Конспект «Особенности текстовых документов и эффективность поиска информации в Интернете»

# Системы на основе индексирования.

Поисковый индекс любого сайта зависит непосредственно от его контента, внешних и внутренних ссылок, наличия изображений, графиков и прочих материалов. Вводя в поисковой строке запрос, интернет-пользователь обращается к индексу. Затем из него на основании данных осуществляется ранжирование результатов поиска, список страниц, которые выстраиваются по мере уменьшения релевантности запросу.

Представьте, что Всемирная Сеть – это большая библиотека. В ней обязательно должен быть специальный каталог, который делает поиск необходимых материалов гораздо проще. Всем книгам, которые находятся в библиотеке, принадлежит свой шифр. Все шифры объединены темами, разделами и прочими параметрами.

Когда человек приходит в библиотеку и просит книгу на определенную тематику (делает запрос), библиотекарь идет к нужному разделу, достает все книги, соответствующие ему, и подбирает для читателя самую подходящую.

По аналогичному принципу работают и поисковики: пользователь делает запрос, поисковая машина достает все соответствующие страницы и выдает наиболее релевантные.

Индексированием занимаются два типа роботов:

* Основные. Изучают контент, содержащийся на страницах интернет-ресурса;
* Быстрые (быстророботы). Анализируют и индексируют новые материалы, которые добавились после обновления сайта.

Чтобы веб-ресурс был проиндексирован в самых популярных поисковых системах, вебмастеру нужно сообщить о своем проекте:

* Добавить сайт на индексацию, самостоятельно заполнив специальную форму поисковика через такие сервисы, как Google Webmaster, Яндекс.Вебмастер и т.п. Такой способ индексации проходит медленно, от двух недель и дольше, потому что проект попадает в очередь.
* Отправить ресурс на индексацию посредством размещения ссылок на других веб-сайтах. Данный метод является самым эффективным, потому что найденные таким образом страницы боты считают полезными, и индексируют их куда быстрее – не более двух недель, а если повезет, то и вовсе за 12 часов.

# Семантически-навигационные системы

Семантический поиск — способ и технология поиска информации, основанная на использовании контекстного (смыслового) значения запрашиваемых фраз, вместо словарных значений отдельных слов или выражений при поисковом запросе. Улучшение результатов поиска при обработке запросов достигается за счет более точной интерпретации поисковых намерений пользователя.

Атрибуты семантического поиска (отличающие его от не семантического поиска) не обязательно являются его преимуществами. Некоторые из атрибутов могут повысить точность поиска за счет использования большого количества времени (или других ресурсов). Соответственно, эти десять атрибутов являются лишь характерными чертами семантического поиска, дающими преимущество только в идеальных условиях.

* Обработка морфологических вариаций.
* Обработка синонимов с правильными значениями.
* Обработка обобщений.
* Обработка концептуального множества.
* Обработка базы знаний.
* Обработка запросов и вопросов, заданных простым языком.
* Возможность определения непрерывного параграфа и наиболее соответствующего предложения.
* Возможность к адаптации и органичному прогрессу.
* Способность работать, не полагаясь на данные статистики, поведение пользователей и других искусственных средств.
* Способность обнаруживать результаты своей собственной деятельности.

Суть семантического поиска заключается не только в форме вопросов, заданных поисковой системе. Поскольку, веб — это набор неструктурированных HTML-страниц, в основе семантического поиска также лежит и базовая информация.

Одной из самых интуитивных и наиболее точных семантических баз данных считается Freebase. Freebase работает не только через текстовый поиск, а что наиболее важно, и через — MQL (Metaweb Query Language). MQL работает подобно JSON (текстовый формат обмена данными), но обладает более широкими возможностями. С его помощью можно составить любой запрос в Freebase, ответом на который будет тот же запрос, но уже с прикрепленными результатами поиска. Powerset, по сути, это тематическая база данных, которая работает с определенной структурированной информацией[5].

Google, в первую очередь, ориентируется на статистическую частоту запросов и почти не принимает во внимание семантику. Особо стоит отметить новую систему SearchMonkey от Yahoo! Эта система ничего не добавляет к найденным результатам, но использует семантические аннотации для более полного, интерактивного и полезного пользовательского интерфейса.

Компании Hakia и Powerset постоянно работают над улучшением возможностей поиска. Их стратегия включает создание структур подобных Freebase с дальнейшим проведением поиска по наиболее релевантным результатам на естественном языке. Основное отличие заключается в том, что Hakia (как и другие) использует технологию для поиска по всей сети, а Powerset замкнул свой поиск только на Wikipedia[5].

# Поисковый образ документа.

Поисковый образ документа - поисковый образ, выражающий основное смысловое содержание документа. Поисковый образ документа содержит:   
- признаки, необходимые для поиска документа по запросу;   
+ идентифицирующие и другие сведения о документе: выходные данные, тип, язык и т.д.

# Поисковый образ запроса

Поисковый образ запроса - поисковый образ, выражающий смысловое содержание информационного запроса.

# Информационно-поисковый язык: алфавит, лексика, грамматика.

Информационно-поисковый язык состоит из алфавита, лексики и грамматики. Алфавит - система знаков, используемая для записи слов. В информационно-поисковых языках могут быть использованы: буквы латинского алфавита; кириллица; цифры; пунктуационные знаки. Лексика (словарный состав) - совокупность слов, входящих в состав языка, называемых также лексическими единицами. Лексическая единица - слово или семантически неделимое словосочетание, выражающее какое-либо понятие. Грамматика - набор правил, по которым из конечного числа элементов определенного типа (например, букв или слов) можно получить язык для выражения содержания документов или запросов или описания фактов с целью последующего поиска. Грамматика подразделяется на морфологию и синтаксис. Морфология - правила построения и изменения слов. Синтаксис - правила построения и изменения соединения слов (построение фраз). Слова любого языка в процессе отображения предметов реального мира вступают между собой в определенные отношения.

# Парадигматические соотношения.

Парадигматические отношения - логические отношения, существующие между лексическими единицами информационно-поискового языка независимо от контекста, в котором эти лексические единицы употребляются. Эти отношения обусловлены предметно-логическими, а не языковыми факторами, т.е. относятся к категории внеязыковых связей. Примеры парадигматических отношений: часть - целое (отдел - организация); род - вид (ценная бумага - акция); причина-следствие; функциональное сходство; ассоциации. Учет парадигматических отношений необходим для правильного выбора и точного употребления слов. Поэтому в семантически развитом информационно-поисковом языке должны быть в явном виде выражены важнейшие отношения между терминами, иначе при отображении текста документа может произойти потеря или искажение смысла документа. Например, при поиске нормативных документов, касающихся термина "акция", для увеличения полноты поиска возможно указание термина "ценная бумага".

# Омонимия

Омонимия (греч.homos‘одинаковый’ и ōnyma‘имя’) – звуковое и / или графическое совпадение языковых единиц, значения которых не связаны друг с другом.

# Полисемия

Полисеми́я (от греч. πολυσημεία — «многозначность») — многозначность, многовариантность, то есть наличие у слова (единицы языка, термина) двух и более значений, исторически обусловленных или взаимосвязанных по смыслу и происхождению.

# Синонимия

Синонимия — полное или частичное совпадение значений языковых выражений.

# Синтагматические отношения

Синтагматические отношения - отношения слов при соединении их в словосочетания и фразы. Линейные логические отношения, которые устанавливаются между словами непосредственно при их использовании в тексте, объединяют эти слова в сочетания и предложения. Для уточнения смысла документа или запроса, помимо ключевых слов, часто необходимо указывать в каких синтагматических отношениях эти слова находятся. Так, фраза "защита окружающей среды от человека" и фраза "защита человека от окружающей среды" имеют совершенно разный смысл, хотя и состоят из одних и тех же ключевых слов. Таким образом, развитый информационно-поисковый язык должен обладать средствами отображения парадигматических и синтагматических отношений.

# **Ключевые слова**

**Ключевые слова или просто ключ** - это набор символов, к которому сводятся поисковые запросы и которые используются для поиска информации среди проиндексированных сайтов поисковой системой.