «pre-coordination») *индексировании*, или классифицировании, тексты в зависимости от их содержания включаются в соответствующий класс (один или несколько), в котором собираются все тексты, имеющие в основном одинаковое смысловое содержание. Каждому такому тексту присваивается индекс этого класса, служащий далее его поисковым образом.

Предкоординация заключается в построении словарного состава информационно-поискового языка до его использования при

[1] Индекс (англ. «index», «code2, «notation», «mark», «symbol») − условный знак (в т.ч. слово, словосочетание, цифра, буквенный или буквенно-цифровой код и т.п.), обозначающий определённое понятие и используемый для записи результатов классифицирования, а также идентификации объектов поиска в информационно-поисковых массивах.

индексировании. Оно характеризуется применением словосочетаний и фраз, выражающих сложные понятия и представлено известными классификационными системами: УДК, МКИ, ГРНТИ, ББК, ТБК и др.

Координатное (посткоординатное) *индексирование* (англ. «coordinate (postcoordinate) indexing») заключается в построении словарного состава путём разделения сложных понятий на составные элементы и последующего объединения полученных лексических единиц ИПЯ при индексировании документов вводимых в информационно-поисковые массивы и запросов путём использования логических операторов и других средств, представляющих его синтаксис.

При координатном индексировании основное смысловое содержание текста выражается перечнем полнозначных слов, выбираемых либо из самого текста или его заглавия, либо из специального нормативного словаря. В первом случае такие лексические единицы называются ключевыми словами, а во втором − дескрипторами.

Информационно-поисковый язык – это искусственный язык, предназначенный для формирования специальных характеристик (индексов, дескрипторов, тезаурусов, ключевых слов и др.) объектов (документов, фактов и т.п.), как правило хранящихся в базе данных, с целью обеспечения поиска с получением результата, релевантного запросу пользователя.

ИПЯ используется для выражения содержания документов и информационных запросов или описания фактов, и обеспечивают поиск в автоматизированных информационных системах, в том числе в АБИС. К видам ИПЯ относят: ключевые слова, дескрипторы, рубрикаторы, тезаурусы.

Ключевое слово, недескриптор (англ. «keyword», «non-descriptor») − это лексическая единица[2], выбираемая из обрабатываемого (индексируемого) текста (вводимых в систему документов и запросов на поиск) а не словаря. Ключевые слова должны составляться на основе специальных правил и построенных на их основе технологических инструкций, обеспечивающих однозначность их понимания и применения.

Дескриптор (англ. «descriptor») − это лексическая единица дескрипторного ИПЯ, выбираемая при индексировании не из обрабатываемого текстового или другого материала, а из специального словаря.

Дескрипторы отличаются от ключевых слов тем, что им придана смысловая однозначность.

Каждое ключевое слово или дескриптор обозначают класс, в который потенциально входят все тексты, где в выражения основного смыслового содержания входит это слово. Логическое произведение классов, которые обозначены всеми словами, выражающими в совокупности основное смысловое содержание текста, образует некоторый сложный класс.

[2] Лексическая единица (ЛЕ) − обозначение отдельного понятия в естественном или специально созданном искусственном языке, например ИПЯ. Лексическая единица может иметь вид слова, устойчивого словосочетания, аббревиатуры, символьного кода и т.п.

Построенный таким способом сложный класс обозначается перечнем ключевых слов или дескрипторов, и этот перечень служит поисковым образом данного текста или выражением на информационно-поисковом языке смыслового содержания запроса.

Разновидностью координатного индексирования является *пермутационное*, или *циклическое, индексирование*, основанное на использовании ключевых слов заглавия текста и заключающееся в том, что все ключевые слова заглавия вместе с контекстом поочерёдно выводятся в поисковую колонку.

В этой колонке ключевые слова даются в алфавитном порядке. На основе координатного индексирования созданы и более сложные информационно-поисковые языки. Основное преимущество координатного индексирования перед классификационным заключается в том, что координатное индексирование не создаёт никаких затруднений при поиске текстов по любому, заранее непредусмотренному сочетанию признаков.

Особым типом индексирования следует считать раскрытие смыслового содержания текста через приводимую вместе с ним библиографию − имена авторов и библиографические описания их работ, на которые ссылается автор данного текста. Такое индексирование служит основой для составления указателей цитированной литературы − эффективного инструмента как для поиска документов, так и для решения других (науковедческих, прогностических и т. д.) задач.

Процесс индексирования включает:

1. Анализ содержания индексируемого материала и выбор из него существенных для его понимания лексических единиц;

2. Формирование перечня ключевых слов, используемых при свободном индексировании;

3. Нормализацию ключевых слов по форме и содержанию при помощи словаря используемого ИПЯ пред- или посткоординатного типа;

4. Избыточное индексирование[3];

5. Заполнение рабочего листа с введением в него грамматических средств.

В зависимости от объекта и содержания процесса индексирования его результатами являются: поисковый образ документа (ПОД), поисковый образ запроса (ПОЗ) или поисковое предписание (ПП).

Элементарной единицей информационного поиска, как правило, является документ. Очевидно, что хранящаяся в различных документах текстовая информация в общем случае является слабо структурированной.

Чтобы можно было находить нужный пользователю документ, последний должен включать некоторые специальные компоненты,

[3] Избыточное индексирование (англ. «redundant indexing») текстов осуществляется путём включения в поисковый образ документа и поискового предписания близких по смыслу лексических единиц ИПЯ для повышения полноты поиска.

позволяющие его идентифицировать. С этой целью формирую поисковый образ документа.

ПОД – характеристика, кратко выражающая основное смысловое содержание документа. Простейшими ПОД являются заглавие документа и фамилия его автора.

Метод, в котором основное смысловое содержание документа выражается в краткой форме, не может обеспечить нахождение всех документов, содержащих требуемую информацию. При этом среди найденных документов попадаются фактически не отвечающие на данный информационный запрос. Эти документы образуют так называемый «*поисковый шум*». При автоматизированном поиске наилучшие результаты достигаются, когда он осуществляется по ПОД и (или) по их рефератам. В другом случае выписанные словосочетания и слова сравниваются с фиксированным словарем. При этом слова, ненайденные в словаре устраняются, а оставшиеся сортируются по алфавиту.

ПОЗ – это сформулированный информационный запрос, являющийся такой же краткой характеристикой, поисковым предписанием.

ПОЗ является формально описанной моделью информационной потребности пользователя, поэтому ПОЗ и ПОД должны соответствовать друг другу.

Информационный поиск – это поиск документов, сведений о них или фактов, соответствующих информационному запросу.

Считается, что первые средства навигации в текстовой информации появились в Библии. В 1247 году (Hugo de St. Caro) было задействовано 500 монахов для составления конкорданса ключевых слов к Библии. Затем, в средние века индексирование стало применяться в журналах (журнальные индексы – Королевское научное общество, 1600 годы).

На первый взгляд кажется, что любой человек, пришедший в библиотеку, без труда может найти ту или иную информацию. Тем не менее, поиск качественной, адекватной запросам информации в реальной жизни не так легок. Поскольку автоматизированная система является инструментом, используемым человеком при поиске, а не интеллектуальным автоматом для поиска информации, эффективность её использования зависит от того, насколько хорошо человек знает природу объектов и свойства инструмента, посредством которого он с этими объектами работает. Обычно информационный поиск производится не по текстам документов, а по кратким характеристикам содержания или определённым внешним признакам документов. В этом случае процедура информационного поиска сводится к сопоставлению ПОД с заданным ПОЗ.

Обычно поисковые процессы включают четыре стадии:

1) формулировка (осуществляется до начала поиска);

2) начало поиска;

3) обзор полученных результатов;

4) модификация поиска (после обзора полученных результатов может потребоваться уточняющий поиск).

Более удобная нелинейная схема поиска информации состоит из следующих этапов:

1) фиксация информационной потребности на естественном языке;

2) выбор поисковых сервисов сети и формализация записи информационной потребности на конкретных информационно-поисковых языках;

3) выполнение созданных запросов;

4) предварительная обработка полученных списков ссылок на документы;

5) обращение по выбранным адресам за искомыми документами;

6) предварительный просмотр содержимого найденных документов;

7) сохранение релевантных документов для последующего изучения;

8) извлечение из релевантных документов ссылок для расширения запроса;

9) Изучение всего массива сохраненных документов;

10) возврат к первому этапу, если информационная потребность не полностью удовлетворена.