ИНСТИТУТ ЭЛЕКТРОННЫХ УПРАВЛЯЮЩИХ МАШИН

Р.И.ПУШКАРСКАЯ

СИСТЕМА АВТОМАТИЧЕСКОГО ИНДЕКСИРОВАНИЯ НА ЯЗЫКИ ПРЕДМЕТНОГО ТИПА

(Специальность № 255 – техническая кибернетика)

АВТОРЕФЕРАТ

диссертации, представленной на соискание
ученой степени кандидата технических наук

Москва – 1969 г.

В области вычислительного машиностроения происходит эволюция, в ходе которой постоянно увеличивается удельный вес математического обеспечения. В процессе этой эволюции меняется и сам характер математического обеспечения - если прежде оно сводилось в основном к набору стандартных программ, то теперь оно все больше превращается в комплекс систем, обеспечивающих проникновение вычислительных устройств, их применение в различных сферах умственной деятельности человека, среди которых обработка информации занимает одно из важнейших мест. Следует ожидать, что эта тенденция в дальнейшем лишь усилится, и что проблемы, связанные с использованием вычислительных устройств и комплексов, приобретут для вычислительного машиностроения еще большее значение.

Среди сфер возможного использования вычислительного оборудования одной из важнейших является сфера автоматической обработки семантической информации. В настоящее время возможны три направления в автоматической обработке семантической информации. Два из них - автоматический перевод с одного естественного языка на другой и автоматическое реферирование - продвинуты еще недостаточно для того, чтобы стать широкой сферой применения вычислительного оборудования третье же - автоматический поиск семантической информации -

представляет собой сферу потенциально чень широкую и уже достаточно подготовленную для производственного использования кибернетических методов и вычислительных машин. В этой, последней, сфере наименее разработанным и наиболее перспективным является автоматическое индексирование, т.е. перевод содержания обрабатываемых документов и текстов на искусственный язык того или иного рода. Без автоматизации процедуры индексирования невозможно создание полностью автоматизированных систем поиска и обработки семантической информации.

В настоящее время известно очень мало работ, посвященных автоматическому индексированию, что объясняется повидимому трудностью задачи – как теоретической, так и технологической. В нашей стране, а возможно и в мире, существует лишь одна, функционирующая в полупроизводственном режиме, система автоматического индексирования – система Пусто-Непусто. Остальные находятся на значительно более ранней стадии реализации. Существенным во всех этих работах является то, что все они ориентированы на дескрипторные информационные языки, допускающие пословный перевод с естественного языка. Так, например, в наиболее разработанной системе автоматического индексирования – Пусто-Непусто-2 – за рамки пословного перевода выходит лишь автоматический анализ трех омонимов и около ста словосочетаний. Ожидается, что даже после завершения этой системы число омонимов и словосочетаний

увеличится не более, чем на порядок и их анализ останется, таким образом, скорее исключением, чем регулярным элементом системы.

В то же время большая часть используемых в мировой информационной практике языков относится к другому типу, который не допускает пословного перевода в процессе индексирования и требует осмысления содержания переводимых текстов. К этим языкам относятся все классификационные языки, а также языки предметные, которые, как известно, имеют в информационной практике наиболее широкое применение. Если работ, посвященных автоматическому индексированию мало, то автоматизация перевода на языки последнего типа вообще до сих пор не рассматривалась. Если исследователи и касались этого вопроса, то только для того, чтобы сказать о невозможности его реализации. В результате для внедрения вычислительных комплексов оказывается закрытой обширная область – область автоматического индексирования информационных языков предметного типа.

Таким образом, настоящая диссертация посвящена проблеме расширения области использования вычислительных машин и представляет собой исследование по математическому обеспечению вычислительного машиностроения, ставящее своей целью исследование возможности автоматического перевода на предметные языки.

В представленной работе проблема автоматического индексирования на языки недескрипторных типов понимается не как абстрактный теоретический вопрос о возможности решения этой проблемы. Целью диссертации является построение действующей модели, эксперимент на ней и создание таким образом предпосылок для разработки практически действующих информационных систем с автоматическим индексированием на языки предметного типа.

Положенная в основу настоящей диссертации общая идея построения автоматической системы перевода на языки недескрипторных типов заключается в следующем:
Имеется поисковая система, в которой используется предметный язык. Необходимо построить систему автоматического индексирования на язык этой поисковой системы. Задача решается построением для системы автоматического индексирования вспомогательной поисковой системы дескрипторного типа с автоматическим переводом документов на язык этой вспомогательной поисковой системы. Документы, подлежащие переводу на предметный язык, вводятся вначале во вспомогательную поисковую систему, переводятся автоматически на ее язык и образуют ее поисковой массив. После этого во вспомогательную поисковую систему в качестве запросов вводятся слова того предметного языка, на который должны быть переведены исходные документы. По этим запросам проводятся поиски и затем для каждого документа собира-

ются те слова предметного языка, на которые, как на запросы, документ был выдан. Объединение этих слов и образует перевод документа на предметный язык. Таким образом, в соответствии с изложенной схемой задача построения системы автоматического перевода на предметный язык сводится к задаче построения специализированной вспомогательной поисковой системы дескрипторного типа с автоматическим индексированием. С этой задачей, вообще говоря, может справиться любая достаточно мощная универсальная поисковая система дескрипторного типа, располагающая автоматическим индексированием, например, система Пусто-Непусто-2. Однако, использование в качестве элемента системы автоматического индексирования столь мощной и громоздкой системы, как система Пусто-Непусто-2, сделает систему автоматического индексирования неприемлемо сложной и дорогой, и такое решение задачи по всей видимости не будет представлять никакого практического интереса. Поэтому главная трудность заключается в построении такой вспомогательной дескрипторной поисковой системы, которая, с одной стороны, была бы достаточно мощной для решения стоящей перед ней задачи, а с другой, была бы достаточно проста для того, чтобы не стать препятствием при практической реализации системы автоматического индексирования. Основным требованием к семантической силе системы индексирования является отсутствие или почти отсутствие потерь, ибо если шум можно рассчитывать в дальнейшем устра-

нять путем каких-либо ухищрений в основной поисковой системе, то потери индексирования ничем не восполнимы. Это требование удалось выполнить путем введения в логику системы автоматического индексирования специального аппарата, называемого в диссертации косвенными признаками.

Таким образом предметом защиты является действующая экспериментальная система автоматического индексирования на информационно-поисковые языки предметного типа.

В настоящей диссертации в качестве предметного языка были выбраны фиксированные заголовки и подзаголовки словника предметного указателя к реферативному журналу "Геофизика". В качестве документов, содержание которых переводилось на этот язык, использовались рефераты из этого же журнала. Задача автоматического индексирования на языки предметного типа решается, таким образом, путем создания системы, позволяющей автоматически распределять тексты рефератов по фиксированному списку заголовков и подзаголовков предметного указателя, т.е. по-существу путем создания системы автоматического составления предметного указателя.

Вопрос автоматического составления предметного указателя ставится уже не впервые, однако специалисты как правило единодушно высказывались о невозможности использовать при автоматическом составлении указателя текста документа.

До сих пор считалось, что при автоматизации процесса составления предметного указателя можно использовать либо только заглавие, выделяя автоматически из него информативные слова и пермутируя их, либо ключевые слова, приписанные тексту человеком. В последнем случае автоматизация составления предметного указателя сводится к автоматической сортировке ключевых слов.

Диссертация состоит из введения, шести параграфов и заключения. Во введении сформулирована и обоснована задача, в § I приводится обзор существующих методов автоматического перевода на предметные языки, в § 2 рассматривается задача перевода на языки предметного типа, § 3 посвящен основной части диссертации - методике разработки вспомогательной поисковой системы, предназначенной для автоматического перевода текстов на естественном языке на языки предметного типа, в § 4 описана реализация системы, в § 5 описан эксперимент и его результаты, в § 6 определены области применения разработанной системы, в заключении даны основные выводы.

Поскольку проблема автоматического индексирования на языки предметного типа была сведена к построению вспомогательной поисковой системы дескрипторного типа, которая по-существу является системой автоматического составления предметного указателя, то основную часть диссертации составляет описание этой вспомогательной системы.

В основу логики вспомогательной поисковой системы дескрипторного типа была положена логика поисковых систем класса "Пусто-Непусто". Оказалось возможным модернизировать ее и свести несколько эшелонов выдачи в один, что позволило выдержать принятую в реферативных журналах форму предметного указателя. Существенным отличием от логики других систем класса "Пусто-Непусто" является введение так называемых косвенных признаков, с помощью которых удалось снизить процент потерь информации. Подробнее о логике системы будет сказано ниже.

Язык системы строился по тому же принципу, что и язык поисковых систем класса "Пусто-Непусто", т.е. слова естественного языка объединялись в классы эквивалентности, называемые дескрипторами. Заголовки и подзаголовки предметного указателя послужили основной базой для выбора дескрипторов.

При создании языка первоначально набирался словник системы (следует отличать от словника предметного указателя), в который вошли существительные и прилагательные как несущие основную смысловую нагрузку текстов. Словник набирался по текстам рефератов раздела "Динамическая и синоптическая метеорология" РЖ Геофизика. В словник вошло 5 000 слов.

На базе словника был составлен словарь дескрипторов, в котором каждый дескриптор имеет свой номер.В ходе работы было проведено два эксперимента, отличающихся друг от друга только словарным составом: в первом варианте в словарь вошел 991 дескриптор, во втором - 303 дескриптора.

В основном дескрипторы состоят из одного слова естественного языка. Например:

БУРЯ - 60

ЗАСУХА - 237

КОНВЕРГЕНЦИЯ - 320

и т.д.

В качестве дескрипторов выбирались и такие словосочетания как

БАКИНСКИЙ НОРД - 37

БАРИЧЕСКОЕ ПОЛЕ - 305

ЛИНИЯ ТОКА - 911

ФУНКЦИЯ ТОКА - 782,

поскольку они образовали устойчивые смысловые образования.

Слова естественного языка, являющиеся синонимами в данной области, по определению дескриптора объединялись в один класс эквивалентности. Например,

КРАТКОСРОЧНЫЙ

НА N ДНЕЙ $/N \leq 2/$

НА S ЧАСОВ $/0 < S \leq 72/$

объединены в один класс с номером 335.

Одно из слов класса эквивалентности выделяется в качестве основного и объявляется дескриптором. Как правило, дескриптору присваивается номер слова, которое в алфавитном порядке встретилось раньше других слов этого класса.

Например, синонимом термина ТАЙФУН являются термины: ТРОПИЧЕСКИЙ ЦИКЛОН, ТРОПИЧЕСКИЙ ШТОРМ, ТРОПИЧЕСКОЕ ВОЗМУЩЕНИЕ, ЦИКЛОНИЧЕСКИЙ ШТОРМ. Каждый из этих терминов имеет свой порядковый номер в словаре

ТАЙФУН – 697
ТРОПИЧЕСКИЙ ЦИКЛОН – 867
ТРОПИЧЕСКИЙ ШТОРМ – 868
ТРОПИЧЕСКОЕ ВОЗМУЩЕНИЕ – 866
ЦИКЛОНИЧЕСКИЙ ШТОРМ – 869

В качестве дескриптора выбран термин ТАЙФУН и всему этому классу приписан номер 697, который и является номером дескриптора.

Помимо таких очевидных синонимов существуют термины, объединение которых в один класс в пределах системы автоматического составления предметного указателя, несмотря на меньшую очевидность, оказывается целесообразным. Например, синонимом термина АНТИЦИКЛОН взяты выражения ОБЛАСТЬ ВЫСОКОГО ДАВЛЕНИЯ, ОБЛАСТЬ ПОВЫШЕННОГО ДАВЛЕНИЯ. Этот класс образует также один дескриптор с номером 22.

Существуют эквивалентные по смыслу термины, отличающиеся друг от друга лишь порядком следования слов. В качестве примера можно привести следующие выражения:

ВЫСОТНАЯ ПЛАНЕТАРНАЯ ФРОНТАЛЬНАЯ ЗОНА
ПЛАНЕТАРНАЯ ВЫСОТНАЯ ФРОНТАЛЬНАЯ ЗОНА

В последнее время часто употребляется просто выражение

ВЫСОТНАЯ ФРОНТАЛЬНАЯ ЗОНА,

синонимичное двум приведенным выше, Таким образом, возникает три синонимичных выражения и, соответственно, три сокращения

Этих выражений:

ВПФЗ, ПВФЗ, ВФЗ,

которые объединены в один класс с номером I25.

В первом эксперименте язык составлялся по типу языка поисковой системы общего вида. Однако, как выяснилось при проведении первого эксперимента, такой способ снижал качество автоматически составленного предметного указателя за счет большего шума и потерь.

Во втором эксперименте основу языка системы составили дескрипторы, входящие в заголовки и подзаголовки словника предметного указателя, или образующие их.

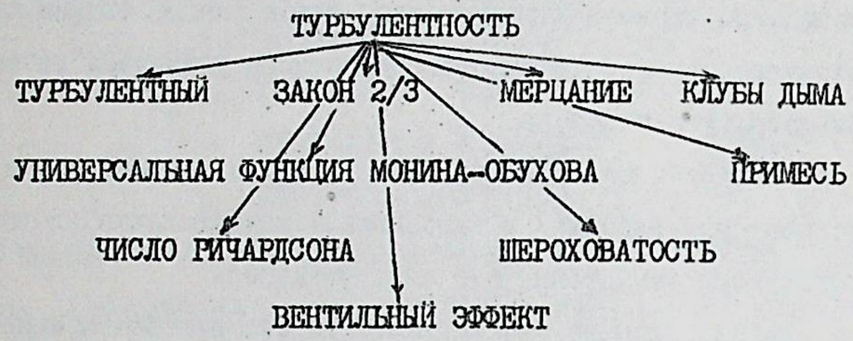
Логика системы автоматического составления предметного указателя состоит из трех элементов: базисных отношений между дескрипторами, критерия смыслового соответствия и косвенных признаков. Базисные отношения между двумя дескрипторами А и В устанавливались в том случае, если требовалось, чтоб данному заголовку, в котором есть дескриптор А, соотносился реферат, в котором встречается дескриптор В. Таким образом, базисные отношения строились по принципу тезаурусов. Наиболее распространенный вид отношений такой:

АРКТИКА
↓
АРКТИЧЕСКИЙ

Дескриптор АРКТИКА называется вышестоящим по отношению к дескриптору АРКТИЧЕСКИЙ, последний же называется нижестоящим по отношению к дескриптору АРКТИКА.

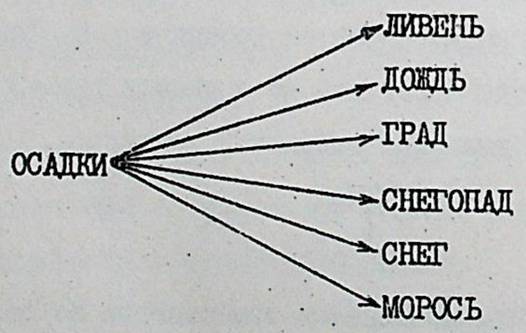
Во втором эксперименте были значительно расширены базисные отношения. Например, в первом эксперименте дескрип-

тору ТУРБУЛЕНТНОСТЬ подчинялся только один дескриптор ТУРБУЛЕНТНЫЙ. Во втором эксперименте дескриптору ТУРБУЛЕНТНОСТЬ подчиняется 9 дескрипторов, отношения между которыми можно выразить схемой:



Такое подчинение означает, что если в реферате упоминается ЗАКОН 2/3, то этот реферат должен быть соотнесен заголовку ТУРБУЛЕНТНОСТЬ и т.д.

Можно привести пример и более очевидного подчинения:



Критерий смыслового соответствия формулируется следующим образом:

Реферат соотносится заголовку (заголовку+подзаголовку), если для каждого дескриптора заголовка (заголовка+подзаголовка) в реферате имеется равный или нижестоящий дескриптор.

Таким образом, при автоматической предметизации множество рефератов, выдаваемых на один заголовок или комбинацию из заголовка и каждого подзаголовка, состоит из одного эшелона.

Для повышения качества системы были разработаны косвенные признаки, также входящие в логику системы и которые не могли быть подчинены непосредственно какому-либо дескриптору, т.к. состояли не из одного дескриптора, а из некоторого набора их. Каждый косвенный признак входит в систему на правах самостоятельного заголовка, однако, рефераты, соотнесенные ему, приписываются заголовку, к которому приписан данный косвенный признак. Во второй вариант системы вошло 23 косвенных признака. В качестве примера можно привести косвенные признаки к следующим заголовкам:

319	КОНВЕКЦИЯ	ВЕРТИКАЛЬНАЯ СТРУЯ 70,686
		ВЕРТИКАЛЬНЫЙ ПЕРЕНОС 70,528
585	ПРОГНОЗ	
	218 долгосрочный	ПРЕЕМСТВЕННОСТЬ СИНОПТИЧЕСКИХ ПРОЦЕССОВ 575,646,602

Процесс индексирования в проведенном эксперименте не был автоматизирован, однако, индексация производилась алгоритмически. Для этого была разработана специальная инструкция. Индексировался полный текст рефератов, за исключением таблиц, рисунков и формул. Все встретившиеся в реферате дескрипторы выписывались в порядке возрастания с исключением повторов. На индексировании было занято 5 человек, причем каждый из них индексировал в среднем за один день по 30 рефератов. Для контроля каждый реферат индексировался дважды.

Ввод информации в ЭВМ производился с помощью перфокарт.

Отдельно каждый заголовок вводился в машину как самостоятельный запрос, затем поочередно совокупность из заголовка и каждого подзаголовка вводилась в качестве самостоятельного запроса. После этого рефераты, вошедшие под каждую совокупность из заголовка и подзаголовка, исключались из списка номеров рефератов, вошедших под заголовок. В результате каждому заголовку и подзаголовку оказывались приписаны в порядке возрастания номера рефератов, им соответствующих.

Вывод информации производился на широкую печать, так что выданный на печать автоматически составленный предметный указатель может быть использован как форма для офсетной печати.

Машинное время, которое было затрачено на проведение эксперимента на машине "Гамма-барабан", составило I8 часов, что в пересчете на быстродействие ЭВМ "Минск-22" составляет 30 минут.

Оценкой автоматической системы индексирования на языки предметного типа является по существу оценка вспомогательной поисковой системы. Для оценки вспомогательной системы были выбраны следующие параметры:

I. потери
2. шум системы
3. среднее количество рефератов, приходящееся на один заголовок,
4. среднее количество заголовков, приходящееся на один реферат,

причем наиболее существенным является последний параметр, т.к. он характеризует глубину индексирования.

Результаты оценки сведены в таблицу № I.

Таблица I

Параметры	I вариант	2 вариант
количество заиндексированных рефератов	589	7I5
количество дескрипторов	99I	303
количество заголовков и подзаголовков	72	I05
потери	2%	I%
шум	20%	I0%
среднее количество заголовков, приходящееся на один реферат	4	4
среднее количество рефератов под одним заголовком	47	38

Для более глубокой оценки было проведено сравнение результатов автоматического и"ручного" составления предметного указателя. Для сравнения совершенно произвольно были выбраны несколько заголовков и подзаголовков. Результаты сравнения приведены в таблице № 2.

Данные в первом столбце приведены в таблице с исключенным шумом.

Таблица 2

заголовки	количество рефератов			
	при автомат. сост.		при ручном сост.	
	выдано	потеряно	выдано	потеряно
Антициклоны	7I	-	II	60
Барическое поле	28	9	I4	23
Вертикальные движения	29	2	7	24
Ветер, геострофический	I5	-	2	I3
Воздушные массы	24	I	I	24
Давление	I30	-	6	I24
Муссоны	I9	-	I4	5
Пассаты	6	-	-	6
Прогноз, численный	43	5	25	23
Смерчи	9	-	3	6
Струйные течения	44	2	20	26
Тайфуны	43	I	22	22
Тропопауза	27	-	7	20
Турбулентность	96	9	59	46
Ураганы	27	I	22	6
Фен	6	-	3	3
Фронты	45	2	II	36
Циклоны	I07	I0	36	8I

Из таблиц видно, что при ручном составлении теряется значительная часть информации. При автоматическом же составлении предметный указатель гораздо полнее. По сведениям Научно-методического отдела ВИНИТИ среднее количество заголовков, приходящееся на один реферат по РЖ Геофизика, составляет I, в системе оно равно 4. Таким образом, вспомогательная система дает большую глубину индексирования. Потери и шум системы невелики по сравнению с потерями и шумом поисковых систем общего вида.

З А К Л Ю Ч Е Н И Е

Предметом диссертации является система, автоматически выполняющая процесс перевода на предметный язык текстов, написанных на естественном языке. Задача решается следующим образом:

– для системы автоматического индексирования на языки предметного типа строится вспомогательная поисковая система дескрипторного типа с автоматическим индексированием на этот язык. Тексты, подлежащие переводу на предметный язык, вводятся во вспомогательную поисковую систему, в которую слова предметного языка вводятся в качестве запросов. В результате функционирования вспомогательной поисковой системы словам предметного языка оказываются сопоставлены некоторые наборы документов, или можно сказать иначе – каждому документу оказываются сопоставлены некоторые наборы слов предметного языка, которые и являются переводом текстов, написанных на естественном языке, на предметный язык. Вспомогательную поисковую систему удалось создать достаточно простыми и экономными средствами за счет расширения базисных отношений, за счет сокращения словаря дескрипторов, а в основном за счет включения в логику вспомогательной поисковой системы специального аппарата, называемого косвенными признаками.

Из диссертации могут быть сделаны следующие выводы:

Выводы:

I. Задача автоматического индексирования на языки предметного типа является разрешимой. Качество индексирования и простота используемой для этого системы позволяет рассчитывать на практическое значение систем такого рода.

2. Косвенные признаки, используемые в настоящей работе, являются очень сильным семантическим средством. Разработанная в настоящей диссертации система обязана своей простотой и высокими результатами индексирования именно косвенным признакам.

3. Принципы, выработанные в результате проведенного исследования и эксперимента имеют более общее значение – они могут быть использованы и для организации автоматического индексирования на классификационные языки.

Результаты диссертации были доложены на ряде семинаров, на секции документалистики Польской Академии наук, на Ш Всесоюзной конференции по информационно-поисковым системам и автоматической обработке информации в Москве в I966 г., на симпозиуме по информации и медико-биологическому прогнозированию в I968 г.

По диссертации опубликованы следующие работы:

I. Пушкарская Р.И., Чернявский В.С. Об автоматическом составлении предметного указателя. НТИ, 1964, № 7.

2. Пушкарская Р.И. Автоматическое составление предметного указателя к реферативным журналам. Труды III Всесоюзной конференции по ИПС и автоматизированной обработке НТИ. М., 1966.

3. Сухаребский Л.М., Пушкарская Р.И., Пушкарский В.Г. Принципы автоматического составления предметного указателя к реферативным журналам. Сб. "Вопросы научного прогнозирования", М., 1968.

4. Пушкарская Р.И. Логика и язык дескрипторной системы автоматического составления предметного указателя. НТИ, 1968, № 12.

5. Пушкарская Р.И. Эксперимент по автоматическому составлению предметного указателя (в печати).

Заказ 147
Тираж 150 экз. бесплатно
Ротапринт ИНЭУМ