**Использование информационно-поисковых тезаурусов в автоматизированных информационных системах**

В настоящее время разработано большое количество информационно-поисковых языков (ИПЯ), различных по назначению и принципам построения. Существуют, например, информационно-поисковые языки, основанные на различных классификациях, так называемые ИПЯ классификационного типа.

ИПЯ классификационного типа широко используются в автоматизированных информационно-поисковых системах (ИПС) промышленного назначения, применяемых, например, при поиске аналогов проектируемых деталей с целью заимствования технологического процесса для их обработки. Однако применение ИПЯ, основанных на той или иной классификации, в автоматизированных документальных ИПС ограничивается тем, что введение новых сложных многоаспектных понятий требует бесконечного расширения классификации, что, конечно, невозможно.

Поэтому в последнее время в автоматизированных ИПС предпочтение отдается так называемым ИПЯ дескрипторного типа. Дескрипторы – это термины естественного языка, выражающие определенные понятия. Словарь дескрипторов с указанными между ними смысловыми отношениями, охватывающий определенную область знания, называется информационно-поисковым тезаурусом. Идея применения ИПТ для информационного поиска документов состоит в описании содержания документов и запросов с помощью дескрипторов, входящих в его состав. На практике эта идея реализуется следующим образом. Текст документа, вводимого в ИПС, уменьшается до объема реферата, в котором выделяются слова, несущие основную смысловую нагрузку, так называемые ключевые слова (КС). При помощи ключевых слов достаточно точно передается содержание документа. После этого каждое слово заменяется близким ему по смыслу дескриптором информационно-поискового тезауруса.

Совокупность терминов тезауруса - дескрипторов, заменивших ключевые слова, образует поисковый образ документа (ПОД). Точно так же на язык дескрипторов переводится и запрос. Поиск документов после этого уже можно поручить машине. ЭВМ с присущей ей быстротой просматривает массивы ПОДов и сравнивает их с поисковым образом запроса (ПОЗ). Цель такого сравнения – выявить дескрипторы, принадлежащие одновременно ПОДу и ПОЗу, т.е. найти пересечение множеств дескрипторов ПОДа и ПОЗа. Чем больше зона пересечения, тем точнее соответствие документа запросу. Минимальная величина зоны пересечения оговаривается принятым критерием смыслового соответствия. Изменяя его, можно варьировать точность и полноту поиска в зависимости от нужд потребителей информации.

Таковы принципы работы ИПС с ИПЯ дескрипторного типа. Качество их работы во многом зависит от совершенства применяемого информационно-поискового тезауруса. Поэтому требования к ИПТ очень высоки. В нем не должно быть, например, терминов, выражающих одно и то же понятие, т.е. синонимов – это привело бы к потерям информации при поиске. Кроме этого, должны быть зафиксированы некоторые отношения между терминами (род – вид, часть – целое и другие), служащие целям повышения точности и полноты поиска.

Тезаурус является необходимым лингвистическим пособием любой механизированной или автоматизированной информационно-поисковой системы, использующей ИПЯ дескрипторного типа.

Первые тезаурусы, удовлетворяющие специфическим задачам информационного поиска, были разработаны в начале 60-х годов нашего века.

В соответствии с тематическим профилем различают многоотраслевые, отраслевые и узкотематические тезаурусы.

Основной целью разработки традиционных информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. По своей сути тезаурус для ручного индексирования является искусственным языком описания, построенным на основе естественного языка. При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной обрасти, которые имеются у профессиональных индексаторов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Именно индексатор должен хорошо понимать всю терминологию, использованную в тексте, - для описания основной темы текста ему понадобится значительно меньшее количество терминов.

При автоматической обработке текстов человека-посредника между текстом и описанием его содержания в виде дескрипторов нет. Есть только автоматический процесс и Тезаурус, который должен содержать и те знания, которые содержатся в традиционных информационно-поисковых тезаурусах, и те знания (насколько это возможно), которые использует индексатор для определения основной темы текста.

Именно поэтому традиционные тезаурусы, разработанные для ручного индексирования, невозможно использовать при автоматическом индексировании

АИ-тезаурус должен включать не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывать широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня.

АИ-тезаурус должен включать в виде отдельных единиц семантически близкие понятия, в отличие от тезаурусов для ручного индексирования, где совокупности близких понятий сводятся к одному, наиболее представительному понятию для уменьшения субъективности индексирования. Таким образом, единицы АИ-тезауруса должны быть значительно ближе к понятийному аппарату предметной области, чем дескрипторы традиционного тезауруса. Синонимические ряды понятий должны быть значительно богаче, чем совокупности вариантов дескриптора в тезаурусе для ручного индексирования, поскольку синонимы должны описывать различные способы выражения данного понятия в тексте для автоматического процесса, а не для человека. Ряды синонимов включают в себя не только существительные и именные группы, а также прилагательные, глаголы, глагольные группы. Значительные усилия должны быть также сделаны, чтобы найти различные многословные синонимы. В тезаурусах для ручного индексирования большинство таких вариантов рассматривались как избыточные, поскольку индексаторы могли узнать такие синонимы в тексте благодаря своим языковым знаниям.

АИ-тезаурусы должны включать средства для описания и разрешения многозначности терминов. Комментарии, пометы, призванные прояснить значение термина и случаи его применения как индексирующего элемента для индексаторов, являются бесполезными в АИ-тезаурусе. Они могут быть использованы лишь как смысловые средства различения понятий при разработке АИ-тезауруса и изучении результатов автоматической обработки.

При этом необходимо решить вопросы, связанные с представлением в тезаурусе многозначных терминов, а именно, как и насколько подробно должны быть описаны различные значения многозначных терминов, чтобы такое описание могло стать базой для эффективного разрешения многозначности терминов в процессе автоматического индексирования. Дело в том, что слишком детальное разбиение значений, не поддержанное мощностью методов разрешения многозначности, ведет к серьезным потерям качества автоматического индексирования

Расширенная понятийная база АИ-тезауруса и ее использование в автоматической обработке текстов существенно увеличивают роль отношений между понятиями - концептуальных отношений. Концептуальные отношения в АИ-тезаурусе должны служить для решения трех основных проблем.

Во-первых, для навигации от нижестоящих понятий к вышестоящим и наоборот, что необходимо при автоматическом расширении запроса и автоматической рубрикации текстов.

Во-вторых, в процессе построения автоматического концептуального индекса необходимо не только обнаружить термины, но и определить их относительную важность для содержания текста, определить, насколько тот или иной термин соответствует основной теме текста. В идеале термины, получившие максимальный вес в результате автоматической обработки текста, должны совпасть с теми терминами, с помощью которых описал бы основную тему текста человек-индексатор. При этом нужно учитывать тот факт, что употребление термина в тексте не является независимым от употребления других терминов. Значимые для текста термины обычно сопровождаются множеством других семантически и тематически близких терминов. Поэтому для качественного определения веса термина в тексте, необходимо распознавать совокупности таких близких по смыслу терминов в тексте, пользуясь для этого концептуальными связями, описанными в АИ-тезаурусе.

В-третьих, концептуальные отношения должны служить для разрешения многозначности терминов в процессе автоматического индексирования.

Для выполнения всех этих функций недостаточно двух типов отношений между дескрипторами, которые приняты в традиционных тезаурусах в общественно-политической области: отношений ВЫШЕ- НИЖЕ, обладающих свойством транзитивности, и отношений АССОЦИАЦИЯ, симметричных нетранзитивных. Легко можно найти примеры отношений, отличных от отношений ВЫШЕ- НИЖЕ, но обладающих транзитивностью или примеры ассоциаций, которые лучше использовать как несимметричные.

Поэтому типы отношений в АИ-тезаурусе должны быть расширены. Такое расширение типов должно производиться не в сторону нарастания типов названий этих отношений, а в сторону описания различных навигационных типов концептуальных отношений.