НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК УКРАИНЫ
ИНСТИТУТ ПРОБЛЕМ МОДЕЛИРОВАНИЯ В ЭНЕРГЕТИКЕ

На правах рукописи

ГРИЦЮК Людмила Ивановна

РАЗРАБОТКА И ИССЛЕДОВАНИЕ СПОСОБОВ СВЕРТЫВАНИЯ И ФОРМАЛИЗОВАННОГО ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ В ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ

05.13.09 - математическое и программное обеспечение вычислительных машин и систем

Автореферат диссертации на соискание ученой степени кандидата технических наук

Киев - 1995

Диссертацией является рукопис

Работа выполнена в Институте проблем моделирования в энергетике НАН Украины

Научный руководитель: кандидат технических наук З.Х.Борукаев

Официальные оппоненти: д.т.н. А.И.Стасик
к.т.н. А.П.Токарь

Ведущая организация: Институт прикладной информатики при государственной администрации г.Киева

Защита состоится "29" июня 1995 г. в 15 часов на заседании специализированного совета Д.01.91.01 в Институте проблем моделирования в энергетике по адресу: 252680, Киев-164, ул. Генерала Наумова,15

С диссертацией можно ознакомиться в библиотеке института

Автореферат разослан "26" мая 1995 г.

Ученый секретарь специализированного совета, кандидат технических наук З.П. Семагина

Актуальность проблемы. На современном этапе развития науки и техники ощутимые достижения в любой сфере исследований возможны только при условии существенного повышения качества и оперативности получаемой информации. В связи с этим большую значимость приобретают вопросы децентрализованного накопления информационных массивов с обеспечением возможности быстрого обмена информационными ресурсами и повышения эффективности средств вычислительной техники, реализующей переработку информации.

Что касается собственно информации, то для ее хранения в базе данных (БД), оперативной обработки и предоставления для неоднородных групп пользователей существенным является способ организации на машиночитаемых носителях. В этой связи приобретают значимость исследования, связанные с формальным представлением информации, что предполагает оперирование теми единицами текста, которые составляют его словесную форму, имеют формальные признаки выделяемости и служат индикаторами содержания, а также изыскание методов и средств для хранения подготовленной подобным образом информации в базе данных.

Применение для реализации поставленных задач имеющихся как специализированных, так и универсальных пакетов программ подразумевает , во-первых, использование формы представления данных, не всегда удобной для исследуемой предметной области и, во-вторых, не всегда позволяет автоматизировать этапы исследования, специфичные для потребностей тематической области. Большинство из известных разработок связаны либо с дополнительной специализацией базового пакета в конкретных приложениях за счет подключения сервисных и функциональных программных модулей, либо с созданием специального инструмента исследования, включающего информационное, алгоритмическое и программное обеспечение, а также человеко-машинную технологию исследования.

Цель работы. Целью диссертационной работы является развитие методов построения концептуальных моделей данных в вычислительных системах и разработка способов свертывания входной информации и средств ее формализованного представления в базе данных с учетом особенностей предметно-тематической области.

Для достижения поставленной цели в диссертации были сформулированы и решены следующие задачи:

- разработать обобщенную концептуальную модель данных путем исследования концептуальных требований пользователей и естественных связей предметной области;

- исследовать информативность различных частей входного документа и, используя концептуальную модель, выбрать метод его свертывания;

- исследовать возможности по отображению информации, подлежащей хранению в базе данных, на машиночитаемых носителях с позиций эффективности поиска и выбрать способ ее формального представления с учетом особенностей тематической области;

- обосновать выбор типа модели данных, построить схему базы данных и выбрать метод управления данными для обеспечения эффективных процессов обработки и хранения информации;

- разработать соответствующее математическое обеспечение для реализации таких функций обработки данных как хранение на машиночитаемых носителях, лексикографическую обработку терминов тематической области и выдачу информации в ответ на запрос пользователей;

- разработать программный комплекс для вычислительных систем, реализующий содержательную переработку информации и основные виды поиска предметной области.

Методы исследования. При решении поставленных задач в диссертационной работе были использованы теория и методы проектирования информационных систем и баз данных, а также методы и средства разработки программного обеспечения для вычислительных машин и систем.

Научная новизна. К основным научным результатам относятся следующие:

- разработана методика построения обобщенной концептуальной модели данных, отражающая информационные требования пользователей и естественные связи предметной области;

- разработана методика построения определяющих матриц для формального представления информации предметно-тематической области;

- на основе анализа существующих информационно-поисковых языков в зависимости от уровня интеграции лексики ес-

тественного языка разработан единый информационно-поисковый язык, обеспечивающий уменьшение семантических ограничений для информации, хранящейся на машиночитаемых носителях;

- разработана новая методика лексикографической обработки терминов тематической области, отличительной особенностью которой является приписывание каждому термину из группы тождественных понятий условного эквивалента той же группы.

Практическая ценность работы. На основе полученных в диссертационной работе результатов разработаны алгоритмы, реализованные в виде комплекса программных модулей в автоматизированной информационной системе (АИС) "ПОИСК", позволяющей хранить информацию на машиночитаемых носителях, производить ее поиск и выводить для просмотра в свернутом виде, получать распечатки требуемых сочетаний реквизитов в соответствии с концептуальной моделью данных, в том числе реализовывать важную для пользователей процедуру автоматического оформления списка литературы.

Решены практические задачи создания и ведения базы данных по реферативному контуру по электронному моделированию, электрическим машинам, фонду описаний изобретений в академических учреждениях и нумерационному фонду описаний изобретений по нескольким подклассам разделов G и H Международной классификации изобретений.

Реализация результатов работы. Диссертационная работа выполнена в ИПМЭ НАН Украины в рамках научно-исследовательских тем "Термин", выполненной в течение 1985-1989 гг. по Постановлению Президиума АН УССР N 537 от 5.12.84 г., "Автоматизация", выполненной в течение 1990-1991 гг. по Постановлению Бюро Президиума АН УССР N 66-Б от 7.03.90 г., и "Интерфейс", выполняемой в течение 1993-1994 гг. по Постановлению ГКНТ N 15 от 1.03.93 г.

Публикации. Основные положения и результаты диссертации нашли свое отражение в шести опубликованных работах.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения и содержит 148 страниц машинописного текста, 40 рисунков и 8 таблиц. Библиография содержит 181 наименование.

СОДЕРЖАНИЕ РАБОТЫ

Во введении показана актуальность, научная новизна и практическая значимость работы, определены цель, задачи и методология исследования.

В первой главе отмечено, что система информационного обслуживания, призванная предоставлять потребителю новую информацию, определяющую стратегию и тактику научного поиска, выбора направления исследования, постановку задачи и методы ее решения, может быть представлена тремя составляющими - документальной, фактографической и концептографической, которые могут функционировать только на основе своих специфичных рядов вторичных документов. Для получения последних используют аналитико-синтетическую переработку информации - способ преобразования исходной информации, включающий совокупность операций по систематизации, анализу и синтезу информации и позволяющий выразить содержание исходного текста в более экономичной форме при сохранении или некотором допустимом уменьшении его информативности, называемый свертыванием.

Рассмотрены виды свертывания информации в зависимости от сферы коммуникаций. Это обусловило появление научного и информационного свертывания; в зависимости от типа и характера потребностей в информации, что обусловило появление двух подходов к свертыванию: метаинформативного, связанного с подготовкой вторичных документов библиографического ряда, и информативного, связанного с подготовкой документов фактографического ряда. Исследованы информативность и функции аннотаций и рефератов как наиболее распространенных способов интеллектуального свертывания в сфере информационного обслуживания.

Обобщены факторы, влияющие на процесс свертывания, и отмечено, что выбор способа свертывания информации определяется:

- назначением подготавливаемых вторичных документов;
- их потенциальной включаемостью в ту или иную сферу информационного обслуживания;
- характером входной информации;

- типом структуры текста;

- спецификой отрасли или предмета, для которых создается вторичный документ.

Отмечено, что особым и наиболее распространенным видом свертывания информации является индексирование - описание содержания и формы сообщения средствами того или иного информационно-поискового языка (ИПЯ). Такой язык предназначен для формализованного описания смыслового содержания документов и составления запросов к системе на поиск и выдачу информации. Проведен анализ ИПЯ в зависимости от уровня интеграции лексики естественного языка, в связи с чем выделены два больших класса языков - посткоординируемые и предкоординированные ИПЯ. Установлено, что посткоординируемые ИПЯ предназначены для такого способа использования, при котором сложные классы строятся из простых сначала при переводе (потенциально), а затем реально при информационном поиске. Рассмотрен наиболее распространенный язык этого класса - дескрипторный и, наряду с его преимуществами и удовлетворительными результатами при использовании для поиска научно-технической информации, отмечены недостатки при использовании для других видов поиска, в частности, для патентного. Рассмотрены системы, близкие по подходу к дескрипторным, но имеющие табличную форму выражения и реализованные в виде матриц; показана модификация матриц с учетом специфики отрасли. Относительно предкоординируемых ИПЯ установлено, что они имеют такую структуру, при которой перевод текста сводится к включению его в предварительно построенные сложные классы. В этой связи рассмотрены перечислительные классификации, в частности: Международная классификация изобретений, используемая в области патентного поиска, фасетные классификации и функционирующие на их основе информационно-поисковые системы (ИПС). Выполнена сравнительная характеристика дескрипторных и классификационных языков с точки зрения эффективности поиска информации путем сравнения коэффициентов полноты (П) и точности (Т):

$$П=ВР/Р, \quad Т=ВР/В,$$

где ВР - количество выданных релевантных документов;

Р - количество релевантных документов в поисковом массиве;

В - общее количество документов, выданных в ответ на запрос.

Отмечено увеличение точности поиска при использовании дескрипторных ИПЯ и повышение полноты выдачи для классификационных языков.

Рассмотрены также другие типы языков, в частности: язык библиографических ссылок, специализированные ИПЯ, которые организации - генераторы БД разрабатывают и используют для узких тематических направлений. При этом отмечено, что качество того или иного информационного языка можно оценить, сравнив их в процессе поиска, проводимого с помощью ИПС, которые в зависимости от характера выдаваемой информации подразделяются на документальные и фактографические. Помимо этой классификации и ее детализации на основе известных функционирующих систем, рассмотрена классификация ИПС по ряду других оснований.

Проведен анализ способов формирования запросов на поиск информации, исследовано влияние субъективного фактора и выработана общая стратегия поиска.

Вторая глава посвящена построению обобщенной концептуальной модели данных, дающей общее представление о предметной области и позволяющей правильно сформулировать задачу по исследованию способов свертывания информации предметной области. Первым шагом на этом пути явилось определение границ предметной области, поскольку решение любой функциональной задачи возможно лишь при наличии необходимых и достаточных знаний не только о ее сути, но и о фрагменте реальности, к которой она относится. В этой связи были проведены исследования по анализу информативности различных видов документов, а также статистические исследования частоты и объема их встречаемости в реально существующих БД, что предопределило выбор патентной информации для дальнейшего представления в БД.

Исследование динамики информационных потребностей пользователей в зависимости от этапов НИОКР показало, что информационное обеспечение должно адаптироваться не столько к плановым, сколько к техническим этапам, поскольку характер информационных потребностей даже на одном и том же плановом этапе может быть различным для различных категорий специа-

листов и определяться условиями конкретной задачи, совокупность которых образует логическую структуру разработки. Исходя из этого, разработана методика исследования и формирования информационных потребностей пользователей, включающая:

- использование таких форм получения информации, как анкетирование и интервьюирование потенциальных пользователей;

- знание проектировщиком БД предметной области, что достигается либо тесным контактом с заказчиком, либо использованием собственного опыта работы в конкретной области;

- изучение форм ведения поиска, характерных для рассматриваемой предметной области;

- изучение традиционных форм учета и отчетности, принятых в предметной области, а именно: картотек по ведению патентного фонда, справочных материалов и отчетов, содержащих сведения о динамике изобретательской деятельности, а также журнала регистрации заявок на изобретения.

Исследование влияния характера информационных потребностей на степень свернутости текста показало, что в случае, когда специалист не может достаточно четко сформулировать свой запрос или ему нужна информация тематического или систематизированного характера, только первичный документ может удовлетворить его потребности. Для выбора последних нужны вторичные документы, служащие для ориентации в тематическом документальном потоке, т.е. документы библиографического ряда. В случае, когда потребности специалиста могут быть удовлетворены определенным фрагментом текста, степень свернутости информации в котором выше или ниже, чем в первичном документе, нужны вторичные документы, служащие для ориентации в информационном потоке, т.е. документы фактографического ряда.

Выделение библиографической составляющей из текста входного документа не вызвало затруднений, поскольку она, как правило, представляется в стандартизованной форме, а для рассматриваемой предметной области идентифицирована ИНИД-кодами и сгруппирована в виде библиографического описания, расположенного на первой странице описания изобретения.

Выделению фактографической составляющей предшествовал анализ структуры частей описания изобретения с позиций ин-

формативности. Было отмечено, что хотя некоторые реквизиты библиографического описания и несут фактографическую нагрузку в виде указания на объект и область применения, эффективность фактографического поиска по ним невысока и информационный шум составляет 65-70%. Вследствие этого акцент переместился на собственно описание изобретения. Были проведены исследования по определению объема информации, подлежащего вводу в базу данных, способного удовлетворить информационные потребности пользователей. Для этого были рассмотрены альтернативные варианты по представлению информации в полнотекстовой и свернутой формах.

Сравнительный анализ полнотекстовых БД показал, что, с одной стороны, хранение полных текстов документов в памяти ЭВМ требует слишком много места и материальных затрат, что расточительно; с другой стороны - попросту дублирует массив документов, работа с которым в процессе автоматизированного поиска требует дополнительных интеллектуальных затрат, хотя для данной предметной области имеются предпосылки для их значительного сокращения, обусловленные высокой степенью структуризации патентных документов.

Рассмотрение методов свертывания информации, в частности реферирования как наиболее распространенного метода, выявило ряд ограничений, вызванных как субъективным фактором (неоднозначностью формализации, сложностью выявления семантической информации), так и невозможностью описывать некоторые факты традиционными средствами, что характерно для функциональных схем, химических структурных формул, блок-схем электронных устройств и объясняется противоречием между стремлением к универсальности при свертывании и высокой приспособленностью технических решений.

Исследование описания, проводимого с учетом особенностей тематической области - электронного моделирования, в которой результаты научных исследований и конструкторских разработок характеризуются сложными структурными построениями, выраженными в виде функциональных схем и описанными в формулах изобретений, показало, что наибольшую фактографическую нагрузку для этой области несет именно формула изобретения. Исследование с учетом остальных факторов, влияющих на процесс свертывания и выявленных в первой главе, также указыва-

ет на формулу изобретения, как квинтэссенцию всего описания: она является единственным критерием определения объема изобретении и имеет большое значение для прогнозирования, поскольку открывает возможности для дальнейших изысканий.

Внимание к формуле с позиций ввода в базу данных объясняется тем, что ее можно отнести к полуформализованному фрагменту первичной информации, поскольку ее составление структурировано, что регламентировано нормативными актами. Дополнительные ограничения на структуру формулы, введенные нами для информации тематической области и состоящие в разграничении состава и связей элементов схем и на протяжении ряда лет используемые при составлении заявок на изобретения, создают дополнительные предпосылки для ее формализации и автоматизации ввода в базу данных.

После описания анализа информации предметной области и формирования ряда запросов по фактографическому полю путем синтеза информации, выделенной при анализе предметной области и информационных потребностей пользователей, была построена обобщенная концептуальная модель данных и структурирована по направлениям поиска.

Исследовано также воздействие рассмотренных категорий информации, используемых в процессе построения концептуальной модели, на характеристики БД и отмечено, что учет одних информационных потребностей пользователей обеспечивает доступ только к текущим приложениям, что значительно сужает возможности проектируемой БД. Включение для анализа информации о предметной области расширяет ее возможности, благодаря использованию незапланированных приложений.

Описанная методика построения концептуальной модели данных представлена в работе в виде схемы.

Третья глава посвящена изысканию способов формализованного представления фактографической информации, выделенной на предыдущем этапе проектирования БД. В связи с этим были проанализированы возможности по отображению в БД полного текстового фрагмента и его графического представления в виде функциональной схемы. Оба указанные способа оказались неприемлимыми. Так, первый способ сопряжен с материальными затратами, аналогичными затратам при использовании полнотекстовых БД. Второй способ, хотя формально и реализуем, требует,

во-первых, дополнительной интеллектуальной обработки текстового материала, связанного с отсутствием нормализации терминов тематической области и заменой на схемах названий элементов и узлов цифровыми обозначениями, и, во-вторых, связан с определенными техническими трудностями ввода графических изображений на машиночитаемый носитель.

Приведенные аргументы указывают на необходимость дополнительной обработки свернутой информации перед введением в базу с целью уменьшения интеллектуальных затрат пользователей при поиске. Предпосылки для этого:

- введение обязательных дополнительных ограничений на структуру частей формулы изобретения;

- включение в формулу выверенных терминов, что обусловлено рядом требований, предъявляемых к ее составлению, в частности, лаконичности, определенности, общности и полноты, обеспечение которых во многом зависит именно от правильного выбора терминов.

Анализ имеющихся информационных языков показал несостоятельность их применения в чистом виде для рассматриваемой тематической области. Это обусловлено тем, что известные языки не могут достаточно точно отражать связи между элементами структуры, особенно тогда, когда структуры включают наборы одинаковых элементов, имеющих разветвленную систему связей между собой и с другими элементами. Так, дескрипторные ИПЯ не в состоянии решить задачу индексирования документов, относящихся к области вычислительной техники, схемотехники, т.е. там, где обычно выделяются особенности построения функциональных схем устройств. Поэтому для отражения сущности изобретения, выраженного в виде функциональной схемы и описанной в формуле изобретения, возникла необходимость в отыскании средств и методов по адекватному отображению в БД структур и включающих подструктур, составляющих суть конкретного технического решения.

Для решения задачи формального представления наиболее пригодным оказался матричный метод обработки информации, обусловленный табличной формой его реализации и способностью обеспечить нужную степень детализации при поиске. Однако использование известных и широко применяемых в процессе патентного поиска аналитических и структурных матриц не дало

положительных результатов. Подобно ряду исследователей, модифицирующих известные матрицы для различных тематических областей, были разработаны и исследованы матрицы для тематической области "электронное моделирование". Их характерным отличием является указание не только входящих в функциональную схему блоков устройства, но и фиксация их входов и выходов по строкам и столбцам матрицы соответственно. Такая матрица стала пригодной для отображения фактографической информации тематической области, поскольку, помимо состава, позволила отражать связи между блоками. Ряд дополнительных усовершенствований, таких как разграничение блоков ограничительной и отличительной частей формулы изобретения, позволяющее выделить новизну, привнесенную в процессе создания изобретения, указание цели изобретения и фиксация различных типов связей между блоками, создали предпосылки для реализации перспективных запросов пользователей.

Фактически вся информация, содержащаяся в формуле изобретения, представлена на матрице в интегрированном виде, что и обусловило выбор названия матрицы - определяющая. В случае, если бы идея изобретения потребовала большей степени детализации, предложенную методику можно было бы использовать для представления информации на уровне узлов и элементов устройства, описанного в формуле изобретения. Схема построения определяющих матриц такова, что позволяет заполнять их путем последовательного "движения" по формуле изобретения, избегая возвратов, чему способствует жесткая структуризация формул изобретения изначально и ввод ограничений на их структуру, осуществленный в процессе исследований и обусловленный особенностями тематической области.

В работе приведена поэтапная реализация процесса построения определяющих матриц, начиная от выделения приоритетных направлений предметно-тематической области до построения результирующей матрицы для абстрактной функциональной схемы; разработано руководство по построению определяющих матриц. Использование определяющих матриц позволило информацию, содержащуюся в формуле изобретения, представить формально, т.е. завершить первый этап формализации и начать разработку математического обеспечения по ее отображению на машиночитаемые носители.

Поскольку назначением БД является не только хранение информации, но и обеспечение связей между различными элементами данных, необходимых для эффективного представления в ответ на запросы, возникает потребность в соответствующем уровне ее проектирования. Для этого были исследованы взаимосвязи выделенных данных, произведено сравнение их структуры со структурными средствами известных моделей данных, что и предопределило выбор реляционной модели для рассматриваемой предметной области. Исследование данных, подлежащих вводу в БД, с позиции теории нормализации - основного понятия реляционной модели, состоящего в группировке элементов данных в ряд отношений и основанного на том, что определенные наборы отношений в процессе модификации обнаруживают лучшие свойства по сравнению с любыми другими наборами, содержащими те же данные, позволило путем последовательного приведения данных от первой нормальной формы к третьей получить схему реляционной модели данных предметной области "Патентный поиск".

Практическая разработка и реализация автоматизированной информационной системы потребовали выбора конкретного инструментального средства, который был осуществлен путем анализа альтернативных средств (параметрически настраиваемых универсальных средств, программного обеспечения на языках высокого уровня и СУБД) и завершился выбором реляционной СУБД FoxBASE.

Четвертая глава посвящена разработке математического и программного обеспечния для ведения формализованных данных первичной информации.

На этапе их ввода в базу данных предусмотрено использование вспомогательного файла, поля которого соответствуют реквизитам входного документа и расположены в порядке их следования в документе. Реализовано перераспределением введенных во вспомагательный файл данных в автоматическом режиме по реальным файлам в соответствии со схемой базы данных. Это позволило:

- облегчить процедуру ввода данных за счет "линейного" ввода требуемых реквизитов в одном цикле работы с массивом описаний;

- отказаться от использования на этапе ввода библиогра-

фической информации широко распространенных промежуточных операций, связанных с введением предмашинных форматов подготовки данных;

- снизить вероятность ошибок оператора за счет визуального контроля и редактирования данных с помощью дисплея.

Что касается фактографической информации, представленной формально с помощью определяющих матриц, то согласно разработанному алгоритму, ее ввод в БД не требует дополнительной интеллектуальной обработки и осуществляется путем занесения в соответствующее поле вспомогательного файла названий элементов на естественном языке. Они вводятся через запятую, с указанием в скобках их количества, если оно превышает единицу. На следующем шаге из вспомогательного файла эта информация затем переносится в файл-тезаурус и файл состава элементов схемы с использованием специально разработанного ИПЯ, состоящего из сочетания буквы латинского алфавита и цифры, символы которого заранее не перечисляются, а строятся динамически в процессе ввода новых данных. Разработанный ИПЯ характеризуется:

- компактностью, поскольку его термины - двух- или трехзначные символы;

- простотой формирования, поскольку осуществляющий эту процедуру алгоритм достаточно прост;

- прозрачностью для пользователя, т.к.процесс индексирования осуществляется автоматическим путем, а пользователь в процессе обращения к БД оперирует привычными для него терминами тематической области;

- неограниченным диапазоном терминов, являющимся функцией цифрового диапазона.

Предложенный язык относится к классу посткоординируемых ИПЯ, а по сути наиболее близок к классу дескрипторных языков. Поэтому было проведено их сравнение, в результате которого выявлены преимущества предложенного языка, а именно:

- отсутствие мнемонической связи между терминами ИПЯ и естественного языка, которым они приписываются, позволило получить преимущества, связанные с явлением устранения субъективного фактора и создало предпосылки для организации процесса индексирования автоматическим путем;

- уменьшение семантических ограничений за счет приписы-

вания каждому термину естественного языка его эквивалента на ИПЯ, в то время как в дескрипторном языке один дескриптор соответствует группе ключевых слов;

- отсутствие зависимости языка от объема и приращения новых знаний.

Что касается вопроса лексикографической обработки терминов, состоящей в нормализации терминов и устранении их синонимии, полисемии, омонимии и т.п., этот процесс, всегда сопровождающий процесс индексирования независимо от вида ИПЯ и особенностей предметно-тематической области, сопряжен с большими интеллектуальными затратами, связанными с привлечением высококвалифицированных специалистов, хорошо владеющих терминологией тематической области. Поэтому потребовался детальный анализ этапов формирования БД, на котором такая обработка являлась бы наиболее рациональной, а именно: этапа аналитико-синтетической переработки информации и построения определяющих матриц, этапа формирования вспомогательного файла, этапа индексирования, связанного с формированием файла-тезауруса и файла состава элементов схемы. Ни на одном из перечисленных этапов проведение лексикографической обработки не было признано целесообразным, поскольку не обеспечивало соблюдения установок, принятых при формировании БД, в частности: сокращения затрат специалистов отрасли, одноразового обращения к каждому экземпляру описания изобретения из обрабатываемого массива, а также обеспечения минимальных потерь при информационном поиске. Отсутствие лексикографической обработки терминов на указанных этапах приводило к расширению файла-тезауруса за счет появления в нем синонимии, однако, это не вызвало изменения намеченной стратегии, поскольку расширение тезауруса для узкотематических областей является незначительным. Им можно пренебречь по сравнению с выгодами, получаемыми за счет устранения потерь введенной информации вообще.

Разработанная методика обработки терминов предусматривает проведение распечатки фрагмента файла-тезауруса после очередного его формирования и направление затем высококвалифицированным специалистам для проведения лексикографической обработки с последующим занесением выявленной синонимии в файл синонимов в автоматизированном режиме. Это дает возмож-

ность пользователю при составлении запроса на поиск оперировать привычными ему терминами, выбираемыми из меню, сформированного исключительно из названий элементов тематической области. Последнее обеспечило достижение однозначности в терминологии для выражения поискового образа документа и запроса, что позволило их максимально приблизить и тем самым увеличить полноту выдачи. В отличие от методики составления классического тезауруса, здесь группы тождественных понятий не заменяются дескрипторами, а каждому из понятий ставятся в соответствие синонимы исключительно из терминов выделенной группы.

Разработано также математическое и программное обеспечение для формирования запроса на поиск и проведения поиска, реализующее диалоговый режим общения и позволяющий избежать ошибок, связанных с предварительным формированием запроса, а также отказаться от услуг инфопосредника.

Представлена схема функциональных возможностей АИС "ПОИСК", реализованная в виде комплекса программных модулей и позволяющая осуществлять основные виды поиска предметной области.

В заключении приведены основные научные результаты и практические выводы по работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Основной результат диссертационной работы состоит в том, что предложен комплексный подход к решению задачи формализованного представления, свертывания информации и построения концептуальных моделей данных в вычислительных системах.

Кроме того получены следующие научные и практические результаты:

1. Показана необходимость свертывания входной информации для удовлетворения информационных потребностей пользователей на любом уровне информационного обслуживания посредством специфичных рядов вторичных документов. Исследована зависимость между подходами к свертыванию информации, обусловленными типом и характером потребностей, и функциями порождаемых вторичных документов. Обобщены факторы, влияющие на

процесс свертывания информации и обусловливающие выбор способа свертывания.

2. Разработана методика исследования и формирования информационных потребностей пользователей и установлено влияние характера потребностей на степень свернутости текста.

3. Путем синтеза информации, выделенной при анализе информационных потребностей пользователей и анализе информации предметной области, построена обобщенная концептуальная модель данных, отражающая семантику объекта и инвариантная по отношению к методам управления данными.

4. Разработана методика формального представления фактографической информации входного документа на основе аналитико-синтетической переработки информаци предметной области.

5. Проведено исследование модели с позиций теории нормализации и построена схема реляционной модели базы данных предметной области "Патентный поиск".

6. Разработан единый ИПЯ, позволяющий уменьшать семантические ограничения для индексируемой информации за счет приписывания каждому термину его эквивалента на ИПЯ.

7. Разработана методика лексикографической обработки терминов тематической области, отличающаяся отсутствием обобщающих терминов для групп тождественных понятий.

8. Разработано математическое обеспечение для реализации функций хранения данных, лексикографической обработки терминов тематической области, формирования запроса на поиск информации и проведения поиска.

9. Разработан программный комплекс, реализованный в автоматизированной информационной системе "ПОИСК", обеспечивающей выполнение основных видов поиска предметной области.

Основные результаты диссертации опубликованы в работах:

1. Борукаев З.Х., Прицюк Л.И. Особенности формирования и ведения базы данных автоматизированной информационной системы "Патенты" // НТИ.сер.1.-1993.-N12.-С.15-18.

2. Борукаев З.Х., Прицюк Л.И.; Скуридин В.П. Патентная служба: поиск ведет ЭВМ // Вопр.изобретательства.-1991.-N3.-с.50-53.

3. Прицюк Л.И. Автоматизация работы патентного подраз-

деления // Интеллектуальная собственность. -1994.- N3-4. - с.35-37.

4. Прицюк Л.И. Некоторые вопросы генерации фактографической информации // Там же.-1994.-N5-6.-с.43-46.

5. Прицюк Л.И. Построение концептуальной модели базы данных автоматизированной системы (тематическая область - электронное моделирование) // НТИ.сер.2.-1994.-N3.-с.1-4.

6. Борукаев З.Х., Прицюк Л.И., Скуридин В.П. Об одном подходе к построению автоматизированной информационно-поисковой системы для патентных подразделений.-Киев,1990.-38с.-(Препр.АН УССР. Ин-т проблем моделирования в энергетике; 90-22).

В работах [1], [2], [6] автору принадлежит:

[1] - разработка программного обеспечения,

[2,6] - разработка конкретных методов обработки входных данных, ориентированных на использование ЭВМ.

GRITSYUK L. I. DEVELOPMENT AND INVESTIGATION OF METHODS OF INFORMATION FOLDING AND FORMALIZED REPRESENTATION IN COMPUTING SYSTEMS

Dissertation for candidate of science degree by speciality 05.13.09 "Software of computing machines and systems". The Institute of Simulation Problems in Power of National Academy of Science of Ukraine. Kyiv. 1995.

During the information storage in database the method of information organization and representation on the machine-readable medium is highly essential. In connection with this, a complex of standardization restrictions of an input documents array of the problem field and mechanism for its formalized representation are developed. The method of database forming is proposed, which implements automatic redistribution of information inputted by a single access to the input document. The computer technology of maintenance of formalized data of initial information is developed, in particular, the technique of dictionary processing of topical field's terms, and unified information retrieval language is developed too, which ensures absolute matching to the terms of natural language.

Грицюк Я.І. Розробка та дослідження засобів згортання та формалізованого подання інформації в обчислювальних системах.

Дисертація на здобуття вченого ступеня кандидата технічних наук в спеціальності 05.13.09 - математичне та програмне забезпечення обчислювальних машин та систем, Інститут проблем моделювання в енергетиці НАН України, Київ, 1995.

При зберіганні інформації в базі даних суттєвим є засіб її організації та подання на носіях, які читаються машинами, в зв'язку з чим розроблено комплекс обмежень по стандартизації вхідного масиву документів предметної області та апарат для її формального подання. Використано засіб формування даних, завдяки якому реалізується автоматичний перерозподіл інформації, введеної шляхом одноразового направлення до масиву вхідних документів. Розроблено комп'ютерну технологію ведення формалізованих даних первинної інформації, зокрема, методику лексикографічної обробки термінів тематичної області та єдину інформаційно-пошукову мову, яка забезпечує абсолютну відповідність терміним природньої мови.

Ключові слова: згортання інформації, формалізоване подання, обчислювальні системи.

Подписано к печати 25.05.1995 г. Формат 60x84/16
Бумага офсетная Усл.-печ.лист.1,0.Уч.-изд.лист 1,0.
Тираж 100. Заказ 255. Бесплатно

Полиграф. уч-к Института электродинамики АН Украины,
252057, Киев-57, проспект Победы, 56.

Грядок Д.І. Розробка та дослідження засобів автоматизованого формалізованого подання інформації в обчислювальних системах.

Дисертація на здобуття вченого ступеня кандидата технічних наук за спеціальністю 05.13.06 – математичне та програмне забезпечення обчислювальних машин та систем. Інститут проблем моделювання в енергетиці НАН України, Київ, 1995.

Для документів, інформації в базі даних суттєвим є засіб подання їх у вигляді, які читаються машиною, в зв'язку з чим розроблено комплекс обмежень по стандартизації вхідного масиву документів предметної області та правил формалізованого подання. Використано засоби формування перевірки даних, реалізується автоматичний переклад кодування інформації. Введено заходи однорідного направлення до масиву вхідних документів. Розроблено комп'ютерну техніку ведення формалізованих даних первинної інформації, зокрема, методику декомпозиційної обробки термінів тематичної області та єдину інформаційно-пошукову мову, яка забезпечує абсолютну відповідність термінам природної мови.

Ключові слова: вторинна інформація, формалізоване подання, обчислювальні системи.

Подписано к печати 25.05.1995 г. Формат 60х84/16
Бумага офсетная. Усл.-печ. лист 1,0. Уч.-изд. лист 1,0.
Тираж 100. Заказ 455. Бесплатно

Полиграф. уч-к Института электродинамики АН Украины,
252057, Киев-57, проспект Победы, 56.