tanulmányok

212/88

MTA Számítástechnikai és Automatizálási Kutató Intézet Budapest

Исследовательский Институт Вычислительной Техники и
Автоматизации Венгерской Академии Наук
Computer and Automation Institute
Hungarian Academy of Sciences

МОДЕЛИ АССОЦИАТИВНЫХ ОБРАЗОВ

Хенчей Густав

Tanulmányok 212/1988
Studies 212/1988

H-1502 Budapest PO Box 63. XI. Kende u. 13-17. Hungary

A kiadásért felelős:

DR. KEVICZKY LÁSZLÓ

Jelen tanulmány a szerző kandidátusi disszertációját tartalmazza, amelyet sikeresen megvédett.

ISBN 963 311 254 0
ISSN 0324-2951

Alfaprint

ОГЛАВЛЕНИЕ

Стр.

ВВЕДЕНИЕ

Задача сопоставления отдельно взятого объекта с множеством объектов с целью разбиения этого множества на две части, - похожих на данный объект и не похожих на него, - возникает очень часто как внутренняя задача в самых разных процедурах анализа информации - сортировки, выявления и описания структуры, организации быстрого поиска, распознавания и классификации, принятия решения в условиях большого числа альтернатив. Процедуры такого типа составляют базовые семантические элементы в АСУ, функционирующей в условиях быстро изменяющейся информационной среды и требующей для обработки больших объемов данных.

До настоящего времени эти процедуры конструировались на базе двух принципов:

а) принятия решения о сходстве объекта из массива с анализируемым объектом как функции только этих двух объектов;

б) принятия указанного решения в виде пороговых логических функций, в которых свободные параметры (пороги) задаются извне конструктором.

Оба эти принципа резко ограничивают возможности использования конструируемых процедур в информационных АСУ интенсивного функционирования.

В таких АСУ обрабатываемый информационный массив все время меняется, так что невозможно зафиксировать неизменную систему порогов, которая была бы пригодна на достаточно больших промежутках времени эксплуатации таких АСУ.

Ясно также, что на практике пользователь часто приспосабливает свои требования к оценке сходства в зависимости от того, какой информацией он располагает. Если информационный массив

действительно содержит объекты, близкие к запрашиваемому, то требования пользователя высокие, и он получает из массива "настоящие" аналоги, которыми с высоким уровнем уверенности он может пользоваться для принятия решения. Наоборот, если обрабатываемый массив косвенно связан с запросом, пользователь вынужден снижать требования к оценке сходства. В этом случае он довольствуется только грубо приближенными ориентирами для принятия искомого решения.

При использовании известных процедур выделения аналогов пользователь вынужден прибегать к многократной обработке массива, варьируя значения порогов в логических пороговых условиях процедур. Все это делает его работу замедленной и ненадежной.

Таким образом, возникла проблема разработки такой процедуры выделения аналогов, которая была бы свободна от настройки пороговых условий (т.е. чтобы пороги в ней выставлялись автоматически) и которая сама могла бы приспосабливать уровень требований к оценке сходства для заданной пары запрос-массив на основе их взаимного сопоставления.

Очевидно, что наиболее прямо было подойти к решению этой проблемы с позиций вариационного подхода, когда выбирается подходящий критерий (функционал), аргументом которого могут быть все возможные разбиения анализируемого массива на две непересекающиеся части. Содержательный смысл такого функционала - оценка качества выделения из массива множества аналогов на данный запрос.

В настоящей диссертационной работе реализован именно этот подход.

Сложность конструирования подходящего критерия заключалась, во-первых, в том, что он должен был быть явно зависимым (и

должным образом) от запроса, и, во-вторых, в том, что он должен был допускать построения эффективного алгоритма поиска искомого экстремального разбиения.

Для разрешения первой сложности в диссертации используется идея векторной оценки сходства на парах векторов - объектов. На основе этой идеи строится алгоритм преобразования обрабатываемого массива в множество векторов оценок сходства его элементов (объектов) с данным запросом. Построенное таким образом пространство обладает важным свойством: каждая координата этого пространства - это оценка сходства любого объекта с данным объектом-запросом. Другими словами, чем большие значения имеют координаты наблюдаемого вектора в этом пространстве, тем ближе он к заданному вектору-запросу (который в этом пространстве, очевидно, имеет максимальные значения всех координат).

Для преодоления второй из указанных сложностей привлечена теория монотонных систем. Два достоинства этой теории определили наш выбор:

- она дает алгоритм точного решения трудной экстремальной задачи разбиения большого массива на две части;

- она дает простой с вычислительной точки зрения алгоритм.

Были приняты во внимание, конечно, и такие особенности теории монотонных систем:

- она дает разбиение обрабатываемого массива на особые две части: одна часть это наиболее близкие в некотором точном смысле объекты, а вторая - остальные;

- она позволяет использовать широкий класс функций, оценивающих сходство между объектами, и тем самым может быть легко привязана к специфике конкретного типа информации.

Описываемая далее конструкция названа моделью ассоциативного образа запроса. Можно сказать, что эта модель реализует адаптивный подход к разработке процедуры выявления множества аналогов в АСУ информационного типа. В этом заключается актуальность и новизна настоящей диссертации.

Для того чтобы сделать модель ассоциативного образа прикладным инструментом исследования пользователя, необходимо было:

I) конкретизировать способ преобразования информационного массива в пространство векторных оценок его элементов с запросом;

2) сконструировать специальные монотонные системы, которые, с одной стороны, были бы гибким средством анализа, а с другой, - давали бы наиболее экономные алгоритмы обработки;

3) построить структурное описание модели ассоциативного образа, которое позволило бы легко его интерпретировать, а если необходимо, то и модифицировать применительно к текущим нуждам принятия решения.

Испытание работоспособности разработанной технологии обработки запроса в АСУ информационного типа проводилось на трех примерах практических задач:

I) модернизации словаря ключевых слов международного журнала ИФАК "Автоматика";

2) оценки распределения потребности в укреплении материально-технической базы школы в школьной сети большого региона;

3) формирования правила индивидуального направления к врачам-специалистам по данным доврачебного анкетного опроса для АСУ поликлиники.

Практический эффект от использования результатов диссертации документально подтвержден прилагаемыми официальными справками и актами.

Статьи автора, отражающие основное содержание диссертационной работы, указаны в библиографии

ГЛАВА I

ПРОБЛЕМА СТРУКТУРНОГО ОПИСАНИЯ ИНФОРМАЦИОННЫХ МАССИВОВ В АСУ

§I.I. Формирование множества аналогов - основной элемент управления информационными системами и процессами

При автоматической обработке информации задача выделения в информационном массиве множества аналогов описания данного конкретного объекта является одним из основных элементов эксплуатации информационных систем и управления информационными процессами. Это верно для самых разных автоматических информационных систем. Однако наиболее полно указанная роль задачи выделения аналогов выявляется в таких автоматических информационно-поисковых системах, где обычно в форме описания задаваемого объекта фиксируется вся семантика запроса к этой системе. Она важна также для разработок диалоговых систем [1-6] и новых систем поддержки интеллектуальных исследований - экспертных систем [7-10].

В связи с успехами компьютеризации автоматические информационные системы стали или становятся необходимым элементом автоматических систем управления различного назначения: традиционных АСУ промышленных предприятий, менее традиционных АСУ предприятий непромышленной сферы (поликлиники, больницы, юридической консультации, магазина, учебного заведения, библиотеки, проектной или исследовательской организации, административной службы и др.), АС комплексного управления социально-эко-

номическим развитием региона (города, района), АСУ региональных или национальных транспортных и энергетических служб и еще многих, многих других.

В каждой такой системе функционирует несколько изменяющихся банков данных, которые используются как в автономном, так и в интегрированном режимах. И все-таки несмотря на действительно огромное разнообразие этих информационных систем, один из аспектов, который их объединяет, состоит в том, что основу их функционирования составляет операция подбора аналогов к заданному объекту из обрабатываемого информационного массива.

В системах такого рода, функционирующих в АСУ на искомую операцию выделения аналогов накладываются дополнительные требования - увеличение скорости выдачи окончательного решения (для этого необходимо не только убыстрение отдельных шагов обработки, но и минимизация взаимодействий с пользователем, т.е. максимально допустимая автоматизация), придания большей универсальности для сокращения числа специальных разработок [11].

Ясно, что модель "аналоговости", на которой строится такая операция, является определяющим моментом, влияющим на обе основные характеристики информационной системы - полноту и точность ответа на запрос.

Чтобы такая модель хорошо передавала смысл реализуемого сравнения, необходимо выполнение ряда условий:

а) исходные признаки, в терминах которых строятся описания анализируемых объектов, должны отражать существенные черты проблемной ситуации проводимого анализа;

б) мера близости, с помощью которой оценивается сходство сравниваемых объектов по выбранным признакам, должна соответствовать содержательным представлениям о сходстве объектов в рассматриваемой проблемной ситуации;

в) используемые признаки и меры сходства должны допускать проведение процедуры сравнения на разном уровне детальности, т.е. на них должна быть задана упорядочивающая структура уточнения описаний объектов;

г) решающие функции, на основе которых объект базы данных принимается как аналог предъявленного объекта-запроса, должны иметь мало свободных параметров, по которым их необходимо настраивать, так как такой процесс настройки, во-первых, требует проведения большого эмпирического исследования, и, во-вторых, должен повторяться всякий раз, когда свойства эксплуатируемого банка данных или потока запросов меняются (своя трудная задача и в том, чтобы уметь диагносцировать эти моменты изменения свойств банка данных и потока запросов).

Простейшие системы выделения множества аналогов - это процедуры, которые последовательно и независимо друг от друга сравнивают запрос с каждым отдельным описанием анализируемого информационного массива по выбранному заранее подмножеству признаков. Если на этом подмножестве признаков данное сравниваемое описание совпадает с запросом, то оно объявляется аналогом. В более современных и более сложных системах вместо полного совпадения используется частичное с заданным заранее набором мест и числом (порогом) возможных несовпадений в пределах выделенного (определяющего) набора признаков.

Часто вместо числа совпадений на заданном подмножестве

признаков используются специальные более сложные меры сходства, учитывающие, например, частоту встречаемости данного признака в обрабатываемом банке данных.

Рассмотрим пример. В [12] описаны процедуры обработки запроса при использовании баз стандартных рентгендифракционных спектров (типа БДJCPDSи БЭРД).

Пусть x - изменяемый в процессе эксперимента параметр съемки спектра (это может быть длина волны, энергия, угол отбора излучения и др.), а y - интенсивность излучения, зарегистрированная в точке x. Тогда

$$y_j = (y_{j1}, y_{j2}, \ldots, y_{jN}) \qquad (1)$$

- это N - мерный вектор, представляющий j-й спектр, а база данных содержит M такого типа векторов $(N, M \sim 10^4)$.

Наиболее важная практическая задача анализа этой базы следующая. Обозначим $g(x)$ - спектр анализируемой смеси (руды, сплава или другой), состоящей из нескольких соединений-компонентов, стандартные спектры которых представлены в БД в виде совокупности векторов $\{y_{j1}, y_{j2}, \ldots, y_{jt}\}$. Вектор $g(x)$ всегда можно представить в виде линейной комбинации [13-15]:

$$g(x) = \sum_{i=1}^{t} c_i y_{j_i}(x), \qquad (2)$$

где c_i - концентрация i-й компоненты в смеси. Задача состоит в том, чтобы в ЕД выделить систему $\{y_{j_1}, y_{j_2}, \ldots, y_{j_t}\}$, с помощью которой можно было бы хорошо аппроксимировать $g(x)$ в соответствии с (2).

Поскольку исходные данные носят статистический характер,

постольку имеются неконтролируемые расхождения между истинными и измеренными положениями и интенсивностями соответствующих линий стандартного и исследуемого спектров. Следствием этого получаем, что задача аппроксимации $g(x)$ в виде (2) формально становится некорректной.

Поэтому на практике она решается не в общем виде и не чисто на формальной основе, а с помощью выделения специальных случаев и с максимальным использованием эвристического учета априорной информации. В частности, вместо сравнения с полной БД обычно используют ее небольшую часть, содержащую вместо 10^4 спектров всего $10^2 \div 10^3$ спектров (выборка минералов, соединений легких металлов, силикатов, фосфатов, окислов, машиностроительных материалов и т.п.). Далее, обычно, вместо самих векторов $y_j(x)$ и $g(x)$ используют их уплотненные представления - совокупности пар $\{x_i, y_i\}$, соответствующие только положениям максимумов на спектрах (для исследуемого образца g имеем $\{x_k, g_k\}$).

Если в пределах заданного окна Δx имеем для i -й линии спектра, что

$$x_i^g - \Delta x \leq x_i^y \leq x_i^g + \Delta x \,, \qquad (3)$$

то принимается соглашение: в стандартном спектре y i-я линия - это линия, которая имеется в исследуемом спектре g . После этого для спектра y подсчитывается отношение: число линий, которые имеются у g , к его общему числу линий (максимумов).

Обозначим это отношение через K . Если $K \geq K_0$, где

K_0 - заранее выбранный порог, то сравниваемый спектр оставляется для дальнейшего анализа. В противном случае он исключается из анализа.

Отобранная таким образом подвыборка (исходная выборка из БД, как говорилось, строится из содержательных соображений - это, например, только силикаты) на заключительном этапе селектируется с помощью более тонких критериев сходства двух спектров:

$$F = K \cdot \left(1 - \frac{\sum_{i=1}^{n} |x_i^g - x_i^y|}{\Delta x \cdot n}\right) \cdot \left(1 - \frac{\sum_{i=1}^{n} |g_i - y_i|}{\sum_{i=1}^{n} g_i}\right), \qquad (4)$$

где Δx и K - это уже определенные величины, а n - число линий, на которых получено выполнение соглашения о совпадении линий.

Окончательная подвыборка, каждый спектр которой прошел условие (4), рассматривается как искомое множество аналогов и используется для минимизации [13]:

$$| g(x) - \sum c_i y_i | . \qquad (5)$$

Обратим внимание на то, что в рассмотренном случае множество отобранных спектров хотя по существу процедуры отбираются как аналоги, выступают в (5) как дополняющие друг друга объекты, которые только совместно (системно) воспроизводят исследуемый спектр $g(x)$.

Другими словами, по самому смыслу требование разделить массив на ассоциированные и неассоциированные с данным запросом в данном случае следует интерпретировать как требование

построения некоторой целостной процедуры над массивом.

Другой типичный пример такого рода дают взаимоотношения конструктора и технолога [16]. Конструктор формулирует пакет требований к материалу, необходимому для создания конструкции. У технолога имеется банк технологий, обеспечивающих производство нескольких тысяч (иногда десятков тысяч) типовых материалов. Необходимо выделить из банка небольшую часть технологий, которые бы конструктор с технологом совместно, на содержательном, а не формальном уровне, могли бы сопоставить с заданным пакетом требований:

а) конструктор - чтобы найти наиболее приемлемые изменения этого пакета требований;

б) технолог - чтобы найти наиболее надежные модификации технологий, приближающие выбранные типовые материалы к заданному пакету требований.

Эта проблема, очевидно, возникает из-за того, что несмотря на большой объем банка типовых материалов, его все же недостаточно для того, чтобы содержать материалы, необходимые для новых конструкций.

При использовании традиционной модели аналоговости выделение искомой части информационного массива из банка типовых материалов - это длительный и кропотливый итерационный процесс, на последовательных этапах которого все более и более смягчаются требования к пакету, предъявляемому технологом, и одновременно расширяются принимаемые конструктором модификации к утвержденным технологиям выбираемых материалов-кандидатов.

Этот процесс взаимных уступок, оказываемых друг другу конструктором и технологом, обрывается, когда находится под-

ходящий вариант как окончательное решение, а точнее, - несколько таких вариантов, так как каждый такой подходящий вариант - это ожидаемый вариант, для которого имеется лишь спрогнозированная модификация твердо установленной технологии производства.

Организация такого процесса многократного отбора из банка разных вариантов моделей аналоговости требует тонкого подбора свободных параметров системы отбора, в результате чего этот диалог между технологом и конструктором затягивается на долгое время. Кроме того, между ними возникают взаимные претензии. Каждый считает, что именно он пошел на чрезмерный компромисс.

Чтобы сократить время этого диалога и сделать взаимодействие между конструктором и технологом объективно-стандартизованным, необходимо иметь такую модель аналоговости, которая бы гарантировала выделение из банка оптимальной в некотором заданном смысле совокупности аналогов, т.е. такой их набор, что любая его модификация может только ухудшить выбранный критерий.

Наконец, имеются традиционные области, где на основе выбора аналогов строится процесс решения. Это медицинская диагностика [17,18] и краткосрочный прогноз погоды [19,20]. И в том и в другом случае изучаемое состояние (больного или атмосферы) сопоставляется с БД ранее собранных состояний, эволюция которых хорошо изучена.

Если удается получить множество аналогов изучаемого состояния так, чтобы оно было малой мощности (из одного-трех пред-

ставителей БД), то диагноз или прогноз соответственно строится на базе знания известной эволюции отобранных аналогов. Если же это множество оказывается большей мощности, то оно используется как основа для качественного обсуждения экспертов.

Важно, что в обоих указанных случаях точно так же, как и в случае спектров, отбор множества аналогов строится на сравнении задаваемого объекта-запроса с каждым объектом БД или ее части, которая фиксируется из содержательных представлений о типе запроса.

Важно также обратить внимание, что и в этих областях с неформализованными процедурами принятия решения за исключением простейших случаев, когда множество аналогов получается равным I÷3, элементы этого множества выступают не как независимые аналоги изучаемого состояния (хотя именно так они отбираются), а как его характеристики, описывающие разные (дополняющие) стороны этого состояния (т.е. системно).

Точно так же используется задача формирования множества аналогов в криминологии [21] , в конструировании машин [22-26] и во многих других областях. Даже в деле библиотечно-информационного обеспечения научных разработок ситуация весьма близкая [27] . Запрос выделяет сферу прагматического интереса исследователя в БД, которая в действующих системах, во-первых, составляется из элементов БД, сравниваемых с запросом независимо (порознь), и, во-вторых, используется далее запрашивающим, конечно, не изолированно, во взаимном сопоставлении (системно).

Сказанное можно резюмировать следующим образом:

а) операция выделения множества аналогов из БД по запросу является одной из основных в эксплуатации БД;

б) используемая модель всегда жестко фиксирована, предполагает неизменным содержание БД и потока запросов (эта жесткость находит выражение, например, в фиксации порогов, участвующих в селекции подвыборок на разных этапах отбора);

в) сравнение запроса с элементами БД проводится отдельно для каждого элемента, хотя в последующем отобранное множество аналогов используется целостно;

г) хотя в каждой конкретной области используется своя специфическая модель аналоговости, подход к построению этих моделей чрезмерно узкий: он сводится к выбору того или иного коэффициента сходства описаний пары объектов.

Указанные три ограничения используемых процедур выделения множества аналогов, которые снижают эффективность эксплуатации СУБД и сужают сферу их применения, обусловлены, как нам представляется, тем, что разработчики СУБД ведут свои исследования в отрыве от другой области обработки больших массивов информации на ЭВМ, которую по предложению Тьюки называют анализом данных , а в СССР - прикладной статистикой.

Именно в этой последней области ведется интенсивная работа по созданию и изучению разных коэффициентов сходства между формализованными описаниями объектов [28]. Именно в ней актуализирован вопрос о создании решающих правил с малым числом свободных параметров (и, в частности, поэтому такие правила стремятся синтезировать без жестко задаваемых порогов [21]). Именно здесь активно изучаются такие подходы к измерению сходства, когда соответствующее значение оказывается функцией не

только пары сравниваемых объектов, но и "окружения (контекста)" функционирования этих объектов [29].

Данное наблюдение послужило отправной точкой выбора направления исследования, которое составило содержание настоящей диссертации.

Главная цель работы была сформулирована как построение модели аналоговости для прикладных СУБД АСУ, свободной от указанных трех ограничений, на базе накопленного опыта в прикладной статистике.

Сразу следует сделать важную оговорку: в прикладной статистике изучение информационного массива ведется с некоторой неизменной позиции (со стремлением описать существенные черты его как такового) в отличие от прикладной СУБД, где каждый запрос создает "свой взгляд" на обрабатываемый массив, и поэтому требуется иметь процедуру-функцию, которая вычленяет существенные черты обрабатываемого массива с точки зрения предъявленного запроса. Поэтому построение указанной модели не может основываться на формальном перенесении моделей прикладной статистики в сферу конструирования процедуры обработки из СУБД. Необходимо модифицировать эти последние модели таким образом, чтобы вместо автономного режима, в котором они работают, иметь управляемый режим, реагирующий на внешнее воздействие (запрос).

Из разных исследований прикладной статистики для наших целей наиболее существенным является тот, который объединяет методы под названием "кластерный анализ". Именно здесь мы можем найти исследования, в которых разрабатываются интересующие нас вопросы: разнообразие коэффициентов сходства, минимизация числа пороговых условий с заранее задаваемыми порогами

в решающих правилах, процедуры сравнения объектов, зависящего от контекста [30,31].

При характеризации этих методов мы разделим их на два типа: методы, основанные на поиске в данных структуры разбиения, и методы, основанные на поиске специальной структуры, когда данные разделяются на две части - однородных (или взаимозаменяемых) и неоднородных объектов.

Главное различие между этими двумя группами методов состоит в том, что по первым методам исходный информационный массив разделяется на части так, что каждая из них содержит однородные объекты. Напротив, по методам второй группы выделяется единственная группа наиболее однородных объектов, так что оставшиеся объекты образуют разрозненное множество изолированных элементов (точнее, как будет видно из дальнейшего, эти методы дают сразу несколько вложенных друг в друга разделений обрабатываемого множества на однородные и неоднородные, отличающиеся степенью однородности однородных частей).

§I.2. Методы построения разбиений информационного массива на подмассивы, состоящие из однородных частей

Если информационный массив, из которого требуется выделить множество аналогов под данный запрос, заранее разбито на несколько однородных частей, то процедуру формирования ответа можно строить как процедуру проверки, какая из этих однородных частей более всего удовлетворяет условиям аналоговости.

Если, например, каждой такой части из однородных объектов поставить в соответствие обобщенный объект, который мог бы служить ее эталоном, то процесс выделения аналогов можно было бы максимально упростить.

Именно, можно было бы в соответствии с выбранным коэффициентом сходства линейно упорядочить эти эталоны по их сходству с запросом, и в качестве искомого множества взять ту часть однородных объектов, эталон которой в наибольшей степени похож на объект-запрос.

Такая организация процедуры, во-первых, не требует заранее какого-либо внешнего порога (выбирается просто лучший эталон), и, во-вторых, она автоматически оказывается контекстнозависимой, так как выделяемая часть является однородной не сама по себе, а в сравнении со всеми объектами обрабатываемого массива.

Как же строятся методы кластеризации, которые обеспечивают поиск разбиения обрабатываемого массива на однородные части? Ответить на этот вопрос детально не представляется возмож-

ным, так как это очень интенсивно разрабатываемый раздел прикладной статистики. Это хорошо показывают регулярные обзоры [32-35], и включенная в них библиография. Поэтому для их характеристики мы рассмотрим лишь несколько достаточно сильно различающихся вариантов.

Поскольку рассматриваемые варианты излагаются, чтобы оценить их эффективность как возможной основы для разработки искомой адаптивной процедуры выделения из БД множества аналогов по заданному объекту, далее большое внимание уделено рассмотрению алгоритмов, с помощью которых решаются кластерные задачи.

Рассматриваемые ниже варианты охватывают только так называемое вариационное направление в кластеризации. В его рамках строится некоторый функционал, оценивающий качество кластеризации из некоторого допустимого класса, и поэтому задача сводится к экстремизации этого функционала.

Итак, W - исходное множество объектов $(|W|=N)$; матрица $A=\|a_{ij}\|$ задает для каждой пары объектов из W значение выбранного коэффициента сходства, определенного на всех парах; $R=\{R_1, R_2, \ldots, R_m\}$ - разбиение на m кластеров, рассматриваемое на W .

Одним из самых широко используемых является следующий функционал [36] :

$$J_1(R)=\sum_{s=1}^{m}\frac{N_s}{N}\left(\frac{1}{N_s^2}\sum_{i,j\in R_s}a_{ij}\right)=\frac{1}{N}\sum_{s=1}^{m}\frac{1}{N_s}\sum_{i,j\in R_s}a_{ij}\ , \qquad (6)$$

где N_s – число объектов в кластере R_s .

Максимизация $J_1(R)$ обычно рассматривается на множестве $\mathcal{R}^m$ всех возможных разбиений W на m непустых кластеров. Она приводит к плотным (с большой средней величиной внутренних связей между объектами) кластерам большого размера за счет малой плотности кластеров малого размера.

Действительно, пусть имеется два ℓ – элементных множества чисел X и Y $(|X|=|Y|=\ell)$. Рассмотрим множества подстановок на X и Y :

$$I=\{I_X : I_X=\{i_1, i_2, \ldots, i_\ell\}\},$$

$$J=\{J_Y : J_Y=\{j_1, j_2, \ldots, j_\ell\}\}.$$

Определим на множестве $I\times J$ функцию

$$f(I_X, J_Y)=\sum_{s=1}^{\ell} x_{i_s}\cdot y_{j_s},$$

где $x_{i_s}\in X$, $y_{j_s}\in Y$, $s=\overline{1,\ell}$.

Максимум этой функции достигается на подстановках $I_X^*=\{i_1^*, i_2^*, \ldots, i_\ell^*\}$ и $J_Y^*=\{j_1^*, j_2^*, \ldots, j_\ell^*\}$, таких, что

$$x_{i_1}^*\geq x_{i_2}^*\geq \ldots \geq x_{i_\ell}^*,$$
$$y_{j_1}^*\geq y_{j_2}^*\geq \ldots \geq y_{j_\ell}^*.$$

Подстановка $(I_X^*, J_Y^*)\to(I_X, J_Y)$, одновременно переставляющая номера i_k и j_k , $k=\overline{1,\ell}$, не уменьшает значение $f(I_X, J_Y)$. Поэтому, полагая, что множество X – множество размеров кластера, а множество Y – множество средних величин коэффициентов сходства между объектами внутри кластеров, можно объяснить факт большой плотности кластеров большого размера за счет малой плотности кластеров малого размера при оптимизации $J_1(R)$.

Заметим, что матрица A для $J_1(R)$ без ограничения общности может считаться симметричной, так как преобразование

$$\hat{a}_{ij} = \frac{1}{2}(a_{ij} + a_{ji})$$

не меняет, очевидно, функционала $J_1(R)$.

В [37] исследуется другой функционал, не инвариантный относительно указанного преобразования:

$$J_2(R) = \sum_{s=1}^{m} \max_{i \in R_s} \sum_{j \in R_s} a_{ij} \quad . \tag{7}$$

В [37] задача кластеризации формулируется как аппроксимационная. Для этого вводится матрица $\tau = \| \tau_{ij} \|$ размера $N \times N$:

$$\tau_{ij} = \begin{cases} \text{I}, \text{ если } & i, j \in R_s \ , \\ 0, \text{ если } & i \in R_s \ , \ j \in R_t \ , \ s \neq t \ . \end{cases}$$

Предложенный в [ ] функционал аппроксимации имеет в этих обозначениях следующий вид:

$$J_3(\lambda, R) = \sum_{i,j=1}^{N} (a_{ij} - \lambda \tau_{ij})^2 .$$

Экстремизация J_3 по λ при фиксированном R дает возможность явно выразить экстремальное λ^* , так что подставляя последнее в J_3 , получаем

$$J_3(\lambda, R) = \frac{\sum_{i,j=1}^{N} a_{ij} \cdot \tau_{ij}}{\sqrt{\sum_{i,j=1}^{N} \tau_{ij}}} \quad . \tag{8}$$

Последняя формула показывает, что $J_3(\lambda^*, R)$ аналогичен $J_1(R)$.

Рассмотрим теперь общую схему алгоритма поиска локального экстремума функционалов типа приведенных примеров, зависящих от разбиения в случае, когда число кластеров заранее фиксировано. Она сводится к следующей обработке.

I. Тем или иным способом выбирается некоторое начальное разбиение.

2. Некоторый (очередной) объект переносится из кластера, в котором он находился к данному шагу, последовательно во все кластеры, начиная с первого. При каждом переносе подсчитывается новое значение функционала $J(R)$ и сравнивается со значением этого функционала до переноса. Если при очередном пробном переносе данного объекта значение функционала возросло, то рассматриваемый объект действительно переносится в новый кластер. Если после пробных переносов во все другие кластеры значение функционала $J(R)$ ни разу не возросло,то рассматриваемый объект остается в том же кластере, в котором он находился до осуществления данного шага. Затем алгоритм переходит к новому шагу, на котором осуществляются пробные переносы следующего объекта.

Алгоритм заканчивает свою работу после того, как просмотр всех объектов не приводит к изменению ни одного кластера.

Рассмотрим, как конкретизируется описанная общая схема алгоритма, если экстремизировать $J_1(R)$. Основной шаг в соответствии с общей схемой состоит в пробном переносе некоторого элемента i^* из кластера R_{s*} в отличный от него кластер R_{t*}. При этом разбиение R переходит в разбиение R'.

Можно показать [36], что разность $J_1(R')-J_1(R)$ значений функционала на этих разбиениях вычисляется по формуле

$$\Delta(i^*, t^*, R^*)=\left[\frac{f_{s^*}(R_{s^*})}{N_{s^*}-1}-\frac{f_{t^*}(R_{t^*})}{N_{t^*}-1}\right]+ \tag{9}$$

$$+\left[\frac{\sum_{j\in R_{t^*}}(a_{i^*j}+a_{ji^*})}{N_{t^*}-1}-\frac{\sum_{j\in R_{s^*}}(a_{i^*j}+a_{ji^*})}{N_{s^*}}\right],$$

где $$f_\alpha(R_\alpha)=\frac{\sum_{i,j\in R_\alpha} a_{ij}}{N_\alpha}.$$

Формула (9) показывает, что приращение функционала значительно экономнее расчета значения самого функционала, так как используется информация лишь об одной строке i^* и об одном столбце i^* матрицы A, а также о двух частных функционалах $f_{s^*}(R_{s^*})$ и $f_{t^*}(R_{t^*})$ и о величинах кластеров N_{s^*} и N_{t^*}.

Приведенная характеризация методов кластеризации, основанных на поиске разбиений множества на однородные кластеры, показывает, что:

I) модель взаимной аналоговости между объектами выделяемых кластеров имеет составную зависимость от базового коэффициента сходства пары объектов, ст меры однородности кластера, составленной как функция от названных коэффициентов, и от функционала, определяемого через эти меры (с одной стороны, такая конструкция делает гибким механизм выбора подходящей модели, а с другой, - требует определенного навыка в ее использовании, и,что более важно для рассматриваемых целей, знания качественной информации о структуре анализируемого массива);

2) реализация метода весьма сложна; даже в случае использования сугубо приближенных алгоритмов требуемый объем вычис-

лений и память позволяют проводить приемлемые расчеты лишь для сотен объектов; тысячи, и особенно, десятки тысяч объектов требуют разработки специальных алгоритмов и резко сужают число доступных для выбора моделей;

3) не только в практических разработках, но и в теоретических исследованиях рассмотренного направления кластерного анализа не было предложено метода, гарантирующего получения глобального оптимума (хотя бы одного разбиения, доставляющего глобальный оптимум) экстремизируемого функционала.

§I.3. Кластерные методы выделения наиболее однородной части информационного массива

Качественное представление об однородности (взаимозаменяемости) предполагает, что все элементы соответствующего множества объектов похожи между собой, и значит, если описание эталонного объекта этого множества можно принять в качестве аналога заданного запроса, то и все порождающее эталон множество можно принять как искомое множество аналогов. Ситуация вполне подобная рассмотренной в предыдущем параграфе, но с той существенной разницей, что в случае отказа принять в качестве аналога эталон получаем сигнал, что в данном информационном массиве нет эталонов для данного запроса.

Такое более жесткое условие функционирования процедуры отбора эталонов компенсируется гарантией, что выделяемое подмножество наиболее однородно в некотором заданном смысле по сравнению со всеми другими подмножествами. По сравнению с описанными в §I.2 настоящей главы методами кластеризации, где необходимо было "делить" однородные объекты по разным кластерам, в методах, которые будут рассмотрены ниже, наиболее однородные объекты собираются в один единственный кластер.

Следует отметить, что такого типа кластерные процедуры имеются в литературе в гораздо меньшем числе, чем те, что рассматривались во втором параграфе настоящей главы.

Рассмотрим два таких метода: аппроксимационный и монотонных систем.

Пусть $A = \|a_{ij}\|$ - матрица значений некоторого коэффициента

сходства между парами объектов ($i,j = \overline{1,N}$, где N - число объектов рассматриваемого информационного массива). Пусть множество H - это подмножество из множества W всех объектов информационного массива. Будем оценивать однородность H среди всех элементов 2^W с помощью функционала [34]:

$$I(H) = \sum_{i,j \in H} (a_{ij} - \bar{a}_H)^2 + \sum_{(i,j) \notin H} (a_{ij} - \bar{a}_{\tilde{H}})^2 , \qquad (10)$$

где $(i,j) \notin H$ означает логическое объединение условий $(i \in H, j \in H)$, $(i \notin H, j \in H)$ и $(i \notin H, j \notin H)$, а средние $\bar{a}_H$ и $\bar{a}_{\tilde{H}}$ выражаются с помощью формул

$$\bar{a}_H = \frac{1}{|H|^2} \sum_{i,j \in H} a_{ij} ,$$

$$\bar{a}_{\tilde{H}} = \frac{1}{|H| \cdot |W \setminus H|} \left(\sum_{\substack{i \in H \\ j \in W \setminus H}} a_{ij} + \sum_{\substack{j \in H \\ i \in W \setminus H}} a_{ij} \right) + \frac{1}{|W \setminus H|^2} \sum_{\substack{i \in W \setminus H \\ j \in W \setminus H}} a_{ij} .$$

Ясно, что минимизация такого функционала на 2^W порождает такое множество H^*, что для большинства $(i,j) \in H^*$, $a_{ij} \approx \bar{a}_{H^*}$. Именно в этом смысле следует понимать представление о том, что его минимизация дает наиболее однородное подмножество в W. Другие функционалы такого типа, предлагаемые для той же цели, рассмотрены в [38].

Все эти постановки задач имеют один общий недостаток: для нахождения их точного решения требуется алгоритм экспоненциальной сложности. Поэтому при практическом использова-

нии функционала (10) для нахождения подмножества однородных объектов ограничиваются применением сугубо приближенных алгоритмов, связанных с поиском локальных экстремумов $I(H)$ (т.е. таких, которые невозможно изменить варьированием размещения только одного объекта без его увеличения).

В [39], например, рассказывается как с помощью приближенного алгоритма минимизации функционала, мало отличающегося от (10), решалась задача проектирования большого вычислительного комплекса для автоматических расчетов синтеза сложных химических реакций. Из этой работы можно заключить, что приближенный характер решения приводил к необходимости проведения очень большого числа разных вариантов расчета, отличающихся выбором начального разбиения, с которого стартовал алгоритм. Не менее важно, что большое число из этих вариантов оказалось необходимым сравнивать на содержательном уровне, что не только отнимало много времени высококвалифицированных специалистов, но и требовало от них предварительной выработки некоторого специального навыка. Такие условия работы, очевидно, мало пригодны, если речь идет о модификации метода к задаче обработки запроса с целью получить для предъявленного образца множества аналогов.

Рассмотрим теперь метод монотонных систем [40-46]. Пусть, как и раньше, W, $|W| = N$ - исходное множество описаний объектов из обрабатываемой БД или выборки, извлеченной из нее (информационный массив). Пусть задана скалярная функция $\mathcal{T}$, которая каждой паре (i, H), где $H \subseteq W$ - произвольное подмножество W, $i \in H$, ставит в соответствие число $\mathcal{T}(i, H)$. В частности, если на элементах $i \in W$ определена матрица $A = \|a_{ij}\|$ коэффициентов попарного сходства, то в качестве $\mathcal{T}(i, H)$ можно

принять любую из функций:

$$\pi_1(i,H) = \sum_{j\in H} a_{ij}, \tag{11}$$

$$\pi_2(i,H) = \max_{j\in H} a_{ij}, \tag{12}$$

$$\pi_3(i,H) = \min_{j\in(W\setminus H)} a_{ij} \tag{13}$$

Число $\pi(i,H)$ называется весом элемента i на множестве H. От такой функции требуется только, чтобы она удовлетворяла условию монотонности

$$\pi(i,H) \geq \pi(i,H') \tag{14}$$

для всех $i \in H' \subseteq H \subseteq W$.

Далее с помощью $\pi(i,H)$ строится скалярная функция $F(H)$ на всех элементах булеана 2^W:

$$F(H) = \min_{i\in H} \pi(i,H). \tag{15}$$

Определение I.I. Ядрами (экстремальными подсистемами) монотонной системы $\langle W, \pi, F\rangle$ называются такие подмножества W, на которых достигается максимум $F(H)$.

Упорядочим элементы множества W произвольным образом. С каждой последовательностью $A = \langle \alpha_1, \alpha_2, \ldots, \alpha_N \rangle$, где $W = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, взаимнооднозначно свяжем последовательность вложенных подмножеств множества W: $H_1 = W$,

$H_2=W\setminus\alpha_1,\ldots,H_{k+1}=H_k\setminus\alpha_k,\ldots,H_N=\alpha_N$. Обозначим ее через $\bar{H}=$ $=\langle H_1,H_2,\ldots,H_N\rangle$.

Определение I.2. Упорядоченная последовательность A элементов W называется определяющей, если в соответствующей ей последовательности $\bar{H}$ подмножеств существует такая подпоследовательность $\bar{\Gamma}=\langle\Gamma_1,\Gamma_2,\ldots,\Gamma_p\rangle$, где $\Gamma_1=W$, что выполняются условия

$$\pi(\alpha_k,H_k)<F(\Gamma_{j+1}),\quad \forall\alpha_k\in\Gamma_j\setminus\Gamma_{j+1},\quad j=\overline{1,p} \qquad (16)$$

и

$$F(L)\leq F(\Gamma_p),\quad \forall L\subset\Gamma_p. \qquad (17)$$

Определение I.3. Множество G, $G\subseteq W$ называется определимым, если существует определяющая последовательность такая, что $\Gamma_p=G$.

В теории монотонных систем построение прикладных алгоритмов основывается на двух центральных теоремах [47].

Теорема I.1. На определимом множестве G функция $F(H)$ достигает глобального максимума. Существует единственное определимое множество. Все множества, на которых достигается глобальный максимум F, лежат внутри определимого множества.

Теорема I.2. Семейство подмножеств из 2^W, на которых $F(H)$ достигает глобального максимума, замкнуто по отношению к операции объединения.

Из этих теорем получаем, что объединение всех ядер (самое большое ядро) - это как раз и есть определимое множество.

Как показано в [42-45], справедливы следующие соотношения, вскрывающие структурный смысл наибольшего ядра:

1) $F(H) < F(G), \quad \forall H \subseteq W, \ |H| > |G|,$ (18)

2) $F(H) \leq F(G), \quad \forall H \subseteq W, \ |H| \leq |G|,$ (19)

3) $F(H) < F(G), \quad \forall H \subseteq W, \ H \setminus G \neq \phi,$ (20)

4) $F(H) \leq F(G), \quad \forall H \subseteq W, \ H \subseteq G,$ (21)

5) $F(H) < F(G), \quad \forall H \subseteq W, \ H \supset G,$ (22)

причем пара 1) и 2) эквивалентна 3) и 4), которая, в свою очередь, эквивалентна паре 4) и 5).

Интересно, что и для множеств Γ_j, $j = \overline{1,(p-1)}$ выполняются аналогичные соотношения:

1) $F(H) < F(\Gamma_j), \quad \forall H \subseteq W, \ H \setminus \Gamma_j \neq \phi, \ j = \overline{2,p},$ (23)

2) $F(H) \leq F(\Gamma_j), \quad \forall H \subseteq \Gamma_j, \ H \supset \Gamma_{j+1}, \ j = \overline{1,(p-1)}.$ (24)

На геометрическом языке это можно перефразировать следующим образом: каждое подмножество Γ_j $(j = \overline{1,p})$ последовательности $\overline{\Gamma}$ является максимумом функции $F(H)$ на теоретико-множественном полуинтервале *) $[\Gamma_1, \Gamma_{j+1})$. Все множества

---

*) Напомним, что $\Gamma_1 = W$.

H, $H \subseteq W$, на которых достигается значение $F(H)$, превышающее величину $F(\Gamma_j)$ для $j = \overline{1, (p-1)}$, лежат внутри множества Γ_{j+1} .

На таком языке естественно задать два определения.

Определение I.4. Окрестностью множества H называется совокупность всех множеств L таких, что $H \subseteq L \subseteq W$, т.е. полуинтервал $[W, H)$.

Определение I.5. Множество H^0 называется строгим локальным максимумом функции $F(H)$, если оно доставляет строгий максимум этой функции в своей окрестности.

Теорема I.3 [48]. Множество H^0 тогда и только тогда является строгим локальным максимумом $F(H)$, когда $H^0 = \Gamma_j$, где Γ_j - это множество из последовательности $\overline{\Gamma}$.

Далее строгие локальные максимумы функции $F(H)$ будем называть квазиядрами.

Рассмотрим сужение монотонной системы $\langle W \setminus \Gamma_j, \pi', F' \rangle$, где для $\forall H \subseteq (W \setminus \Gamma_j)$, $i \in H$:

$$\pi'(i, H) = \pi(i, H \cup \Gamma_j) . \qquad (25)$$

Можно показать [47], что ядром этой системы является множество $G' = \Gamma_{j-1} \setminus \Gamma_j$.

Пусть $A = \langle \alpha_1, \alpha_2, \ldots, \alpha_N \rangle$ - произвольная, но фиксированная определяющая последовательность.

Тогда, как показано в [42-45], верны следующие факты:

I. Для любого Γ_j, $j = \overline{2, p}$ из последовательности $\overline{\Gamma}$ справедливо неравенство

$$\max_{\alpha_n \in W \setminus \Gamma_j} \pi(\alpha_n, H_n) \le \max_{\alpha_k \in \Gamma_{j-1} \setminus \Gamma_j} \pi(\alpha_k, H_k) , \qquad (26)$$

а значит, в частности, и

$$\max_{\alpha_n \in W \setminus G} \pi(\alpha_n, H_n) \le \max_{\alpha_n \in \Gamma_{p-1} \setminus G} \pi(\alpha_n, H_n) . \qquad (27)$$

2. Для любого k, $k \in W$ такого, что $k \notin \Gamma_j$ для некоторого $j = \overline{2, p}$, т.е. для $k \in W \setminus \Gamma_j$, верно, что

$$\pi(k, \Gamma_j \cup k) < F(\Gamma_j), \ \forall k \in W \setminus \Gamma_j , \ \forall j = \overline{2, p} \qquad (28)$$

а значит, в частности, и

$$\pi(k, G \cup k) < F(G), \ \forall k \in W \setminus G . \qquad (29)$$

Зафиксируем натуральное число n $(n < N)$ и рассмотрим задачу: найти подмножество G_n множества W такое, что

$$F(G_n) = \max_H F(H), \ \forall H \subseteq W, \ |H| > n \qquad (30)$$

Теорема I.4. [48]. Для любого Γ_j, $j = \overline{1, (p-1)}$ из $\overline{\Gamma}$ и любого H, $H \subseteq W$ такого, что $|H| > |\Gamma_j|$, верно

$$F(H) < F(\Gamma_j) . \qquad (31)$$

Для любого Γ_j, $j = \overline{1, (p-1)}$ из $\overline{\Gamma}$ и $H \subseteq W$ такого, что $|\Gamma_j| \ge |H| > |\Gamma_{j+1}|$, имеем

$$F(H) \le F(\Gamma_j) . \qquad (32)$$

Пусть $T \subseteq W$. Рассмотрим задачу: найти G_T такое, что

$$F(G_T) = \max_{T \subseteq H \subseteq W} F(H). \qquad (33)$$

Теорема I.5.[48]. Для любого $T \subseteq W$ и такого Γ_s из $\overline{\Gamma}$, что $T \subseteq \Gamma_s$, но $T \not\subseteq \Gamma_{s+1}$, имеем

$$F(H) < F(\Gamma_s), \quad T \subseteq H \subseteq W, \quad H \setminus \Gamma_s \neq \emptyset, \qquad (34)$$

$$F(H) \leq F(\Gamma_s), \quad T \subseteq H \subseteq W, \quad H \subseteq \Gamma_s. \qquad (35)$$

Для того, чтобы описать общий алгоритм поиска ядра монотонной системы, удобно задать следующую вспомогательную процедуру [48].

Процедура Слой (h). Предполагается, что известно множество H , функция $\pi(i,H)$ на его элементах и значение порога h .

Процедура является итерационной, в которой каждая итерация состоит из двух шагов.

Шаг I. Определяется множество $H' \subseteq H$ такое, что

$$\pi(i,H) \leq h , \quad \forall i \in H'. \qquad (36)$$

Шаг 2. Для $\forall i \in H'$ определяется $\pi(i,H')$.

Процедура Слой (h) есть последовательность поочередного применения шагов I и 2, последняя итерация которой определяется условием, что на очередном применении первого шага множество H' оказывается пустым (т.е. что очередное множество H на этом шаге удовлетворяет условию

$$\pi(i,H) > h, \quad \forall i \in H. \qquad (37)$$

Чтобы выделить множество, удовлетворяющее условию (37), обозначим его специальным образом через E .

Алгоритм построения, определяющий последовательности [42-45,48].

Полагаем $\Gamma_1 = H = W$. Вычисляем $\pi(i,W)$ для $\forall i \in W$ и находим

$$h_1 = \min_{i \in W} \pi(i,W) \quad \text{и} \quad h_2 = \max_{i \in W} \pi(i,W). \qquad (38)$$

Выбираем некоторое $h: h_1 < h < h_2$. Для $H = W$ и выбранного h используем процедуру Слой (h). Возможен один из двух вариантов ее исхода.

I. $E = \phi$. Строится новое значение для правой границы порога h_2 , а именно: полагаем $h_2 = h$, после чего вычисляется новое (меньшее) значение порога h ; начиная с предыдущего H , с этим новым порогом вновь запускается процедура Слой (h), и этот процесс уменьшения порога продолжается до тех пор, пока не случится второй вариант.

2. $E \neq \phi$. E объявляется очередным элементом последовательности $\bar{\Gamma}$, т.е. $\Gamma_{j+1} = E$. Вычисляется $F(\Gamma_{j+1})$ (заметим, что $F(\Gamma_{j+1}) > h$) и используется процедура Слой $\left(F(\Gamma_{j+1})\right)$ на $H = \Gamma_{j+1}$.

Возможны два случая:

а) $E = \phi$; полагаем $\Gamma_{j+1} = \Gamma_p$;

б) $E \neq \phi$, переходим к выполнению шага - множество

$H := E$, $h_1 := F(H)$, $h = \frac{1}{2}(F(H) + h_2)$, и вновь процедура Слой (h) на E.

В [47] показано, что множество Γ_ρ, получаемое в результате описанного алгоритма, есть определимое множество, т.е. наибольшее ядро.

Описанные элементы теории монотонных систем показывают, что:

а) для основной задачи нахождения в информационном массиве наибольшего однородного подмассива можно строить весьма общий точный алгоритм полиномиальной сложности; притом получающееся решение отличается целым рядом интересных с прикладной точки зрения свойств (единственность, максимальность по мощности);

б) по ходу поиска наиболее однородного подмассива выявляется описание целостной структуры агрегированного представления о сходстве объектов во всем информационном массиве: структура квазиядер, экстремальная характеризация теоретико-множественных разностей соседних квазиядер.

Указанные свойства с учетом того, что теория монотонных систем дает очень широкие возможности построения разных моделей сходства *), показывают важные преимущества этого аппарата по сравнению с другими методами кластерного анализа в

*) В [49] мы показываем, как в практических случаях можно генерировать большое разнообразие монотонных систем. Однако, поскольку далее мы развиваем модели, опирающиеся всего на две конкретные системы, результаты из [50] не вошли в настоящую диссертацию.

деле его использования для построения моделей процедур выделения аналогов для запросов в КБ.

Однако, как отмечалось, чтобы приспособить методы кластеризации для указанных целей, необходимо преобразовать их таким образом, когда каждый запрос по своему перестраивает задачу выделения искомого наиболее однородного (и похожего на запросный объект) подмассива.

Как будет видно из дальнейшего, предлагаемый в диссертации подход к проблеме сделать управляемым запросом результат кластеризации, практически не зависит от конкретной специфики выбираемого метода. На этой независимости настройки обрабатываемого массива на запрос и процедуры выделения подмножества однородных описаний объектов в диссертации строится ряд процедур для характеризации выделяемого множества аналогов, которые всегда необходимы в практических работах.

§ I.4. Цели диссертации

Резюмируя характеристику состояния дел с разработкой моделей основной процедуры в СУБД, выделения в БД множества аналогов на запросный объект, можно утверждать, что

I) существующие модели не отвечают современным требованиям придания таким моделям свойства адаптивности;

2) в области кластерного анализа накоплен большой набор методов выделения подмножеств объектов, которые взаимно можно рассматривать в качестве аналогов; этот набор интенсивно пополняется;

3) вопрос о приспособлении методов кластерного анализа для целей СУБД сводится к

а) параметризации метода кластеризации (точнее, придания ему должной зависимости от поступающего запроса);

б) резкого увеличения быстродействия метода анализа однородности в информационном массиве;

4) одним из наиболее пригодных для целей приспособления к СУБД является специальный метод кластерного анализа, составляющий содержание так называемой теории монотонных систем.

В связи со сказанным, главной целью диссертации является создание адаптивной модели выделения в БД множества аналогов запросного объекта на базе использования теории монотонных систем. Такую модель далее будем называть моделью контекстно-зависимого ассоциативного образа, чтобы отличать ее от традиционных моделей аналогов.

Достижение этой цели предполагается достичь решением ряда научно-методических задач. Это следующие задачи.

I. Создание схемы преобразования обрабатываемого массива к виду, когда в описании каждого объекта акцентированы существенные свойства запроса и элеминированы несущественные его свойства (вопрос, как делить свойства на существенные и несущественные, решается с помощью заранее фиксированной процедуры).

2. Выбор конкретной монотонной системы и построение с ее помощью простейшей контекстнозависимой модели ассоциативного образа (запроса).

3. Разработка итерационной процедуры поочередного использования алгоритма выделения подмножества однородных описаний объектов (ядра монотонной системы) и алгоритма акцентирования существенных свойств запроса в анализируемых объектах. Замысел такой итерационной процедуры состоит в том, чтобы минимизировать требование на задание разных априорных знаний о степени существенности разных признаков, в терминах которых описываются объекты и запрос.

4. Формирование набора характеристик для создания гибкой возможности пользователю манипулировать выделяемым ассоциативным образом (возможностями его суждения и расширения в связи с внешними содержательными представлениями).

5. Разработка нового метода кластерного анализа на базе созданной модели ассоциативного образа. Решение этой задачи дает принципиальный путь к конструированию средств активного самоструктурирования и реструктурирования описания семантики эксплуатируемой БД с последующим построением быстрых процедур обработки запроса на базе данных о такой текущей структуре обрабатываемой БД.

6. Апробация разработанного метода построения модели ассоциативного образа на модельных данных и в решении практических задач АСУ.

ГЛАВА 2

МОДЕЛЬ АССОЦИАТИВНОГО ОБРАЗА В МЕТОДАХ АГРЕГИРОВАНИЯ

§2.I. Общая схема разработки модели

Далее предполагается, что в n - мерном пространстве имеется множество из N векторов и один специально выделенный вектор. Требуется разделить множество на два непересекающихся подмножества, одно из которых состоит из похожих на выделенный вектор векторов, а другое - из непохожих.

В данной главе разбиваемое множество векторов интерпретируется как информационный массив, в котором необходимо выделить аналоги для предъявленного специального вектора-запроса.

В первой главе отмечалось, что поток запросов может существенно видоизменяться. Существенно изменяются и информационные массивы, которые хранятся в базах данных. Притом, и это даже важнее с точки зрения рассматриваемого вопроса, изменения в потоке запросов происходят в значительной степени независимо от изменений в анализируемом информационном массиве. Следствием этого являются неконтролируемые резкие снижения качества обработки запроса, необходимость регулярно проводить диагностику соответствия потока запросов содержанию анализируемых массивов, и перестройку режимов обработки запросов.

В первой главе отмечалось, что разнообразие режимов обработки запроса определяется главным образом разнообразием используемых коэффициентов, которые служат для количественной

оценки похожести между векторами. Поэтому, если такой коэффициент выбран, то далее используется логическая процедура: сравнение коэффициента с порогом [51-60]. Когда условие сравнения выполнено, соответствующий вектор из массива принимается как искомый аналог. В противном случае, он не принимается в качестве аналога.

Для дальнейшего изложения важно выделить две лимитирующие характеристики отмеченных разных режимов обработки запросов:

I) множество аналогов - это множество тех векторов информационного массива, каждый из которых независимо от других удовлетворяет выбранному условию;

2) пороги, используемые в условиях, не подстраиваются под индивидуальный запрос, а выбираются заранее и жестко фиксируются. Когда вместо сравнения с порогом используется правило "отобрать не более k векторов, наиболее похожих на запрос", вместо порога число k становится лимитирующим параметром процедуры [61-68].

Предлагаемый в настоящей диссертации подход делает эти характеристики нелимитирующими. Разные подмножества из информационного массива оцениваются целостно. Вместо независимого сравнения запроса и вектора(независимого от данного информационного массива) используется целостное сравнение разных разбиений информационного массива: разбиение на подмножество аналогов и дополнение к этому подмножеству.

В соответствии с целями и задачами диссертации, сформулированными в первой главе, развиваемый нами подход базирует-

ся на использовании аппарата монотонных систем (точнее, на модификации этого аппарата, связанный с идеей отображения обрабатываемого массива на запрос и последующего анализа преобразованного таким образом массива методом монотонных систем). В настоящем параграфе мы излагаем схему этого подхода в общем виде, включая не только процедуру выделения искомого множества аналогов, но и сопровождающие процедуры, необходимые пользователю для активного манипулирования при окончательном построении искомого множества и для выработки аргументов (уверенности), что оно выбрано должным образом с предметной точки зрения.

Открывая рассмотрение общей схемы сразу отметим важное методологическое преимущество аппарата монотонных систем: он не требует, как это имеет место в обычных методах кластеризации, предварительного выбора базового коэффициента сходства пар объектов. Исходно необходимо задаться функцией $\mathcal{T}(i, H)$, связывающей элемент с множеством. В следующем параграфе, в котором наш подход конкретизируется с прикладной ориентацией, мы существенно воспользуемся этим преимуществом аппарата монотонных систем.

Поскольку $\mathcal{T}(i, H)$ интерпретируется как оценка похожести элемента i с элементами из H, решение задачи поиска максимума (I5) - это такое подмножество $G \subseteq W$, что наименее похожий на элементы из G элемент i^* имеет высокую оценку $\mathcal{T}(i^*, G) = F(G)$ похожести.

Для того, чтобы приспособить теорию монотонных систем к нуждам нашей задачи разбиения информационного массива векторов на подмножества аналогов данного запроса и тех, которые

не являются его аналогами, необходимо уметь так конструировать оценку $\mathcal{T}(i, H)$, чтобы она существенно зависела от предъявленного запроса, и чтобы эта зависимость имела направление, хорошо отражающее содержательный смысл решаемой задачи.

В настоящей диссертации не ставится вопрос об анализе разных возможных способов конструирования такой оценки. Разрабатывается и исследуется всего один такой способ (точнее, одно параметрическое их семейство). Он назван моделью ассоциативных образов.

Идея способа, как отмечалось уже в первой главе, состоит в том, чтобы при предъявлении любого вектора в качестве запроса, автоматически формировалась специфическая именно для этого запроса процедура преобразования исходного информационного массива, и чтобы конструируемая далее монотонная система в качестве основы имела преобразованный массив.

При этом, если говорить на интуитивном уровне строгости, то, очевидно, что используемое преобразование должно быть направлено на то, чтобы в каждом векторе исходного массива выделить свойства, наиболее характерные для запроса. Другими словами, вместо традиционной логической оценки вектора "похож-не-похож" на запрос, предлагается на предварительном этапе использовать коэффициент похожести (на запрос) как компоненту нового искомого вектора - образа, отображения исходного вектора на запрос. В таком виде эта компонента оказывается мерой ассоциативной связи обрабатываемого вектора с запросом.

Чтобы новый вектор был разносторонней характеристикой связи исходного вектора, предлагается использовать одновременно не один, а несколько качественно различных коэффициентов оценки сходства этого вектора с запросом. При этом, конечно,

достигается и разносторонняя характеризация запроса (его представления в исходном векторе).

Итак, конструирование преобразования состоит в выборе набора коэффициентов - количественных оценок сравнения. Если этот выбор сделан, и запрос предъявлен, каждый вектор исходного информационного массива может быть однозначно преобразован в новый вектор коэффициентов сходства с запросом.

Если затем на новом массиве построена монотонная система $\langle W, \pi, F \rangle$ и решена задача нахождения максимума $F(H)$ на 2^W, то получаемое в качестве решения подмножество G как раз и выбирается в качестве подмножества аналогов запроса, т.е. в качестве его ассоциативного образа.

Перейдем теперь к описанию предлагаемой модели в точных терминах. Исходный информационный массив векторов будем обозначать через X $(|X|=N)$, а его элементы (вектора) через x_i, $i=\overline{1,N}$. Специальный вектор запроса обозначим через e.

Предположим, что имеется априорная информация, позволяющая выделить в исходном n - мерном пространстве некоторое число m специальных подпространств $R=\{R_1, \ldots, R_m\}$. Каждое такое подпространство имеет свою автономную роль в анализе сходства анализируемых векторов. Это дает основание конструировать искомый вектор оценок ассоциативных связей исходного вектора с запросом в виде составного вектора, каждый подвектор которого соответствует своему выделенному подпространству.

Пусть семейство R выделенных подмножеств признаков таково, что в произвольном векторе $x \in X$ легко выделяются соотве-

тствующие части вектора*) $x \xrightarrow{R} \{z^1, z^2, \ldots, z^m\}$.

Аналогично, и в векторе запроса e можно выделить набор $\{e^1, \ldots, e^m\}$ его частей, соответствующих выделенным подмножествам $R_1, \ldots, R_m$.

Пусть $\zeta = \{\zeta_1, \ldots, \zeta_k\}$ - это выбранный набор различных коэффициентов сходства между векторами. Тогда каждая пара (z^i, e^i) $i=\overline{1,m}$ порождает свой k - мерный вектор оценок $\{\zeta_1(z^i, e^i), \ldots, \zeta_k(z^i, e^i)\}$.

В результате исходный n - мерный вектор x при заданном запросе e преобразуется в mk - мерный вектор:

$$y = \{\zeta_1(z^1, e^1), \ldots, \zeta_k(z^1, e^1), \zeta_1(z^2, e^2), \ldots, \zeta_k(z^2, e^2), \ldots, \zeta_1(z^m, e^m), \ldots, \zeta_k(z^m, e^m). \quad (39)$$

Требуемая монотонная система задается над множеством из N векторов - образов указанного отображения X в построенное mn - мерное пространство Y.

Компоненты построенных векторов имеют специфический смысл. Любая из величин $\zeta_s(z^t, e^t)$ есть частная оценка сходства векторов x и e, и поэтому чем она больше, тем больше оснований считать x элементом ассоциативного образа e.

Поэтому функция $\pi(i, H)$ должна быть монотонной не только по H, как того требует теория, но и в направлениях роста любой компоненты $\zeta_s(z^t, e^t)$, как оцениваемого вектора y_i, индекс "i" которого фигурирует в $\pi(i, H)$, так и любого другого вектора, входящего как элемент в H. Следующие типы функций

---

*) Следует помнить, что в данном случае R - любое семейство подмножеств, а не разбиение.

удовлетворяют этим требованиям по построению:

$$\mathcal{T}_1(i,H) = \sum_{t=1}^{m\times k} \left| a\cdot y_i^t - b\cdot \min_{j\in H}\{y_j^t\}\right|^q, \qquad (40)$$

$$\mathcal{T}_2(i,H) = \sum_{t=1}^{m\times k} \left| a\cdot \max_{j\in H}\{y_i^t\} - b\cdot y_i^t\right|^q, \qquad (41)$$

где a,b,q - положительные константы, y^t - t-я компонента вектора y .

Легко видеть, что при $q=1$

$$\mathcal{T}_1(i,H) = a\cdot\rho(y_i) - b\cdot\rho(Y^H), \qquad (42)$$

$$\mathcal{T}_2(i,H) = a\cdot\rho(Y_H) - b\cdot\rho(y_i), \qquad (43)$$

где использованы обозначения

$$\rho(y_i) = \sum_{t=1}^{m\times k} y_i^t, \qquad (44)$$

$$\rho(Y^H) = \sum_{t=1}^{m\times k} \min_{j\in H}\{y_j^t\}, \qquad (45)$$

$$\rho(Y_H) = \sum_{t=1}^{m\times k} \max_{j\in H}\{y_j^t\}. \qquad (46)$$

Функции (42),(43) замечательны тем, что на их основе строятся так называемые P - монотонные системы, для которых решение задачи (18-22) достигается не сложнее, чем за $c\cdot N$ число шагов вычисления функции $\mathcal{T}(i,H)$. Именно этот тип функций получил наибольшее распространение в прикладных исследованиях, использующих теорию монотонных систем [41-45] .

Итог проведенных построений сводится к следующему: предложена процедура, которая любой паре (e, W) элемент $*$ $*$ множество ставит в соответствие подмножество

$$G(e) \subseteq W : (e, W) \to G(e). \qquad (47)$$

Подмножество $G(e)$ является лучшим из всех возможных подмножеств W образом "e" в W в смысле преобразования (39) и критерия (15).

В заключение параграфа сделаем три общих замечания, касающиеся использования отображения (47) в сложных схемах анализа.

I. Если в качестве элемента "e" в (47) взять произвольный элемент $i \in W$, то получим $G(i)$. Поэтому выполнение процедуры (47) для всех $i \in W$ выявляет внутреннюю структуру сходства между элементами W. Такая структура в виде семейства подмножеств $\{G(i),\ i=\overline{1,N}\}$ может быть использована для детального анализа найденного для внешнего по отношению к W элемента "e" ассоциативного образа $G(e)$. В частности, можно оценить степень специфичности $G(e)$ с помощью следующего процесса:

а) определим подсемейство $\{G(i),\ i \in G(e)\}$,
и построим подмножество $G_1(e) = \bigcup_{i \in G(e)} G(i)$;

б) пусть на "k"-м шаге аналогичных построений найдено $G_k(e) = \bigcup_{i \in G_{k-1}(e)} G(i)$, тогда $G_{k+1}(e) = \bigcup_{i \in G_k(e)} G(i)$;

в) шаг k^* - шаг останова,*) если
$G_{k^*}(e) = G_{k^*-1}(e)$; соотношение между подмножествами

*) Очевидно, что указанный процесс сходится за конечное число шагов, так как W - конечно, а оператор $G_k = T(G_{k-1})$ - изотопный.

$G(e)$, $G_{k*}(e)$ и W можно рассматривать как качественную характеристику специфичности $G(e)$ в W (очевидно, что $G(e) \subseteq G_{k*}(e) \subseteq W$; поэтому количественно о специфичности можно судить по мощности $|G_{k*}(e) \setminus G(e)|$ разности $G_{k*}(e) \setminus G(e)$) .

2. Пусть W_1 и W_2 - два подмассива массива W .Тогда можно для каждого элемента из W вычислить его ассоциативные образы в W_1 и W_2 : $G_1(i)$ и $G_2(i)$. После этого величиной $|G_1(i) \cap G_2(i)|$ можно характеризовать степень проявления связи W_1 и W_2 на элементе i и на ее базе строить различные функции связи пары подмножеств из данного множества. Функции такого типа в нашем случае необходимы для целей сопоставления ассоциативных образов пары запросов или ассоциативных образов одного запроса, но составленных на базе двух существенно отличных отображений обрабатываемого массива в запрос. Заметим также, что такого рода функции важны и для разработки моделей кластеризации (в частности, они интенсивно используются, в так называемых агломеративных методах [69-74]).

Последнее замечание показывает, что модель ассоциативных образов может не только эксплуатировать и приспосабливать для своих нужд методы кластерного анализа, но и сама включаться в разработку новых методов кластеризации.

3. Рассмотрим случай, когда коэффициент сходства $\zeta(x,e)$ в (39) взят таким образом, что он измеряет различие между сравниваемыми векторами. Тогда описанная процедура (39) будет выделять в W " антиассоциативный " образ элемента e - подмножество элементов, которые в наибольшей степени непохоже на e .

Если предполагается, что e слабо связан с элементами из W, то может оказаться, что важнее, прежде всего, удалить насколько это возможно "антианалоги" элемента e.

В этой связи особенно полезным оказывается свойство решения (18-22), которое дают алгоритмы, предложенные в [75]: это решение, как отмечалось в первой главе, оказывается наибольшим по включению подмножеством, которое отвечает экстремуму (15), т.е. удаленной оказывается максимально возможная часть из "антиобразов".

В дальнейшем мы часто будем предпочитать рассматривать дополнение к такому антиобразу как искомый ассоциативный образ по сравнению с обычным ассоциативным образом, так как оно гарантирует выделение подмножества наименьшей мощности. Это важно, когда метод анализа строится как многоэтажная процедура.

Чтобы описанной схеме придать прикладное значение необходимо выбрать некоторую конкретизацию типов данных, которые фигурируют в рассматриваемом анализе. В этом выборе мы в настоящей работе руководствовались следующими соображениями.

Семантика шкал, по которой определяются значения анализируемых векторов может быть разнотипной. Значениями компонент этих векторов могут быть числа, номинальные и ранговые показатели, дерерья - иерархические классификации, произвольные бинарные отношения и множества более сложной структуры.

Чтобы обрабатывать такие данные будем далее использовать подход, развитый в [60]. В соответствии с ним исходная система разнотипных шкал преобразуется в систему булевых данных. Очевидно, что такой переход всегда возможен. Причем всегда

имеется возможность регулирования, какие свойства исходных шкал требуется наследовать в булевых преобразованных данных. Так, в частности, если исходная шкала номинальная, то ей можно поставить во взаимнооднозначное соответствие требуемый набор булевых переменных. Можно поступить и иначе - представить эту шкалу меньшим числом булевых переменных, агрегируя в одну булевую переменную несколько значений исходной номинальной шкалы.

Числовые данные можно представить булевыми переменными таким образом, чтобы сохранить информацию о линейном порядке числовой шкалы, или, наоборот, задать преобразование, которое элиминирует эту информацию [76].

Такая возможность регулировки преобразования разного типа шкал в булевые рассматривается в [60] как важный фактор дополнительных возможностей анализа обрабатываемого массива.

Далее, мы не будем касаться конкретных процедур преобразования исходного информационного массива и запросов в булевый стандартный вид. Предполагается, что конкретный способ преобразования выбран, преобразование проведено и дальнейшая работа ведется в однотипных булевых данных.

С позиций нашей основной цели, - разработки модели выделения ассоциативного образа для СУБД АСУ, - булевые данные имеют три достоинства:

I) они наиболее просты; на их языке наиболее выразительно и понятно формулируются предлагаемые процедуры анализа;

2) они допускают конструирование наиболее эффективных процедур вычислений;

3) в большом числе практических задач используются как исходные именно булевые данные.

В силу сказанного было решено намеченную конкретизацию реализовывать именно на булевых данных.

Выше неоднократно обращалось внимание на то, что модель ассоциативного образа, если предполагается ее прикладное использование,не может быть ограничена только формализмом определения множества аналогов. Поскольку это последнее множество контекстно-зависимое необходимо предоставить пользователю возможность синтезировать конечный результат, используя как это множество, так и информацию о его внутренней и внешней структурах.

Под этими структурами мы понимаем следующее.

Внутренняя структура - это упорядочение элементов множества аналогов запроса, основанное на оценке роли этих элементов в том, что данное множество выбрано как множество аналогов. Кроме этого упорядочения в описании внутренней структуры желательно иметь стратификацию этого множества - разбиение его на части, упорядоченные как целостные образования по оценке роли в образовании множества аналогов.

Для того, чтобы строить такое описание внутренней структуры можно, очевидно, воспользоваться языком оценки специфичности множества аналогов - $\overline{G}(e)$, который был охарактеризован выше.

В следующем параграфе, где строится основная конкретизация нашей модели для тех же целей предлагается другой более простой, учитывающий специфику конкретизации язык построения новой (суженной) монотонной системы на $\overline{G}(e)$.

Следует обратить внимание на то, что пользователь при описании закономерностей чаще оперирует не сравнениями объектов, а связями между признаками. Поэтому при разработке языка описания внутренней структуры необходимо кроме указанного выявления отношений между элементами и подмножествами множества аналогов иметь также язык описания связей между признаками, которые акцентируются, во-первых, уже выделением самого множества аналогов, и, во-вторых, указанным агрегированием его элементов.

Для различения этих языков мы будем называть описание внутренней структуры на первом языке поверхностным, а на втором - глубинным.

Глубинный уровень, как будет видно, требует усложнения исходной модели ассоциативного образа. Поэтому в следующем параграфе параллельно с разработкой основной (не усложненной) модели мы даем описание только поверхностного уровня описание структуры множества аналогов.

Для разработки описания внешней структуры необходимо, прежде всего, построить механизм вычисления функции, с помощью которой любой объект, доступный описанию в выбранной системе признаков, можно было оценить по степени его "удаленности" от построенного множества аналогов (не только те, что имеются в наличии в обрабатываемом массиве, но и любые те, которые можно сконструировать искусственно.

Как и при построении описания внутренней структуры множества аналогов во внешнем описании целесообразно выделять два уровня - поверхностный и глубинный, т.е. уровень анализа отношений между элементами исходного информационного массива,

не попавшими в множество аналогов, с этим множеством, и уровень анализа связей между признаками, которые предопределяют эти отношения.

Повидимому, к задачам построения внутренних и внешних структур выделяемого множества аналогов можно подойти с разных сторон.

Нам представилось целесообразным решить их, оставаясь в рамках единного аппарата монотонных систем.

Для реализации такого единообразного подхода в диссертации используется два приема: (I) построения композиции монотонных систем, т.е. создание таких конструкций из них, когда для вычисления весовой функции $\pi(i,H)$ одной системы необходимо уже иметь решение (ядро) другой; (2) построение монотонных систем одновременно не только на множестве элементов информационного массива, но и на множестве, описывающих их признаков, а также на более сложных, составленных из этих двух, множествах.

Содержательный анализ этих приемов послужил основой и для решения другой задачи диссертации - разработать новые модели кластерного анализа на базе созданных моделей ассоциативного образа.

§2.2. Ассоциативные образы на множествах, связанных с булевыми матрицами

Булевые данные - важная разновидность информационных массивов. Они сами достаточно часто используются в практике, но еще чаще к этому типу данных сводят произвольные номинальные данные, ранговую и иерархическую информацию, и даже числовую информацию, когда ее используют в загрубленном виде [75,77-79].

Информационный булевой массив будем представлять в виде матрицы $\Phi = \|\varphi_{ip}\|$, у которой N строк и n столбцов. Строка матрицы соответствует одной записи анализируемого массива. Тот факт, что объект, который соответствует i-й записи, имеет p-й признак (или свойство) фиксируется равенством $\varphi_{ip} = 1$. Если наоборот, $\varphi_{ip} = 0$, то это означает, что p-й признак отсутствует у i-го объекта.

Внешний объект - запрос $e = (e^1, \ldots, e^n)$ устроен аналогично: $e^p = 1$ означает, что в запросе имеется p-й признак, а $e^p = 0$ означает обратное, что не имеется.

В данном параграфе общая модель ассоциативного образа, которая разработана в предыдущем параграфе, конкретизируется для таких булевых данных.

Как следует из описания общей модели первый шаг ее построения состоит в выборе двух элементов:

1) некоторого семейства $R = \{R_1, \ldots, R_m\}$ наборов из исходных признаков

и

2) набора k различных коэффициентов ζ похожести, каждый

элемент которого можно вычислить для любых двух предъявленных векторов одной размерности.

Выбор семейства R может быть осуществлен двумя способами: на основе смыслового анализа исходного множества признаков специалистом, или автоматически, - на основе анализа статистических связей между признаками на данном конкретном информационном массиве. Возможен и комбинированный вариант. Для реализации второго способа можно воспользоваться какой-либо из известных процедур кластерного анализа [32-33]. Для выбора коэффициента похожести также имеется много возможностей [28, 89].

Учитывая отсутствие специфики в этих выборах в настоящем параграфе использована лишь их тривиальная реализация: полагаем, что в качестве групп разбиения выступают все единичные признаки, а в качестве коэффициента двух булевых векторов по одной выделенной координате используется функция совпадения "1" на x и e. В результате оказывается, что преобразование (39) приобретает вид:

$$y^t = x^t \& e^t, \quad \forall\, t = \overline{1,n}. \qquad (48)$$

Очевидно, что это преобразование обладает следующим свойством изотонности:

$$e_1 \subseteq e_2 \Rightarrow y(x, e_1) \subseteq y(x, e_2), \quad \forall x, \qquad (49)$$

где векторы e и y интерпретируются как множества признаков, которые у них имеются. В дальнейшем мы будем часто прибегать к такой интерпретации булевых векторов.

В качестве функции $\mathcal{T}(i,H)$, будем использовать простейшую функцию:

$$\mathcal{T}(i,H) = a \cdot \sum_{t=1}^{n} \max_{j \in H} y_j^t - b \cdot \sum_{t=1}^{n} y_i^t . \qquad (50)$$

Она соответствует использованию формулы (41) из общей модели. В интерпретации теоретико-множественных обозначений (50) имеет вид:

$$\mathcal{T}(i,H) = a \cdot |Y_H| - b \cdot |y_i| , \qquad (51)$$

где

$$Y_H = \bigcup_{j \in H} y_j ,$$

что совпадает с известными функциями, успешно примененными в [41,42] для изучения организационных систем. Главное отличие нашей функции от использованных в [42] состоит в том, что в [42] они служат характеристикой взаимозависимости между элементами изучаемого множества, а в данном случае они как от параметра зависят от внешнего вектора запроса e и тем самым выступают измерителем влияния запроса на структуру связей изучаемого информационного массива.

Решение задачи (II-2) в данном случае, как показано в [42], достигается следующим простым алгоритмом.

Алгоритм

Шаг I. Вычислить набор из N чисел $|y_j|$, $j = 1,2,\ldots,N$ и упорядочить их в порядке убывания:

$$|y_{i_1}| \geq |y_{i_2}| \geq \ldots \geq |y_{i_N}| . \qquad (52)$$

Шаг 2. Используя найденную на Шаге I последовательность индексов $I = \{ i_1, \ldots, i_N \}$ построить последовательность множеств:

$$Y_{H_1} = \bigcup_{j \in I} y_j ,$$

$$Y_{H_2} = \bigcup_{j \in I \setminus i} y_j , \qquad (53)$$

$$\ldots$$

$$Y_{H_N} = y_{i_N}$$

и вычислить их мощности $|Y_{H_1}|, |Y_{H_2}|, \ldots, |Y_{H_N}|$.

Шаг 3. На основе формулы (51) определить последовательность чисел $\pi(i_1, H_1), \pi(i_2, H_2), \ldots, \pi(i_N, H_N)$. (54)

Шаг 4. Определить

$$\pi(i^*, H_*) = \max_{s \in I} \pi(i_s, H_s), \qquad (55)$$

причем, если максимумов несколько, то в качестве i^* выбрать тот, который соответствует минимальному номеру s в последовательности I. Найденное H_* доставляет искомое решение задачи (18-22), т.е. $H_* = G$.

Содержательный смысл формул (54) и (55) указывает, что найденное решение G – это множество "антианалогов" запроса e в множестве W. Поэтому далее везде для последующего ана-

лиза (т.е. в качестве множества аналогов) используем его дополнение $W \setminus G = \bar{G}$.

Отметим некоторые свойства G .

I. Для любого $e \neq 0$

$$e \in W \Rightarrow e \in \bar{G} \,. \tag{56}$$

2. Если $e_1, e_2 \in W$ и $e_1 \leq e_2$ $(e_1 \neq 0)$, то в силу свойства изотонности преобразования (47)

$$e_1 \in \bar{G}(e_2) \tag{57}$$

а поэтому, очевидно,

$$\bar{G}(e_1) \cap \bar{G}(e_2) \neq 0. \tag{58}$$

3. Если $|e_2| = |e_1| + 1$ и дополнительная равная "I" координата e_2^{α} запроса e_2 такова, что $x_i^{\alpha} = 0$, $\forall i \in W$, то $\bar{G}(e_2) = \bar{G}(e_1)$; аналогично, если $e_2^{\alpha} = 0$, а $x_i^{\alpha} = 1$, $\forall i \in W$, то также их множества аналогов совпадают.

4. Если в упорядочении величин

$$|y_{i^*}| > |y_{i+1}| > \ldots > |y_{i_N}|$$

все неравенства строгие, то $G(e)$ - единственный максимум функции (15). Это свойство непосредственно следует из установленного в [43] факта, что любое решение (18-22) может быть получено как правый сегмент в (52) за счет перестановки только равных элементов:

Важным дополнительным анализом аналогов e в W является построение ассоциативного образа с заменой преобразования (48) совпадения " 1 " -x координат e и x преобразованием

$$y^t = \bar{x}^t \wedge \bar{e}^t, \qquad (59)$$

где $\bar{u} = I - u$. В этом случае в массиве W в качестве ассоциативного образа запроса e выделяется подмножество, элементы которого аналогичны e по отсутствующим признакам.

Целесообразность выявления такого рода аналогов по отсутствию отмечаются в исследованиях по автоматическому формированию понятий [81] и в работах по распознаванию образов [82].

Для целей обработки запросов в информационных системах использование ассоциативных образов с преобразованием ( ) должно носить факультативный, разъясняющий характер. Такую же роль может играть ассоциативный образ, основанный на использовании функции булевой эквивалентности

$$y^t = x^t \sim e^t. \qquad (60)$$

В целом, касаясь вопроса варьирования преобразованием (48), следует отметить логический характер возникающих здесь интерпретаций. В этой связи определенный интерес может иметь и использование той или иной функции импликации. Еще большие возможности вариации преобразованием $y(x, e)$ возникают, если допустить возможность в рассматриваемых булевых векторах "пропусков" или ошибочных значений.

В заключении данного параграфа опишем внутреннюю структуру выделенного ассоциативного образа $\bar{G}(e)$. Для этого построим на нем новую монотонную систему с функцией

$$\pi^{\nu}(i, H) = a \cdot |y_i| - b |Y^H|, \qquad (61)$$

где

$$i \in H \subseteq \bar{G}(e), \quad a \quad Y^H = \bigcap_{j \in H} y_j$$

Решение задачи (18-22) с такой функцией $\pi(i, H)$ при использовании преобразования (48) можно интерпретировать как "ядро" ассоциативного образа $\bar{G}(e)$. Обозначим его через $J(\bar{G}(e))$.

Кроме того, что оно само по себе выделяет в $\bar{G}(e)$ "наиболее сильных" аналогов запроса e, оно дает возможность естественным образом ввести линейный порядок на элементах $\bar{G}(e)$ по их силе быть аналогом e:

$$\Pi'(j_1, J(\bar{G}(e))) \geq \Pi(j_2, J(\bar{G}(e))) \geq \ldots \geq \Pi(j_g, J(\bar{G}(e))), \qquad (62)$$

где использованы следующие обозначения

$$g = |\bar{G}(e)|,$$

$$\Pi(j, J(\bar{G}(e))) = \Pi(j, J(\bar{G}(e)) \cup \{j\}).$$

Система чисел в (62) показывает не только упорядочение элементов из $\bar{G}(e)$, но и взаимное сходство в оценке "аналогичности" e, шкалу которой дает функция (61). Пользуясь этой шкалой можно оценить степень аналогичности e для отброшенных как антианалоги элементов из $G(e)$, а также и для произвольного нового вектора, который даже не входит в W. Функцию $\Pi(j, J(\bar{G}(e)))$ можно рассматривать как своего рода "размытое" множество аналогов e в W, которая в отличие от стандартных форм задания размытых множеств, определенных только на базовом множестве W, известна на всем множестве всех возможных булевых векторов заданной размерности W.

Таким образом, построение монотонной системы (61) на ассоциативном образе $\bar{G}(e)$ в качестве структуры этого образа дало:

а) его ядро $J(\bar{G}(e))$,

б) упорядочение $\bar{G}(e)$, определенное по (62), и

в) размытый ассоциативный образ, заданный с помощью функции $\Pi(j, J(\bar{G}(e)) \cup \{j\})$ на множестве всех возможных булевых векторов размерности $2^{|W|}$.

В соответствии с определениями предыдущего параграфа конструкции а) и б) - это внутренняя его структура, а конструкция в) - внешняя, причем их поверхностный уровень.

§2.3. Подпространство признаков, характерное для ассоциативного образа запроса

В рамках общей модели ассоциативного образа из §2.I в §2.2 для булевых данных была предложена схема обработки запроса e на массиве W, которая в качестве результата выдает

I) $\bar{G}(e)$ - множество аналогов e в W (ассоциативный образ e);

2) $J(e)$ - ядро $\bar{G}(e)$ $(J(e) \subseteq \bar{G}(e))$;

3) $\Pi(j, J)$ - размытый ассоциативный образ e в W.

Все эти конструкции выделяют информацию, которая содержится в W об e с помощью только оценки элементов W, но без явного указания, какие из признаков играют основную роль в выявленных связях между W и e.

Чтобы уметь строить такое указание, формализуем качественное представление о роли признаков, описывающих заданное множество $\bar{G}(e)$. При этом мы хотим иметь такую формализацию, которая бы позволяла разделять исходное множество признаков на две части: характерные для описания $\bar{G}(e)$ и нехарактерные. Естественно пытаться ее строить в рамках того же метода теории монотонных систем, которая позволила разделять W на $G(e)$ и $\bar{G}'(e)$.

Обозначим множество всех признаков через $\mathcal{P}(|\mathcal{P}|=n)$, а его подмножество, выделяемое единичными координатами запроса e, - через $P(e)$. Характерные признаки для $\bar{G}(p)$, можно пытаться выделять и в $\mathcal{P}$ и в $P(e)$. Каждый из этих вариантов несет свой содержательный смысл.

Однако для всех наших дальнейших построений более адекватным является анализ $P(e)$, так как процедура построения $\bar{G}(e)$ опиралась исключительно на информацию из этого подмножества признаков. Далее для сокращения обозначений везде, где это не приведет к путанице, будем писать P вместо $P(e)$ и $\bar{G}$ вместо $\bar{G}(e)$.

Через $\Phi(\bar{G},P)$ будем обозначать матрицу, которая получается из Φ выделением элементов, стоящих в Φ на пересечении строк из $\bar{G}$ и столбцов из P, а через $\Phi(P,\bar{G})$ – транспонированную к ней матрицу:

$$\Phi(P,\bar{G}) = \Phi^{T}(\bar{G},P). \qquad (63)$$

Рассмотрим теперь монотонную систему, у которой базовое множество $W=P$, а значения функции $\pi(i,H)$ вычисляются по информации из матрицы $\Phi(P,\bar{G})$ аналогично формуле (51):

$$\pi(i,H) = a\cdot|Y_H(P,\bar{G})| - b\cdot|y_i(P,\bar{G})|, \qquad (64)$$

где $i\in H\subseteq P$, $y_i(P,\bar{G})$ – подмножество элементов $\bar{G}$ $(y_i\subseteq\bar{G})$, у которой имеется i-й признак $(i\in P)$, а

$$Y_H(P,\bar{G}) = \bigcup_{j\in H} y_j(P,\bar{G}).$$

Обозначим решение задачи ( ) для этой системы через $F(e)$, а соответственно через $\bar{F}(e)$ – искомое множество характерных в $P(e)$ признаков для ассоциативного образа $\bar{G}(e)$. Полученное решение можно интерпретировать таким образом, что в W аналогами e являются элементы из $\bar{G}(e)$, причем из признаков $P(e)$, которые имеются в e, характерными для выделенных аналогов являются элементы из $\bar{F}(e)\subseteq P$.

В связи с построенной конструкцией возникает вопрос: получили бы мы ту же пару $(\bar{G}(e), \bar{F}(e))$, если бы действовали в обратном порядке: сначала на всем множестве W выделили бы из $P(e)$ часть признаков, которые "типичны" для большинства элементов из W (соответствующая процедура, очевидно, строится с помощью решения задачи (18-22) для монотонной системы (64), где только $\Phi(P,\bar{G})$ заменяется на $\Phi(P,W)$).

Найденное решение $\bar{F}_W(e)$ используется далее для нахождения $\bar{G}_W(e)$ по матрице $\Phi(W,\bar{F}_W(e))$ с монотонной системой, определяемой формулой (64).

Ответ на этот вопрос, очевидно, отрицательный: вообще говоря, эти две процедуры дают разные ответы, т.е. $(\bar{G}(e),\bar{F}(e)) \neq$ $\neq (\bar{G}_W(e),\bar{F}_W(e))$.

Вместе с тем, поскольку признаковая характеризация выделяемого множества аналогов важна для содержательного анализа этого множества, представляется целесообразным изыскать такой способ обработки запроса, который бы одновременно порождал пару "множество аналогов и характерные для него признаки" как целостное образование.

Ниже предлагается такого рода способ, который основан на тестировании всех возможных подмножеств $P(e)$ с определением в каждом из них своего специфического ассоциативного образа, выбираемого из всего W . На множестве таких пар "подмножество признаков и связанный с ним ассоциативный образ" строится единая оценочная функция опять с применением аппарата монотонных систем.

Рассмотрим матрицу $\Phi(W,Q)$, где Q - подмножество $P(e)$, и определим на W монотонную систему, функция $\pi(i,H)$ кото-

рой вычисляется на основе этой матрицы следующим образом:

$$\pi(i,H)=a\cdot|Y_H(Q)|-b\cdot|y_i(Q)|, \qquad (65)$$

где $y_i(Q)$ – подмножество признаков из Q, которые имеются у i-го элемента из H, а $Y_H(Q)=\bigcup_{j\in H} y_j(Q)$. Пусть $G(Q)$ – наибольшее подмножество W, на котором достигается максимум $F_Q(H)$:

$$F_Q(H)=\min_{i\in H}\pi_Q(i,H),$$

т.е.

$$F_Q(G(Q))=\max_{H\subseteq W}F_Q(H).$$

На множестве всех возможных пар (β,Q), где $\beta\in Q\subseteq P(e)$, определим новую функцию

$$P(\beta,Q)=F(G(Q\setminus\{\beta\})). \qquad (66)$$

Из того, что $|Y_H(Q\setminus\{\beta\})|\geq|Y_H(Q\setminus(\{\beta\}\cup\{\beta'\}))|$ для всех $\beta,\beta'\in Q$ и $|y_i(Q)|$ не зависит от H при всех Q, сразу следует, что

$$P(\beta,Q)\geq P(\beta,Q\setminus\{\beta'\}) \qquad (67)$$

для всех β,β' $(\beta\neq\beta')\in Q\subseteq P(e)$, т.е. что $P(\beta,Q)$ задает монотонную систему на $P(e)$. Решение экстремальной задачи на ней порождает пару $(Q^*,G(Q^*))$, которую как раз и выбираем как искомую пару: "характерное множество Q^* признаков и выделяемый на его основе ассоциативный образ $\overline{G}(Q^*)$" для запроса e в информационном массиве W. Содержательный смысл этой пары следующий: удаление любого из признаков $\beta\in Q^*$ приводит к силь-

ному изменению ассоциативного образа $\bar{G}(Q^*)$. Или, другими словами, набор Q^* признаков и ассоциативный образ $\bar{G}(Q^*)$ элементов из W - наиболее согласованная в указанном смысле слова пара.

Хотя по замыслу построения пара $(Q^*, \bar{G}(Q^*))$ играет ту же роль, что и пара $(\bar{G}(e), \bar{F}(e))$, которая отбирается последовательно, они описывают информацию в W с точки зрения запроса e существенно по-разному: в $(\bar{G}(e), \bar{F}(e))$ в качестве признаков, характеризующих $\bar{G}(e)$, используются те, каждый из которых имеется у многих элементов из $\bar{G}(e)$. Это, в частности, означает, что построение ассоциативных образов из W на $\bar{F}(e)$ и на $\bar{F}(e) \setminus \{з\}$ даст близкий результат для большинства $з \in \bar{F}(e)$. Напротив, ассоциативный образ $\bar{G}(Q^*)$ - это критический образ в том смысле, что удаление любого $з \in Q^*$ приводит к ассоциативному образу $\bar{G}(Q^* \setminus \{з\})$, существенно отличному от $\bar{G}(Q^*)$.

В указанном смысле следует считать целесообразным при обработке e на W получать и $(\bar{G}(e), \bar{F}(e))$ и $(Q^*, \bar{G}(Q^*))$. Однако, следует иметь ввиду, что вычисление $(Q^*, \bar{G}(Q^*))$ требует заметно больше времени, чем вычисление $(\bar{G}(e), \bar{F}(e))$.

На протяжении всего изложения §§2.1 и 2.2, а теперь и настоящего параграфа главное внимание уделялось разработке "большого запаса" различных конструкций, которые должна выдавать система обработки информационного массива W по запросу e . Это внимание было обусловлено следующей целью:

ассоциативный образ, который вырабатывает система, должен быть не только "правильным", (т.е. удовлетворяющим в конечном итоге потребителя), но и аргументированным или, другими словами, чтобы правильность (или неправильность) ответа пользователь

мог легко уяснить себе без привлечения других исследований по анализу W .

Эта цель вполне соответствует известному современному требованию выдачи не только правильного решения, но и объяснения решения (и даже его обоснования), которое предъявляют разработчики экспертных систем к таким системам [65-70,83,84].

В предыдущих параграфах каждый раз, когда строилась та или иная конструкция ассоциативного образа, мы рассматривали возможность ее многократного использования для детализации структуры W с точки зрения e .

В данном случае рассмотрим их с точки зрения построенной признаковой характеризации ассоциативного образа.

Во-первых, точно так же, как $\bar{G}(e)$ было построено на базе анализа строк матрицы $\Phi(W,P)$, возможно строить ассоциативный образ и по матрице $\Phi(\bar{G}(e),\ \bar{F}(e))$. Это даст множество $\bar{G}_1(e)$. После этого строится матрица $\Phi(\bar{F}(e),\bar{G}_1(e))$. Этот процесс последовательной редукции исходной матрицы $\Phi(W,P)$ поочередно по строкам и столбцам можно продолжать до некоторого k , после которого он естественным образом оборвется. В результате получается две последовательности вложенных друг в друга множеств:

$$\bar{G}(e) \supseteq \bar{G}_1(e) \supseteq \bar{G}_2(e) \supseteq \ldots \supseteq \bar{G}_k(e)\ ; \qquad (68)$$

$$\bar{F}(e) \supseteq \bar{F}_1(e) \supseteq \bar{F}_2(e) \supseteq \ldots \supseteq \bar{F}_k(e)\ . \qquad (69)$$

Наличие сходства между первой из этих последовательностей и внутренней структурой $\bar{G}(e)$, определяемой по $\Pi(j,J)$, построенной в §2.2, если оно будет выявлено, может рассматриваться как свидетельство устойчивости внутренней структуры $\bar{G}(e)$.

Основанием для такого суждения является тот простой факт, что эти два описания структуры $\bar{G}(e)$ созданы на разной основе, что означает малую вероятность их сходства. Поэтому, если тем не менее это сходство будет обнаружено, то оно будет свидетельствовать о наличии сильных внутренних связях в организации $\bar{G}(e)$.

Аналогично описанному процессу детализации $(\bar{G}(e), \bar{F}(e))$ можно построить процесс детализации $(Q^*, \bar{G}(Q^*))$. Вместо матрицы $\Phi(W, P)$, по которой определялась пара $(Q^*, \bar{G}(Q^*))$, возьмем матрицу $\Phi(\bar{G}(Q^*), Q^*)$ и на ее основе найдем $(Q_1^*, \bar{G}(Q_1^*))$, затем по матрице $\Phi(\bar{G}(Q_1^*), Q_1^*)$ найдем $(Q_2^*, \bar{G}(Q_2^*))$ и т.д. до некоторого останавливающего этот процесс k , который даст $(Q_k^*, \bar{G}(Q_k^*))$

Установление тех или иных связей в последовательностях пар, стартующих с $(\bar{G}(e), \bar{F}(e))$ и с $(Q^*, \bar{G}(Q^*))$ не только дает еще один аргумент в пользу устойчивости внутренней структуры $\bar{G}(e)$, но и может помочь на чисто формальном уровне рассмотреть тонкое соответствие, которое должно существовать между типичными признаками образа $\bar{G}(e)$ (элементами $\bar{F}(e)$) и его критическими признаками (элементами Q^*).

ГЛАВА 3

ПРИМЕНЕНИЕ МОДЕЛИ АССОЦИАТИВНОГО ОБРАЗА В ЗАДАЧАХ ВЫЯВЛЕНИЯ СТРУКТУРЫ СЛОЖНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

§3.I. О целесообразности использования модели ассоциативного образа в разработках новых методов кластерного анализа

Хорошо известно [36], что в области изучения структуры сложных эмпирических данных методы кластерного анализа занимают главное место.

До недавнего времени они широко использовались только в работах исследовательского характера *). В работах по проектированию, разработке, построению, испытанию и совершенствованию режимов эксплуатации практических систем они стали применяться относительно недавно [26].

Литература последних лет [4, 40] убеждает, что область создания и эксплуатации информационных систем и, в частности, АСУ различного профиля, может получить важные новые возможности развития за счет внедрения методов кластеризации в дело решения вопросов анализа и синтеза структуры информационных систем и протекающих в них процессов. Действительно, в любой из таких систем и элементы их структуры и события, составляющие содержание протекающих процессов, являются взаимозависимыми, причем так, что эти связи трудно прослеживаются и из-за

---

*) В самых разных областях: техники, медицины, экономики, организации производства и т.д. [34,35,38].

большого числа элементов (событий) и из-за сложной структуры связей, которая к тому же часто является изменчивой во времени. Очевидно, значение знаний о структуре этих взаимозависимостей на всех этапах анализа и синтеза таких систем [11]. Именно на выявление основных свойств таких структур нацелены имеющиеся и разрабатываемые методы кластеризации. Однако примеров использования этих методов в выявлении указанных структур мало [12,15,61,69,83,85-88].

Мы полагаем, что имеются две главные причины, затрудняющие широкое внедрение методов кластеризации в этой области:

I) существующие методы кластеризации опираются на функции, оценивающие сходство-различие пары элементов вне зависимости от множества, в котором эти элементы взаимодействуют и от которого поэтому зависит их сходство; тот очевидный факт, что данную пару элементов в одном случае следует признать как пару сильно похожих элементов, в другом случае необходимо различать, - такие функции принципиально не могут учесть;

2) результаты, которые обычно предоставляются программами кластерного анализа, - это чаще всего лишь перечни состава кластеров, иногда эти перечни сопровождаются результатами элементарной статистической обработки внутри кластеров, еще реже для выделенных кластеров предлагается математическая модель, позволяющая новые не участвующие в обработке элементы кластеризовать (отнести к нужному кластеру); лишь совсем недавно в связи с намечающейся интеграцией работ по кластерному анализу и по созданию так называемых экспертных систем стала осознаваться проблема необходимости, чтобы результат кластеризации сопровождался развернутым его объяснением, а еще лучше обосновы-

вающими аргументами. Сверх этого необходимо иметь специальные диалоговые средства обсуждения этих результатов с пользователем с целью активного осознания им своего уровня доверия к результату и получения возможности его коррекции без привлечения специальных исследований собранного информационного массива или для выработки решения, что такие исследования необходимы [2,7,7о,89-93].

С учетом сказанного, становится актуальным попытаться построить новый метод кластеризации на базе поставленной во второй главе задачи нахождения в информационном массиве подмножества аналогов (ассоциативного образа) любого заданного элемента. Первая причина: тесная содержательная связь этой задачи с задачей, которую решают методы кластеризации (построение кластеризации - это именно процесс анализа структуры аналогов для каждого элемента кластеризируемого множества). Вторая причина более существенна для данного рассмотрения решения задачи поиска ассоциативного образа, представленное в той же второй главе, базируется именно на "контекстной" оценке сходства объектов (принадлежность вектора x из информационного массива X к заданному вектору e определяется с учетом глобальной информации о связях x и e " на фоне " X), и, кроме того, оно помимо состава ассоциативного образа $K(e)$ вектора e в массиве W включает специальное описание и внутреннего строения $K(e)$, и характеристики "его места" в X, и, наконец, определение структуры признакого пространства, объясняющего происхождение найденного $K(e)$. Другими словами, это решение полностью удовлетворяет двум указанным выше требованиям, в несоблюдении которых мы видим основные трудности внедрения методов

кластеризации в область анализа и синтеза информационных систем.

Таким образом, возникает новая задача настоящей диссертации - предложить метод кластеризации, удовлетворяющий указанным требованиям, опираясь на модель ассоциативного образа и на то, что сама,модель, как это уже отмечалось, удовлетворяет указанным требованиям). Решение этой задачи составляет содержание следующего параграфа настоящей главы. В последнем параграфе мы рассматриваем специальную задачу кластеризации - выявление одновременно системы кластеров и макроструктуры, определяющей взаимозависимости между ними. Эта задача особенно часто возникает при анализе функционирования сложных информационных комплексов [94]. В §3.3 показывается, что в этом случае(специальной задачи кластеризации) перспективно разрабатывать новые методы, опирающиеся на предлагаемую в диссертации модель ассоциативного образа.

§3.2. Теоретико-множественная структура ассоциативных образов и ее использование в задаче кластерного анализа

Основная цель настоящего параграфа – рассмотреть новые возможности в разработке кластерных процедур, которые дают непосредственно только ассоциативные образы кластеризуемых объектов, без использования сопровождающей детальной информации о строении этих образов. Эти дополнительные возможности будут очевидны после указанного рассмотрения, и поэтому мы на них останавливаться не будем.

Итак, речь пойдет о задаче кластеризации элементов множества $W=\{1,2,\ldots,N\}$, каждый "i" ($i=\overline{1,N}$) из которых в исходном представлении описан своим вектором $x_i=\{x_i^1, x_i^2, \ldots, x_i^n\}$ значений признаков из множества $\{x_i^t,\ t=\overline{1,n}\}$. Этот первоначальный массив $X=\|x_i^j\|_N^n$ исходных данных в соответствие со схемой анализа, описанной в §2.2, преобразуется в массив $Y^i=\|y_s^t\|_N^{m\times k}$ коэффициентов похожести каждого элемента $s\in W$ с некоторым выделенным элементом "i" также принадлежащим W ($i\in W$). После этого для выделенного элемента "i" строится его ассоциативный образ $K(x_i)$. Повторяя эту процедуру N раз все время для нового элемента из W мы получаем семейство ассоциативных образов в W всех элементов i из этого W. Обозначим это семейство через K:

$$K=\{K(x_1), K(x_2), \ldots, K(x_N)\}. \qquad (70)$$

Рассмотрим три из различных возможных типов бинарных отношений, которые непосредственно устанавливаются на элементах из

W с помощью построенного семейства K :

I. Отношение "элемент x элемент" ("согласованы") задается матрицей $A = \|a_{ij}\|_N^N$, где

$$a_{ij} = \begin{cases} \text{I}, \text{ если для всякого } \ell = \overline{1,N} \quad \text{имеем} \\ \quad [\, i \in K(x_\ell) \Rightarrow j \in K(x_\ell)\,] \,; \\ 0, \text{ в противном случае} \end{cases}$$

2. Отношение "образ x образ" ("пересекаются") задается матрицей $B = \|b_{ij}\|_N^N$, где

$$b_{ij} = \begin{cases} \text{I}, \text{ если } K(x_i) \cap K(x_j) \neq \phi \,, \\ 0, \text{ в противном случае} \,. \end{cases}$$

3. Отношение "элемент x образ" ("находится β ") задается матрицей $C = \|c_{ij}\|_N^N$, где

$$c_{ij} = \begin{cases} \text{I}, \text{ если } \quad i \in K(x_j) \,, \\ 0, \text{ в противном случае} \,. \end{cases}$$

Можно интерпретировать матрицу A как матрицу смежности ориентированного графа, построенного на W как на множестве вершин. Тогда разбиение этого графа на компоненты связности - один из вариантов кластеризации элементов из W . Другой аналогичный по способу построения, но не по результату, вариант получается, если вместо A использовать для тех же цепей матрицу B (ей соответствует, очевидно, уже неориентированный граф).

Рассмотрим теперь матрицу C . Разобьем ее строки на непересекающиеся группы в соответствие с правилом:

i - я и j - я строки объединяются в одну группу, если они совпадают, т.е. если вектора $a_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$ и $a_j = \{a_{j1}, a_{j2}, \ldots, a_{jn}\}$ равны $(a_i = a_j)$.

Обозначим построенную таким образом кластеризацию через $S = \{S_1, S_2, \ldots, S_L\}$, где L - число ее кластеров $(1 \leq L \leq N)$.

Будем интерпретировать элементы семейства K ассоциативных образов как свойства, наличие (или отсутствие) которых можно легко проверить у любого элемента из W (свойство есть, если этот элемент входит в соответствующий образ). В этих терминах построенную кластеризацию можно охарактеризовать следующим образом: для любой пары элементов i, j , принадлежащих любому кластеру S_ℓ из S в наборе свойств K нет ни одного, которое бы их различало (считается, как обычно, что свойство различает два элемента, если у одного из них оно имеется, а у другого отсутствует).

Интересно, что кластеризацию S можно получить и как компоненты связности графа, определенного на W с помощью матрицы $A' = \|a_{ij}\|_N^N$, являющейся прямым аналогом матрицы A :

$$a'_{ij} = \begin{cases} \text{I, если для каждого } \ell = \overline{1,N} \text{ верно, что} \\ \quad [i \in K(x_\ell) \Rightarrow j \in K(x_\ell)] \,\&\, [j \in K(x_\ell) \Rightarrow i \in K(x_\ell)], \\ 0, \text{ в противном случае.} \end{cases} \qquad (71)$$

Указанная связь между кластеризацией S и семейством K ассоциативных образов не исчерпывает отношений между ними. Цель дальнейшего изложения состоит в том, чтобы показать, как на основе анализа чисто теоретико-множественных соотношений между подмножествами из S и из K можно стандартным способом строить кластеризацию, которая существенно более тонко отражает струк-

туру ассоциативных связей между элементами W. С точки зрения этого построения кластеризация S выступает как исходная.

Построение состоит в

I) формировании кластеризации, которую будем называть базовой,

2) выявлении структуры базовой кластеризации в виде нескольких слоев и агрегирования, уровни которого упорядочены по этапам процесса агрегирования.

Формирование базовой кластеризации сводится к нахождению для каждого S_ℓ, $\ell=\overline{1,L}$ подмножества $T_\ell=\bigcup_{i\in S_\ell} K(x_i)$. Ясно, что $S_\ell\subseteq T_\ell$. Семейство $T=\{T_1,T_2,\dots,T_L\}$ будем называть базовой кластеризацией. Для каждого кластера T_ℓ, $\ell=\overline{1,L}$ множество S_ℓ будем называть его ядром, а множество $T_\ell\setminus S_\ell$ – оболочкой (размытой частью кластера). Очевидно, что базовая кластеризация, вообще говоря, имеет пересекающиеся кластеры (последнее не имеет место, если, в частности, $T_\ell=\bigcup_{i\in S_\ell} K(x_i)=S_\ell$).

Под структурой кластеризации T будем подразумевать граф $\Gamma(T,V)$, у которого элементы T (кластеры T_ℓ, $\ell=\overline{1,L}$) – это вершины, а дуги определяются по правилу: $v_{q\ell}=(T_q,T_\ell)$, если $S_q\subseteq T_\ell$. Содержательный смысл импликации, индуцированной дугой $v_{q\ell}$, заключается в суждении: "из факта принадлежности элемента кластеру T_q следует, что он принадлежит также и кластеру T_ℓ.

Назовем структуру $\Gamma(T,V)$ корректной, если для каждого $\ell=\overline{1,L}$ выполняется условие: из того, что $v_{q\ell}\in V$, следует, что у ℓ-ой вершины имеется петля $v_{\ell\ell}\in V$. Другими словами, импликация $v_{q\ell}$ в корректной структуре имеет место тогда и только тогда, когда ядро S_ℓ находится внутри своего класте-

ра T_ℓ . По построению структура $\Gamma(T,V)$ базовой кластеризации всегда является корректной.

Определение. Граф $\Gamma(T',V')$ называется агрегированной структурой корректного графа $\Gamma(T,V)$, если:

I) T' - это разбиение T как множества на подмножества (кластеры следующего уровня), каждое из которых соответствует полному подграфу графа $\Gamma(T,V)$;

2) $v_{q\ell} \in V'$ означает, что в T_q' имеется вершина $T_\alpha \in T$, а в T_ℓ' - вершина $T_\beta \in T$, такие, что $v_{\alpha\beta} \in V$ в $\Gamma(T,V)$.

При этом T' называется агрегированной кластеризацией.

Определим в агрегированной кластеризации T' ядро и оболочку макрокластера T_q' как объединение соответственно ядер и оболочек всех входящих в T_q' исходных кластеров из T . Тогда легко показать, что граф $\Gamma(T',V')$ - это вновь корректный граф.

Поскольку построение нового (агрегированного) корректного графа может привести к возникновению на нем новых полных подграфов, которых ранее не было в графе $\Gamma(T,V)$, то повторное применение операции агрегирования теперь уже к $\Gamma(T',V')$ может дать, вообще говоря, новую информацию в виде следующего уровня агрегированную структуру - граф $\Gamma(T'',V'')$. Такой процесс последовательного агрегирования имеет естественное условие останова - ситуацию, когда дальнейшее применение указанной операции агрегирования невозможно.

Описанный процесс - новая агломеративная процедура, которая начинается с базовой кластеризации T , имеющей L кластеров и заканчивается максимально агрегированной кластеризацией T^* с числом кластеров L^* $(L^* \le L)$.

Важно, что на каждом уровне в полученной в результате этой

процедуры структуре имеется кластеризация, которая представлена в двух формах: ядерной, когда кластеры - ядра не пересекаются, и размытой - с пересечениями между оболочками кластеров, причем ядерные формы образуют свое агломеративное дерево агрегации, а размытые - свое. Связь этих двух деревьев описывается отношением, которое формализовано структурой $\Gamma(\tilde{T}, \tilde{V})$, где обозначение волна" $\sim$ "над символами T и V подчеркивает, что речь идет о некотором, вообще говоря, промежуточном уровне агрегирования.

Описанный агломеративный процесс существенно зависит от того, каким способом разбивается текущий граф $\Gamma(\tilde{T}, \tilde{V})$ на полные подграфы. А разных вариантов такого допустимого разбиения (агрегирования) $\tilde{T}$ и соответственно $\tilde{V}$ может быть несколько.

Эта многовариантность может послужить и на пользу дела - для представления пользователю некоторых вариантов агрегирования с целью, чтобы окончательную кластеризацию он выбрал самостоятельно.

§3.3. Алгоритм обработки данных, представленных в форме графа информационных связей

В области разработки и исследований сложных информационных систем с начала шестидесятых годов традиционным языком описания является язык теории графов. В [95] было показано, что данные, представленные в форме графа, несут существенную специфику с точки зрения постановок задач кластеризации такой информации. В указанных работах были предложены и первые такие специальные постановки задач, и первые алгоритмы их решения.

Главная специфика возникающих в этой области задач кластеризации включает:

1) необходимость различать два описания вершины графа - описание структуры входящих в нее дуг и описание структуры ее исходящих дуг,

2) необходимость не только кластеризации вершин графа по двум указанным описаниям, но и построение агрегированного графа, вершинами которого выступают кластеры, т.е. наряду с кластеризацией представления макроструктуры связей между кластерами.

Таким образом, специфику этих задач можно представить как требование объединять в кластеры не столько вершины, которые сильно связаны между собой, как это делается в общих методах кластеризации, сколько объединять в кластеры вершины, у которых "похожи" структуры и входящих и исходящих из них дуг. Такая переформулировка требований к методам решения задач кластеризации вершин информационного графа непосредственно подсказывает способ применения для этих целей разработанной во вто-

рой главе модели ассоциативных образов.

Действительно, если обозначить через $M(\Gamma)=\|m_{ij}\|_N^N$ матрицу смежности анализируемого графа:

$$m_{ij}=\begin{cases}1, & \text{если в графе } \Gamma \text{ есть дуга } (i,j)\\ 0, & \text{в противоположном случае,}\end{cases} \qquad (72)$$

то в качестве простейшего описания множества вершин этого графа можно взять расширенную матрицу

$$\Psi=\|\psi_{ij}\|_N^{2N}=(M(\Gamma),M^T(\Gamma)),$$

где через $M^T(\Gamma)$ обозначена матрица, транспонированная к $M(\Gamma)$. Первые N компонент строки этой матрицы - это элементы матрицы $M(\Gamma)$, т.е. описание структуры исходящих дуг соответствующей вершины графа, а вторые N компонент ее - это элементы матрицы $M^T(\Gamma)$, т.е. описание структуры входящих дуг той же вершины.

Далее на базе модели ассоциативного образа, предложенной во второй главе, мы строим алгоритм кластеризации вершин информационного графа по их описаниям, представленным в виде строчек матрицы Ψ .

Будем называть запрос максимальным, если все компоненты соответствующего ему вектора e равны "I". В нашем случае размерность пространства, в котором задан информационный массив (матрица Ψ) и вектор запросов, равна $2N$.

Вырожденным ассоциативным образом запроса e будем называть такой образ, при построении которого полагаем, что $y_i=x_i$, т.е. когда компоненты исходных векторов трактуются как некоторые коэффициенты сходства их с запросом e .

В предлагаемом ниже алгоритме искомая кластеризация множества W вершин графа Γ строится как последовательная про-

цедура исчерпания этого множества вырожденными ассоциативными образами максимальных запросов.

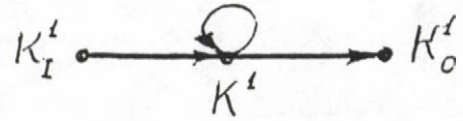
На основе информации из матрицы Ψ с помощью процедуры, описанной в §2.3, отыскивается вырожденный ассоциативный образ максимального запроса. Обозначим его через K^1 $(K^1 \subseteq W)$.

Далее с помощью процедуры выделения подмножества типичных признаков для K^1, описанной в §2.4, находим в W подмножество F вершин, каждая из которых имеет много дуг, ведущих в K^1. При этом элементы F несут пометку типа дуг, которые связывают их с K^1 - входящие или исходящие из K^1 (в соответствие с блочным представлением матрицы Ψ через матрицы $M(\Gamma)$ и $M^T(\Gamma)$. Другими словами, каждый элемент из W в F имеет знак, который указывает, что он рассматривается только как "вход" или только как "выход" из K^1, т.е. один и тот же элемент W может встретиться в F дважды. Поэтому выделим из F отдельно подмножество F_I "входов" в K^1 и отдельно подмножество F_o "выходов" из K^1. Вообще говоря, $F_I \cap F_o = \Phi$.

Теперь мы готовы, чтобы построить первые три кластера вершин графа Γ :

$$K_i^1 = F_i \setminus K^1, \quad K^1, \quad K_o^1 = F_o \setminus K^1. \qquad (73)$$

В качестве макроструктуры, которая по построению должна связывать эти три кластера, естественно рассмотреть следующий трехблочный комплекс:



Блок K^1 этого комплекса интерпретируется как блок "переработки информации", блок K_I^1 - "поставщик" информации для K^1 (его справочная), а блок K_O^1 - "потребитель" результатов обработки, которая проводится в K^1 .

Глава 4

ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ МОДЕЛИ АССОЦИАТИВНОГО ОБРАЗА, ЕЕ ИСПЫТАНИЕ И ИСПОЛЬЗОВАНИЕ ПРИ КОНСТРУИРОВАНИИ БЛОКА ДИСПЕТЧЕРИЗАЦИИ В АСУ ПОЛИКЛИНИКИ

§ 4.I. Организация экспериментального исследования

Модель ассоциативного образа, представленная в гл. 2 и 3, - это весьма общий подход и к задачам конструирования СУБД АСУ, и к задачам обработки сложных данных вообще. Он является общим не только потому, что используемый в нем аппарат слабо специфицирован (например, допускает использование произвольных монотонных по второму аргументу функций оценки $\pi(i,H)$ взаимозависимости элемента и множества). Он общий и из-за большого разнообразия своих структурных особенностей (разнообразия процедур описания внутреннего и внешнего строения множества аналогов для данного запроса).

Реализация этого подхода в достаточно полном объеме - это разветвленный пакет программ, осуществление которых есть большое самостоятельное научное исследование.

Для апробации модели, которая есть главная цель настоящей главы, необходимо реализовать программно лишь серию процедур и сконструировать из них комплекс, приспособленный для экспериментального исследования модели. Выполнение такой работы потребовало построения плана этого экспериментального исследования. Этот план был составлен с учетом целей и задач диссертации, сформулированных в § I.4. Его изложению как раз и посвящен настоящий параграф.

Отправной точкой разработки явилось представление, что необходимо создать несколько различных модельных эксперимен-

тальных ситуаций, каждая из которых, с одной стороны, могла бы служить прототипом массовых прикладных исследований, а с другой стороны, позволяла бы тем или иным образом применить разработанный аппарат ассоциативных образов и оценить перспективность такого применения (его содержательную адекватность и методическую работоспособность). В основе этого представления лежала установка: новизна предлагаемой модели ассоциативного образа должна вызывать сильные затруднения в решении вопросов, как и где ее использовать на практике. И главная трудность не в том, чтобы уметь априори оценить ожидаемый эффект, а в том, что привычные приемы применения стандартных методов выявления эмпирических закономерностей не пригодны, чтобы утилизировать эту новую модель.

Таким образом, было решено сначала придумать, как и где в принципе можно было бы применить модель, чтобы результат этого применения было легко оценить, интерпретировать, продемонстрировать. Затем необходимо было каждую такую виртуальную ситуацию эксперимента конкретизировать, т.е. найти совершенно определенный экспериментальный материал и приспособить к нему схему апробации нашей модели.

В соответствии с целями диссертации, поставленными в § I.4, область внедрения предложенной модели - разработка АСУ поликлиники. Ее мы также рассматривали и как сферу апробации модели. Такого рода АСУ только-только создаются. Поэтому исследование в этом направлении прежде всего должно ответить на вопрос - каков главный алгоритм функционирования такой системы.

Ниже описываются три совершенно разные виртуальные ситуации проведения экспериментального исследования по анализу массивов информации с указанием, как в них можно использо-

вать нашу модель. Это описание по необходимости носит качественный, ориентировочный характер. Ниже описываются характеристики проведения экспериментов.

I. Классификация ключевых слов журнала ИФАК "Автоматика". Этот журнал Международной федерации по автоматическому управлению (Изд. Пергамон Пресс, Лондон) использует список ключевых слов, состоящий из 565 терминов (Приложение I). В 1983г. Комитет по публикациям ИФАК принял решение о необходимости усовершенствования этого списка и попросил Технический комитет ИФАК по терминологии и стандартизации рассмотреть модернизации терминов автоматизации. Задача состоит в следующем:

- исключить термины, которые "устарели";
- объединить в общие термины те, которые покрывают слишком узкие области;
- детализировать термины, которые в процессе развития науки стали слишком обобщенными и созрели до разделения на подтермины.

В этой работе участвуют десятки специалистов по терминологии и выдающиеся ученые по автоматическому управлению. Во время Всемирного конгресса ИФАК в 1984г. была собрана информация о научных интересах более 1000 участников на основе этого же списка ключевых слов. Поэтому возникла идея использовать этот массив информации и способствовать вышеупомянутой терминологической работе.

В результате анализа этого массива информации с помощью методов из гл. 2 была получена классификация ключевых слов и выдвинуты предложения по модификации списка терминов.

При появлении новых направлений и областей научных исследований имеется только ограниченный круг специалистов, которые разрабатывают эти направления и области, причем в доволь-

но узком спектре возможных приложений.

Если в дальнейшем новое направление развивается, то происходит разветвление дерева исследований и появляется ряд подтерминов. При широком развертывании данной области эти подтермины поднимаются до ранга терминов. Мы видели такое явление в таких областях (терминах), как адаптивное управление, САПР или искусственный интеллект. На настоящем этапе развития эти термины уже не могут быть ключевыми словами списка, который используется журналом ИФАК "Автоматика", и должны быть детализированы.

Повышенный интерес специалистов по терминам часто объясняется вышеупомянутым фактом. Однако надо отметить, что причиной такого сверхинтереса может быть и настоящая популярность, показывающая высокое значение и перспективность данной области.

Если специалисты по терминологии проводят глубокий анализ самых популярных терминов, то они могут исключать из их списка те ключевые слова, которые требуют дальнейшего разбиения. После такой фильтрации мы получим список областей автоматического управления, вызывающий самый высокий интерес специалистов.

По просьбе Программного комитета следующего Всемирного конгресса ИФАК, который будет проводиться в 1990г., мы провели дополнительный анализ и выбрали основные области научных исследований, по которым рекомендуется пригласить докладчиков пленарного заседания Конгресса и провести специальные заседания за круглым столом.

Руководители Международной федерации по автоматическому управлению заявили о необходимости создания автоматизированной системы управления для постоянного усовершенствования тер-

минологии в области автоматизации, которая способствует созданию терминологических словарей ИФАК и выбору тематики научных мероприятий, организованных Федерацией.

2. Выявление социально-экономических закономерностей. Социально-экономические исследования, которые проводятся все в более увеличивающихся масштабах различными общественными и государственными организациями, опираются на анализ больших баз сильно изменяющихся комплексных данных. Очень важно, что эти базы быстро изменяющейся информации являются в каждый момент исследования уникальными и крайне слабо или совсем не изученными. Поэтому в практике социально-экономических исследований часто необходимо решать задачу быстрого проведения ориентировочного анализа с целью выявления грубых закономерностей в этих больших уникальных меняющихся данных.

Типичный подход состоит в том, что выделяются два поднабора признаков - объясняющий и объясняемый [94]. По первому весь банк сортируется на кластеры так, что каждый кластер соответствует одной из возможных комбинаций значений первого набора признаков (каждый признак имеет конечное множество значений). В объясняемом наборе признаков выделяются одна-две комбинации значений, для которых у исследователя имеется гипотеза об их распределениях по множеству комбинаций первого набора признаков. Эту гипотезу он проверяет стандартной процедурой сортировки. В то же время исследователь рассматривает конкретную объясняемую комбинацию значений признаков из второго набора лишь как представитель некоторого более широкого подмножества объектов. После построения распределения интересующей комбинации по комбинациям объясняющего набора он строит искомое распределение на подмножестве объектов, на которых встречается комбинация значений признаков второго набора. Если встре-

чающаяся комбинация отличается от заданной им комбинации значений этих же признаков на малую величину в единицах заранее выбранной оценки сходства таких комбинаций, то проверяемая гипотеза принимается [16,72,73,84,91,96-102].

Таким образом, задача выявления закономерностей оказывается тесно связанной с задачей выявления множества аналогов данного запроса. Поэтому у нас имеется возможность в саму постановку обсуждаемой задачи ввести существенное изменение. Именно, вместо искомого распределения для множества традиционных аналогов построить распределение для ассоциативного образа выбранной им комбинации. Если материал исследования выбран так, чтобы результаты можно было просто оценить с содержательной точки зрения, то получаем третью возможность апробировать предлагаемую модель ассоциативного образа.

3. Схема работы АСУ поликлиники. В настоящее время, когда компьютеризация проникает практически во все отрасли медицины, открывается возможность сделать диспансерное обслуживание населения не только вскрывающим латентную заболеваемость, но и прогнозирующим ее возникновение и развитие.

В последние годы для совершенствования этого обслуживания и особенно первых его этапов во всем мире разрабатываются анкетно-компьютерные методы доврачебного осмотра здорового населения.

По самому замыслу такого рода опросы нацелены на общенациональный охват населения.

Вместе с тем, это не просто опросы, а и заключения предупреждающего характера о том, к каким врачам данному обследованному следует обратиться.

Ясно, что такого рода механизированные обследования могут играть лишь роль селектора, выделяющего группу населения с

повышенным риском заболеваемости. Их важнейшее свойство, которое выделяет их из других методов, состоит в том, что обследуемый сразу получает варианты решения, к каким врачам ему следует обратиться.

Такие компьютерные системы доврачебного осмотра работают на основе грубых принципов:

I) ответ на один вопрос анкеты - "есть" или "нет" соответствующей жалобы (есть также ответ "не знаю");

2) каждая жалоба связывается только с одним врачом-специалистом;

3) рекомендация, к каким врачам обратиться, составляется по правилу - ко всем тем, в списках жалоб которых больной выбрал не менее чем заранее фиксированное число n_0.

Первый принцип касается только типа анкеты, а второй и третий - метода ее обработки.

С современных воззрений теории распознавания указанный метод, конечно, не выдерживает критики. Во-первых, он игнорирует фундаментальный факт, что жалобы можно оценивать только совместно, а не изолированно каждую. Во-вторых, он противоречит общепринятым в медицине представлениям, что одна и та же жалоба может иметь отношение к разным врачам. Наконец, в нем не учитывается содержание жалоб, а оценивается только их число, причем пороговое число, с которым сравнивается достигнутое, одинаково для всех специалистов.

Хорошо известно, как строятся решающие правила в подобных ситуациях в соответствии с теорией распознавания образов [103]. Такой подход можно было бы использовать и в данном случае, например, в следующем виде.

Обучающая выборка - анкеты с ответами обследованных и заключениями экспертов-врачей, которые изучили каждую такую ан-

кету и указали, куда следует направить соответствующего опрошенного.

После того как такая выборка сформирована, на ее базе решается столько задач обучения распознаванию, сколько врачей-специалистов имеется в общем списке возможных для направления. При этом каждый такой раз обучающая выборка разбивается на два кластера обследованных: один кластер - это те обследованные, которым рекомендовано пойти к данному врачу, остальные - которые не получили такую рекомендацию.

Вместе с тем, если последовательно придерживаться представления, что жалобы одного индивида - это структурное образование, отражающее статус и состояние его здоровья, то изолированное прогнозирование каждого врача-специалиста вызывает возражение.

Нам представилось перспективным рассмотреть поставленную задачу под углом зрения разработанной модели ассоциативного образа.

С этой точки зрения ситуация характеризуется следующей информацией. Накоплен специальный массив данных - результатов опросов по заданной анкете на представительной выборке обслуживаемой популяции. Для каждого человека, попавшего в выборку, квалифицированный врач-эксперт дал заключение о целесообразных дополнительных обследованиях этого человека у других специалистов-врачей.

Результаты анкетирования вновь опрошенного человека система рассматривает как запрос, для которого она строит ассоциативный образ в накопленном массиве. В итоге оказывается, что для данного проанкетированного лица в выборке целенаправленно будет отобрана специальная подвыборка аналогов (по системе жалоб).

Поскольку для каждого элемента этой выборки известны не только результаты анкетирования, но и экспертное заключение о целесообразных направлениях к врачам-специалистам, представляется рациональным использовать эту вторую информацию о множестве аналогов в качестве базы для автоматического формирования искомого решения.

На данном множестве аналогов, прежде всего, следует определить два списка врачей-специалистов. Первый включает специалистов, рекомендованных экспертами каждому из лиц из сформированного множества аналогов. Второй включает всех специалистов, которые встретились в рекомендациях хотя бы у одного лица, попавшего в множество аналогов. Первый список обозначим через $X^{\cap}$, а второй - через $X^{\cup}$. Ясно, что рекомендация - список X для данного человека, к результатам анкетирования которого подобрано множество аналогов, должен удовлетворять условию

$$X^{\cup} \supseteq X \supseteq X^{\cap} \qquad (75)$$

Вспомним, что речь идет, с одной стороны, о выборке рекомендательного списка врачей для данного конкретного опрошенного, а с другой стороны, - что опрашиваемый это представитель потока людей, которых обслуживает данная конкретная поликлиника со своей структурой интенсивностей обслуживания. Последняя, очевидно, характерна для относительно небольшого интервала времени. Поэтому возникает задача связать формирование рекомендаций на запрос с информацией о загрузках врачей на этот период времени.

Пусть $u(x,t)$ - среднее время ожидания к врачу x в момент времени t. На момент t_0 получения результатов анкетирования данным человеком имеем два ряда чисел:

$$U^{\cap} = \langle u(x_1, t_0), u(x_2, t_0), \ldots, u(x_{L_{\cap}}, t_0) \rangle,$$
$$U^{\cup} = \langle u(x'_1, t_0), u(x'_2, t_0), \ldots, u(x_{L_{\cup}}, t_0) \rangle,$$

которые соответственно определяют структуру занятости врачей из $X^{\cap}$ и из $X^{\cup}$ (первый ряд, очевидно, является частью второго).

Заранее фиксируется критическое время ожидания $T_{кр}$, превышение которого определяет, что человек не может быть обслужен врачом в день обследования, и поэтому ему следует рекомендовать посетить врача на следующий день.

Через $U^{\cap}(T_{кр})$ и $U^{\cup}(T_{кр})$ обозначим ряды чисел, которые получаются из построенных рядов $U^{\cap}$ и $U^{\cup}$ соответственно вычеркиванием элементов, которые превышают порог $T_{кр}$.

В дальнейшей работе сначала участвует список врачей $X^{\cap}(T_{кр})$, если ряд $U^{\cap}(T_{кр})$ оказался непустым. Если же этот ряд пуст, то список врачей $X^{\cup}(T_{кр})$, характеристики занятости которых представлены в $U^{\cup}(T_{кр})$. Если же и ряд $U^{\cup}(T_{кр})$ оказался пуст, то используется первоначальный список $X^{\cap}$ (когда $X^{\cap} = \phi$, берется весь $X^{\cup}$).

Один из этих списков в зависимости от указанных условий и в указанной последовательности предпочтений предъявляется проанкетированному человеку как результат анкетирования. Список сопровождается следующими пояснениями.

Из результатов анкетирования следует, что Вам желательно получить консультацию по Вашим жалобам у одного из врачей предлагаемого списка. Желательно, чтобы Вы получили ее сегодня (это пожелание указывается в случаях $X^{\cap}(T_{кр}) \neq \phi$ и $X^{\cup}(T_{кр}) \neq \phi$). Кроме того, укажите, к какому из них подойдете.

После посещения выбранного врача необходимо подтвердить системе факт его посещения. Тогда от системы Вы получите или новую рекомендацию, какого специалиста и когда в эти дни Вам следует посетить, или указание срока, через который требуется вновь посетить поликлинику для нового профилактического осмотра.

Предполагается, что посещение врачей заканчивается одним из трех видов указаний: указанием по текущей терапии, направлениями на инструментальные анализы, направлениями к другим врачам-специалистам. Указания, которые делает врач, он сообщает обследуемому и системе. Система далее ведет больного в соответствии с тактикой мероприятий, предложенных врачом. Она играет роль диспетчера, формируя расписание, определяющее когда и в какой последовательности следует выполнять предписания врача.

Если же врач ничего не рекомендовал (т.е. с его узкоспециальной точки зрения человек здоров), то система вновь на базе списков $X^{\cap}$, $X^{\cup}$ и числа $T_{кр}$ (это число могло сильно измениться) формирует скорректированный список врачей, рекомендованный ею для прохождения консультаций данным обследуемым.

§ 4.2. Характеристики экспериментальных данных и особенности программной реализации

Для первой серии экспериментов был использован материал исследования, проведенного лично автором настоящей диссертации.

В 1984г. был проведен опрос всех участников Конгресса ИФАК о сферах их интересов. Для этого был использован список ключевых слов журнала ИФАК "Автоматика". Результаты опроса были представлены в виде булевой матрицы объект-признак, где объ-

ектами были ключевые слова, а признаками служили участники Конгресса. Таким образом была построена матрица φ_{ip} $(i=1,2,\ldots,565, p=1,2,\ldots 1124)$, элементы которой принимали значение "I", если i-й термин попал в сферу интереса p-го участника Конгресса.

Вторая серия экспериментов базировалась на файле данных о распределении признаков, определяющих качество благоустройства школьных зданий Владимирской области РСФСР. Эти данные были любезно представлены автору отделом моделирования Главного вычислительного центра Министерства просвещения СССР, который заинтересовала наша разработка с точки зрения ее использования в расчетах, требующих получения обобщенных и оперативных оценок уровня социально-экономического развития школьной сети в разных регионах республики. Размер файла - более 800 записей, каждая длиной в IO признаков.

Наконец, для проведения третьей группы экспериментов, которые направлены на практическое внедрение, - экспериментальную отработку одной из моделей нового типа АСУ - АСУ поликлиники городского региона - было проведено специальное обследование. Это обследование было проведено лабораторией № 38 Института проблем управления АН СССР. Оно включало создание специальной компьютерной системы доврачебного осмотра. В ее основу была положена специальная модификация вопросника, разработанного в Корнельском университете, которая учитывала специфику советской популяции опрашиваемых. В разработке этой модификации принимал участие Институт психологии АПН СССР и методическая служба Министерства здравоохранения СССР.

На созданной системе было опрошено 59 человек, которые были здоровы на момент опроса. Результаты опроса были представлены терапевтам высокой квалификации, занимающимся специаль-

но исследованиями по вопросам организации диспансерного обслуживания.

В соответствии со схемой предстоящей обработки, которая была описана в предыдущем параграфе и разработана диссертантом, врачи-специалисты на основе информации о результатах опроса сформировали для каждого опрашиваемого рекомендации: к каким врачам ему следует обратиться (или отмечали, что на период обследования для данного человека консультация врачей не требуется).

Эти два типа данных - результаты опроса по вопроснику Корнельского университета и рекомендации врачей - служили исходной информацией для построения экспериментальных схем функционирования создаваемой АСУ.

Схемы обработки и их программная реализация, на основе которых проводилась стробация модели ассоциативного образа. В отличие от идеализированного описания модели, представленного в гл. 2 и 3 диссертации, где все конструкции строились на учете только наличия свойств (т.е. значений "I" соответствующих признаков) в обрабатываемых векторах, в практически реализованных схемах была предусмотрена возможность строить модель ассоциативного образа и на основе анализа отсутствия этих свойств (т.е. значений "0" соответствующих признаков).

Кроме того, в отличие от гл. 2 и 3, где все построения опираются на нахождение ядра монотонной системы с весовой функцией

$$\pi(i,H) = \alpha \cdot |Y_H| - \beta |y_i|,$$

в созданном программном комплексе можно использовать для тех же целей ядра другой монотонной системы с весовой функцией

$$\pi(i,H) = \beta |y_i| - \alpha |Y^H|,$$

где

$$Y^H = \bigcap_{j \in H} y_j$$

Основанием для такого решения послужили работы [42], которые показали, что совместное использование этих двух типов монотонных систем существенно упрощает анализ конечного результата, так как легко позволяет оценить результат на устойчивость (по сходству ответов, даваемых порознь этими двумя монотонными системами).

В реализованной схеме наряду с ассоциативными образами (их четыре с учетом, что отдельно ведется обработка на значениях "I" и на значениях "0" с использованием $\pi(i,H)$ и $\tilde{\pi}(i,H)$) строились и их внутренние и внешние структурные характеристики. Пусть $Q(e)$ - обозначение, общее для любого из указанных ассоциативных образов. На его основе строились следующие ряды образов:

I) $Q_{11}(e) = \bigcup_{i \in Q(e)} Q(i), \ldots, Q_{1k}(e) = \bigcup_{i \in Q_{1k-1}(e)} Q(i)$,

2) $Q_{21}(e) = \bigwedge_{i \in Q(e)} Q(i), \ldots, Q_{2k}(e) = \bigwedge_{i \in Q_{2k-1}(e)} Q(i)$)

3) $e_{11}(Q(e)) = \bigcup_{i \in Q(e)} y_i, Q(e_{11}), \ldots, e_{1k}(Q(e_{1k-1})) = \bigcup_{i \in Q(e_{1k-1})} y_i, Q(e_{1k})$,

4) $e_{21}(Q(e)) = \bigwedge_{i \in Q(e)} y_i, Q(e_{21}), \ldots, e_{2k}(Q(e_{2k-1})) = \bigwedge_{i \in Q(e_{2k-1})} y_i, Q(e_{2k})$,

5) $Q(e), F(Q), Q(F(Q)), \ldots$

6) $F(e), Q(F), F(Q(F)), \ldots$

Обратим внимание, что образы из рядов I, 5, 6 - это те, которые были введены в гл. 2, а образы из рядов 2,3, 4 - их модификации. Эксперименты показали, что при содержательном анализе удобно иметь разные варианты характеризации "места" ассоциативного образа в обрабатываемом информационном масси-

ве. Они же показали, что в указанных рядах особое значение имеют крайние образы, т.е. первые и последние, так как именно они наиболее контрастно демонстрируют возможности эволюции ассоциации на данном массиве.

Отметим еще, что построение всех экспериментов было организовано в виде серий многовариантных расчетов с последующим выбором "подходящих" вариантов на содержательном уровне. Этому способствовали не только уже указанные возможности двух монотонных систем, анализа и "0" и "I" значений в векторах, но и параметричность используемых весовых функций $\Pi(i,H)$ и $\tilde{\Pi}(i,H)$.

Однако, чтобы не усложнять анализ результатов в зависимости от параметров этих функций, в программе установлена связь между α и β : $\alpha = 1-\beta$. В итоге каждый эксперимент строится как набор из I0 расчетов со значениями $\alpha = 0,1; 0,2; \ldots; 0,9$

В проведенных экспериментах мы сознательно пошли на выдачу результатов в виде большого разнообразия "проекций" для анализа на качественном уровне, поскольку предлагаемая в диссертации модель является принципиально новой, относительно которой отсутствует какой-либо опыт использования.

Между тем следует подчеркнуть, что во всех описанных экспериментах имеются совершенно очевидные эмпирические приемы, позволяющие автоматически вычленять в создаваемом большом разнообразии проекций наиболее существенные. В частности, ясно, что в построенных рядах наиболее существенная информация заключается в различии первых и последних членов этих рядов. Ясно,также, что на базе описанных схем экспериментов можно конструировать новые процедуры решения известных задач: первая серия дает инструмент специалистам терминологии для усо-

вершенствования набора ключевых слов; вторая дает проверку рабочих гипотез об обеспеченности школьной сети (разведочного анализа по терминологии Тьюки), и, наконец, третья - демонстрирует возможности использования модели ассоциативного образа как блока в системе автоматического управления потока в организации массового многопрофильного обслуживания населения.

В заключение данного раздела отметим, что развитая в работе схема использования модели ассоциативного образа для решения задач кластерного анализа, изложенная в гл. 3 и в статье [104], была применена Л.С.Дульневым для создания методики совершенствования информационно-технологической структуры проектной организации, функционирующей в условиях внедрения САПР [105].

§ 4.3. Эксперименты

Первый эксперимент. В данном исследовании наряду с практической задачей классификации ключевых слов решался еще и методический вопрос об использовании свободного параметра α в формуле (65) для целей управления расслоением обрабатываемого информационного массива.

Как было сказано в § 4.2, наша задача на содержательном уровне описания заключалась в поиске такого разбиения набора ключевых слов журнала ИФАК "Автоматика" на классы, чтобы эти классы можно было упорядочить по уровню популярности их среди специалистов. Для проведения обработки использовались охарактеризованные в § 4.2 данные опроса участников XI конгресса ИФАК.

Простейший вариант решения этой задачи, очевидно, заключался в том, чтобы:

I) подсчитать для каждого ключевого слова число ученых, ко-

торые выбрали его как относящийся к сфере их интересов;

2) упорядочить ключевые слова по этой величине, и затем ось ее значений разбить на интервалы любым из известных алгоритмов кластеризации.

Однако такое решение не учитывает важное, явно не формулируемое требование, чтобы слова в каждом искомом классе должны быть согласованы по составу выбравших их ученых. В этой связи для решения поставленной задачи было решено использовать разработанную модель ассоциативного образа.

Была сконструирована специальная процедура, которая почти так же проста, как указанная выше с вычислительной точки зрения. Однако она, во-первых, выделяет классы похожих слов с учетом состава выбравших их ученых, и, во-вторых, число классов, на которое разбивается весь набор, определяется автоматически.

Процедура. Параметр α, который фигурирует в формуле пробегает все возможные значения с заданной точностью (в нашем случае это были 0.1, 0.2,..., 0.9). При каждом из этих значений находится ассоциативный образ запроса, состоящего из всех "1". Эти образы для разных соседних значений α могут совпадать, так что выделяются группы соседних значений α, на каждом из которых ассоциативные образы совпадают.

После этого интервал двух соседних значений, для которых получились разные ассоциативные образы, разбивается на ряд подинтервалов. Процедура построения ассоциативных образов повторяется для найденных более детальных значений α.

Затем вновь фиксируются все пары соседних значений α, дающие разные ассоциативные образы, и процедура дробления повторяется на подинтервале.

Такой процесс порождения все новых более детальных значений α, для которых строятся ассоциативные образы, заканчивается, когда наибольшая разница между значениями, дающими разные образы, оказывается во много раз (более 10) меньше наименьшего по длине интервала значений, соответствующего одному и тому же ассоциативному образу.

В результате получается некоторая структура разбиения оси α на неравномерные по длине интервалы, а каждому интервалу сопоставляется свой ассоциативный образ. Число интервалов и ассоциативных образов при этом оказывается сформированным автоматически.

Важно, что так сформированное семейство ассоциативных образов структурно организовано. Обозначим элементы этого семейства через $\bar{G}_1, \bar{G}_2, \ldots \bar{G}_L$, причем так, что соответствующие им коэффициенты упорядочены в ряд:

$$\alpha_{i_1} > \alpha_{i_2} > \ldots > \alpha_L ,$$

Тогда, очевидно, имеем цепь

$$\bar{G}_1 \subset \bar{G}_2 \subset \ldots \subset \bar{G}_L$$

из которой легко строится искомая классификация T:

$$T_1 = \bar{G}_1 , T_2 = \bar{G}_2 \setminus \bar{G}_1 , \ldots , \bar{G}_L \setminus \bar{G}_{L-1} , W \setminus \bar{G}_L = G_L$$

По смыслу выделенных $\bar{G}_i$ в построенной классификации, первый класс можно назвать классом самых популярных ключевых слов, а $(L+1)$-й класс - классом редко используемых ключевых слов. Все остальные классы содержат слова разного уровня популярности, лежащего между этими крайними классами.

Для практических целей часто удобно так построенную классификацию на $(L+1)$-й класс загрузить до классификации на три класса - популярные слова, промежуточные и редкие, объединяя

в них группы соседних классов.

Эта процедура была использована на описанном массиве. Результаты этого использования зафиксированы в Приложении I, содержащем перечень всех ключевых слов. Для наглядности наиболее популярные слова представлены в табл. 1.

Таблица показывает, что они никак не могут служить в качестве ключевых, так как являются характеристиками целых направлений исследований. Однако на их основе за счет детализации каждого такого слова можно сконструировать искомые слова. Но это уже работа специалистов по терминологии автоматики.

Второй эксперимент. Использовался массив из 846 I0-мерных булевых векторов, характеризующих качество всех школ Владимирской области на конец I986г. В табл. 2 дан перечень атрибутов, по которым оценивается качество. Из таблицы видно, что наряду с обычными характеристиками, отражающими разные точки зрения на развитие школ на современном этапе (наличие спортивного или актового залов), в перечне фигурируют характеристики, отражающие острые социально-экономические проблемы укрепления материально-технической базы школы (наличие канализации, водопровода и т.п.).

Указанные особенности перечня табл. 2 объясняют важность взятого для анализа банка данных для решения разных задач планирования управления и учета материального хозяйства школьной сети.

В эксперименте, поставленном на этом информационном массиве, мы руководствовались следующими дополнительными представлениями.

Анализируемая сеть школ является территориально распределенной. Одна ее часть располагается в сельской местности, другая - в городах районного и областного подчинения, третья - в

Таблица I

Самые популярные термины в списке ключевых слов журнала ИФАК "Автоматика"

I.	Адаптивное управление
2.	Автоматизация в авиации и в космонавтике
3.	Искусственный интеллект
4.	Химическое производство
5.	САПР
6.	Применение вычислительной техники
7.	Управление при помощи ЭВМ
8.	Математическое обеспечение
9.	ЭВМ в автоматизации
I0.	Анализ систем автоматизации
II.	Синтез систем автоматизации
I2.	Теория управления
I3.	Прямое цифровое управление
I4.	Обучение
I5.	Иерархические системы
I6.	Человек и автоматизация
I7.	Идентификация
I8.	Автоматизация в промышленности
I9.	Фильтр Кальмана
20.	Большие системы
2I.	Линейные системы
22.	Системы человек-машина
23.	Микропроцессоры
24.	Моделирование
25.	Системы с многими переменными
26.	Нелинейные системы управления
27.	Нелинейное программирование
28.	Оптимальное управление
29.	Оптимизация
30.	Оценка параметров
3I.	Автоматизация энергетических систем
32.	АСУТП
33.	Роботы
34.	Оценка состояния

областном центре, в городе Владимире.

В нашем банке из 846 школ было 563 сельских и 277 городских. Причем из последних 47 школ размещены в городе Владимире.

Естественно было предположить, что эти три разные ее части имеют разную как по уровню, так и по структуре обеспеченность материальной базы.

Таким образом, возникает задача выявить эти различия, и если они есть, то оценить их с точки зрения удовлетворительности современным требованиям, которые общество предъявляет к качеству школьного обслуживания. Дополнительная задача - определить очередность укрепления материальной базы по указанным трем территориально-административным частям школьной сети.

Постановку и решение поставленных двух задач мы моделировали на указанном банке данных в виде построения и анализа ассоциативных образов двух групп запросов.

Первая группа включала II запросов, из которых главный - это запрос, определяющий наличие всех IO характеристик качества табл. 2 (запрос № I в табл. 3), а остальные IO запросов - это варианты главного, отличающиеся от него отсутствием одного из этих качеств (запросы №№ 2-IO в табл. 3).

Цель построения соответствующих ассоциативных образов можно определить следующими вопросами:

I) сколько в абсолютном и в процентном отношении имеется хорошо обеспеченных школ в выделенных трех частях школьной сети;

2) насколько структуру различий по этим величинам (какова между сельскими школами, школами малых городов и школами областного центра) можно считать устойчивой.

Для ответа на первый вопрос использовался главный запрос. Его ассоциативных образ служил определением хороших школ. Он

Таблица 2

Характеристика качества школьных заданий

№№ п/п	Характеристика
I.	Здание типовое
2.	Электрическое освещение
3.	Центральное отопление
4.	Водопровод
5.	Горячая вода
6.	Канализация
7.	Спортивный зал
8.	Актовый зал
9.	Библиотека
IO.	Столовая

Таблица 3

№ запроса	Запросы: Номера характеристик из табл. 2										P_i^G: Школы			$\frac{P_i^G}{P_i^W} \cdot 100$: Школы		
	I	2	3	4	5	6	7	8	9	I0	Сельской местности	Малых городов	Владимира	Сельской местности	Малых городов	Владимира
I	I	I	I	I	I	I	I	I	I	I	II4	II9	47	20	52	I00
2		I	I	I	I	I	I	I	I	I	II5	92	33	20	40	70
3	I		I	I	I	I	I	I	I	I	I38	95	33	24	4I	70
4	I	I		I	I	I	I	I	I	I	II5	92	33	20	40	70
5	I	I	I		I	I	I	I	I	I	9I	94	33	I6	4I	70
6	I	I	I	I		I	I	I	I	I	9I	92	33	I6	40	70
7	I	I	I	I	I		I	I	I	I	I08	I25	35	I9	54	74
8	I	I	I	I	I	I		I	I	I	96	93	33	I7	40	70
9	I	I	I	I	I	I	I		I	I	I38	94	33	24	4I	70
I0	I	I	I	I	I	I	I	I		I	I54	I23	35	27	53	74
II	I	I	I	I	I	I	I	I	I		I4I	94	33	25	4I	70

разделялся далее на три части по принадлежности его элементов (школ) соответственно к группе сельских, малых городов и областного центра школ.

Для ответа на второй вопрос использовались 10 модификаций главного запроса, ассоциативные образы которых обрабатывались аналогичным образом. Обработка производилась для того, чтобы оценить, насколько устойчивой сохраняется картина различий между разными типами школ в этих ассоциативных образах по сравнению со структурой школ ассоциативного образа главного запроса.

Ответы на оба вопроса дает табл. 3 , в которой столбцы группы e_i соответствуют разным качествам по перечню табл. 2 . Первая строка табл. 3 соответствует главному запросу, а остальные 10 строк соответствующим вариантам главного запроса.

Три столбца группы P_i^G табл. 3 дают структуры распределения ассоциативных образов в абсолютных значениях, а три столбца группы $\frac{P_i^G}{P_i^W} \cdot 100$ - те же структуры, но в процентах хорошо видно, что:

1) имеется резкий контраст в качестве обеспеченности школ в выделенных трех типах (более чем в 1,5 раза лучше обеспечены школы города Владимира, по сравнению со школами малых городов, которые в свою очередь более чем в 2 раза качественнее обеспечены, чем сельские школы);

2) диспропорции обеспеченности, выявленные по главному запросу, устойчивы. Тот факт, что при более дифференцированном рассмотрении (при сокращении набора анализируемых качеств) уровень обеспеченности школ города Владимира как бы сокращается (что не очевидно), объясняется тем, что ассоциативный образ строится над всем банком данных: более частное рассмотрение де-

лает контраст школ города Владимира по сравнению с сельскими школами и школами малых городов менее выраженным по сравнению с рассмотрением полного перечня качеств.

Первая группа экспериментов проводилась с функцией $\pi(i,H)$, вычисляемой по формуле (50) при $\alpha = 0.3$.

Вторая группа расчетов строилась для выявления остроты нуждаемости укрепления материальной базы школ, т.е. в определенном смысле для целей, противоположных целям первой группы.

Во второй группе мы использовали два запроса:

I) выделить ассоциативный образ гипотетической школы, в которой нет канализации, водопровода, центрального отопления и электричества;

2) выделить ассоциативный образ гипотетической школы еще большего неблагополучия, у которой дополнительно к тому, что отсутствует в школе первого запроса, еще отсутствует горячая вода и столовая.

Второй запрос в этих расчетах по отношению к первому играл ту же роль, что и модификации к главному запросу в первой группе расчетов, т.е. служил для оценки устойчивости получаемого результата. Кроме того, во второй группе экспериментов для оценки устойчивости выявляемой структуры нуждаемости использовался и другой прием - расчеты проводились при двух существенно разных значениях α (0.3 и 0.7).

Результаты расчетов второй группы представлены в табл. 4 Таблица показывает, что контраст в распределении нуждаемости по указанным жизненно важным характеристикам между выделенными тремя типами школ заметно выше, чем по общему уровню обеспеченности. Табл. 4 , как нам представляется, подсказывает следующую схему распределения материальных ресурсов:

Таблица 4

α	№ запроса	Запросы: Номера характеристик из табл. 2										P_i^G: Школы			$\frac{P_i^G}{P_i^W} \cdot 100$: Школы		
		I	2	3	4	5	6	7	8	9	I0	Сельской местности	Малых городов	Владимира	Сельской местности	Малых городов	Владимира
0,3	I	I				I		I	I	I	I	468	94	I0	83	34	24
	2	I						I	I	I		4I7	74	2	74	27	4
0,7	3	I				I		I	I	I	I	324	36	0	48	I3	0
	4	I						I	I	I		252	26	0	45	9	0

наиболее целесообразно (а может быть даже необходимо) дать приоритет в распределении ресурсов для сельских школ, причем главным образом придания этим школам жизненно важных качеств 2-4 и 6 по нумерации табл. 2 . Очевидно, что эта подсказка опирается на представление, что размещение школьной сети не будет подвергнуто резкому изменению, так как ясно, что если стратегия модернизации будет опираться на свертывание сети в сельской местности, то указанная рекомендация неадекватна. И тем не менее даже в последнем случае наш расчет оказывается полезным. Он указывает, что если свертывать в сельской местности школы, то это, конечно, прежде всего те, которые не обеспечены качествами 2-4 и 6 из перечня табл. 2.

В заключение сделаем важное техническое замечание. В обрабатываемом массиве представлены данные не разных школ, а разных школьных зданий, так что около 5% школ в нем представлены несколькими векторами. Мы не располагали информацией о принадлежности зданий школам. Поэтому приведенный анализ численно не совсем точен. Однако на качественном уровне, имея в виду, что школ с несколькими зданиями менее 5%, и то, что наши результаты устойчивы, полученные выводы представляются корректными.

С методической точки зрения указанные технические неточности, очевидно, не могут повлиять на выявленную в этом эксперименте возможность использовать модель ассоциативного образа для поддержки решений в АСУ с большими банками информации.

Третий эксперимент. Схема использования модели ассоциативного образа в АСУ поликлиники, как она описана в § 4.2, является сугубо предположительной уже потому, что в настоящее время не существует общепринятых требований к разработке таких АСУ и отсутствует опыт эксплуатации таких систем.

Вместе с тем, в этой схеме выделены некоторые элементы, которые имеют весьма веское основание найти широкое практическое внедрение в АСУ поликлиники.

Важнейший из них - это, конечно, использование безврачебного вопросника. И вот в деле совершенствования использования такого вопросника в § 4.2 предложено задействовать нашу модель ассоциативного образа.

Главное назначение модели - совершенствование правила выделения группы специалистов-врачей, к которым целесообразно направить ответившего на вопросы.

В § 4.2 обсуждалось построение правила выделения группы специалистов, основанное на удовлетворении двух противоречивых требований - (I) адекватности выделяемой группы состоянию ответившего на вопросы и (2) соответствия ее структуре занятости работающих на данный момент врачей.

На современном этапе разработки АСУ поликлиники экспериментально исследовать в полном объеме правильность такого построения не представляется возможным.

Главное, как мы полагаем, состоит в том, чтобы экспериментально оценить принципиальную целесообразность использования нашей модели. И в этом плане следует отказаться от учета занятости работающих врачей и поставить эксперимент, определяющий эффективность создаваемого правила выделения необходимой группы специалистов. Далее описываются результаты такого эксперимента.

По вопроснику из Приложения 2 было опрошено 59 человек, из которых было 39 женщин и 20 мужчин. Средний возраст опрошенных - 32,4 года.

Каждый из опрошенных, кроме того, был осмотрен терапевтом, по заключению которого он получил рекомендацию, кого из специ-

алистов ему следует посетить. Всего в рекомендациях терапевта было использовано шесть разных специалистов. В табл. 5. указана статистика частоты данных врачом-терапевтом рекомендаций по врачам-специалистам.

В соответствии с § 4.2 в качестве модели правила выделения необходимых для посещения специалистов использовалась следующая процедура.

Для данного человека, который ответил на вопросы вопросника (это может быть любой человек - как один из 59 обследованных, так и любой иной), его ответы фиксируются в виде булевого вектора. Каждая компонента этого вектора соответствует одному вопросу вопросника, а ее значение "I" означает, что ответивший согласен с наличием соответствующего симптома у него ("О", наоборот, означает, что ответивший не находит этого симптома у себя).

Этот вектор мы рассматриваем как запрос, а в качестве информационного массива, из которого выделяется ассоциативный образ, рассматривается указанный выше массив из 59 аналогичных булевых векторов, фиксирующих ответы указанной группы опрошенных лиц.

В результате выделения ассоциативного образа с данным человеком, который только ответил на вопросы вопросника, мы связываем группу моделей, для которых, с одной стороны, известны рекомендации терапевта (к каким специалистам-врачам следует обратиться), а с другой стороны, известно, что они отвечали на вопросы того же вопросника похожим образом. Последнее обстоятельство дает основание формируемую группу специалистов-врачей, рекомендуемую посетить данному человеку, определять из врачей, которые были рекомендованы представителям выделенного ассоциативного образа этого человека. В соответствии с

Таблица 5

Распределение рекомендаций по врачам-специалистам

№№ п/п	Наименование специалиста	Число рекомендаций
I	Терапевт	2
2	Кардиолог	6
3	Гастроэнтеролог	3
4	Эндокринолог	I
5	Дерматолог-венеролог	I0
6	Стоматолог	43

§ 4.2 мы фиксируем два варианта правил (омирования такой группы:

а) в качестве рекомендуемой к посещению данным человеком группы врачей-специалистов взять всех врачей, которые были рекомендованы терапевтом хотя бы одному из представителей ассоциативного образа;

б) взять в искомую группу только тех врачей, которых терапевт рекомендовал всем представителям ассоциативного образа.

Забегая вперед, отметим, что второй из указанных вариантов имеет устойчивую, резко преувеличенную склонность признавать ответивших на вопросы практически здоровыми, т.е. отказывать им в рекомендациях посетить врачей-специалистов. Причина этого эффекта, как будет показано, состоит в том, что рекомендации врача-терапевта сугубо индивидуальны, и на относительно больших по мощности группах людей ассоциативного образа по каждому рекомендованному врачу имеются различия в рекомендациях.

В этой связи главное внимание следует уделить именно первому варианту.

Ясно, что если проводимый эксперимент по оценке правила выделения группы врачей-специалистов, рекомендованных данному человеку, проводится на людях, для которых известны рекомендации врача-терапевта, то желательно, чтобы правило дало рекомендации, близкие к рекомендациям терапевта. При этом представление о близости имеет два свойства: прежде всего, важно не пропустить в формальном правиле рекомендаций тех врачей, которые рекомендованы терапевтом; важно и другое, что рекомендация не была тривиальной, включающей всех доступных специалистов.

В проведенном эксперименте в качестве "испытуемых" фигурировали те же 59 ответивших на вопросы, которые составляют и базовый информационный массив. В табл. 6 для каждого из них указаны две группы врачей-специалистов: указанная врачом-терапевтом и указанная описанным правилом (первый вариант) на базе ассоциативного образа. Табл. 6 показывает, что ни в одном случае нет пропуска рекомендаций, данных врачом-терапевтом. Однако некоторая гипертрофированная склонность давать ответ "может быть болен всеми болезнями" видна, как говорится, невооруженным глазом. Вместе с тем, правило, результаты которого представлены в табл. 6 , нельзя считать тривиальным: вместо 59x6 = 354 рекомендаций, которые дало бы тривиальное правило, имеем 139 рекомендаций, т.е. на 61 % меньше.

Заметим, что табл. 6 получена при использовании монотонной системы, функция $\Pi(i,H)$ которой вычисляется по формуле (41) с $\alpha = 0{,}9$. Варианты этой таблицы для случаев $\alpha = 0{,}7; 0{,}5$ представлены соответственно в табл. 7 и 8 .

Отметим, что второй вариант правила (рекомендации, которые имеют все представители ассоциативного образа) в 58 из 59 случаев дал ответ "практически здоров". Чтобы наглядно увидеть, что фактически отсутствует возможность выбора правила, которое было "промежуточным" между "все" и "ничего", как можно соответственно назвать первый и второй вариант, в табл. 9 представлены частоты встречаемости рекомендаций по специалистам по всем построенным ассоциативным образам (при $\alpha = 0{,}9$).

Относительно небольшой уровень эффективности полученного правила, по-видимому, не является индивидуальной характеристикой нашего эксперимента, а есть принципиальное ограничение, связанное с использованием безврачебного вопросника.

Таблица 6

Рекомендации врача и машинного правила

№ лица	Номер специалиста из табл. 5											
	Рекомендации врача						Рекомендации машинного правила					
	I	2	3	4	5	6	I	2	3	4	5	6
I	2	3	4	5	6	7	8	9	I0	II	I2	I3
I					I	I		I	I		I	I
2			I			I			I			I
3		I						I	I			I
4						I		I	I		I	I
5						I				I		I
6	I				I	I	I				I	I
7						I		I			I	I
8					I						I	I
9						I					I	I
I0		I				I		I				I
II						I		I				I
I2												I
I3			I			I			I		I	I
I4					I						I	I
I5						I						I
I6						I						I
I7						I	I	I	I		I	I
I8						I					I	I
I9	I	I				I	I	I			I	I
20						I					I	I
2I										I		I
22						I						I
23						I		I			I	I
24						I					I	I
25						I						I
26						I		I			I	I
27												I
28												I
29				I		I		I		I		I

Продолжение табл. 6

I	2	3	4	5	6	7	8	9	I0	II	I2	I3
30						I					I	I
3I					I	I					I	I
32					I	I					I	I
33												
34		I	I			I		I	I		I	I
35						I					I	I
36		I			I	I	I	I			I	I
37						I					I	I
38												I
39								I				I
40												I
4I					I		I				I	I
42						I		I			I	I
43						I	I	I			I	I
44		I				I		I			I	I
45					I	I		I		I	I	I
46					I	I		I	I		I	I
47						I	I	I	I		I	I
48						I					I	I
49						I					I	I
50							I				I	I
5I						I						I
52										I		I
53						I		I		I	I	I
54						I						I
55											I	I
56						I					I	I
57						I				I		I
58												I
59						I	I					I

Таблица 7

Рекомендации врача и машинного правила

№ лица	Номер специалиста из табл. 5											
	Рекомендации врача						Рекомендации машинного правила					
	I	2	3	4	5	6	I	2	3	4	5	6
I	2	3	4	5	6	7	8	9	I0	II	I2	I3
I					I	I		I	I	I	I	I
2			I			I	I	I	I	I	I	I
3		I						I	I	I		I
4						I		I	I		I	I
5						I	I	I		I		I
6	I				I	I	I	I		I	I	I
7						I		I	I		I	I
8					I			I		I	I	I
9						I		I	I	I	I	I
I0		I				I	I	I	I	I	I	I
II						I		I	I		I	I
I2								I			I	I
I3			I			I	I	I	I	I	I	I
I4					I			I			I	I
I5						I		I	I		I	I
I6						I			I		I	I
I7						I	I	I	I		I	I
I8						I		I	I		I	I
I9	I	I				I	I	I	I	I	I	I
20						I		I			I	I
2I								I		I		I
22						I	I	I	I			I
23						I		I	I		I	I
24						I	I	I	I	I	I	I
25						I	I	I	I			I
26						I		I			I	I
27							I	I	I	I		I
28								I				I
29					I		I	I	I	I		I

Продолжение табл. 7

I	2	3	4	5	6	7	8	9	I0	II	I2	I3
30						I		I			I	I
3I					I	I	I	I	I		I	I
32					I	I		I	I	I	I	I
33									I		I	I
34		I	I			I	I	I	I	I	I	I
35						I	I	I			I	I
36		I			I	I	I	I	I	I	I	I
37						I	I	I	I	I	I	I
38								I	I		I	I
39								I				I
40							I		I	I	I	I
4I					I		I	I	I		I	I
42						I		I	I	I	I	I
43						I	I	I	I		I	I
44		I				I	I	I	I	I	I	I
45					I	I	I	I	I	I	I	I
46					I	I	I	I	I		I	I
47						I	I	I	I	I	I	I
48						I	I	I	I		I	I
49						I	I	I	I	I	I	I
50							I	I			I	I
5I						I		I	I		I	I
52									I	I		I
53						I	I	I	I	I	I	I
54						I		I	I		I	I
55							I	I	I		I	I
56						I	I	I	I		I	I
57						I	I	I	I	I	I	I
58								I		I	I	I
59						I	I	I	I	I	I	I

Таблица 8

Рекомендации врача и машинного правила

№ лица	Номер специалиста из табл. 5											
	Рекомендации врача						Рекомендации машинного правила					
	I	2	3	4	5	6	I	2	3	4	5	6
I	2	3	4	5	6	7	8	9	I0	II	I2	I3
I					I	I	I	I	I	I	I	I
2			I			I	I	I	I	I	I	I
3		I					I	I	I	I	I	I
4						I	I	I	I	I	I	I
5						I	I	I	I	I	I	I
6	I				I	I	I	I	I	I	I	I
7						I	I	I	I	I	I	I
8					I		I	I	I	I	I	I
9						I	I	I	I	I	I	I
I0		I				I	I	I	I	I	I	I
II						I	I	I	I	I	I	I
I2								I	I	I	I	I
I3			I			I	I	I	I	I	I	I
I4					I			I	I	I	I	I
I5						I	I	I	I	I	I	I
I6						I		I	I	I	I	I
I7						I	I	I	I	I	I	I
I8						I	I	I	I	I	I	I
I9	I	I				I	I	I	I	I	I	I
20						I	I	I	I	I	I	I
2I								I	I	I	I	I
22						I	I	I	I	I	I	I
23						I	I	I	I	I	I	I
24						I	I	I	I	I	I	I
25						I	I	I	I	I	I	I
26						I		I	I	I	I	I
27							I	I	I	I	I	I
28								I	I	I	I	I
29					I		I	I	I	I	I	I

Продолжение табл. 8

I	2	3	4	5	6	7	8	9	I0	II	I2	I3
30						I	I	I	I	I	I	I
3I					I	I	I	I	I	I	I	I
32					I	I	I	I	I	I	I	I
33							I	I	I		I	I
34		I	I			I	I	I	I	I	I	I
35						I	I	I	I	I	I	I
36		I			I	I	I	I	I	I	I	I
37						I	I	I	I	I	I	I
38							I	I	I	I	I	I
39								I		I	I	I
40							I	I	I	I	I	I
4I					I		I	I	I	I	I	I
42						I	I	I	I	I	I	I
43						I	I	I	I	I	I	I
44		I				I	I	I	I	I	I	I
45					I	I	I	I	I	I	I	I
46					I	I	I	I	I	I	I	I
47						I	I	I	I	I	I	I
48						I	I	I	I	I	I	I
49						I	I	I	I	I	I	I
50							I	I	I	I	I	I
5I						I	I	I	I	I	I	I
52								I	I	I	I	I
53						I	I	I	I	I	I	I
54						I		I	I	I	I	I
55							I	I	I	I	I	I
56						I	I	I	I	I	I	I
57						I	I	I	I	I	I	I
58							I	I	I	I	I	I
59						I	I	I	I	I	I	I

Таблица 9

Рекомендации машинного правила с учетом частности

№ лица	Номера специалистов по табл. 5					
	I	2	3	4	5	6
I	2	3	4	5	6	7
I		I	I		I	4
2			I			I
3		I	2			4
4		I	2		I	6
5				I		2
6	I				I	3
7		I			2	4
8					I	I
9					I	I
I0		2				3
II		I				2
I2						2
I3			I		2	3
I4					2	5
I5						3
I6					I	4
I7	I	2	2		4	8
I8					2	7
I9	I	I			3	4
20					I	I
2I				I		2
22						7
23		I			2	4
24					3	3
25						4
26		I			I	2
27						2
28						3
29		I		I		2
30					I	I
3I					I	3

Продолжение табл. 9

I	2	3	4	5	6	7
32					I	I
33						
34		I	I		2	5
35					2	4
36	I	2			3	6
37					3	5
38						5
39		I				7
40						2
4I	I				2	6
42		I			I	3
43	I	I			3	5
44		I			2	4
45		I		I	3	9
46		2	I		2	5
47	I	I	I		2	4
48					2	I
49					I	6
50	I				2	4
5I						5
52				I		8
53		I		I	2	6
54						4
55					2	2
56					I	5
57				I		9
58						5
59	I					5

Действительно, обратим внимание, что в эксперименте мы отступили от правила § 4.2, что врач строит рекомендации исключительно тоже по ответам человека на вопросы вопросника. В нашем случае врач-терапевт, дававший рекомендации, имел полную свободу применять свои знания при составлении и оценке состояния осматриваемого больного. Его рекомендации, очевидно, более точны, по сравнению с теми, которые он же мог бы дать руководствуясь только ответами на вопросы вопросника. Тот факт, что несмотря на это существенное ужесточение условий эксперимента по проверке правила назначения рекомендаций, правило получилось нетривиальным, говорит и о том, что целесообразно использовать в нем нашу модель ассоциативного образа, и о том, что примененный вопросник позволяет выявлять состояние человека без существенных искажений.

В заключение сделаем два замечания.

I. Предложенный способ построения рекомендаций существенно зависит от того, кем сформулированы рекомендации для обследуемых, данные о которых взяты в качестве информационного массива, - какова квалификация этого врача, насколько хорошо он знает обследуемый контингент и т.п. Формируемое правило - это модель именно этого специалиста.

2. Формируемое правило в равной или в большей мере наследует особенности выбранного контингента обследуемых - статистика симптомов на них и ее соответствие статистике тех, для которых рекомендации будут строиться по правилу, могут решающим образом влиять на эффективность правила.

Эти два вопроса, связанные с формированием базового информационного массива (выбор контингента обследованных и выбор врача, который проводит обследование), - важная предметная работа, которая выходит, очевидно, за рамки данной диссертации.

ЗАКЛЮЧЕНИЕ

По диссертационной работе можно сделать следующие выводы.

I. Предложена схема преобразования, которая вектора исходных признаков информационного массива заменяет на векторные оценки сходства соответствующих объектов с запросом.

Схема позволяет легко варьировать типом преобразования, что конкретно продемонстрировано для булевых данных.

2. Выделен специальный двухпараметрический класс монотонных систем, позволяющий использовать наиболее экономичный в вычислительном отношении алгоритм построения ассоциативного образа запроса; этот класс монотонных систем, кроме того, дает ассоциативные образы запросов наиболее легкой интерпретации.

3. Создан алгоритм, позволяющий для выделенного ассоциативного образа определить в множестве обрабатываемых признаков подмножество наиболее информативных в заранее заданном смысле.

4. Построено формальное описание внешней и внутренней структур ассоциативного образа. Сконструированы процедуры выявления таких структур для заданного конкретного ассоциативного образа. Выявление такой структуры позволяет пользователю манипулировать предлагаемым многовариантным решением с целью его лучшего приспособления поставленной практической задаче.

5. Для данных, характеризуемых только качественными признаками взаимного сходства, построен новый алгоритм кластеризации, дающий в качестве решения два набора кластеров: набор непересекающейся классификации и одновременно классификацию

с пересекающимися классами. Число классов в выстраиваемых классификациях определяется автоматически.

6. Модель ассоциативного образа применена для решения трех прикладных задач (совершенствования словаря ключевых слов для журнала ИФАК "Автоматика"; анализа и распределения материально-технических ресурсов по школьной сети крупного региона; разработки правил принятия решений для АСУ поликлиники при доврачебном исследовании).

Литература

I. Хаббард Дж. Автоматизированное проектирование баз данных. - М.: Мир, 1984. - 292 с.

2. Забежайко М.И. и др. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез. - НТИ, серия 2, № 10, 1987. - С. 1-14.

3. Попов С.В., Шохин В.А. Нетрадиционные методы автоматического документального поиска. - НТИ, серия 2, № 12, 1987. - С. 16-18.

4. Розенман М.И. Тенденции развития больших диалоговых ИПС. - НТИ, серия 2, № 6, 1987. - С. 6-9.

5. Olmstead M., Labreche S. DIALOG version 2, Questel. Plus: a comparison, Online. - V. 10, N 5, 1986. - P. 31-35.

6. Hawkins D.T., Levy R. Front end software for online database searching. - Online, V. 9, N 9, 1985. - P. 25-30.

7. Поспелов Г.С., Разин А.И. Основные тенденции развития современных экспертных систем. - НТИ, серия 2, № 2, 1987.

8. Attias R. DARS Substructure Search System. - J. Chem. Inf. Comput. Sci. - V. 23, N 3, 1983. - P. 102-108.

9. Expert Systems in Goverment Symposium (ed. Karna K.N.). - Washington: IEEE Comp. Soc. Press, 1986. - 466 p.

10. Weiss S.U. and Kulikowski C. Expert - a system for developing consultation models. - Proc. 6th IJCAI, 1979. - 490p.

11. Эпштейн В.Л., Сеничкин В.И. Языковые средства архитектора АСУ. - М.: Энергоиздат, 1982. - 199 с.

12. Диурс П., Айзенаэр Т. Распознавание образов химии. - М.: Мир, 1977. - 230 с.

13. Нахмансон М.С. Информационный поиск в базах спектральных данных и метод изучения особенностей их структуры. - М.:

I4. Васильев Е.К., Нахмансон М.С. Качественный рентгенофазовый анализ. - Новосибирск: Наука, I986. - I95 с.

I5. Рентгенофазовый качественный анализ материалов с помощью ЭВМ и система АРФА. - Заводская лаборатория. - Т. I285, № 5. - С. 23-29.

16. Scriabin M. A cluster-analytic approach to facility layout. - Manag. Sci., V. 31, N 1, 1985. - P. 33-49.

17. Hiroshi Tsumura, Tetsuo Wasano, Hideaki Masuzawa, Hiroshi Miyake. Patient-Oriented Multidisciplinary Medical Consultation System. - Medinfo 86, R. Salamon, B.Blum, M.Jorgensen (Editors), Elsevier-Science Publishers B.V. (North Holland), IFIP-IMIA, 1986. - P. 289-293.

18. Miller R., Pople H. and Myers J. INTERNIST - A computer-based diagnostic consultant for general internal medicine. - The New England Journal of Medicine, 1982. - P. 8-15.

I9. Распознавание образов и медицинская диагностика. (Под ред. Нейморка Ю.И.). М.: Наука, I972. - 328 с.

20. Казакевич Д.И. Основы теории случайных функций и ее применение в гидрометеорологии. Л.: Гидрометеоиздат, I97I. - 267 с.

2I. Верхатен и др. Распознавание образов: состояние и перспективы. - М.: Радио и связь, I985. - I03 с.

22. Аветисян Д.А. и др. Автоматизация проектирования строительных и технических объектов. - М.: Наука, I986. - I35 с.

23. Волков Б.А. Оценка качества проектных решений при вариантном проектировании. - Проектирование и инженерные изыскания. - № 5, I983. - С. I5-I7.

24. Кисель И.М. Из опыта разработки и эксплуатации САПР-ХИМ в Минхимпроме и в Минудобрений СССР. - Проектирование и инженерные изыскания. - № 6, I984. - С. 22-24.

25. Коренков И.П. САПР: принципы построения и структура. М.: Высшая школа, 1986. - 127 с.

26. Дульнев Л.С. Методы исследования и разработки организационно-технических структур САПР в проектных организациях строительного профиля. (Дис. на соиск. учен. степ. по спец. 05.13.12). - М.: МПУ, 1987. - 252 с.

27. **Burgin R. Microcomputer software use in libraries. - Libr. Software Rev., V. 5, N 6, 1986. - P. 332-336.**

28. Раушенбах Г.В. Меры близости и сходства в социологии. - В кн.: Анализ нечисловой информации в социологических исследованиях. (Отв. ред. Андреенков В.Г., Орлов А.И., Толстова Ю.Н.). - М.: Наука, 1985. - С. 169-203.

29. Карп В.П., Кунин П.Е. Метод направленного обучения в переборной схеме М.М.Бонгарда и онкологическая диагностика. - В кн.: Моделирование обучения и поведения. - М.: Наука, 1975, С. 7-17.

30. Мелс Т.Э. Описания класса общих корреляционных коэффициентов для случайных элементов на конечном множестве. - Тр. ВЦ Тартусского университета. Вып. 26, 1972. - С. 3-18.

31. Амиров И.Ш. Неединственность оценки коэффициентов взаимозависимости. - Мат. методы в экономике и международные отношения. Вып. 3. - М.; 1974. - С. 17-31.

32. Дорофеюк А.А. Алгоритмы автоматической классификации (обзор). - Автоматика и телемеханика, № 12, 1971. - С.22-31.

33. **McCormick R. A review of classification. - J.Royal Statistical Soc., 1971. - P.134.**

34. **Hartigan J.A. Clustering Algorithms. - New York: Wiley, 1975. - 402 p.**

35. Канал Л. Обзор систем для анализа структуры образов и разработки алгоритмов классификации в режиме диалога. - В кн.: Распознавание образов при помощи цифровых вычислительных машин. - М.: Мир, 1974. - С. 124-143.

36. Браверман Э.М., Мучник И.Б. Структурные методы обработки эмпирических данных. - М.: Наука, 1983. - 464 с.

37. Sanchez-Mazas A., Excoffier L., Langaney A. Measure and representation of the genetic similarity between populations by the percentage of isoactive genes. - Theoria - Segunda Epoca - Ano II - Curso, N 4, 1986-1987. - P.143-154.

38. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. - М.: Статистика, 1974. - 240 с.

39. Darc workshop manual. - S.I.: Telesystemes Questel, 1985. - 246 p.

40. Выханду П.Л. Средства для создания оптимальных запросно-ориентированных баз данных. - Тез. сем. "За ускорение научно-технического прогресса", Таллин, 1986. - С. 17-18.

41. Мучник И.Б., Шварцер Л.В. Субмодулярные функции и монотонные системы в задачах агрегирования. - Автоматика и телемеханика., №5 и № 6, 1987. - С. 135-148, 138-147.

42. Кузнецов Е.Н. Анализ структуры матрицы связей с помощью построения на ней монотонной системы. - А и Т, № 7, 1980 - с. 128-136.

43. Мучник И.Б., Чкуасели Н.Ф., Шварцер Л.В. Лингвистический анализ булевых матриц с помощью монотонных систем. А и Т, № 4, 1986. - С. 113-124.

44. Выханду Л.К. О некоторых методах упорядочения объектов и признаков в системах данных. - Тр. Таллинского политехни-

ческого ин-та, Таллин, № 428, 1980. - С. 28-33.

45. Выханду Л.К., Выханду П.Л. Быстрый поиск на битматрицах.- Тр. Таллинского политехнического ин-та, Таллин, № 554, 1983. - С. 49-60.

46. Кузнецов Е.Н., Мучник И.Б., Шварцер Л.В. Локальные преобразования монотонных систем. - А и Т, № 12, 1985. - С. 85-95; № 1, 1986. - С. 116-125.

47. Муллат И.Э. Экстремальные подсистемы монотонных систем. - А и Т, № 5, 1976. - С. 130-139; № 8, 1976. - С. 169-178; № 1, 1977. - С. 143-152.

48. Кузнецов Е.Н., Мучник И.Б., Шварцер Л.В. Монотонные системы и их свойства. - В кн.: Анализ нечисловой информации в социологических исследованиях. (Отв. ред. А.И.Андреенков, А.И.Орлов, Ю.Н.Толстова). - М.: Наука, 1985. - 220 с.

49. Кузнецов Е.Н. и др. Монотонные системы на матрицах данных.- MTA SZTAKI Kozlemenyck Budapest, 31/1984. - P. 153-158.

50. Hencsey G. Selection of similar elements from information-description sets. - E, N 52, 1988.

51. Мучник И.Б. Анализ структуры экспериментальных графов. - А и Т, № 9, 1974. - С. 62-80.

52. De Soete G., Desorbo W.S., Carroll J. Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. - J. of Classification, V. 2, N 2/3, 1985.

53. Art D., Gnanadesikan R., Kettenring J.R. Data-based metrics for cluster analysis. - Utilitas Mathematica, 21A, 1982. - P. 75-99.

54. Cunningham J.P. Free Trees and Bidirectional Trees as a Representation of Psychological Distance. - Journal of Mathematical Psychology, 17, 1978. - P. 165-188.

55. DeSarbo W.S., Rao V.R. Constrained Classification: the Use of a priori Information in Cluster Analysis. - Psychometrika, 49, 1984. - P. 187-217.

56. Jamby M., Lebeaux M.O. Cluster Analysis and Data Analysis. - Amsterdam: North Holland, 1983. - 769 p.

57. Maronn R., Jacovkis P.M. Multi-variate clustering procedure with variable metrics. - Biometrics, 30, 1974. - P. 499-505.

58. Васильев В.И. Распознающие системы: Справочник. - Киев: Наукова Думка, 1983. - 419 с.

59. Dubes R., Jain A.K. Models and methods in cluster validity. - Proc. IEEE Comput. Soc. Conf. Pattern Recognition and Image Process, Chicago, 1978. - P. 148-155.

60. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания и классификации. - Проблемы кибернетики, М.: Наука, вып. 33, 1978. - С. 5-68.

61. Погорельский Ю.Э. Выполнение тематических заданий информационными и изобретательскими службами предприятий. - НТИ, серия 1, № 8, 1987. - С. 18-20.

62. Погорелко К.П. Использование расширенного набора поисковых функций в документальной ИПС по онкологии. - НТИ, серия 2, № 4, 1987. - С. 18-21.

63. Аветисян Д.О. Проблемы информационного поиска. - М.: Финансы и статистика, 1981. - 295 с.

64. Salton G., Fox E.A., Wu H. Extended boolean information retrieval. - Communications of the ACM, V. 26, N 12, 1983. - P. 1022-1036.

65. Fuju S., Tsukamoto Y., Fujii M., Kaneda Y., Matsuo M., Yamasaki R. Three Dimensionalization of Cerebral Arteries from Cineangiograms. - Medinfo 86, R.Salamon, B.Blum, M.Jorgensen (Editors), Elsevier-Science Publishers B.V. (North-

Holland), IFIP-IMIA, 1986.

66. Development of a Computer-Aided Instruction System for Check and Prevention of Cardio-vascular Disorders (CAUTION). - P. 285-288.

67. Michio Kimura, Kihachiro Shimizu, Fumito Tsuchiya, Teruo Koyama, Shigekota Kaihara. Steps Toward Feasible Consultation Systems: The Knowledge Based Antibiotic Medication Counselling System ANTICIPATOR. - P. 276-281.

68. Fujimasa I., Nakajima M., Mabuchi K., Chinzei T., Abe Y., Hyakuna Y., Atsumi K. Development of an On-Line Real-Time Computed Thermography System (CTS). - P. 709-711.

69. Barr A. and Feigenbaum E.A. The Handbook of Artificial Intelligence. - V. 2, Palo Alto CA: Henristech Press, 1982.

70. Di Eugenio B. Correction of incorrect assumptions and generation of wide answers in FIDO system. - AICA, Annual Conference Proc., Palermo, Sept., 1986.

71. Holley R.P. Classification in the USA. - Int. Classif., V. 13, N 2, 1986. - P. 73-78.

72. Gordon A.D. Links between clustering and assigment procedures. - Proc. in Computational Statistics, 7th Symposium, Rome, Italy, 1986. - P. 149-156.

73. Nishisato S. Generalized forced classification for quantifying categorical data. Data Analysis and Informatics. - IY Proc. of the Fourth International Symposium, Versailles, France, 9-11 Oct., 1985 (Amsterdam, Netherlands: North - Holland, 1986). - P. 351-362.

74. Колычев П.М. К классификационной проблеме. - НТИ, сер. 2, № II, 1987.

75. Баскакова Л.В., Журавлев Ю.И. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств. - Ж. выч. мат. и матем. физики. Т. 2I, № 5, I98I. - С. I264-I275.

76. Бонгард М.М. Проблемы узнавания. - М.: Наука, I967. - 4I2 с.

77. Горелик А.Л. Современное состояние проблемы распознавания. М.: Радио и связь, I985. - I60 с.

78. Журавлев Ю.И. Экстремальные задачи, возникающие при обосновании эвристических процедур. - В кн.: Проблемы прикладной математики и механики. - М.: Наука, I97I. - С. 67-75.

79. Журавлев Ю.И. Непараметрические задачи распознавания образов. - Кибернетика, № 6, I976. - С.93-I03.

80. Лейбкинд Р.А., Рудник Б.П. Моделирование организационных структур (классификационный подход). - М.: Наука, I98I. - I4I с.

8I. Goldberg S.I. Diagnostics on the basis of the informative spase of the antisyndromes.- Problems of Control and Information Theory. - Budapest, V. 13, 1989. - P. 401-411.

82. Goldberg S.I., Skripochenko O.A. Neakness estimations for objects with complex structure. - Problems of Control and Information Theory. - Budapest, V. 15, 1986. - P. 231-238.

83. Reggia T.A. Artificial Intelligence and Medical Decision Making. - MEDINFO 83. - P. 475-479.

84. Maitree K., Tanaka M., Ichikawa T. A knowledge-based system organization for image data retrieval. - Mem. Fac. Eng. Hirosima Univ., V. 9, N 2, 1986. - P. 11-22.

85. Dempster A.P. Upperand lower probabilities induced by multivalued mapping. - Annals of Mathematical Statistics, 38,

86. Tsumura H., Wasano T., Masuzawa H., Miyake H. and Higashida M. Doctors: Penden Optimal Clinical Treatment Order Request System. - WGAI of Information Processing Society of Japan, 1984. - P. 35-49.

87. Held T.P., Carlis T.V. MATCH - a new high-level relational operator for pattern matching. - Commun. ACM, V. 30, N 1, 1987. - P. 62-75.

88. Thuan H. Contribution to the theory of relational databases. - Tanulmanyok Magy. Tud. Acad. Szami-tastech. & Autom. Kut. Inter. (Hungary), N 184, 1986. - P. 1-156.

89. Medeiros C.B., Tompa F.W. Understanding the implications of view update policies. - Algorithmica, V. 1, N 3, 1986. - P. 337-360.

90. Abendroth T.W. Evaluation of diagnostic test performance. - Proc. of the Tenth Annual Symposium on Computer Applications in Medical Care, 1986. - P. 411-420.

91. Anderson T.G. Modeling systems under indirect observation: a structural model of physician use of a computer-based hospital information system. - Proc. of the 2nd European Simulation Congress, Antwerp., Belgium, Sept., 1986. - P. 769-774.

92. Bearman D. Archival and bibliographic information networks. - J. Libr. Adm., V. 7, N 2, 1987. - P. 99-110.

93. Yamazaki M. Aspect of materials databases. - Joho Kanri (Japan), V. 29, N 9, 1986. - P. 743-755.

94. Миркин Б.Г. Анализ качественных признаков и структур. - М.: Статистика, 1980. - 319 с.

95. Эпштейн В.Л. О приложении теории графов для описания и анализа схемы потоков информации в управляющих системах.-

Автоматика и Телемеханика, № 8, 1965. - С. 1403-1409.

96. Martin J.M. From medical data to health knowledge. - Methods Inf. Med., V. 26, N 1, 1986. - P. 3-12.

97. Tompkins W.J. Biomedical computing using personal computers. - IEEE Eng. Med. & Biol. Mag., V. 5, N 3, 1986. - P. 61-64.

98. Szilas A. Methods of computer application and principles of data evaluation in the complex screening examinations at Pecs. - Inf. Elektron. (Hungary), V. 21, N 4, 1986. - P. 238-244.

99. Willett P., Winterman V. Implementation of non-hierarchic cluster analysis methods in chemical information systems. - J. Chem. Inf. & Comput. Sci., V. 26, N 3, 1986. - P. 109 - 118.

100. Hagmann R.B. An observation on database buffering performance metrics. - Proc. of Very Large Data Bases, 1986. - P. 289-293.

101. Hepfer C. Using dBase III to prepare a subject index to statistical resources. - Libr. Software Rev., V. 5, N 5, 1986. - P. 284-287.

102. Rajendran P.P. Enriched title-based keyword index generation using dBase II. - Micro-comput. Inf. Manage., V. 3, N 4, 1986. - P. 297-314.

103. Фомин В.Н. Математическая теория обучаемых опознающих систем. - Л.: ЛГУ, 1976. - 341 с.

104. Хенчей Г. Ассоциативно-структурный анализ булевых данных с применением монотонных систем. - Автоматика и Телемеханика. - № 2, 1987. - С. 137-141.

105. Дульнев Л.С., Хенчей Г. Метод агрегирования информационно-технологической структуры САПР. - Киев, 1987. - 12 с.

A TANULMÁNYOK SOROZATBAN 1988-BAN MEGJELENTEK

203/1988 KNVVT EG-25 Problems and tools of the integration of information systems. Proceedings 1987.
Edited by: Rumjana Kirkova, Tibor Remzső, Ferenc Urbánszki

204/1988 Csetverikov Dimitrij: Digitális texturavizsgálat néhány uj módszere

205/1988 Hernádi Ágnes: Uj eszközök a fogalmi modellezésben

206/1988 The second Hungarian workshop on image analysis
Edited by: Dimitrij Csetverikov, Géza Álló

207/1988 Suzanne Márkusz - Gábor Márkus: Logic Puzzless and Logic Programming I
/Logikai fejtörők - logikai programozás I/

208/1988 Proceedings of the 5th International Meeting of Young Computer Scientists /IMYCS'85/
Edited by: E. Csuhaj-Varju, J. Demetrovics, J. Kelemen

209/1988 Галя Младенова Ангелова: Синтаксические и семантические структуры реляционных языков запросов

210/1988 Publications'1987 - Publikációk'1987
Edited by/Szerkesztette: Petróczy Judit

211/1988 Eszenszki József - Kas Iván - Palotási András Podmaniczky András - Szücs Miklós - Vörös Károly Zalán Frigyes - Alexander Mihajlovics Klocskov Valerij Alexandrovics Plahov:
Tanulmány a számitógépes, raszteres mikrofilm lap készités elvi és gyakorlati kérdéseiről