АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

№ 4 1977

СИСТЕМЫ АВТОМАТИЗАЦИИ НАУЧНЫХ ИССЛЕДОВАНИЙ

УДК 681.325.67 : 547

И. Н. АНИКЕЕВА, Т. С. ВАСЮЧКОВА, Ю. П. ДРОБЫШЕВ, С. П. СОКОЛОВ

(*Новосибирск*)

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ ПОИСКА ХИМИЧЕСКИХ СОЕДИНЕНИЙ ПО СТРУКТУРНЫМ ФРАГМЕНТАМ

Описываемая система является частью комплексной машинной системы для идентификации химических соединений по данным спектроскопии молекул *. Информационную базу системы составляет машинный каталог, содержащий описание графов структур химических соединений. Информационно-поисковая система (ИПС) «Структура» реализована на БЭСМ-6 и предназначена для поиска в каталоге органических соединений по их молекулярным структурам или фрагментам структур.

Совместное использование этой системы и ИПС «Спектр»**, включающей в себя спектральную информацию, позволяет исследовать различные связи между структурами химических соединений и их спектрами: устанавливать спектроструктурные корреляции, определять соответствие синтезированных структур спектральным данным и т. д.

Кодирование структур и машинный каталог. Основными требованиями при выборе способа кодирования химических структур были простота и надежность этой процедуры.

Каждое соединение представлено в машинной памяти порядковым номером, брутто-формулой и описанием структуры. Используется принятое в НИОХ СО АН СССР представление структуры, основанное на таблице типовых структурных фрагментов, содержащей около 100 единиц. Таблица может быть пополнена. Структура представляется графом, вершины которого суть типовые структурные фрагменты, а ребра — химические связи определенных типов (одиночная, двойная, дробная и т. д. — всего семь типов). Сначала производится произвольная нумерация вершин графа. Затем составляется его топологическое описание, которое состоит из ряда строк по числу вершин графа. Начальный элемент каждой строки есть одна из вершин графа (образующий фрагмент), а следующие — фрагменты, связанные с образующим. Элементы строки содержат присвоенный порядковый номер, описание фрагмента (брутто-формулу) и тип связи. Таким образом, в первоначальном описании каждый фрагмент упоминается $K+1$ раз, где K — число связанных с ним других фрагментов. Каждая связь упоминается дважды.

* Дробышев Ю. П., Коптюг В. А. Комплексная машинная система для решения структурных задач методами молекулярной спектроскопии.— «Автометрия», 1972, № 4, с. 109—118.

** Там же.

Первоначальное описание содержит значительную избыточность и после ввода в ЭВМ подвергается канонизации и сжатию. Операция канонизации заключается в упорядочении образующих фрагментов (т. е. строк) и связанных фрагментов внутри строк в соответствии со старшинством согласно таблице типовых фрагментов. Таким образом, запись структуры при любой нумерации вершин начального графа приводится к одной и той же форме.

Для сжатия первоначального описания структуры, во-первых, информация из буквенно-цифровой (десятичной) формы переводится в цифровую (восьмеричную), а во-вторых, каждая связь в описании упоминается только один раз. Сжатое описание составляется следующим образом. Поскольку фрагменты уже упорядочены (принцип произвольный), у каждого j-го фрагмента имеется $s(j)$ ему предшествующих и связанных с ним фрагментов. В качестве образующих фрагментов берутся только фрагменты, для которых $V(j)-s(j)>0$. Нетрудно установить, что каждый j-фрагмент будет присутствовать в новом описании только $\delta+s(j)$ раз, где $\delta=\begin{cases}1, V(j)>s(j);\\ 0, V(j)=s(j);\end{cases}$ $V(j)$ — валентность фрагмента.

Эти операции позволяют увеличить плотность хранения информации примерно в три раза. Канонические описания структур образуют машинный каталог структур, хранящийся на магнитных лентах или дисках. На одном диске можно разместить 70—80 тыс. структур.

Общая схема поиска. Для поиска в машинном каталоге соединений по входящему в него сложному структурному фрагменту (блоку) в систему вводится запрос, содержащий описание блока, составленное по указанным выше правилам. Существует дополнительное условие: нумерация вершин графа (или типовых фрагментов) должна начинаться с 1 и не содержать пропусков.

После поступления запросов проводится их полный синтаксический и семантический анализ. Синтаксический анализ заключается в проверке правильности кодирования запросов. На этом этапе выявляются ошибки, возникающие из-за употребления неверного кода или пропуска звеньев графа, или неверного его описания; в частности, если в описании вершина A связана с вершиной B связью типа m, то в описании должно быть отмечено, что вершина B связана с A также связью типа m. При обнаружении ошибки система сообщает об этом абоненту с указанием типа ошибки. Классификация возможных ошибок с соответствующими сообщениями приведена в таблице.

Правильно составленные запросы переводятся в каноническую форму и передаются программе поиска. Процесс поиска заключается в последовательном сравнении запроса со структурами из машинного каталога. Сравнение проводится в полной канонической форме. Для этого запись каждой структуры из каталога восстанавливается до полного канонического описания. Как обычно, анализируемая структура считается релевантной, если в нее вкладывается целиком запрашиваемый блок, т. е. его граф является подграфом структуры соединения. Для ускорения поиска анализ на релевантность проводится в два этапа: на первом используются общие характеристики структур такие, как: число строк описания, их длины, количества связей определенного типа; на втором проводится более детальный анализ, заключающийся в идентификации типовых фрагментов и связей.

Главная трудность возникает в случае, когда в соответствующих строках запроса и анализируемой структуры имеются одинаковые фрагменты, особенно строки. Здесь приходится прибегать к взаимным перестановкам одинаковых элементов (и строк), причем число проверяемых комбинаций растет очень быстро с ростом таких элементов, что резко снижает характеристики системы в отношении объема памяти и времени обработки запроса.

Номера п/п	Текст	Комментарий
1	В СТРУКТУРЕ *K* ЗОНЫ НОМЕРА *N* И ЗАПРОСЕ *L* СОДЕРЖИТСЯ БОЛЕЕ 6 ОДИНАКОВЫХ СВЯЗАННЫХ ФРАГМЕНТОВ	*K* — номер структуры, *N* — номер зоны, *L*—номер запроса. Нарушено ограничение 4
2	ДЛИНА ПАКЕТА ЗАПРОСОВ ПРЕВЫШАЕТ МАКСИМАЛЬНУЮ	Вводятся слишком большие запросы. Необходимо уменьшить число запросов в пакете
3	ИМЯ АРХИВА НА ЛЕНТЕ НЕ СОВПАДАЕТ С ЗАПРАШИВАЕМЫМ АРХИВОМ	Либо неверно пробито имя архива на 10-й перфокарте, либо заказана бобина не с тем архивом
4	НЕВЕРЕН ДИАПАЗОН ПОИСКА. КОНЕЧНАЯ ЗОНА В ЗАПРОСЕ БОЛЬШЕ, ЧЕМ В АРХИВЕ	
5	НЕВЕРЕН ДИАПАЗОН ПОИСКА. НАЧАЛЬНАЯ ЗОНА В ЗАПРОСЕ МЕНЬШЕ, ЧЕМ В АРХИВЕ	
6	НЕВЕРНОЕ ОПИСАНИЕ ФРАГМЕНТА В ЗАПРОСЕ. ЗАПРОС ОТБРАКОВАН	Возможны следующие ошибки в запросе: фрагмент не описан как образующий; для образующего фрагмента описаны не все с ним связанные фрагменты; неверно указан вид связи между фрагментами
7	НЕВЕРНЫЙ ЗАПРОС. ОШИБОЧНЫЙ СИМВОЛ ЛИБО НАРУШЕНЫ КОЛИЧЕСТВЕННЫЕ ОГРАНИЧЕНИЯ	Возможны следующие ошибки: неверный код фрагмента; неверный вид связи; нет в начале запроса признака на полное совпадение; неверно пробит номер вершины; пропуск символа; число вершин в запросе >99; число связанных фрагментов с одним образующим >8
8	ОШИБКА АРХИВА	Имеется ошибка в структуре архива
9	ЧИСЛО ЗАПРОСОВ >12	

В данном варианте системы возможности анализа ограничены для структур, описания которых содержат более пяти связанных одинаковых фрагментов. Однако подобные структуры встречаются довольно редко. О встрече таких структур система сообщает абоненту. Это не означает, что такие структуры не могут анализироваться при ином запросе.

Результатом поиска является некоторое множество релевантных структур, идентификаторы которых (номер в соответствующем каталоге) выдаются на печать. В случае необходимости абонент может получить распечатку релевантных структур.

Алгоритм поиска. Основной рабочей программой системы является программа поиска химических соединений в машинном каталоге по заданному набору структурных фрагментов (запросу), в частности по полной структуре.

Поиск требуемого химического соединения осуществляется последовательным просмотром указанной области каталога и попарным сравнением графа запроса и графа структуры соединения из каталога

(в дальнейшем будем называть эти графы просто «запрос» и «структура»).

Процесс поиска выполняется в два этапа. Вначале исследуются общие характеристики каждой пары запрос — структура. При этом проверяются следующие соотношения:

1) число вершин в запросе не превышает числа вершин в структуре;

2) число вершин данного типа (имени, поскольку графы именованы) в запросе не превышает числа вершин этого же типа в структуре;

3) число связей данного типа в запросе (дуг графа) не превышает числа связей этого же типа в структуре;

4) для каждой вершины в запросе и в структуре подсчитывается число связанных с нею вершин. Для вершины A в запросе должна найтись вершина B того же типа в структуре, такая, что число вершин, связанных с A, не превосходит числа вершин, связанных с B. При этом разным вершинам запроса должны быть сопоставлены разные вершины структуры.

Если хотя бы одно из этих соотношений не выполняется, то это означает, что данная структура не может включать в себя граф заданного запроса. Следует отметить простоту, естественность, эффективность упомянутых четырех критериев, по которым ведется фильтрация структур на I этапе поиска. Эта процедура выполняется в процессе подготовки структуры к обработке на II этапе, когда она перекодируется в каноническую форму с упорядочением строк и фрагментов внутри строк. Поэтому фильтрация не требует дополнительного времени. При опытной эксплуатации системы обнаружено, что примерно в 94% случаев * несоответствие запроса и структуры выясняется на I этапе. Это дает большую экономию машинного времени, поскольку основное время работы программы приходится на второй этап.

Если все четыре критерия удовлетворены, то наступает II этап поиска — проверка взаимно-однозначного соответствия всех вершин графов запроса и структуры, выполняемая методом упорядоченного перебора: вершине A в запросе сопоставляется вершина B того же типа в структуре, строится множество W_A всех путей в графе запроса, выходящих из вершины A, и множество W_B всех путей в графе структуры, выходящих из вершины B. Если не выполняется условие $W_B \subseteq W_A$, то вершины A и B не соответствуют друг другу и следует взять вместо B другую вершину того же типа.

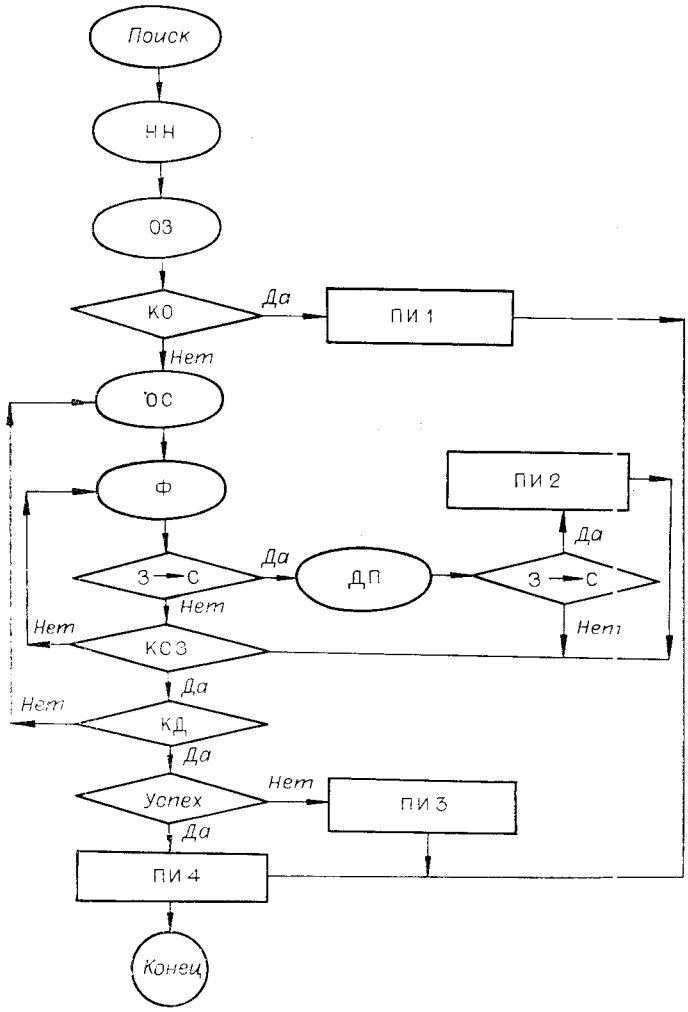
Если найдено такое соответствие вершин графов запроса и структуры, что для всех пар (A, B) выполняется отношение $W_B \subseteq W_A$, то граф запроса является подграфом структуры и данная структура релевантна запросу.

Наибольшие трудности на II этапе возникают в случаях, когда какая-либо вершина связана с несколькими вершинами одного типа. При этом приходится рассматривать все перестановки таких вершин. Перебор большого числа подобных вариантов может вызвать превышение выделенных ресурсов ЭВМ. Поэтому введено ограничение на допустимое число однотипных вершин, связанных с основной (не более пяти).

Это не означает, что при нарушении ограничения анализ не будет проведен. Более того, может быть даже получен положительный результат, т. е. установление взаимно-однозначного соответствия между запросом и структурой и в этом случае, но проведение полного анализа не гарантировано. О превышении ресурсов, находящихся в распоряжении системы, последняя информирует абонента с указанием пары запрос — структура, для которой это имело место.

---

* Цифра относится к объему каталога в 3 тыс. структур.



Функциональная схема обработки запроса:

НН — начальная настройка таблиц; ОЗ — чтение и отбраковка запросов; ОС — обработка очередной структуры с ленты; Ф — фильтрация очередного запроса; ПИ1 — печать на АЦПУ текста ВСЕ ЗАПРОСЫ ОТБРАКОВАНЫ; ПИ2 — печать на АЦПУ текста ЗАПРОС НОМЕР *L* СОДЕРЖИТСЯ В СТРУКТУРЕ *K*, затем печать текста структуры и запроса; ПИ3 — печать на АЦПУ текста В УКАЗАННОМ ДИАПАЗОНЕ ЗАПРАШИВАЕМЫХ СТРУКТУР НЕТ; ПИ4 — печать на АЦПУ текста ПРОГРАММА ПОИСК СТРУКТУРЫ РАБОТУ ЗАКОНЧИЛА; КО — конец отбраковки; З → С — запрос содержится в структуре; КСЗ — конец списка запросов; КД — конец диапазона поиска в архиве; Успех — поиск завершен успешно; ДП — детальный поиск на соответствие вершин, смежности и наименований.

На рисунке приведена функциональная схема обработки запроса.

Результатом поиска является некоторое множество структур, релевантных введенной группе запросов. На печать выдается информация о соответствии структур запросам. При этом последовательно печатаются:

оповещение о работе программы;

сообщения об ошибках при введении запросов, если таковые имели место (см. таблицу);

тексты неверных запросов, место ошибки выделяется символом «*»;

тексты правильных запросов;

номера просмотренных зон каталога;

при успешном поиске — фраза ЗАПРОС № СОДЕРЖИТСЯ В СТРУКТУРЕ №№.

Структура идентифицируется своим номером в каталоге, а запрос — номером в группе. По желанию абонента на печать могут быть выведены описания найденных релевантных структур.

Технические характеристики. Опытная проверка системы была проведена на базе БЭСМ-6 с каталогом на магнитной ленте объемом около 3 тыс. структур. Информационная емкость ленты 35÷40 тыс. структур. Требуемый объем МОЗУ 24 листа. Одновременно обрабатываются от 1 до 12 независимых запросов. При этом должны учитываться следующие ограничения: 1) число фрагментов в запросе не более 99; 2) число фрагментов, связанных с одним образующим, не более 8; 3) число запросов в пакете не более 12; 4) вершины запроса должны нумероваться подряд, начиная с единицы.

При данном объеме каталога время работы процессора менее 1 мин, например при 9 запросах поиск среди 3 тыс. структур занял 40 с.

Кроме программы поиска в библиотеку входят программы проверки, пополнения, исправления и распечатки массивов.

Работа над ИПС «Структура» выполнялась в тесном контакте с лабораторией НИОХ СО АН СССР, руководителю которой — чл.-кор. АН СССР В. А. Коптюгу — авторы выражают благодарность за помощь.

Поступила в редакцию 24 марта 1977 г.

УДК 651.011.55 : 681.32 : 543.51

Б. Г. ДЕРЕНДЯЕВ, С. А. НЕХОРОШЕВ, Л. М. ПОКРОВСКИЙ, В. И. СМИРНОВ, Г. П. УЛЬЯНОВ

(*Новосибирск*)

АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ДАННЫХ МАСС-СПЕКТРОМЕТРИЧЕСКИХ ЭКСПЕРИМЕНТОВ НА БАЗЕ ЭВМ «МИНСК-32» В РЕЖИМЕ НИЗКОГО РАЗРЕШЕНИЯ

Использование масс-спектрометрии в химии до недавнего времени сдерживалось большими затратами времени на представление масс-спектрометрических данных в удобной для исследователя численной форме. При работе на спектрометрах, не связанных с ЭВМ, оператор должен преобразовать масс-спектр, представленный на бумаге в координатах «аналоговое значение интенсивности — время», в новые координаты — «значения интенсивности — массовые числа», проводя в случае необходимости одновременное усреднение результатов измерений по данным нескольких экспериментов. В силу приборных нестабильностей, нелинейности развертки, широкого динамического диапазона изменения интенсивностей сигналов, большого объема спектральных данных и ряда других факторов полная обработка результатов масс-спектрометрического эксперимента является весьма трудоемкой. Большие непроизводительные затраты времени могут быть устранены путем создания автоматизированного комплекса, состоящего из спектрометра и ЭВМ.

В литературе к настоящему времени описано несколько типов таких систем [1—6], осуществляющих обработку масс-спектров, зарегистрированных в условиях низкого, среднего и высокого разрешения. При