4
П-71 19

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной
и прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

/19/

Э. И. КОРОЛЕВ, А. В. КУЗНЕЦОВ

ЯЗЫК и ЛОГИКА ДЕСКРИПТОРНЫХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

Москва — 1971

ПЯ-институт языка

ИРЯ

институт русского языка
им. в. в. виноградова



www.ruslang.ru

66157.

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР
Проблемная группа по экспериментальной и
прикладной лингвистике

Предварительные публикации

Ф.И.Королев, А.В.Кузнецов

ЯЗЫК И ЛОГИКА ДЕСКРИПТОРНЫХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

Москва - 1971

4
П-71 19
БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР
66157

Для многоаспектного поиска информации на больших массивах, насчитывающих сотни тысяч документов, в настоящее время применяются электронные вычислительные машины, обладающие большой памятью, мощной логикой и быстродействием. Автоматизированные информационно-поисковые системы (ИПС) созданы в электротехнике ("Пусто-Непусто-2"), приборостроении ("Реферат"), вычислительной технике ("Бит"), химии ("Аргон"), радиоэлектронике ("Сетка"), легкой промышленности ("Кристалл") и других областях науки и техники.

В настоящем кратком сообщении описываются язык и логика автоматизированной ИПС по оптике, реализованной на ЭВМ "Минск-32" *. Особое внимание обращается на лингвистические аспекты ИПС — способы организации словаря-тезауруса, критерий смыслового соответствия запроса и документа, алгоритм автоматического перевода с естественного языка на дескрипторный и др.

I. Основные принципы построения ИПС.

1. Подобно большинству развитых отечественных ИПС (см. таблицу) описываемая поисковая система является дескрипторной. Основные парадигматические отношения между ключевыми словами (синонимия, родо-видовые и другие ассоциативные отношения) зафиксированы в словаре-тезаурусе. Наличие словаря является большим преимуществом дескрипторной ИПС по сравнению с системами, работающими на естественном языке (ИПС "Кристалл"[6] и некоторые другие системы, построенные на его основе). Словарь составляется квалифицированными специалистами по тематике, что позволяет зафиксировать в нем сложные связи между словами того или иного подъязыка науки и техники и затем наиболее полно и единообразно использовать эти сведения при индексировании запросов и документов.

---

* Аналогичная информационно-поисковая система создается в Лаборатории машинного перевода и прикладной лингвистики МГПИИЯ им. М.Тореза на материале документов Общего отдела Госплана СССР.

2. Как показано в таблице, в большинстве существующих дескрипторных ИПС применяется ручной поиск по словарю-тезаурусу (ср. ИПС "Ордината", "Сетка" [16], "Реферат" [20], "Аргон" [11,12], "Агроэлектрин" [9]): при индексировании документов ключевые слова вручную заменяются их дескрипторными номерами, а родовые и/или видовые дескрипторы вводятся в поисковый образ документа (ПОД) и/или в поисковый образ запроса (ПОЗ) ("избыточное индексирование").

Ручное использование словаря имеет следующие недостатки:

1) не позволяет вводить большой массив информации, так как требует значительных расходов времени и привлечения большого числа индексаторов-специалистов;

2) порождает субъективизм при индексации, что в значительной мере снижает единообразие заиндексированного материала и практически исключает возможность коррекции ошибок при индексации;

3) не позволяет в достаточной степени объективно и полно отразить базисные отношения при избыточном индексировании. Это в особенности относится к базисным отношениям, имеющим сильную разветвленность. Избыточное индексирование практически приводит к тому, что в ПОЗ или ПОД учитываются только самые верхние или самые нижние уровни иерархии.

Тем самым в указанных ИПС, по существу, автоматизируется только процесс собственно поиска, что значительно снижает их роль как автоматизированных систем информации. Для сравнения отметим, что развитые зарубежные ИПС, реализованные на ЭВМ, обычно используют автоматический принцип индексирования - ср. хотя бы известные ИПС "Синтол" [7,14] и "Смарт" [8].

По-видимому, единственным исключением среди отечественных систем является ИПС "Пусто-Непусто" - 4 и 2 [2-4,16,18], где применяется полуавтоматический перевод с естественного языка на дескрипторный при индексации документов и автоматически учитываются базисные отношения.

В описываемой системе базисные отношения учитываются

автоматически – при формировании машиной условия поиска.

Сложнее обстоит дело с процессом автоматического индексирования документов. В простейших алгоритмах автоматического индексирования (например, в ИПС класса "Пусто-Непусто") ПОД создается набором всех дескрипторов, входящих в текст документа. Если документ представляет собой реферат (а они бывают достаточно большими), то ПОД насчитывает несколько десятков, а то и более сотни дескрипторов, что ведет к увеличению информационного шума (который в ИПС "Пусто-Непусто" может достигать 50% выдачи [18]). Проблема сокращения ПОД является очень сложной, и все исследования по ней пока не выходят за рамки экспериментов.

Исходя из вышесказанного, в описываемой ИПС применяется сочетание автоматического индексирования с автокодированием.

Большую часть ретроспективного фонда составляют документы, представляющие собой библиографические описания (в основном подробные заголовки статей из отечественной и зарубежной периодики). Такие документы индексируются автоматически. Расширенные заголовки вторичных документов, как правило, достаточно полно отражают их содержание. Введение их в ИПС преимущественно библиографической информации сокращает время автоматического индексирования и существенно снижает потенциальную вероятность "поискового шума", прямо пропорциональную числу дескрипторов в поисковом образе документа.

Меньшую часть фонда составляют рефераты и аннотации. В этих материалах вручную выделяются ключевые слова, которые затем автоматически переводятся в коды дескрипторов.

3. Синтагматические отношения между дескрипторами в ИПС не учитываются, т.е. поисковый образ документа представлен неупорядоченным множеством дескрипторов. Грамматика не используется по следующим причинам:

1) Введение грамматики, как известно, повышает точность выдачи, но зато (и порой существенно) снижает полноту, которая, несомненно, является более важной характеристикой ИПС, чем точность.

Для повышения полноты в ИПС с грамматикой обычно требуется дополнительное применение мощного аппарата логических выводов и механизма синонимического преобразования граммати-

ческих структур ([5,11,15]), что представляет собой сложную задачу, решение которой совершенно излишне для ИПС документального типа [5] .

2) Кроме того, введение элементов грамматик, в отличие от развитой грамматики с автоматическим словарем [5] , делает практически невозможным осуществить процесс автоматического индексирования, о важности которого говорилось выше (ср., например, ИПС "Аргон" [11,12] , "Кристалл" [6] и "Реферат" [20] , где указатели роли и связи определяются вручную).

3) Существующее опасение, что информационный шум сильно возрастает при многократном увеличении массива документов [1] , нам кажется слишком преувеличенным. При детальных и конкретных запросах и относительно небольших ПОД информационный шум в ИПС без грамматики, по нашему мнению, должен быть вполне приемлемым даже на больших массивах документов.

4. В отличие от большинства существующих отечественных систем, работающих преимущественно в режиме избирательного распределения информации, данная ИПС с самого начала эксплуатируется также в режиме ретроспективного поиска. Предполагаемый массив документов системы - 300-500 тыс. документов. Для такого массива необходима программа, сокращающая прямой перебор всего массива поисковых образов документов, и рациональная организация памяти ЗУакт ЭЦВМ. В настоящее время создается дуальная программа поиска, сочетающая преимущества прямой и инверсной организации памяти. Пока же на небольших экспериментальных массивах используется программа, работающая по прямому принципу, которая дает вполне удовлетворительные результаты времени поиска благодаря техническим преимуществам ЭВМ "Минск-32" перед другими машинами.

5. Важным требованием к ИПС является простота и непосредственность результатов выдачи информации. С этой точки зрения потребителю желательно выдавать не номера искомых документов или иные шифры, а непосредственно сам документ. В качестве промежуточного результата в данной ИПС предусмотрена выдача библиографических описаний документов, хранящихся в ЭВМ (автор, заголовок документа и его выходные

данные). Если документ представляет собой реферат, то заголовок дополняется ключевыми словами из текста реферата. С технической точки зрения это не представляет никакой трудности, так как эти сведения все равно необходимо вводить в машину при автоматическом индексировании документа. В то же время пользователь не должен тратить дополнительное время на расшифровку кода документа или поиск его по номеру. К тому же исключаются многие сложные проблемы хранения и пополнения справочно-информационного фонда, возникающие при нумерации документов, подлежащих вводу в ИПС (дублирование массивов и т.п.).

II. Словарь-тезаурус.

Тезаурус предназначен для обеспечения полноты поиска информации путем учета важнейших парадигматических отношений - синонимии и базисных отношений.

1. Ключевые слова, эквивалентные по смыслу (в пределах принятого критерия смыслового соответствия и конкретного употребления описываемой лексемы), объединены в классы условной эквивалентности в дескрипторном словаре.

В настоящее время словарь насчитывает более 5000 ключевых слов по общим разделам оптико-электроники и некоторым смежным областям (электротехника, математика, вычислительная техника), объединенных в более 2000 дескрипторных групп.

Распределение ключевых слов по дескрипторным группам производилось с привлечением специалистов по соответствующим разделам тематики *.

Как уже было сказано, учёт синонимии необходим для того, чтобы повысить полноту выдачи информации. Поэтому слова, тождественные по смыслу или имеющие общий смысл, должны обязательно иметь одни и те же или общие дескрипторные коды. Исходя из этого, слова, обозначающие сложные понятия, должны вклю-

---

* Над составлением словаря, а также индексированием документов и запросов работала группа специалистов под руководством А.Ф.Фишберга. Термины для приводимых ниже примеров взяты из физического энциклопедического словаря, тт. 1-5, М., 1960-1966гг.

ческих структур ([5,11,15]), что представляет собой сложную задачу, решение которой совершенно излишне для ИПС документального типа [5] .

2) Кроме того, введение элементов грамматик, в отличие от развитой грамматики с автоматическим словарем [5] , делает практически невозможным осуществить процесс автоматического индексирования, о важности которого говорилось выше (ср., например, ИПС "Аргон" [11,12] , "Кристалл" [6] и "Реферат" [20] , где указатели роли и связи определяются вручную).

3) Существующее опасение, что информационный шум сильно возрастает при многократном увеличении массива документов [1] , нам кажется слишком преувеличенным. При детальных и конкретных запросах и относительно небольших ПОД информационный шум в ИПС без грамматики, по нашему мнению, должен быть вполне приемлемым даже на больших массивах документов.

4. В отличие от большинства существующих отечественных систем, работающих преимущественно в режиме избирательного распределения информации, данная ИПС с самого начала эксплуатируется также в режиме ретроспективного поиска. Предполагаемый массив документов системы - 300-500 тыс. документов. Для такого массива необходима программа, сокращающая прямой перебор всего массива поисковых образов документов, и рациональная организация памяти ЗУакт ЭЦВМ. В настоящее время создается дуальная программа поиска, сочетающая преимущества прямой и инверсной организации памяти. Пока же на небольших экспериментальных массивах используется программа, работающая по прямому принципу, которая дает вполне удовлетворительные результаты времени поиска благодаря техническим преимуществам ЭВМ "Минск-32" перед другими машинами.

5. Важным требованием к ИПС является простота и непосредственность результатов выдачи информации. С этой точки зрения потребителю желательно выдавать не номера искомых документов или иные шифры, а непосредственно сам документ. В качестве промежуточного результата в данной ИПС предусмотрена выдача библиографических описаний документов, хранящихся в ЭВМ (автор, заголовок документа и его выходные

данные). Если документ представляет собой реферат, то заголовок дополняется ключевыми словами из текста реферата. С технической точки зрения это не представляет никакой трудности, так как эти сведения все равно необходимо вводить в машину при автоматическом индексировании документа. В то же время потребитель не должен тратить дополнительное время на расшифровку кода документа или поиск его по номеру. К тому же исключаются многие сложные проблемы хранения и пополнения справочно-информационного фонда, возникающие при нумерации документов, подлежащих вводу в ИПС (дублирование массивов и т.п.).

II. Словарь-тезаурус.

Тезаурус предназначен для обеспечения полноты поиска информации путем учета важнейших парадигматических отношений - синонимии и базисных отношений.

1. Ключевые слова, эквивалентные по смыслу (в пределах принятого критерия смыслового соответствия и конкретного употребления описываемой лексемы), объединены в классы условной эквивалентности в дескрипторном словаре.

В настоящее время словарь насчитывает более 5000 ключевых слов по общим разделам оптико-электроники и некоторым смежным областям (электротехника, математика, вычислительная техника), объединенных в более 2000 дескрипторных групп.

Распределение ключевых слов по дескрипторным группам производилось с привлечением специалистов по соответствующим разделам тематики *.

Как уже было сказано, учёт синонимии необходим для того, чтобы повысить полноту выдачи информации. Поэтому слова, тождественные по смыслу или имеющие общий смысл, должны обязательно иметь одни и те же или общие дескрипторные коды. Исходя из этого, слова, обозначающие сложные понятия, должны коди-

---

* Над составлением словаря, а также индексированием документов и запросов работала группа специалистов под руководством А.Ф.Фишберга. Термины, почти приводимых ниже примеров взяты из физического энциклопедического словаря. тт. 1-5, М., 1960-1966гг.

роваться сочетанием дескрипторов, обозначающих составляющие простые понятия [17] : КРУПНОМАСШТАБНЫЙ = КРУПНЫЙ + МАСШТАБНЫЙ, ЭЛЕКТРООПТИЧЕСКИЙ = ЭЛЕКТРО + ОПТИЧЕСКИЙ, ИНТЕРФЕРОМЕТР = ИНТЕРФЕРЕНЦИЯ + ИЗМЕРЕНИЕ, АВТОКОЛЛИМАТОР = АВТО + КОЛЛИМАТОР.

Игнорирование этого принципа влечет за собой, кроме потерь информации при поиске, неограниченное увеличение объема дескрипторного словаря, что затрудняет его автоматическое использование - ср. словарь ИПС "Ориентата", насчитывающий 9000 дескрипторов.

С целью повышения полноты поиска в словарной статье ключевого слова помещаются также все его дериваты - как синтаксические (ВАКУУМ - ВАКУУМНЫЙ, ВОЗДУХ - ВОЗДУШНЫЙ, ВОЛОКНО - ВОЛОКОННАЯ, ГОЛОВКА - ГОЛОВНОЙ), так и смысловые (МАШИНА - МЕХАНИЧЕСКИЙ, ДВА - ПАРНЫЙ, РАССЕЯНИЕ - ДИСПЕРГИРУЮЩИЙ). В частном случае, в статью глагола входят также названия приборов, отличающихся от глаголов только суффиксами существительных: ИНТЕГРАЦИЯ - ИНТЕГРАТОР, ТАБУЛИРОВАНИЕ - ТАБУЛЯТОР, ПЕРЕДАВАТЬ - ПЕРЕДАТЧИК, ПРЕРЫВАТЬ - ПРЕРЫВАТЕЛЬ, ИЗЛУЧАТЬ - ИЗЛУЧАТЕЛЬ и т.д.

Сочетания и их сокращения также кодируются составляющими их дескрипторами: ОКГ = ОПТИЧЕСКИЙ + КВАНТОВЫЙ + ГЕНЕРАТОР, ГМ = ГИРОСКОПИЧЕСКИЙ + МАЯТНИК, ИС = ИНТЕГРАЛЬНАЯ + СХЕМА, ВЧ = ВЫСОКАЯ + ЧАСТОТА, ЭВМ = ЭЛЕКТРОННАЯ + ВЫЧИСЛИТЕЛЬНАЯ МАШИНА, ФЭУ = ФОТО + ЭЛЕКТРОННЫЙ + УМНОЖИТЕЛЬ.

Если бы, стремясь к большой точности выдачи, мы закодировали подобные сочетания одним номером, то потери увеличились бы, в частности, из-за того, что в текстах документов один из членов сочетания, как правило, опускается, так как ясен специалисту из контекста: квантовый генератор = оптический квантовый генератор, вычислительная машина = электронная вычислительная машина, оптический преобразователь = электронный оптический преобразователь.

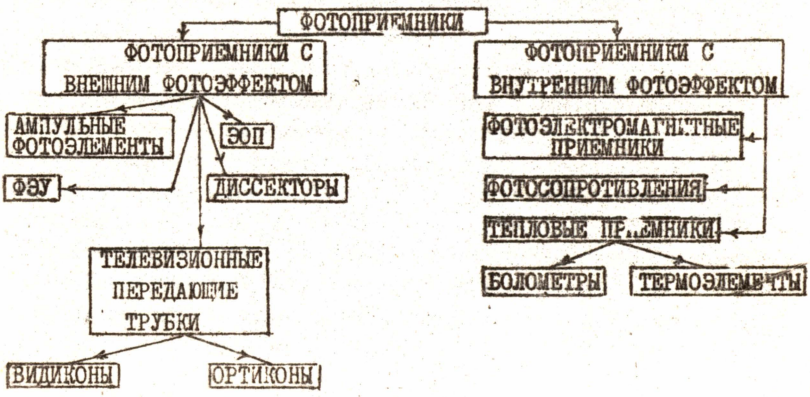
Одним номером кодируются лишь те сочетания, члены которых всегда присутствуют в тексте и стоят рядом, образуя единство: Млечный путь, Большая Медведица, удельный вес, коэффициент полезного действия, железная дорога и некоторые другие. Таких сочетаний сейчас насчитывается всего несколько десятков.

2. Базисные отношения между дескрипторами отражены в таблицах, введенных в память ЭВМ отдельно от дескрипторного словаря, что дает возможность использовать их автоматически при формировании условий поиска и не учитывать базисных отношений при индексировании документов.

В настоящее время составлено 20 таких непересекающихся таблиц, насчитывающих около 1000 рубрик, отражающих базисные отношения (в основном родо-видовые и часть-целое). Таблицы являются сильно разветвленными: более 200 рубрик имеют подчиненные рубрики.

Таблицы составлены специалистами с учетом существующих классификаций, отраженных в различных исследованиях по описываемым областям науки и техники. Степень дробности таблиц целиком определяется конкретной тематикой и реальными информационными потребностями.

Таблица базисных отношений имеют еще одну особенность, вытекающую из критерия смыслового соответствия описываемой ИПС (см. ниже) и принципов кодирования ключевых слов в дескрипторном словаре. В соответствии с ними, на запрос, содержащий родовой дескриптор, будут автоматически выданы документы, содержащие видовые дескрипторы, представляющие собой сочетание родового дескриптора с дескриптором, обозначающим вид, – на запрос о полупроводниковых триодах выдаются документы о полупроводниковых триодах на арсениде галлия. Следовательно, в таблицу достаточно помещать только такие дескрипторы, у которых видовые дескрипторы графически не повторяют родовых:



Каждая рубрика таблицы может состоять из одного или нескольких дескрипторов, содержащихся в дескрипторном словаре. В отличие от жесткой иерархии, классификация рубрик таблиц свободна: каждая рубрика таблицы может иметь сколько угодно нижележащих рубрик и подчиняться более чем одной рубрике. Например, рубрика ПЕНТОДЫ подчиняется рубрикам ЭЛЕКТРОННЫЕ ЛАМПЫ И ГЕНЕРАТОРНЫЕ ЭЛЕКТРОПРЕОБРАЗОВАТЕЛЬНЫЕ ЭЛЕКТРОННЫЕ ПРИБОРЫ, рубрика ГЕРМАНИЙ подчиняется рубрикам ПОЛУПРОВОДНИКИ и МОНОКРИСТАЛЛЫ. Фактически, с помощью таблиц можно выразить любые базисные отношения между рубриками, обладающие транзитивностью.

Все рубрики таблиц имеют сквозную нумерацию, причем у номера вышележащей рубрики указываются номера всех подчиненных рубрик, стоящих на один уровень ниже. Например, номер рубрики ФОТОПРИЕМНИКИ (см. выше) связан с номерами рубрик ФОТОПРИЕМНИКИ С ВНЕШНИМ ЭФФЕКТОМ и ФОТОПРИЕМНИКИ С ВНУТРЕННИМ ЭФФЕКТОМ, номер рубрики ТЕПЛОВЫЕ ПРИЕМНИКИ связан с номерами рубрик БОЛОМЕТРЫ и ТЕРМОЭЛЕМЕНТЫ и т.д.

Кроме базисных отношений, в таблицах иногда отражены отношения синонимии: одинаковый номер имеют рубрики, стоящие на одном иерархическом уровне и имеющие один и тот же смысл, хотя и состоящие из разных дескрипторов: КАНАЛЬНЫЙ ТРАНЗИСТОР = ПОЛЕВОЙ ТРАНЗИСТОР; КИНЕСКОП = ПРИЕМНАЯ ТЕЛЕВИЗИОННАЯ ТРУБКА; СЕЛЕКТИВНЫЕ ПРИЕМНИКИ = ФОТОННЫЕ ПРИЕМНИКИ; СВЕТОДЕЛИТЕЛИ = ЛУЧЕРАСЩЕПИТЕЛИ.

Ш. Индексирование запросов и критерий смыслового соответствия.

В рассматриваемой ИПС поисковый образ запроса составляется вручную. Этот этап работы является более важным, чем индексирование документов, так как от тщательности составления ПОЗ прямо зависят полнота и точность поиска. Поэтому в индексировании запросов обязательно участвуют квалифицированные специалисты, не только хорошо знакомые с языком и логикой системы, но и в совершенстве знающие тематику.

При индексировании запроса в нем выделяются ключевые слова, которые затем заменяются дескрипторными кодами по

дескрипторному словарю. Таким образом, в ПОЗ дескрипторных ИПС синонимия ключевых слов в основном устранена, в то время как ПОЗ систем, работающих на естественном языке, необходимо дополнять синонимами. При включении дескрипторов в ПОЗ индексатору приходится быть очень осторожным, так как увеличение ПОЗ повышает точность выдачи, но зато снижает полноту. В качестве примера рассмотрим запрос "Данные об интегральной чувствительности фотоэлектронных умножителей". Если в ПОЗ включить все ключевые слова, которые есть в дескрипторном словаре, то потеряются документы, в которых говорится, в частности, об интегральной чувствительности фотоумножителей. Эти документы отвечают на данный запрос, хотя и не содержат дескриптора фото, опущенного в контексте.

Если дескриптор или их сочетание образуют рубрику таблицы базисных отношений, то указывается номер этой рубрики. Дескрипторы ПОЗ могут быть связаны различными логическими отношениями, из которых в настоящее время используются самые распространенные - конъюнкция ($\wedge$) и дизъюнкция ($\vee$). Так, например, запрос "Усилители постоянного тока с низким дрейфом нуля" (1) представляет собой конъюнкцию дескрипторов: УСИЛИТЕЛИ $\wedge$ ПОСТОЯННЫЙ $\wedge$ ТОК $\wedge$ НИЗКИЙ $\wedge$ ДРЕЙФ $\wedge$ НУЛЬ. Другой запрос "Новые оптические материалы, используемые в зарубежных лазерах" (2) представляются в виде: ОПТИЧЕСКИЕ МАТЕРИАЛЫ $\wedge$ ЗАРУБЕЖНЫЕ $\wedge$ ЛАЗЕРЫ, где конъюнкцию составляют рубрики таблицы ОПТИЧЕСКИЕ МАТЕРИАЛЫ, ЗАРУБЕЖНЫЕ СТРАНЫ и дескриптор ЛАЗЕРЫ. Запрос "Применение подшипников и смазок в условиях вакуума"(3) представляется в виде: ВАКУУМ $\wedge$ (ПОДШИПНИКИ $\vee$ СМАЗКИ).

Если в запросе указывается рубрика таблицы базисных отношений, то при формировании условия поиска ЭВМ сама выписывает все подчиненные ей рубрики и связывает их отношением дизъюнкции. Тогда, например, запрос (2) будет иметь следующий вид: ЛАЗЕРЫ $\wedge$ (ОПТИЧЕСКИЕ МАТЕРИАЛЫ $\vee$ НЕОДИМ $\vee$ РУБИН ГРАНАТ $\vee$ АРСЕНИД ГАЛЛИЯ $\vee$ КАДМИЙ ...)$\wedge$ (ЗАРУБЕЖНЫЙ $\vee$ США $\vee$ АНГЛИЯ $\vee$ ФРАНЦИЯ $\vee$...).

В принципе обращение к таблицам базисных отношений очень легко автоматизировать. Этого не сделано, в частности,

потому, что потребителя не всегда интересует информация по всем рубрикам таблицы базисных отношений, что выясняется специалистом при уточнении запроса у потребителя.

В некоторых случаях запросу можно дать другую формулировку, сохраняющую его смысл. Используя указанные логические отношения, в запрос можно включать любые его перифразы, что также увеличивает полноту поиска. Например, запрос "Новые конструкции высокотемпературных черных тел" допускает перифразу, где вместо слов "черные тела" употребляются слова "черные излучатели". Поисковый образ этого запроса с перифразой будет иметь следующий вид: (КОНСТРУКЦИИ ∧ ВЫСОКАЯ ∧ ТЕМПЕРАТУРА ∧ ЧЕРНОЕ) ∧ (ТЕЛО ∨ ИЗЛУЧАТЕЛЬ).

Сказанное позволяет перейти к формулировке критерия смыслового соответствия системы. В данной ИПС документ отвечает на запрос только в том случае, если ПОД содержит все дескрипторы, входящие в конъюнкцию ПОЗ, и любой из дескрипторов, входящих в его дизъюнкцию. Тогда на запрос (1) будут выдаваться документы типа. "Транзисторные усилители постоянного тока с низким дрейфом нуля", "Аналоговый усилитель постоянного тока с очень малым дрейфом нуля", "Стабильный усилитель постоянного тока с двухфазовым выходом, имеющий высокий коэффициент усиления и низкий дрейф нуля"; на запрос (3) документы "Требования к подшипникам, рассчитанным на применение в вакууме", "Применение гидростатических подшипников в условиях глубокого вакуума", Расчет газовой смазки для вакуума", "Испытание на долговечность смазки, наносимой на шарикоподшипники, используемые в вакууме".

Если запрос обрабатывается с использованием таблицы базисных отношений, то на него выдаются документы, содержащие дескрипторы, которые стоят ниже дескрипторов рубрики, указанной в запросе. Так, на запрос (2) будут выданы документы типа: "Особенности работы французского оптического квантового генератора на рубине при температуре 78°К", "Зарубежные лазерные диоды на основе арсенида галлия", "Ухудшение параметров оптических материалов стекол английских ОКГ".

Данный критерий смыслового соответствия предполагает выдачу только релевантных документов, полностью соответствующих характеру запроса. При необходимости получить документы, косвенным образом связанные с запросом, последний расширяется с привлечением потребителя информации путем введения в запрос родовых дескрипторов и/или сокращения в нем числа имеющихся дескрипторов.

IУ. Машинные эксперименты.

Для отладки критерия смыслового соответствия ИПС и проверки словаря-тезауруса был проведен ряд экспериментов по поиску на ЭВМ "Минск-32" . Эксперименты проводились на двух массивах: вначале на 4000, а затем на I0000 документов, заиндексированных вручную. Всего было обработано около 500 реальных запросов.

Анализ результатов машинной и ручной выдачи позволил подвести некоторые предварительные итоги.

Полнота и точность ИПС составляют соответственно 87% и 85%.

В число причин, вызывающих потери, входит недостаточно гибкое варьирование запроса, отсутствие или неразработанность надлежащей рубрики таблицы базисных отношений, ошибки в индексации и кодировании документов (ключевое слово не включено в ПОД, приписан неправильный номер дескриптора) и т.д. По мере эксплуатации системы при приобретении опыта в индексировании запросов и пополнении словаря-тезауруса полнота выдачи должна возрастать.

Показатель полноты системы является относительно высоким. Это объясняется прежде всего квалифицированным индексированием запросов, поисковые образы которых составлялись с обязательным участием специалистов по тематике. Повышение полноты обеспечивает также детальная разработанность тезауруса и в первую очередь дескрипторного словаря. Эксперименты подтвердили правильность описанного принципа кодирования ключевых слов "по частям", когда слова, связанные по смыслу, имеют общие дескрипторные коды.

Например, на запрос "Лазерные дальномеры" документ, где говорится об измерении расстояний с помощью лазера, выдается только потому, что ключевые слова "лазер" и "лазерный" являются одним дескриптором, а "дальномер" кодируется сочетанием дескрипторов дальность (= расстояние, дистанция) и измерение (= измерять, -мер, -метр). Еще пример. На запрос "Малошумящие усилители на транзисторах" выдаются документы "Применение биполярных транзисторов в низкошумящих спектрометрических усилителях", "Избирательный усилитель с низким уровнем шумов на транзисторах", так как ключевое слово "малошумящий" кодируется сочетанием дескрипторов мало (=малый, низко-, низкий) и шумящий (= шум). Кроме того, к повышению полноты выдачи ведет автоматический учет весьма разветвленных базисных отношений, когда при формировании условия поиска ЭВМ перебирает иногда до 50-60 рубрик таблиц базисных отношений, что очень трудно сделать вручную.

Причинами информационного шума являются прежде всего ошибки индексации и дескрипторного словаря (когда одним дескриптором считаются слова, которые на самом деле оказываются не вполне синонимичны). Интересно отметить, что отсутствие грамматики вызывает лишь пятую часть шума, причем эта величина остается почти неизменной как на массиве 4000 документов, так и на большем массиве 10000 документов.

Высокий показатель точности объясняется прежде всего тщательностью ручного индексирования документов, которое производилось высококвалифицированными специалистами. В результате поисковый образ документа был небольшим (8-10 дескрипторов) и, следовательно, вероятность ложных связей между дескрипторами, вызывающих шум, невелика. При автоматическом индексировании точность выдачи должна быть ниже, но, по-видимому, не намного, так как ПОД увеличится приблизительно в полтора раза.

У. Автоматическое индексирование документов.

Алгоритм автоматического индексирования предусматривает включение слова документа в ПОД, если оно содержится в одном из трех словарей описываемых ниже.

В первом словаре содержатся основы ключевых слов с приписанными им одним или несколькими дескрипторными кодами.

В настоящее время словарь содержит около 5000 основ, расположенных по углубленному алфавитному принципу, причем основа, включающая другую основу, располагается первой (барит- (овый) - бари-(й), электрон-(ный) - электр-(ический), вентиля-(тор, ция), - вентил-(ь, ьный)) [I8,I9] .

Ключевые слова, имеющие чередования в основе или различное написание, в словаре представлены разными основами (имеющими, естественно, один и тот же дескрипторный код): угол - угл-(а), аммиак - аммиач-(ный), балк-(а) - балочн-(ый), Допплер - Доплер, кассегрен - кассегрейн-(овский). Путем увеличения длины основ и использования их частотности удалось существенно сократить число омонимичных основ. Сейчас их в словаре немногим более десятка: дорог-(ой) - дорог-(а), инди-(й) - инди-(йский), устан-(овка) (действие) - устан-(овка) (механизм) и нек. другие. Омонимичным основам пока приписываются оба дескрипторных кода - в расчете на то, что омонимия разрешится за счет других дескрипторов ПОД.

Отдельно от словаря основ существует словарь сокращений: ОКГ, кпд, CO_2 (углекислый газ), In(индий), МОП (металл-окисел-полупроводник), ИК (инфракрасный).

В настоящее время словарь насчитывает более сотни сокращений (включая латинские обозначения элементов и соединений).

В третий словарь включено несколько десятков словосочетаний (в основах): Млечн- Пут-, железн- дорог-, Ван- Аллен-, Полярн- сияни-.

Слово документа вначале проверяется на полное вхождение в словарь сокращений.

Если слово не является сокращением, то проверяется, не включает ли оно какую-либо основу из словаря словосочетаний. В положительном случае осуществляется проверка остальных основ словосочетания на вхождение в соседние слова текста.

Если слово не входит в словосочетание, то проверяется, не включает ли оно какую-либо основу из словря основ. Так же, как в ИПС "Пусто-Непусто-2" [I8] , в данной системе существует алгоритм разделения сложных слов. Число сложных слов в документах велико, и введение всех их в словарь при-

вело бы к сильному увеличению его объема. Поэтому в словаре содержатся лишь самые частотные сложные слова, а также простые слова, которые могут быть частями сложных. Каждое слово документа, содержащее больше некоторого эмпирически подобранного числа символов, потенциально считается сложным. Поэтому при совпадении максимальной части слова с основой словаря остаток слова вновь ищется в словаре. Если остаток в словаре не находится, то у него отбрасывается одна буква и процесс повторяется. Предположим, в словаре содержатся основы ДВУ-, ЛУЧ-, ПРЕЛОМ-. Тогда слово "двулучепреломление" будет индексироваться следующим образом. Вначале оно получит код основы ДВУ-, затем ЛУЧ-. Поскольку для остатка -ЕПРЕЛОМЛЕНИЕ не найдётся основы, у него будет отброшена одна буква. Остаток - ПРЕЛОМЛЕНИЕ будет последним, для которого в словаре найдётся основа.

В ходе автоматического индексирования предусматривается ввод в ПОД параметрической информации (цифры, имена собственные), которая отсутствует в дескрипторном словаре и должна быть представлена в поисковом образе документа в естественном (текстовом) виде.

В настоящее время алгоритм автоматического индексирования программируется на ЭВМ "Минск - 32."

VI. Организация поискового массива и реализация ретроспективного поиска на ЭВМ.

При выборе варианта реализации алгоритмов информационного поиска на ЭВМ предполагалось решить две основные проблемы:

1) сокращение объема поискового массива до размеров, соответствующих возможностям ЗУ акт. и обеспечивающих нормальные условия для эксплуатации системы;

2) сокращение времени поиска до приемлемого при работе, организованной по принципу диалога с машиной, причем не допускалось решение, уменьшающее полноту и точность выдачи.

Эти проблемы являются взаимно связанными и, как будет видно из дальнейшего, решение первой создает основу для

решения второй.

Сокращение физического объема поискового массива достигается группировкой связанных между собой документов. В каждую группу должно входить около 30 документов, связь между которы и заключается в наличии большого количества общих дескрипторов в поисковых образах [*] . Для каждой группы образуется заголовок, представляющий собой совокупность дескрипторов, входящих в ПОД группы. В заголовок входит около 30 дескрипторов. Различаются дескрипторы "родовые" для данной группы – входящие в большинство ПОД, и "видовые" – входящие в небольшое число документов группы. Эти дескрипторы снабжаются соответствующим признаком в заголовке. Хотя термины "родовой" и "видовой" используются здесь условно, они помогают понять предлагаемый способ группировки информации.

Первоначально заголовки групп составляются вручную и содержат ключевые слова разделов тематики. Причем список заголовков должен быть полным (т.е. охватывать всю тематику), заголовки групп должны быть достаточно точными и учитывать синонимию.

При поступлении нового документа в информационный фонд с помощью специальной программы выполняется анализ и принимается решение о включении этого документа в одну из групп. Критерием является минимизация памяти, необходимой для хранения записи ПОД в массиве. При этом, если поисковый образ поступившего документа содержит новый для выбранной группы дескриптор, то заголовок группы дополняется этим дескриптором. В связи с автоматической корректировкой заголовков групп не нужно абсолютной точности при первоначальном составлении заголовка, но требуется такая точность, которая позволила бы правильно выполнять группировку документов. Для этого, практически необходимо около 10 дескрипторов.

---

[*] Все параметры, названные в этом разделе, определены при выборочном анализе, проведённом вручную. В процессе работы над системой они будут уточняться и корректироваться.

www.ruslang.ru

66157

БИБЛИОТЕКА
ИНСТИТУТА РУССКОГО ЯЗЫКА
ИНСТИТУТА ЯЗЫКОЗНАНИЯ АН СССР

Ввиду того, что по мере новых поступлений в информационный фонд и быстрого увеличения объема информации по отдельным разделам тематики соответствующие группы будут расти, необходимо предусмотреть дополнение списка заголовков групп новыми заголовками и корректировку поискового массива. Эта процедура выполняется полуавтоматически. Программа осуществляет контроль за величиной групп и при достижении сигнального уровня выдаёт сообщение о необходимости разделения некоторой группы на две. Индексатор формирует заголовок новой группы и, возможно, корректирует старый заголовок, после чего повторно работает программа группировки информации.

Группировка информации в поисковом массиве позволяет ввести внутригрупповые коды дескрипторов, которые короче внешних кодов в два раза. Кроме этого различие "родовых" и "видовых" дескрипторов позволяет исключением кодов "родовых" дескрипторов сократить общее количество записи ПОД приблизительно в 2 раза. Если же в какой-либо ПОД не входит один из "родовых" дескрипторов, то он, напротив, должен быть указан в записи ПОД, и этот факт интерпретируется программой при поиске как отрицание вхождения данного дескриптора в поисковый образ документа.

Сократив объем массива записей ПОД, мы одновременно с этим получаем массив заголовков групп, который значительно меньше массива записей ПОД, содержит ценную информацию об этом массиве и позволяет реализовать алгоритм двухступенчатого поиска. Сначала осуществляется поиск на массиве заголовков групп, причем используется тот же самый критерий смыслового соответствия на вхождение дескрипторов запроса в дескрипторы заголовка группы с учетом логических отношений. Затем на записях ПОД затронутых групп отыскиваются документы, отвечающие на запрос.

Поиск на массиве заголовков и массиве записей ПОД осуществляется независимо, поэтому при прямой организации массива записей ПОД возможна инверсная организация массива

заголовков.

В результате сокращения объема поискового массива 500 тыс. ПОД можно расположить на 0,5 магнитной ленты ЭВМ "Минск-32". Ожидаемое среднее время полной обработки поступившего запроса на массиве 500 тыс. документов – 10 минут, причем около 6 минут будет использовано на поиск и распечатку библиографических описаний и всего лишь 4 минуты на поиск документов. При улучшении средств ввода, вывода и передачи информации характеристики системы можно будет, по-видимому, значительно улучшить.

ЛИТЕРАТУРА

1. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики, М., 1968.
2. Бернштейн Э.С. Формализованный язык и критерий смыслового соответствия информационно-поисковой системы "Пусто-Непусто-4". НТИ, 1963, № 12.
3. Бернштейн Э.С., Лахути Д.Г., Чернявский В.С. О критериях оценки поисковых систем. "ТИ, 1964, № 3.
4. Лахути Д.Г. Семантическая поисковая система "Пусто-Непусто-2". Сб. "Системы управления непромышленными предприятиями". М., 1969.
5. Леонтьева Н.Н. О создании информационного языка на базе семантического анализа текста. М., ИРяз., 1970.
6. Юпатов Е.П., Корочкова И.Д., Тарасов В.И. Отраслевая автоматизированная система информационного обеспечения "Кристалл-легпром". М., 1970.
7. Кросс Р.К., Гардэн Ж.К., Леви Ф. Синтол – универсальная модель системы информационного поиска. М., 1968.
8. **Smart, Scientific Report Computation Laboratory Harvard University, 1964, No 7, 1965, No 9, No 10, 1966.**
9. Шаумян С.К., Мовшон Л.М. Аппликативная грамматика и информационные языки. НТИ, 1970, № 2.
10. Информационно-поисковая система "БИТ". Киев, 1968.
11. Иванова Н.И., Маргаритов В.Б. Введение элементов грамматики в дескрипторный информационно-поисковый язык по химии и химической промышленности. НТИ, 1968, № 10.
12. Керимов С.К., Иванова Н.И. Машинная реализация ИПС "Аргон-1". НТИ, 1970, № 3.
13. Черный А.И. Общая методика построения тезаурусов. НТИ, 1968, № 5.
14. **Cros R.S., Gardin J.C., Levy F. L'automatisation des recherches documentaires, un modèle général: Le SINTOL. Paris, 1964.**

I5. Колковский А.К., Мельчук И.А. О семантическом синтезе. Проблемы кибернетики, вып. I9. М., I967.

I6. Информационно-поисковые системы и автоматизированная обработка Научно-технической информации. Труды III Всесоюзной конференции, т. I - информационно-поисковые системы. М., I967.

I7. Королев Э.И. и др. Дескрипторный и тезаурусный словари для ИПС по материально-техническому снабжению. М., I969.

I8. Чернявский В.С., Лахути Д.Г., Федоров Е.Б. К вопросу об автоматическом построении алгоритма автоматического индексирования для систем типа "Пусто-Непусто". НТИ, I969, № I0.

I9. Мельчук И.А. Алгоритм морфологического анализа. Проблемы кибернетики, № 5, I962.

20. Автоматическая отраслевая система научно-технической информации "Реферат". М., I970.

№№ п/п	Шифр ИПС	ИПЯ: Д-дескрипторный ЕЯ-естественный УДК	ИПЯ: Наличие грамматики (Г)	Логика: Б-базисные отношения КСС-критерий смыслового соответствия	Объем дескрипторного словаря	Величина ПОД	Вид индексирования документов
I.	"Пусто-Непусто-2" 1965-1970	Д	-	Б + КСС 2 эшелона	3600 дескр.	от 10 дескрипт. до 130-140 дескрип.	Полуавтоматич.
2.	"Реферат" 1966-1969	Д	Г	Б + КСС 7 эшелонов	8000 дескр.	10-15 дескр.	Ручное
3.	"Сетка" 1965-1969	Д	-	Б + КСС "И", "ИЛИ", "НЕ"	2500 дескр.		Ручное
4.	"Ордината" 1966	Д	-	Б + КСС "И", "ИЛИ", "НЕ"	19000 дескрип.	8-10 дескр.	Ручное
5.	"Аргон" 1969	Д	Г	Б +	25000 дескр.	15-20 дескр.	Ручное
6.	"Агроэлектрин" 1969	Д	Г	Б +	5000 дескр.		Ручное
7.	"Кристалл" 1968	ЕЯ	Г	Б - КСС -3 эшелона	-	15 ключевых слов	автокодирование
8.	"Орбита" 1970	УДК	Г	Б + КСС 3 эшелона	-		Ручное

Таблица

ТОРЫХ АВТОМАТИЗИРОВАННЫХ ИПС

Вид использования базисных отношений	Среднее время обработки 1 запроса	Параметры ИПС: полнота	Параметры ИПС: точность	Принцип построения и функционирования, К-контур	Технические средства реализации
Автомат.	1 мин на массиве 50 тыс. док.	65-70% (1-й эшелон)	50%	ИРИ-2к. прямой РП - 2к. прямой	"Минск-22"
Ручное	При пакете 300 запросов 5 часов на 200 тыс.док.	С граммат. 75% Без граммат. 60%	5% 20%	- " -	"Минск-22"
Ручное	8-10сек.	88%	86%	ИРИ-2к. РП-2к. прямой	"Урал-4" "Минск-22"
Ручное	8 мин.на массиве 36 тыс.	80%	60%	ИРИ -2 к. РП - 2к. инверсный	"Минск-22"
Ручное		64%	36%		"Минск-22"
		80%	75,8%		"Минск-22"
Ручное		72% (для ИРИ)	72%	ИРИ - 2к. прямой РП - 2к. прямой	"Минск-22"
Автомат.				ИРИ и РП-1к.	счетно-перфорационные машины