Тезаурус

Часто в поисковый образ документа включаются только те слова, которые зафиксированы в ***информационно-поисковом тезаурусе.***

Тезаурус – словарь лексических единиц информационно-поискового языка, основанный на лексике естественного языка, отображающий семантические отношения и предназначенный для организации поиска информации путем индексирования документов и/или запросов. Лексическими единицами тезауруса являются ***дескрипторы***. Дескриптор ставится в однозначное соответствие группе ключевых слов естественного языка, отобранных из текста определенной предметной области. Например, в качестве дескриптора может быть выбрано любое (предпочтительно наиболее часто используемое или короткое) ключевое слово или словосочетание или же цифровой код. Многозначному слову естественного языка соответствует несколько дескрипторов, а нескольким синонимичным словам и выражениям – один дескриптор. Тезаурус учитывает семантические связи между словами: антонимы, синонимы, гипонимы, гиперонимы, ассоциации.

Синонимы – слова (словосочетания), разные по написанию, но одинаковые (в рассматриваемой предметной области) по значению: *внедорожник = джип*. ***Антонимы*** - слова с противоположным значением: *быстрый - медленный*. ***Гипоним*** - термин, являющийся частным случаем другого, более общего понятия. ***Гипероним*** - термин, наоборот, являющийся общим для ряда других, частных понятий.

Солдат = гипоним (*военный*); *человек*= гипероним (*военный*)

В Государственном стандарте на "Тезаурус информационно-поисковый одноязычный" определены следующие типы связей:

- род-вид: *средства передвижения - самолет, автомобиль;*
- часть-целое: *двигатель* - часть *автомобиля*;
- причина-следствие: *закончилось горючее – остановился двигатель*;
- сырье-продукт: *резина – автомобильная шина*;
- административная иерархия: *командир корабля - штурман*;
- процесс-субъект: *поездка на такси - шофер*;
- процесс-объект: *поездка на такси - пассажир*;
- функциональное сходство: *самолет - вертолет*;
- свойство - носитель свойства: *высокая проходимость - джип*;
- антонимия;
- синонимия.

Ассоциативное отношение является объединением других отношений, не входящих в иерархические отношения или в отношения синонимии (то есть любые виды связей между словами, возможно весьма специфичные).

Словарная статья (на неформальном уровне) могла бы выглядеть так:

внедорожник = джип

ГИПОНИМ: джип Cherokee
ГИПЕРОНИМ: автомобиль
АССОЦИАЦИЯ: авторалли

Тезаурус и грамматика задают ***информационно-поисковый язык***. Грамматика содержит правила образования производных единиц языка (семантических кодов, синтагм, предложений).

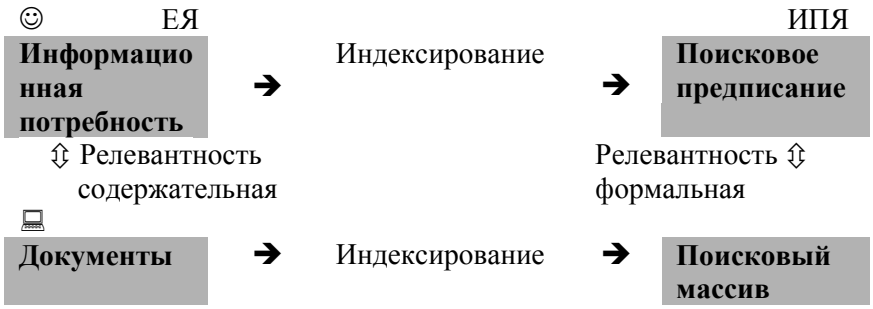
На основании тезауруса и правил грамматики формируются поисковые образы документа и запроса (поисковое предписание). ***Поисковое предписание*** – текст на информационно-поисковом языке, содержащий признаки документов, затребованных пользователем в запросе.

Поисковый образ документа – текст на информационно-поисковом языке, поставленный в однозначное соответствие документу и отражающий его признаки, необходимые для поиска его по запросу. Кроме поисковых признаков, раскрывающих содержание документа или, как минимум, определяющих его тему, поисковый образ документа обычно содержит также идентифицирующие и некоторые дополнительные сведения (выходные данные, тип документа, его язык и т.д.). Поисковые предписания формируются при поступлении запросов, а поисковые образы документов могут создаваться как при пополнении системы новыми документами, так и при поиске ответа на запрос. В системах, где потоки информации велики и часто обновляемы, нет необходимости тратить ресурсы на индексирование, и за поисковый образ документа часто принимается сам документ или же его название.

Релевантность

Целью ИПС является выдача документов, ***релевантных*** (семантически соответствующих) запросу (по-английски relevant – относящийся к делу). Различают релевантность ***содержательную*** и ***формальную***. Релевантность содержательная трактуется как соответствие документа информационной потребности, определяемое неформальным путем, а релевантность формальная – как соответствие, определяемое алгоритмически путем сравнения поискового предписания и поискового образа документа на основании применяемого в информационно-поисковой системе ***критерия выдачи***.

Критерий выдачи – формальное правило, совокупность признаков, по которым определяется степень формальной релевантности поискового образа документа и поискового предписания и принимается решение о выдаче/невыдаче некоторого документа в ответ на информационный запрос.

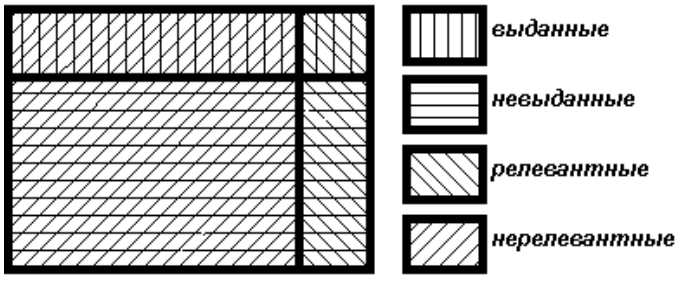


В автоматизированных системах поиск основан на формальной релевантности, содержательная релевантность в них определяется, например, путем экспертных оценок и используется для получения данных об ***эффективности информационного поиска в системе*** (качестве ее работы). В качестве критерия выдачи может быть выбрано полное совпадение поисковых образов документа и запроса, включение множества ключевых слов запроса во множество ключевых слов документа, пересечение этих множеств и др.

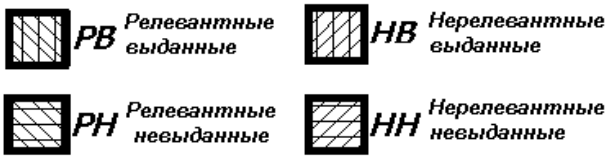
Критерий «на совпадение» хорош, когда необходима точность, например, выбираются лекарства для лечении определенной болезни (пусть их будет немного, зато все подходящие). В большинстве случаев используется критерий «на пересечение».

Дескрипторам могут быть приданы весовые коэффициенты в зависимости от их важности. При поиске коэффициенты дескрипторов, обнаруженных в документах, суммируются, и документы выдаются в зависимости от значения этой суммы (например, если она превысила некоторый порог). Таким образом, если указать, что наиболее важными являются характеристики *скорость* и *престижная марка*, а не *цена*, то можно получить сведения только о слишком дорогих автомобилях. При использовании весов также может применяться ***эшелонированная выдача*** – отобранные документы предъявляются пользователю не в произвольном порядке, а по степени релевантности (по убыванию сумм весов), право окончательного выбора релевантных документов – за пользователем.

Идеальная ИПС должна выдавать документы, содержательно релевантные запросу, и ничего кроме них. Однако на практике это обычно не достигается, наблюдаются молчание ИПС (невыдача некоторого количества релевантных документов) и шум (выдача лишних документов). Массив документов разделяется на **выданные** и **невыданные** – по одному критерию, и на **релевантные** и **нерелевантные** – по другому.



Таким образом, для каждого запроса получаем 4 группы документов:



Соотношение количества документов в каждой из этих групп определяет эффективность информационного поиска. Для оценки эффективности используют следующие характеристики:

$$\text{Полнота выдачи} = \frac{Рв}{Рв+Рн} \times 100\%$$

$$\text{Точность выдачи} = \frac{Рв}{Рв+Нв} \times 100\%$$

$$\text{Информационный шум} = \frac{Нв}{Рв+Нв} \times 100\ \%$$

В идеальной ИПС Рн=Нв=0 и поэтому полнота и точность= 100%, а шум = 0 (найдены все документы и ни одного лишнего). В реальных системах коэффициент полноты достигает 70%, а коэффициент точности поиска колеблется в очень широких пределах, иногда снижаясь до 10%. Величины этих коэффициентов зависят от целого ряда факторов: как внутренних свойств собственно поисковой системы (объема и характеристик информационного массива, информационно-поискового языка, критерия выдачи), так и от многих "внешних" условий: степени специфичности информационных запросов, способности пользователя правильно сформулировать свою информационную потребность на естественном языке и правильно построить конкретный запрос, а также от субъективного представления пользователя о том, что такое нужная ему информация. Из-за ошибок и неточностей, возникающих на каждом из этапов работы как пользователя, так и системы, результаты могут сильно отличаться от того, что хотел получить пользователь, обращаясь к ИПС.

Существует понятие ***устойчивость поиска*** – характеристика изменения полноты и точности при малых (семантически незначительных) изменениях запроса. Средние значения полноты и точности для конкретной системы обычно вычисляют путем тестирования ее на эталонной базе документов.

В зависимости от требований к количеству и качеству выдаваемой ИПС информации выбираются разные критерии выдачи. Если важно не упустить нужную информацию (патентная экспертиза), нужно повысить полноту, если надо сократить объем выдаваемой информации (библиотека), следует улучшить точность.

Известна обратная зависимость между полнотой и точностью поиска в одной системе (при использовании одного и того же информационно-поискового языка), т.е. повышение точности ведет к увеличению шума и, наоборот, при уменьшении шума снижается точность. Улучшить оба эти показателя одновременно можно, только внося изменения в информационно-поисковый язык, делая грамматику и тезаурус более лингвистически развитыми. При этом достижение максимально возможной полноты поиска связано с огромными сложностями. Последние 5-10% требуют такого же усложнения языкового аппарата системы, как и предыдущие 90-95%, что влечет за собой увеличение трудоемкости обработки входной информации и времени поиска.

Лингвистические аспекты информационного поиска

Повышению эффективности ИПС помогает более детальная обработка текста документа. Так, существуют системы, которые для простоты в качестве поискового образа документа принимают его название, однако оно в силу разных обстоятельств не всегда формально отражает содержание текста. Также большое значение имеет применение программ, производящих лингвистически содержательную обработку текстов на естественном языке (учитывающую морфологию, синтаксис). Только с их помощью можно установить, являются ли похожие слова (почти все буквы одинаковые) формами одного слова или же это совершенно разные слова, в соответствие которым поставлены разные семантические единицы.

Более примитивные приемы могут подвести разработчика ИПС. Так, если система не учитывает никакие правила русского языка и работает с шаблонами (типа text*.exe), то при поиске документов, связанных с *бальными танцами*, в качестве ключевого слова-шаблона придется выбрать *бал** (иначе можно пропустить эту информацию, высказанную словами *танцевать на балах*). Тогда в результате поиска могут быть найдены тексты, содержащие слова: *балл, балет, балык, Бальмонт, Бальзак, Балтийское море, балкон* и др.

Все эти слова будут отсеяны, если в качестве ключевого слова будет задано прилагательное *бальный* и система сможет распознавать его во всех его формах. Еще один способ уменьшения шума и повышения точности – введение в информационно-поисковый язык аппарата работы с однокоренными словами. В нашем примере при задании ключа-корня *бал* выданными оказались бы только документы, содержащие разные формы слов *бал* и *бальный*. Однако и в этом случае может быть выдана лишняя информация о *салонах бального платья, владельцах бальных залов*, музыкантах и официантах, *обслуживающих балы*. С помощью синтаксического анализа можно более точно определять словосочетания (например, распознавать их не только когда слова стоят друг за другом, но и когда они разделены рядом других слов). В приведенном примере в системе с синтаксическим компонентом можно было бы вести поиск документов со словосочетаниями *бальный танец* и *танцевать на балу*. Конечно, и это не обеспечивает 100% точности (например, ничто не запрещает выдачу сообщений об *учителях бальных танцев*), однако понятно, что количество выданных документов значительно сократится.

Морфологический поиск – возможность поисковой системы искать слово в документах не только в строго заданном виде, но и во всех его морфологических формах.

Поиск по ключевым словам – поиск документов, которые содержат указанные пользователем ключевые слова.

Поиск по словосочетаниям – поиск документов, которые содержат в точности указанное пользователем словосочетание.

Поиск с расстоянием – поиск, при котором пользователь указывает, на каком расстоянии между собой должны располагаться ключевые слова в документе. При этом под расстоянием понимается количество слов между двумя выделенными словами.

Развитые информационно-поисковые языки допускают использование логических связок: *пассажирский*=NOT(*грузовой*), *сверхзвуковой самолет* = (*самолет*) AND (*скорость* > *М*). В перспективе – возможность описания на информационно-поисковом языке смысла целой фразы (который не всегда складывается из смыслов входящих в нее слов) и возможность формулировки соответствующих семантически сложных запросов.

Часто, даже при поиске нетекстовой информации (например, аудио- и видео-) работа на самом деле ведется с описаниями на естественном языке. Например, для организации поиска фотографий необходимо снабдить каждую из них набором словесных характеристик (типа: "портрет, профиль, полный рост, мужчина"; "пейзаж, лес, осень" и т.п.).

Задачи, связанные с информационным поиском

Применение компьютеров не только ускоряет создание и обработку документов, но и чрезвычайно стимулирует рост их количества и объема. Пользователи регулярно сталкиваются с необходимостью быстро просматривать большой объем документов и выбирать из них действительно нужные. Эта задача возникает при работе с текстовыми массивами, с электронной почтой, при поиске в Интернете.

Для облегчения поиска в больших информационных массивах используются специальные методы и соответствующие программные средства и технологии.

Сократить количество просматриваемых документов может помочь их ***категоризация*** – разбиение массива документов на классы (по **категориям/тематическим рубрикам**). При этом могут учитываться как чисто внешние показатели документа (объем, расширение имени соответствующего файла и т.п.), так и его содержательные характеристики (название, фамилия автора, ключевые слова), которые позволяют отнести текст к той или иной тематической рубрике.

Под термином ***рубрицирование*** понимается сопоставление документу/тексту одной или нескольких рубрик. Так, например, очередное новостное сообщение может быть отнесено к рубрике «политика» или «спорт», или же к двум рубрикам сразу («бизнес» и «строительство»).

Совокупность рубрик может быть составлена заранее (разработчиками информационной системы) или строиться автоматически по входному потоку документов.

Частный случай рубрицирования – ***фильтрация***, т.е. распределение совокупности/потока документов по двум категориям (допустимые/нужные и недопустимые/ненужные) и отсеивание или блокирование документов второй из этих групп.

Оценить степень соответствия найденного ИПС документа без детального изучения его текста помогают **рефераты**/аннотации, содержащие краткое изложение содержания документа. Соответствующий раздел – неотъемлемая часть научной статьи в серьезных научных журналах, а в крупных организациях, особенно государственных, правила делопроизводства предписывают сопровождать каждый документ таким кратким описанием (или набором ключевых слов).

К сожалению, автоматические методы не настолько совершенны, чтобы полноценный реферат можно было построить путем генерации предложений текста. Однако уже сейчас возможно ***автоматическое реферирование*** – составление более или менее информативных и связных рефератов заданного объема – путем выбора информативных предложений из исходного текста (***квазиреферирование*)**, а также автоматическое составление списка ключевых слов.

В качестве ключевых слов система может выбирать слова, наиболее часто встречающиеся в тексте (и являющиеся при этом информативными, т.е. не предлоги, союзы и проч.), либо использовать для отбора какие-либо синтактико-семантические признаки. Фрагмент: *"Определение. Интегралом ... называется ..."* позволяет можно заключить, что *интеграл* – ключевое слово).

При реферировании из текста отбираются предложения, в наибольшей степени характеризующие его содержание. Такими могут считаться, например, предложения, содержащие ключевые слова (чем больше, тем лучше), либо отобранные по некоторым особым признакам. Размер реферата (коэффициент сжатия) или количество ключевых слов задается пользователем. Результатом работы такой системы может являться некоторый новый текстовый документ (реферат или набор ключевых слов) или же данный документ, в котором ключевые слова или наиболее информативные предложения выделены по тексту.

Термины, связанные с поиском в Интернете

Поисковая машина - в Интернете - специальный веб-сайт, на котором пользователь по заданному запросу может получить ссылки на сайты, соответствующие этому запросу.
Поисковая машина состоит из трех компонентов:

- поискового робота,
- индекса системы,
- программы, которая обрабатывает запрос пользователя, находит в индексе документы, отвечающие критериям запроса, и выводит список найденных документов в порядке убывания релевантности.

Индекс системы (в Интернете) компонент поисковой системы; информационный массив, в котором хранятся специальным образом преобразованные текстовые составляющие всех посещенных и проиндексированных роботом веб-страниц и текстовых файлов.

Поисковый робот - компонент поисковой системы; программа, которая посещает веб-страницы, считывает (индексирует) полностью или частично их содержимое и далее следует по ссылкам, найденным на данной странице. Робот возвращается через определенные периоды времени и индексирует страницу снова. Вся информация заносится роботом в индексы поисковой системы.

Пример фактографической ИПС: "Сотрудники ООО **ОО-ОО-ОО"**

№	ФИО	Пол	Возраст	Должность	Адрес	Телефон	Зарплата
1	Мальков	М	55	Ген. Директор	<a1>	<n1>	<z1>
2	Абрамовский	М	66	Гл. Бухгалтер	<a2>	<n2>	<z2>
3	Соловьин	М	33	Нач. Охраны	<a3>	<n3>	<z3>
4	Сойер	М	44	Шофер	<a4>	<n4>	<z4>
5	Грациозная	Ж	22	Секретарь	<a5>	<n5>	<z5>

Возможные ЗАПРОСЫ к ИПС:

1.Вся информация о Соловьине.
2.Адреса всех шоферов.
3.Кто (ФИО) проживает в Москве.
4.Кто получает зарплату > N у.е.
5.Телефон сотрудника, фамилия которого начинается "Голов".

О проблеме полномочий.