1959 г. Сентябрь *Т. LXIX, вып. 1*

УСПЕХИ ФИЗИЧЕСКИХ НАУК

НАУЧНАЯ И ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ КАК ОДНА ИЗ ЗАДАЧ КИБЕРНЕТИКИ

Г. Э. Влэдуц, В. В. Налимов, Н. И. Стяжкин

За последние годы в физике, ее технических приложениях и в областях, смежных с физикой, стали широко применяться механизированные методы накопления и переработки информации. Поэтому мы сочли целесообразным рассмотреть эту проблему как в ее техническом, так и в теоретическом аспектах в журнале, посвященном вопросам физики.

§ 1. ЭКСПОНЕНЦИАЛЬНЫЙ ХАРАКТЕР РАЗВИТИЯ НАУКИ

Статистический анализ развития науки показывает, что количество публикуемых работ, число журналов, количество работников, занятых на исследовательских работах, и ассигнования на эти работы увеличиваются по экспоненте. На рис. 1 приведена кривая[1,2], показывающая рост суммарного числа публикаций в реферативном журнале Physics Abstracts, начиная с 1900 г. Нарушение экспоненциального хода кривой наблюдалось только в годы второй мировой войны, после окончания которой рост публикаций продолжался синбатно первоначальной кривой. Аналогичная закономерность имеет место для роста числа публикаций в химии[3] и биологии[4]. На рис. 2 показан рост числа научных и реферативных журналов[2]. (По оси ординат данные нанесены в логарифмическом масштабе.) Здесь также наблюдается экспоненциальный характер роста, причем в настоящее время общее число научных журналов*) приближается к 100 000, а число реферативных журналов достигает величины, близкой к тремстам[2]. Такой же характер роста наблюдается, по данным зарубежных работ, для ассигнований на исследовательские работы[5]. Построить непосредственно кривую роста числа научных работников не представляется возможным из-за отсутствия соответствующих статистических данных. Однако на основании ряда косвенных показателей — роста числа студентов, оканчивающих соответствующие учебные заведения, и числа «известных ученых», указываемых в различного рода словарях и справочниках, а также роста числа публикаций, патентных заявок, ассигнований на научную работу и пр.— можно утверждать, что здесь также имеет место экспоненциальный характер роста.

Анализ кривых роста для показателей, характеризующих развитие науки, позволяет сделать вывод о том, что во всех упомянутых случаях имеет место экспоненциальный ход развития, выполняющийся с точностью

*) Сведения о числе научных журналов, приводимые различными авторами, варьируют в широких пределах, так как до сих пор нет общепринятого строгого определения для понятия «научный журнал». В реферативных журналах, издаваемых Институтом научной и технической информации АН СССР, реферируются статьи 11—12 тысяч научных журналов. В это число не входят журналы по гуманитарным наукам.

до 1%[2]. Параметры экспоненты для различных показателей варьируют в сравнительно узких пределах так, что за интервал в 10—15 лет все показатели удваиваются.

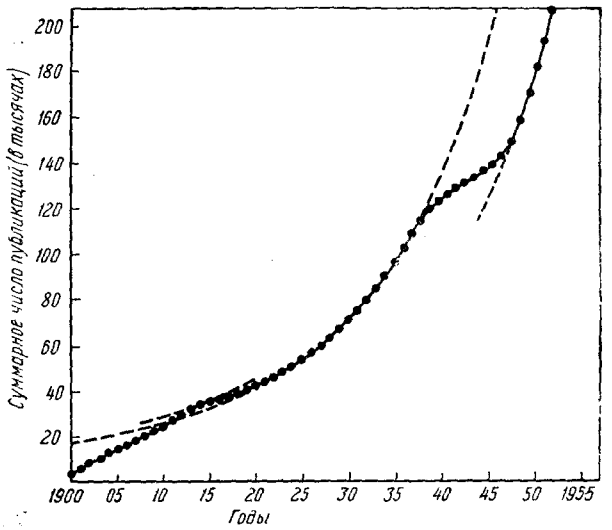


Рис. 1. Рост суммарного числа публикаций в реферативном журнале Physics Abstracts, начиная с 1900 г. (по оси ординат отложены накопленные суммы)[1,2].

С экспоненциальными кривыми можно сделать мысленный эксперимент, экстраполируя их. Экстраполяция в историческое прошлое приводит к разумным результатам. Оказывается, что ординаты почти всех кривых достигают значения, равного единице, приблизительно в 1700 г., т. е. в эпоху, связанную с деятельностью И. Ньютона. Например, экстраполяция кривой роста научных журналов приводит к значению ординаты, равной единице, в 1700 г. В действительности первые журналы начали выходить в 1665 г., при этом они в первые годы образовывали небольшую группу, выпадающую из общей закономерности. Эпоха Ньютона может рассматриваться как начало того исторического периода крупных научных открытий, который продолжается и по настоящее время. Экспоненциальный характер развития науки по большинству показателей прослеживается на протяжении 200—250 лет.

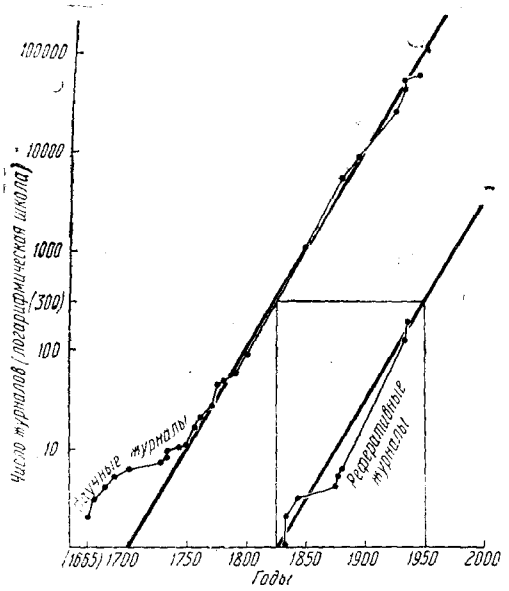


Рис. 2. Рост числа научных и реферативных журналов [2].

Такая закономерность в развитии отдельных показателей наблюдается только для достаточно широких областей науки, таких как физика, химия,

биология. Если же мы возьмем какую-нибудь узкую область науки, то здесь вначале имеет место экспоненциальный характер развития, а затем, после того, как потенциальные возможности развития данной дисциплины оказываются исчерпанными, рост числа публикуемых работ становится линейной функцией времени. Это иллюстрируется рис. 3, где приведена кривая роста для суммарного числа публикаций по теории матриц и определителей[1]. Первая работа в этой области появилась около 1750 г. Начиная с 1800 г., когда общее число опубликованных работ равнялось 10, строго выполнялся экспоненциальный характер роста вплоть до 1880 г., после

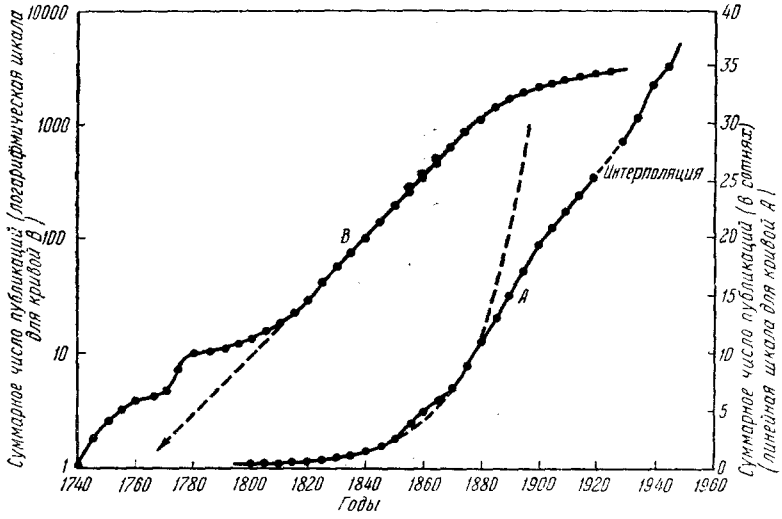


Рис. 3. Рост числа публикаций по теории определителей и матриц[1].

которого наблюдался уже линейный характер роста публикаций. Утверждение об экспоненциальном характере развития науки вполне согласуется с весьма вероятными предположениями достаточно общего характера. Экспоненциальный характер развития науки может быть описан зависимостью

$$y = ae^{kt} \qquad (k > 0),$$

которая является решением дифференциального уравнения

$$\frac{dy}{dt} = ky,$$

где производная $\frac{dy}{dt}$ означает скорость роста интересующих нас показателей, т. е. увеличение их за единицу времени. Таким образом, экспоненциальный ход развития является следствием того,что относительная скорость роста остается постоянной величиной

$$\frac{dy}{y\,dt} = \text{const}.$$

Кажется вполне естественным полагать априори, что при отсутствии ограничивающих факторов скорость роста должна определяться достигнутым уровнем: каждая новая научная концепция должна вызвать определенное количество новых научных работ — развивающих и углубляющих или опровергающих ее.

При этом естественно, что константа варьирует в узких пределах для различных показателей, характеризующих развитие науки, и мы по всем показателям получаем удвоение приблизительно за один и тот же отрезок времени в 10—15 лет. Легко подсчитать, что удвоению за период в 10—15 лет соответствует постоянная относительная скорость роста в 5—7% в год. Несколько быстрее происходит рост ассигнований на исследовательские работы — по данным[5] в США относительная скорость роста ассигнований равна 10% в год. Когда рост числа публикаций оказывается линейной функцией времени, то это значит, что абсолютная скорость роста остается величиной постоянной, не зависящей от достигнутого уровня.

Если мысленный эксперимент продолжать и экстраполировать экспоненциальные кривые в будущее, то мы неизбежно придем к абсурду. После 10-кратного удвоения, т. е. через 100—150 лет, количество публикаций и число научных сотрудников должно будет увеличиваться в тысячу раз, а через 200—300 лет — в миллион раз. Рост числа научных работников происходит быстрее, чем прирост населения на земном шаре. По последним данным, опубликованным Организацией объединенных наций (газета «Известия», 1959, 2.VI, № 129), относительная скорость роста населения на земном шаре составляет 1,6% в год, что соответствует удвоению населения примерно за 45 лет (ранее[6] принималось, что скорость роста равна $\sim$1%, что соответствовало удвоению населения за 70 лет). Поэтому естественно, что кривая роста числа научных работников должна достигнуть насыщения: экспонента должна перейти в S-образную кривую. В биологии для описания подобной ситуации обычно пользуются так называемой логистической кривой, изображенной на рис. 4. Аналитическое выражение для логистической кривой

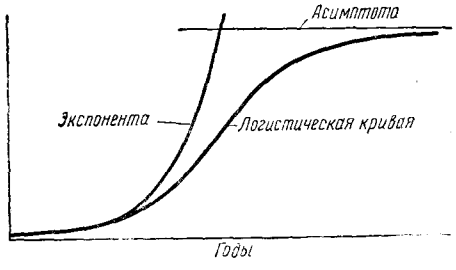


Рис. 4. Переход экспоненты в логистическую кривую при наличии фактора, сдерживающего рост [2].

$$y = \frac{b}{1 + ae^{-kbt}} \qquad (k > 0)$$

является решением дифференциального уравнения

$$\frac{dy}{dt} = ky(b - y) \qquad (0 < y < b).$$

В этом случае рост ограничен, так как b является максимальным значением величины y. Относительная скорость роста

$$\frac{dy}{y\,dt} = k(b - y)$$

здесь уже не остается величиной постоянной — она оказывается линейной функцией y. Чем выше становится достигнутый уровень интересующего нас показателя, тем ниже оказывается скорость роста. Уменьшение скорости роста становится заметным только при достаточно большом значении y. В начальные моменты времени, когда $y \ll b$, логистическая кривая практически совпадает с экспонентой, как это показано на рис. 4.

Основываясь на статистическом анализе процесса развития науки, можно высказать ряд соображений о дальнейших путях ее развития.

Количество научных работ достигло к настоящему времени такого уровня, что стала очевидной недостаточность реферативных журналов как средства передачи первичной информации. Здесь существенным является также то, что отдельные научные дисциплины постепенно теряют свои четкие контуры. Активно работающий исследователь не может ограничиться информацией в какой-то одной области — он должен обращаться к смежным дисциплинам. Наиболее интересные работы сейчас в ряде случаев стали появляться на стыке наук. Количество внутренних межнаучных связей также, по-видимому, растет по экспоненте. Сейчас оказались связанными между собой даже такие, казалось бы, чуждые друг другу в методологическом отношении области знаний, как биология и математика. Аналогичное положение имеет место и в технических исследованиях. Руководитель лаборатории крупного металлургического предприятия должен следить не только за литературой в области металлургии и металловедения, но также и за развитием химии и некоторых разделов физики и математики. Успех работы лаборатории в значительной степени определяется тем, насколько внедряются новые физические и математические (статистические) методы для исследования технологического процесса и контроля качества продукции. В США недавно проводилось статистическое изучение баланса рабочего времени химика-исследователя [7]. В качестве объекта для изучения были взяты химики, так как они составляют наиболее многочисленную группу научных и технических работников: Американское химическое общество насчитывает 80 000 членов. Результаты обследования показали, что в общем балансе времени исследователя на долю научной информации в среднем падает 33,4%. Минимальное значение этой величины составляет 15,7%, максимальное — 61,4%. Под научной информацией понимаются все процессы, связанные с обсуждением научных и технических проблем, поисками и чтением литературы, письменными и устными сообщениями. Если содержание информационной работы определять более широко, включив сюда дополнительно обдумывание и планирование эксперимента, с одной стороны, и служебную и административную информацию, с другой стороны, то оказывается, что химик-исследователь тратит на информационную деятельность в среднем 49,8%. Минимальное значение этой величины составляет уже 20,0%, а максимальное — 94,5%. Таким образом, уже сейчас исследователь тратит в среднем 50% своего времени на информационную деятельность. В связи с экспоненциальным характером роста числа публикаций, очевидно, будет расти и время, затрачиваемое на информационную работу, и химик-исследователь окажется не в состоянии заниматься экспериментальной работой, если не будут найдены новые более эффективные средства для информационной службы. Отсюда, очевидно, то большое значение, которое приобретают сейчас задачи механизации научной и административной информационной деятельности.

В связи с разработкой информационных машин в ближайшее время должна существенно измениться практика публикования работ. Среди научных работников, особенно зарубежных, часто раздаются высказывания о том, что «достаточно новых журналов». Между тем существующие журналы не в состоянии публиковать все поступающие к ним материалы. Как у нас в Советском Союзе, так и за рубежом, задержка в публикации статей иногда доходит до двух лет. В некоторых наших журналах, например в «Заводской лаборатории», около 50% поступающих рукописей отвергается только из-за отсутствия места в журнале, причем средний объем публикаций в этом журнале за последние 10 лет уменьшился вдвое. По данным [8] в США 48,5% докладов, заслушанных на конференциях, появилось в печати только в виде кратких сообщений.

При таком положении дел значительная часть полученной информации оказывается практически недоступной для широкого круга исследователей. Наметившаяся сейчас тенденция к сокращению объема публикуемых статей, по-видимому, будет развиваться и дальше, по крайней мере для работ экспериментального характера. При такой системе публикаций можно будет избежать потери информации только в том случае, если она будет заноситься в долговременную память информационных машин *).

Переход на новые способы информации требует разработки строго стандартных способов компактного свертывания результатов эксперимента и строгой количественной оценки того элемента неопределенности, который связан с каждым экспериментом и обусловлен неизбежными инструментальными и методическими ошибками, ограниченностью экспериментального материала и пр. Ранее, когда редакции журналов не ограничивали объема публикаций, автор мог представлять результаты эксперимента любым удобным для него способом. Читатель получал известное представление о надежности результатов работы на основании пространных описаний условий эксперимента, способов обработки экспериментального материала и пр. Теперь, когда объем публикаций сократился и по каждому, даже узкому вопросу, имеются сотни, а иногда и тысячи статей, такая непосредственная оценка результатов экспериментальных работ становится невозможной. Читатель не может отличить хороших работ от плохих. Наличие невыявленных плохих работ заставляет иногда сомневаться в результатах хороших работ. Стандартизация способов представления результатов эксперимента стала необходимой уже на том уровне свертывания информации, который имеет место сейчас при опубликовании кратких статей. Еще более важной эта задача станет при машинных способах обработки информации, когда информация будет подвергаться дальнейшему свертыванию.

Если в ближайшем будущем кривая роста для числа работников, занятых на исследовательских работах, должна перейти из экспоненты в логистическую кривую, то, естественно, возникает вопрос о том, сохранится ли экспоненциальный характер роста для количества научных работ. Для удовлетворения стремления человека к познанию природы, с одной стороны, и для удовлетворения растущих материальных потребностей, с другой стороны, по-видимому, необходимо, чтобы научные работы продолжали расти по экспоненте так, как это было на протяжении последних 200—250 лет. До сих пор всегда существовала корреляционная связь между ростом валовой продукции и увеличением ассигнований на исследовательские работы, а следовательно, и количеством исследовательских работ, как это, например, видно на рис. 5, где приведены данные, характеризующие рост валовой продукции и ассигнований на исследовательские работы в США. Очевидно, что экспоненциальный характер развития исследовательских работ может сохраниться в дальнейшем только в том случае, если часть интеллектуального труда будет передана машинам. Машины должны облегчить интеллектуальную деятельность человека в такой степени, чтобы общее количество усилий, направленных на исследовательские

*) Одновременно с этим должна повышаться роль обзорной литературы. По статистическим данным [9] уже теперь химики, работающие в США, пользуются обзорными статьями, публикуемыми в 50 различных журналах. Некоторые из этих журналов публикуют обзорные статьи только периодически или эпизодически, одновременно с оригинальными работами. Такое большое количество публикаций обзорного характера, объясняется тем, что обзоры пишутся, исходя из конкретных запросов той или иной группы работников. Например, журнал Analytical Chemistry четвертый (апрельский) номер издает в двух тетрадях, одна из которых посвящена обзорным статьям по самым разнообразным вопросам, представляющим интерес для химика-аналитика.

работы, продолжало расти по экспоненте даже тогда, когда рост числа исследователей будет происходить по какой-то кривой, близкой к логистической.

Экспоненциальный рост числа научных работ будет неизбежно приводить к увеличению срока обучения, если здесь не будут найдены новые пути. Во времена Гаусса к творческой работе приступали в возрасте до двадцати лет. В настоящее время эта величина находится где-то около 25 лет. Только за последние 20—30 лет срок обучения увеличился на два-три года. Дальнейшее увеличение продолжительности образования вряд ли может быть желательным, если считать, что максимум производительности

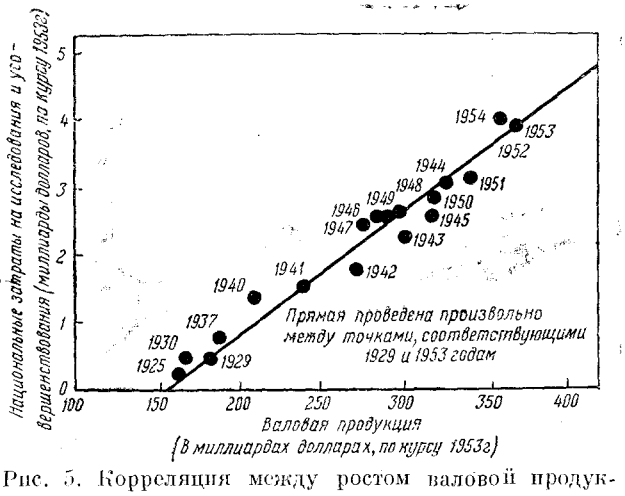


Рис. 5. Корреляция между ростом валовой продукции и затратами на исследования и усовершенствования в США [5].

труда научного работника, измеряемый числом публикуемых работ *), приходится на возраст около 35 лет, как это показано **) на рис. 6 [10, 11].

Развитие информационных машин должно будет коренным образом изменить систему образования. Учащиеся будут освобождены от необходимости запоминать колоссальное количество фактического материала, который сейчас отягчает сознание каждого исследователя.

До последнего времени задача информации в науке и технике ограничивалась преодолением пространственного, временного и языкового

*) Основываясь на статистических данных, Шокли [11] считает, что число публикуемых работ может быть мерой творческой активности научных работников. Имеет место нормально-логарифмическое распределение научных работников по числу публикуемых работ. Логарифм числа публикуемых работ рассматривается как характеристика интеллектуальных способностей работника. Творческая продуктивность работников, работающих в одной и той же лаборатории, различается в десятки раз. Способность к творческой работе варьирует в значительно более широких пределах, чем способности, которые могут быть испытаны при помощи обычных психофизиологических тестов. Этому интересному явлению Шокли дает следующее объяснение: если, например, для создания научной концепции необходимо обладать способностью к ассоциированию четырех идей, то, как это следует из комбинаторики, работник, обладающий способностью к ассоциированию шести идей, будет иметь в 15 раз больше потенциальных возможностей для творческой работы, чем работник, умеющий ассоциировать только четыре идеи. Небольшая вариация в способностях к ассоциированию идей приводит к очень большой разнице в творческих возможностях. Не все выводы Шокли представляются достаточно убедительными. Тем не менее идея применения статистических методов для изучения процессов творческой деятельности безусловно заслуживает внимания.

**) Здесь имеются в виду публикации по творческим работам.

барьеров. В настоящее время, в связи с экспоненциальным характером развития науки, возник новый барьер, обусловленный тем, что невооруженное сознание человека не может непосредственно воспринимать и перерабатывать современные мощные потоки информации. Ниже будут рассмотрены некоторые попытки преодоления этого барьера.

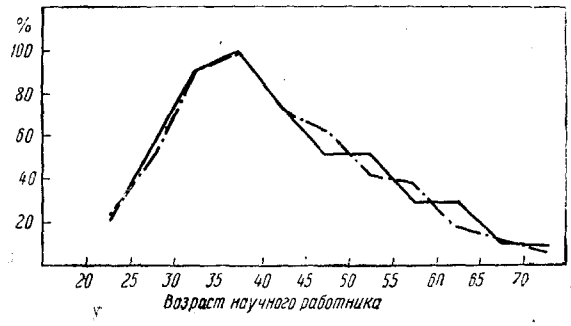


Рис. 6. Распределение числа публикуемых работ по возрасту научного работника. Сплошная линия относится к данным по 14 различным странам, не включая СССР, Англию, Францию, Германию и США. Прерывистая линия относится к данным по США[10, 11].

§ 2. МАШИННО-СТАТИСТИЧЕСКИЕ МЕТОДЫ СВЕРТЫВАНИЯ ТЕКСТА СТАТЕЙ

Машинно-статистический метод реферирования статей, предложенный в [12], может служить примером попытки механизировать один из весьма трудоемких и утомительных процессов интеллектуальной деятельности. Актуальность постановки этой задачи станет очевидной, если учесть, что Институт научной и технической информации АН СССР в настоящее время пользуется услугами 12 000 внештатных референтов. Через 100—150 лет число внештатных референтов должно будет достигнуть 12 000 000, если сохранится экспоненциальный рост числа публикаций и не будут найдены механические способы для обработки и передачи информации.

Машинно-статистический метод реферирования основан на изучении частоты появления слов в статье и их распределения. Среди множества слов, которыми пользуется автор статьи, необходимо найти те, которые являются «значимыми» для передачи данной информации. Предполагается, что значимость слов определяется частотой их появления в реферируемой работе. Затем необходимо выделить значимые фразы, которые могли бы наилучшим образом представлять содержание статьи. Значимость фраз определяется тем, насколько тесно ассоциированы в ней значимые слова.

Практически реферирование может производиться следующим образом: текст статьи с помощью того или иного читающего устройства переносится на магнитную ленту и вводится в вычислительную машину. Машине задается следующая программа: прочесть закодированный текст и выделить индивидуальные слова, отметив положение каждого слова в статье и положение каждой фразы и каждого параграфа, в которых встречаются данные слова. Одновременно машина должна отбросить слова общего характера: местоимения, предлоги, артикли и пр., создающие «шум». Оставшиеся слова располагаются в алфавитном порядке и машина идентифицирует те из них, которые служат для обозначения одного и того же понятия. Эта операция производится путем побуквенного сличения слов. Если в двух словах оказывается по шесть одинаковых букв, то они признаются неразличимыми. Так, например, слова: differ, differentiate, different, differently, difference и differential рассматриваются как символы, служащие для обозначения одного и того же понятия. Такой способ идентификации слов не является безупречным — ошибка идентификации оценивается в 5%; она не может существенно исказить результаты статистического анализа.

Наконец, после проведения всех указанных выше операций, подсчитывается частота появления слов. Значимыми признаются слова, частота появления которых оказывается выше некоторой величины, выбранной на основании серии предварительных экспериментов.

Значимость фраз устанавливается следующим образом: в каждой фразе отмечается положение значимых слов и участки фразы, содержащие значимые слова, берутся в квадратные скобки, как это показано на рис. 7. Затем подсчитывается число значимых слов в квадратных скобках, это число возводится в квадрат и делится на общее число слов, содержащихся в квадратных скобках. Полученное таким образом частное рассматривается как «вес», характеризующий значимость фразы. Реферат составляется путем механического соединения фраз, обладающих наибольшими весами.

Приведенный выше способ оценки веса фразы может быть обоснован, исходя из известного в семантике *) соотношения, согласно которому множественность значения слова, а следовательно, и его смысловой вес, пропорциональны корню квадратному из частоты его появления [13].

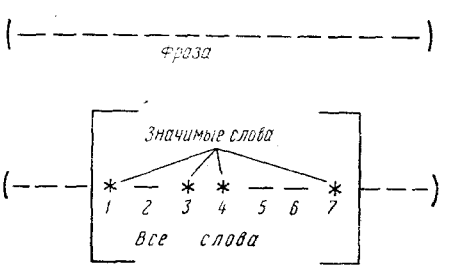


Рис. 7. Оценка статистического веса фразы. Часть фразы, содержащая значимые слова, ограничена квадратными скобками. Вес фразы равен квадрату числа значимых слов, деленному на общее число слов в квадратных скобках: $p=\frac{4^2}{7}=2{,}3$ [12].

Первые опыты машинно-статистического реферирования дали положительные результаты.

Интересно отметить, что машинно-статистическое реферирование можно рассматривать как моделирование одного из процессов реферативной работы. Если редактору реферативного журнала попадется статья, написанная на каком-нибудь труднодоступном языке, и в этой статье нет графиков, формул и прочих общедоступных информационных символов, то для получения некоторого представления о содержании статьи, редактор, просматривая текст, выбирает фразы с наиболее часто встречающимися словами и затем переводит их дословно, пользуясь словарем.

Очевидным недостатком реферата, написанного на основании статистического анализа текста, является его некоторая догматичность. Читателю предлагаются отдельные, не связанные между собой фразы, вырванные из текста. В то же время бесспорным преимуществом такого реферата является его объективность. Общепринятый сейчас процесс реферирования является своего рода искусством. Рефераты, написанные разными референтами-профессионалами, часто существенно отличаются друг от друга. Референты неизбежно вносят известный элемент субъективности в свою работу, обусловленный особенностями их научного мировоззрения, селективностью их интересов и пр. В то же время рефераты, написанные авторами статей, часто оказываются совершенно неприемлемыми, так как они не обладают искусством свертывания информации. Неточность в передаче содержания статьи при помощи рефератов особенно бросается в глаза в небольших рефератах-аннотациях, для составления которых, по-видимому, особенно эффективными могут быть машинно-статистические методы.

*) Лингвистическая семантика—дисциплина, занимающаяся изучением смыслового значения слов. Задачи лингвистической семантики впервые были сформулированы М. Бреаль (M. Bréal) в 1883 г.[13]. Интенсивное развитие этой дисциплины началось лишь за последние годы. Слово семантика происходит от греческого глагола σημαινω—имеет значение.

Статистические методы могут также применяться для машинного индексирования текста. Здесь может быть предложено несколько вариантов. Самый простой из них — выписывание наиболее часто встречающихся слов — вряд ли может быть признан удовлетворительным. Содержание статьи не может быть закодировано набором отдельных слов, так как научные понятия чаще всего определяются не отдельными словами, а их сочетаниями, создающими семантическое поле. Например, существительное «полоса» в сочетании с различными определяющими словами может служить для обозначения множества понятий: «оптическая полоса поглощения», «энергетическая полоса» и т. д. Поэтому работа машины должна программироваться так, чтобы она выписывала наиболее часто встречающиеся существительные и связанные с ними определяющие слова. Интересный опыт в этом отношении был проведен в [14], где было предложено использовать при программировании некоторые особенности английского языка. В качестве единицы индексирования были выбраны «предложные словосочетания», которые по определению являются единицами выражения мысли, состоящими из предлога, сказуемого или местоимения и связанных с ним определений. Длина «предложного словосочетания» варьирует от двух до семи слов и в среднем равна четырем словам. Машине задается такая программа, чтобы она идентифицировала все предлоги (их число не превосходит 50) и выписывала первые четыре слова, следующие за предлогом, если раньше не встретится новый предлог, или пунктуация. Например, в предложении «Within *the scope* of *natural English language*, an infinite number of *different sentence structures* is possible» машина выпишет все напечатанное курсивом. Сочетание этого приема с оценкой частоты появления слов в тексте дает возможность выписать единицы индексирования, так как это показано в таблице I для первых трех слов,

Таблица I

Пример автоматической координации терминов как свойства «предложного словосочетания»

Electron	Electron energy level Electron energy Valence electrons Electron waves Free electrons
Energy	Energy band structure Energy gap Energy spectrum Discrete energy levels Forbidden energy region Kinetic energy Energy curves
Band	Band theory Conduction band Valence band Energy band structure

встретившихся с максимальной частотой при индексировании одной из статей, посвященной физике твердого тела. Программа составляется так, чтобы для индексирования было отобрано из текста статьи 0,5% слов.

Машинно-статистическое индексирование моделирует процесс просматривания текста, обычно предшествующий чтению статьи. Прежде чем

принять решение о целесообразности прочтения статьи, научный работник сканирует ее глазами, читая отдельные короткие предложные словосочетания (которые можно охватить одним взглядом) с наиболее часто встречающимися словами.

Машинно-статистические методы свертывания текста пока еще не вышли за пределы экспериментирования. Первые опыты, проведенные в этом направлении, бесспорно указывают на перспективность исследований в этом направлении. Нужно отметить, что все эксперименты проводились на серийных вычислительных машинах IBM 704 и IBM 650.

Это новое направление техники информационной службы базируется на статистическом изучении семантических и морфологических закономерностей языка. Вероятностные методы исследования уже давно привлекали внимание лингвистов и только в самое последнее время эти методы стали находить практическое применение, что, по-видимому, поведет и к дальнейшему стимулированию теоретических работ в этом направлении.

§ 3. ДОКУМЕНТАЦИЯ ПРИ ПОМОЩИ ПЕРФОРИРОВАННЫХ КАРТ

Обилие публикаций заставляет стремиться к разработке таких систем документации, которые бы дали возможность быстро получать исчерпывающую информацию по большому числу разумно поставленных вопросов.

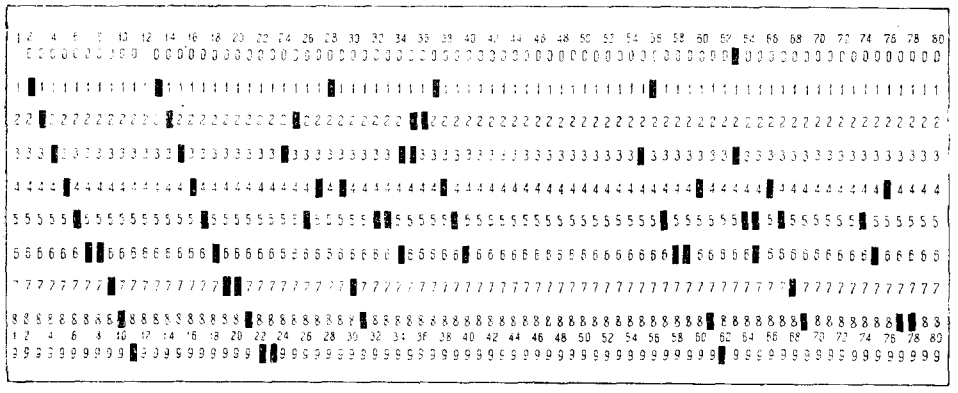


Рис. 8. Стандартная перфокарта с внутренней перфорацией.

Такая система документации возможна только при помощи таблиц с множеством входов. Многомерные таблицы могут быть построены при помощи перфокарт. Каждую перфокарту можно рассматривать как точку многомерного пространства, а стопку перфокарт как таблицу в многомерном пространстве с множеством входов.

На рис. 8 показана стандартная перфокарта с внутренней перфорацией. Эта карта, размером 83×188 *мм* имеет 80 колонок и 10 горизонтальных рядов, перенумерованных числами от 0 до 10. Кроме того, на перфокарте могут еще перфорироваться два верхних не переномерованных ряда. Информация на перфокарты заносится с помощью заранее разработанных кодов. Сортировка перфокарт по заданным признакам производится с помощью специальных машин со скоростью 250—650 штук в минуту.

В [15] подробно рассмотрено применение перфокарт для документации в молекулярной спектроскопии. В левой части перфокарты кодируются положения полос поглощения. Например, если в третьей колонке пробита цифра 4, то это означает наличие полосы поглощения при 3,4 μ. На карту

наносятся все полосы поглощения, которые на 20% и больше отличаются от фона. В правой части перфокарты кодируется информация о химическом составе молекул, ее строении и некоторых физических свойствах соединения. Например, в колонке № 44 кодируются сведения о фрагментах структуры, содержащих азот, в колонке № 48 — сведения о фрагментах, содержащих азот — кислород и т. д. С помощью перфокарт можно решить ряд информационных задач: устанавливать корреляции между спектрами поглощения и строением молекул, находить группу веществ, обладающих набором заданных спектральных признаков или группу спектров, отвечающих заданным элементам строения молекул, устанавливать состав неизвестных смесей по их спектрам и т. д. В [16] описывается применение перфокарт с внутренней перфорацией для идентификации неизвестных веществ, находящихся в порошкообразном состоянии, с помощью рентгеноструктурного анализа. В левой части карты кодируются межплоскостные расстояния, соответствующие дифракционным максимумам, в правой части — информация о химическом составе соединения.

Своеобразной разновидностью перфокарт является Eastman Kodak Minicard (США) (рис. 9) и французская система Filmorex System, в которых информация кодируется на пленке по двоичной системе в виде черных и белых пятен. Вместе с кодированной информацией на карте помещается реферат, написанный обычным образом, рисунки, диаграммы и пр. Фотоэлектронный селектор в Filmorex System прочитывает и выбирает карточки со скоростью 36 000 штук в час; в системе Eastman Kodak Minicard сортировка карт производится со скоростью 60 000 в час.

В приведенных выше примерах каждая перфокарта служила для документации информации, относящейся к какому-либо одному веществу. В некоторых системах перфокарта используется для документации, относящейся к какому-нибудь одному признаку или свойству. Например, в [17] описывается система документации для ядерной физики, в которой каждая карта используется для документации сведений об изотопах, обладающих каким-нибудь одним признаком (стабильность, период полураспада, характер распада, естественная радиоактивность и т. д.). В этом случае используются две серии карт, различным образом окрашенных,— одна для легких, другая для тяжелых ядер. Каждый изотоп занимает на карте определенное место, которое перфорируется тогда, когда изотоп обладает тем свойством, для индексирования которого служит карта. На рис. 10 показана перфокарта с перфорированными участками, соответствующими всем изотопам с числом нейтронов больше 77, для которых период полураспада более 100 лет. Такая система документации отличается крайней простотой, она удобна также тем, что коллекция карт легко может быть пополнена для любого нового свойства, если появится необходимость введения его в систему индексирования.

Широкое применение получила так называемая «Peek-a-Boo»-система*), в которой каждая карта служит для индексирования какого-нибудь одного понятия. На карте размером 20×30 *см* может быть перфорировано 1800 маленьких отверстий. Перфорируются на карте отверстия, соответствующие номерам рефератов, содержащих термин, для индексирования которого служит данная карточка. Если число рефератов больше 1800, то вводится вторая серия карт и т. д. Рефераты составляются в телеграфном стиле и содержащаяся в них информация переносится на перфокарты с помощью заранее составленного словаря, состоящего иногда из нескольких тысяч терминов с перекрестными ссылками для синонимов. При поисках информации отбираются карточки, соответствующие сочетанию понятий, с

*) Дословный перевод: игра в прятки.

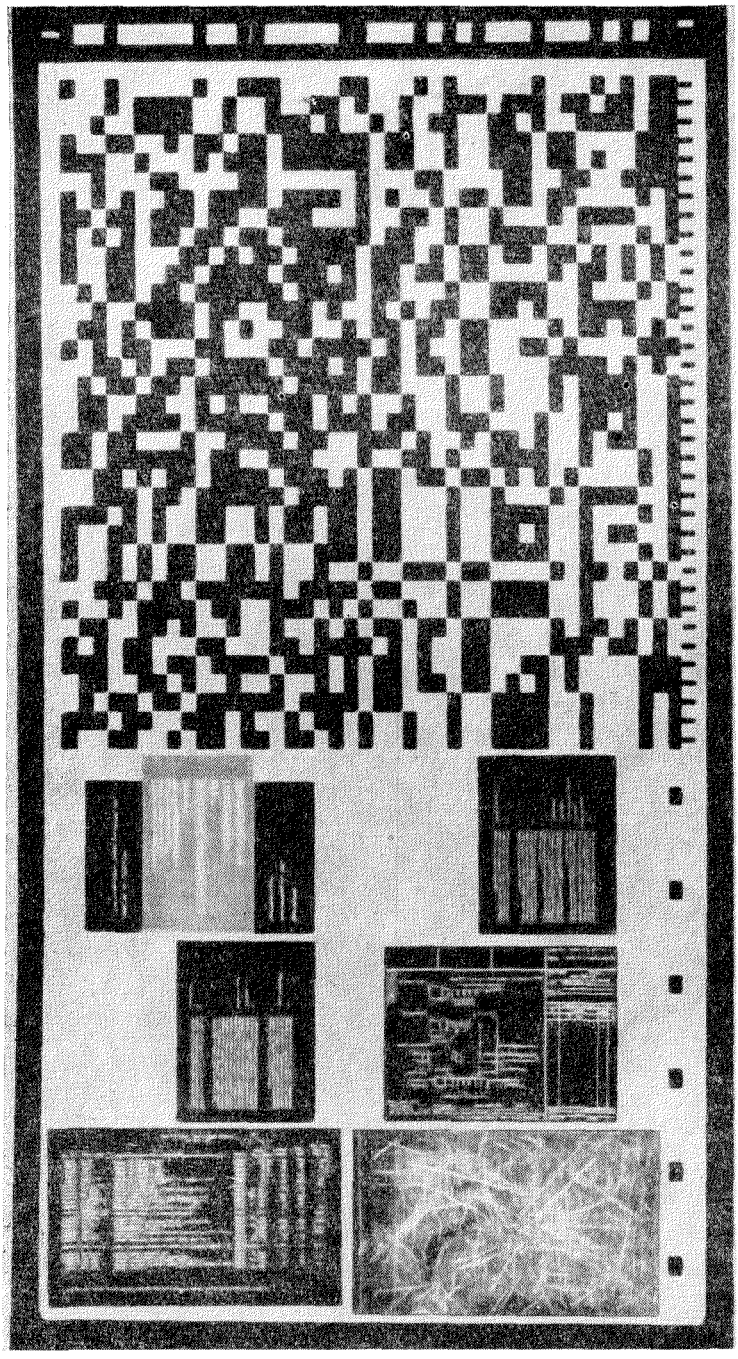


Рис. 9. Миникард системы Кодак, увеличено 7 ×. На верхней части пленки нанесена кодированная информация, на нижней части—информация, представленная обычным образом (реферат, чертежи, карта и т. д.).

помощью которых задается вопрос. Отобранные карточки просматриваются на просвет или сканируются световым лучом. Отверстия, общие всем карточкам, указывают номера рефератов, содержащие информацию по поставленному вопросу. В Национальном бюро стандартов (США) эта система

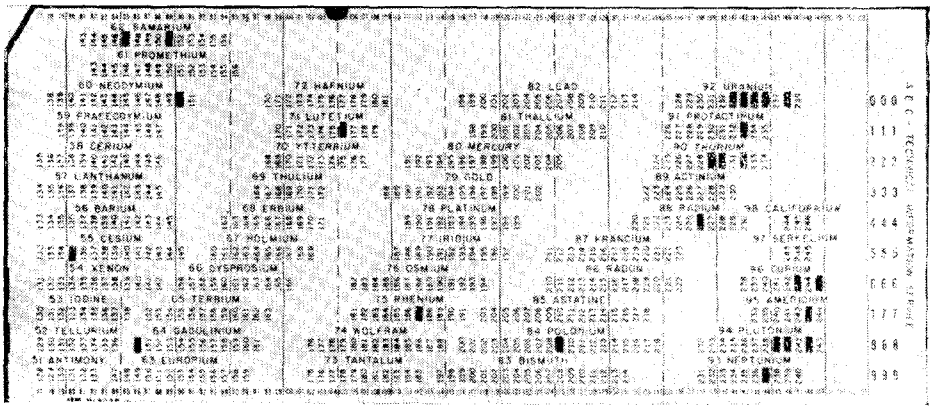


Рис. 10. Перфокарта для документации сведений по ядерной физике. На карте нанесены изотопы тяжелых ядер (с числом нейтронов больше 77) и перфорированы участки, соответствующие тем изотопам, для которых период полураспада больше 100 лет [17].

документации применяется для индексирования сведений, относящихся к приборам и измерениям.

Перфокартами удобно пользоваться для организации внутрилабораторной информации. Например, в аналитической лаборатории фирмы Esso Standart Oil Co[18] используются перфокарты, показанные на рис. 11. При

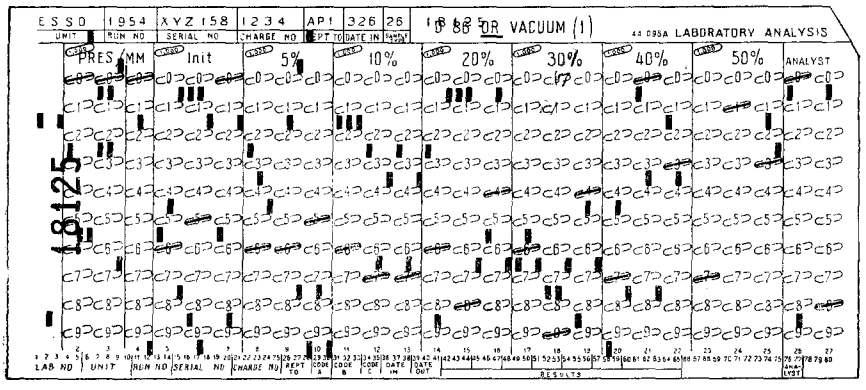


Рис. 11. Один из образцов перфокарт, используемых для нанесения и передачи информации в больших аналитических лабораториях. Аналитик наносит информацию на карту, зачеркивая карандашом участки в соответствии с кодом. Затем на карте соответствующие участки автоматически перфорируются [18].

поступлении пробы в лабораторию на перфокарте кодируется: описание образца, его номер, дата поступления и группа аналитиков, ответственная за выполнение анализов. Непосредственно на рабочем месте аналитик наносит на перфокарту результаты анализа, условия выполнения анализа, точность и пр. Заполнение перфокарты аналитиком производится путем зачеркивания карандашом, дающим токопроводящую черту, переномеро-

ванных участков, расположенных в 26 колонках. Такая система заполнения перфокарты дает возможность избавиться от всех предварительных «черновых» записей, отнимающих много времени. Заполненная перфокарта поступает в табуляторную комнату, где перфорируется на специальной машине, имеющей щупы с контактными щеточками для обнаружения зачеркнутых участков. В табуляторной комнате с помощью перфокарт печатают всякого рода сводки, необходимые для управления большой лабораторией. С помощью перфокарт легко найти анализ по любому признаку: номеру пробы, описанию пробы или условий проведения анализа, результатам анализа и т. д. Такая организация работы облегчает последующую статистическую обработку архивного материала, позволяет легко следить за точностью и правильностью работы лаборатории при помощи

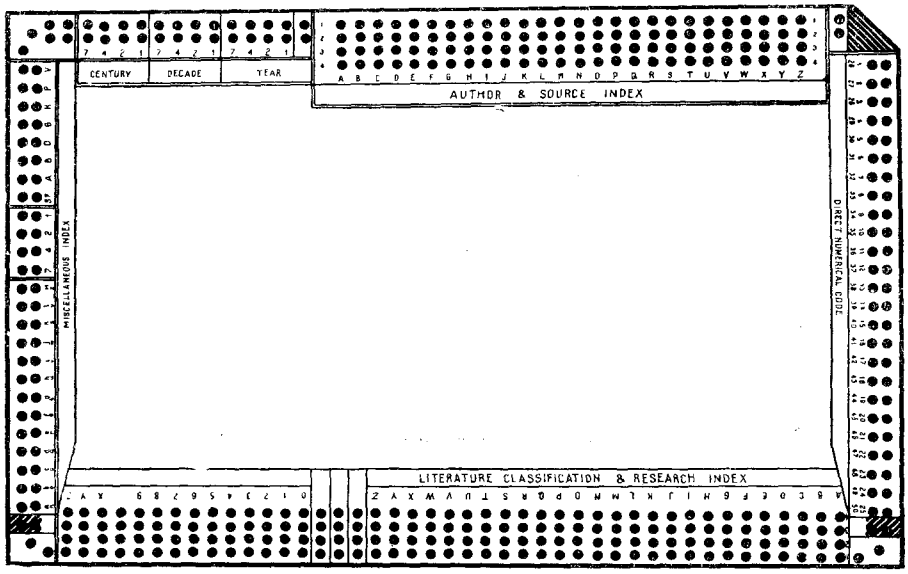


Рис. 12. Один из образцов перфокарт с внешней перфорацией. В средней (неперфорированной) части может наноситься некодированная информация—текст реферата, графики и пр.[19].

повторного анализа проб и стандартных образцов, которые направляются для анализа вместе с текущими пробами и т. д.

Наряду с перфокартами для внутренней перфорации широко используются также перфокарты с внешней перфорацией. На рис. 12 показана одна из таких перфокарт. Существуют системы перфокарт, в которых имеется только два ряда перфораций, и системы, в которых имеется по 10 рядов перфораций с одной или с двух сторон и т. д. Кодированная информация наносится на перфокарты путем прорезывания лунок, соединяющих отверстия так, как это показано на рис. 13. Сортировка карт производится обычно вручную, при помощи спиц: карточки с вырезанными лунками извлекаются из общей стопки. В некоторых случаях применяются простые механические приспособления, позволяющие довести скорость сортировки до 20 000 штук в час. Кроме кодированной информации на такие карты может наноситься также обычная информация: графики, рисунки, текст реферата и т. д. Эта система документации используется обычно в тех случаях, когда требуется заполнить не более 10 000 карт.

Перфокарты с внешней информацией применяются для документации молекулярных спектров в системе DMS (Великобритания, ФРГ)[15],

в масс-спектроскопии (кодируется молекулярный вес, температура кипения, элементы, входящие в соединения, ионные массы), в ядерной физике, когда на одной карте хотят поместить всю информацию о свойствах того или иного изотопа.

Кроме перечисленных выше, можно указать еще следующие области применения перфокарт в практике зарубежных работ: 1) Биология, где перфокарты применяются больше, чем в какой-нибудь другой области. В качестве примера укажем, что Dow Chemical Co для координации работы биологов и химиков использует систему перфокарт с внутренней перфорацией. С помощью специального кода на перфокарты занесены результаты 75 различных испытаний для 11 000 веществ. Для хранения этой информации потребовалось около 200 000 перфокарт. 2) Астрономия. Ликская обсерватория Калифорнийского университета при составлении каталога двойных звезд использует в качестве вспомогательного средства перфокарты с внутренней перфорацией. Все измерения двойных звезд для северной полусферы неба, появившиеся в печати, или имеющиеся в рукописях, занесены более чем на 80 000 перфокарт. 3) Метеорология. В 1948 г. Британская метеорологическая служба начала индексировать на перфокартах с внутренней перфорацией данные о верхних слоях атмосферы. Обсерватории наносят наблюдения на перфокарты, которые затем посылаются в центральную организацию для машинной сортировки, статистической обработки, табулирования и пр. Было подсчитано, что для документации всей информации, поступающей с Британских островов, потребуется в год более 130 000 перфокарт. 4) Геология. В этом случае перфорации соответствуют геологическим объектам и эпохам, географическим районам, дате составления отчета или публикации. 5) Неорганическая химия. Гмелиновский институт неорганической химии широко использует перфокарты для подготовки своих справочных изданий. Перфокарты с внешней перфорацией применяются для кодирования статей по авторам и для индексирования материала по различным частным вопросам (если в данной области число работ находится в пределах 500—5000). Система «Peek-a-Boo» используется для кодирования неорганических комплексов. Перфокарты с внутренней перфорацией для машинной сортировки используются при кодировании сведений о составе соединений и минералов. Пользуясь этой системой, можно, например, отобрать все минералы, содержащие германий, или все неорганические соединения, содержащие фосфаты и т. п. 6) Органическая химия. Этот вопрос будет подробно рассмотрен в следующем разделе. 7) Библиотечное дело. Перфокарты употребляются как вспомогательные средства при подготовке материала для каталогизации, для регистрации движения литературы при межбиблиотечном обмене и передачи сведений о поступивших книгах в смежные библиотеки, статистического учета и пр. 8) Библиографические перфокарты. В центральной части перфокарты печатается реферат, по краям карты кодируется содержание реферата и библиографические данные (первые буквы фамилии автора, название журнала, дата выхода и пр.). Такие библиографические системы употребляются в самых разнообразных областях науки и техники.

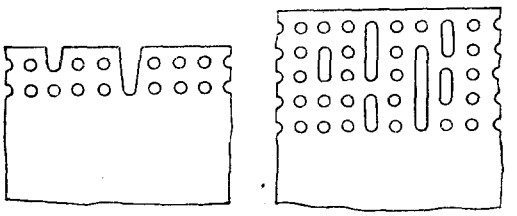


Рис. 13. Нанесение кодированной информации на перфокарты с внешней перфорацией путем прорезания лунок различной формы.

Подробное описание различных систем перфокарт, техники, работы с ними, теории кодирования и области их применения дано в монографии, изданной в 1958 г. под редакцией Кейси, Перри, Берри и Кента. Библиография, приведенная в книге, кончается 677-м наименованием [19]. В первом издании этой книги, вышедшем в 1951 г., библиография состояла всего из 276 наименований; такой рост числа публикаций за истекшие 7 лет свидетельствует о том большом значении, которое приобрели перфокарты в информационной службе.

С применением перфокарт имели место и эксперименты с отрицательным результатом. В конце 1947 г. библиотека Исследовательского центра по атомной энергии в Харуэлле начала индексирование литературы по широкому кругу вопросов на перфокартах с внутренней перфорацией. После двух лет экспериментирования эта система документации была оставлена [20,21]. Здесь встретились, с одной стороны, трудности технического характера, обусловленные недостаточно хорошей организацией работы, а с другой стороны, и трудности принципиального характера. Научный работник при информационных поисках не всегда может четко сформулировать вопрос, особенно, если поиски ведутся в областях,смежных с его специальностью. Визуальное просматривание библиографических карточек, содержащих рефераты или аннотации, часто оказывается частью творческого процесса. При таких поисках исследователь может найти ответ на вопрос, который не был поставлен; при машинных поисках информации с помощью перфокарт этого сделать нельзя [22]. Применение перфокарт как средства передачи информации, по-видимому, ограничивается отдельными частными задачами, возникающими там, где есть большое количество однотипного материала, при документации которого заранее можно сформулировать стратегию поисков.

§ 4. ДОКУМЕНТАЦИЯ В ОБЛАСТИ ХИМИИ

Успехи, достигнутые в настоящее время в отношении механизации поисков некоторых важнейших видов химической информации, представляют несомненный интерес как для всех смежных наук, имеющих дело в той или иной степени с исследованием свойств веществ, так и для общей теории методов документации. Для решения некоторых информационных задач из области химии были впервые применены современные быстродействующие электронные вычислительные машины.

Необходимость в применении сложных вычислительных устройств здесь была обусловлена, с одной стороны, громадным количеством фактического материала, а с другой стороны тем, что при решении одной из основных информационных проблем химии (поисках соединений по структурным признакам) приходится иметь дело с задачей нелинейного характера.

Среди всех отраслей естественных наук химия занимает одно из первых мест по количеству накопленной информации. Такое положение объясняется прежде всего преимущественно эмпирическим характером химических сведений, т. е. отсутствием достаточно надежных дедуктивных правил, на основе которых можно было бы получить нужные нам сведения из некоторых сравнительно немногочисленных основных фактов. Это эквивалентно отсутствию достаточно эффективных средств свертывания химической информации и приводит в конечном счете к тому, что в химии критерию целесообразности хранения удовлетворяют сведения, полученные в результате менее трудоемкой исследовательской работы, чем это имеет место, например, в различных отраслях физики.

Сказанное легко иллюстрировать следующим примером. Если произвести физический эксперимент, состоящий в измерении характеристик

некоторой электрической схемы, собранной из определенного числа известных нам элементов такого типа, как, например, сопротивления, конденсаторы и т. д., то очевидно, что вопрос о целесообразности охранения полученных при этом данных даже не возникает. Все результаты измерений могут быть предсказаны на основе простых физических законов. В отличие от этого, химик-органик, например путем смешения какого-либо из многочисленных органических оснований с каким-либо из не менее многочисленных органических кислот, после ряда несложных операций может получить новое химическое соединение — соль. Причем важно отметить, что все физические константы полученной соли, начиная с ее плотности, точки плавления, коэффициента рефракции и т. д. безусловно представляют интерес с точки зрения их хранения и распространения, поскольку пока нет практически никакой возможности вывести эти сведения из значений констант исходных веществ. Даже простой процесс смешения двух жидкофазных химических соединений будет характеризоваться целым рядом таких величин, как, например, теплота смешения, объемный коэффициент сжатия, упругость паров над смесью и т. д., которые также не могут быть вычислены; в силу этого обстоятельства данный простой процесс может явиться источником большого количества научной информации, предназначаемой для хранения и распространения *). Можно еще отметить, что при смещении нескольких компонентов, ни один из которых не обладает инсектицидными свойствами, не исключена возможность получения полезного инсектицидного состава.

Наиболее убедительным свидетельством степени насыщенности химии (и, в частности, прежде всего органической химии) конкретными «несвертываемыми» **) сведениями служит само число индивидуальных объектов ее исследования, т. е. число описанных в настоящее время химических соединений, составляющее не менее 600 000 [23]. Это число возрастает быстрыми темпами вследствие появления в химической литературе ежегодно сведений о получении не менее 20 000 новых соединений. Еще более многочисленны процессы взаимного превращения химических соединений друг в друга, т. е. особенно велико множество известных химических реакций, число которых достигает многих миллионов.

Необходимо отметить, что эти внушительные цифры отнюдь не свидетельствуют, как можно было бы думать, о хорошей разработанности соответствующих областей знаний. Отвлекаясь от технических трудностей, с которыми приходится сталкиваться при получении макромолекулярных соединений в чистом виде, можно утверждать, что множество всевозможных химических соединений бесконечно хотя бы потому, что число атомов в молекулах соединений практически ничем не ограничено. Однако даже если иметь в виду только молекулы, содержащие сравнительно небольшое число атомов, то нетрудно убедиться, что число известных в настоящее время химических соединений данного типа составляет незначительную долю общего числа подобных соединений, которые должны быть способными к существованию. Например, общее число соединений, состав молекулы которых описывается брутто-формулой $C_{11}H_{10}N_2O_2$, равняется 125 (по данным формульного указателя к справочнику Бейльштейна [24]), в то время как по самым скромным оценкам ожидаемое число соединений этой брутто-формулы составляет много десятков тысяч.

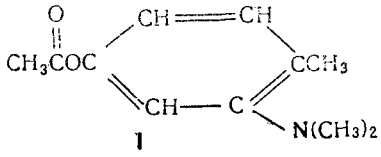
---

*) Мы не рассматриваем здесь вопрос о ценности данных видов информации.

**) Здесь имеется в виду несвертываемость информации в классическом понимании, т. е. невозможность представить наблюдаемые явления с помощью функциональных связей. Вопрос о применении методов теории вероятностей и математической статистики для свертывания информации будет рассмотрен в следующем параграфе.

Большой объем сведений, с которыми приходится иметь дело в химии, послужил причиной того, что в этой отрасли науки на сравнительно раннем этапе развития встал вопрос о разработке особых приемов и средств, облегчающих накопление и передачу информации, что и привело к основанию в 1830 г. первого в истории науки реферативного журнала «Chemisches Zentralblatt». Это же обстоятельство послужило стимулом развития системы указателей химических реферативных журналов в значительной мере более подробных, чем это принято в других реферативных журналах. Например, объем ежегодных указателей к журналу Chemical Abstracts составляет около 20% от объема всех выпусков журнала за год, что по количеству печатных листов представляет собой примерно в 3—4 раза больше, чем соответствующий объем указателей журнала Physics Abstracts.

Благодаря совершенной системе указателей сравнительно легко решается важная одномерная задача поиска информации об определенном заданном химическом соединении; эта задача имеет практическое значение



I

Дайсон: B6.C.[X.C2],4.N[C]2,2

Виссвессер: 1N1 R1eOV

ГКД: J¹C³OC³MMCJ¹C³N³J¹J¹MX4

Рис. 14. Примеры линейных шифров структурной формулы органического соединения I [25].

как для химиков, так и для широкой категории исследователей, занимающихся изучением физико-химических, физических, биологических и других свойств веществ.

Трудности, которые все же встречаются при решении этой задачи, связаны с отсутствием в химии общепринятой и однозначной системы номенклатуры соединений и в особенности органических соединений. Наряду с названиями, характеризующими тем или иным способом строение молекул соединений, в химии используется и много различных исторически сложившихся названий, не связанных со структурой соединений, причем для одного и того же соединения разные авторы в различных местах пользуются самыми разнообразными названиями. С другой стороны, индивидуальность большинства органических соединений вполне однозначно задается структурной формулой, построенной в соответствии с классическими представлениями о валентности или, в случае необходимости, дополнениями, учитывающими стереохимические представления. Поэтому, естественно, возникла мысль об использовании общепринятого и физически осмысленного языка структурных формул для целей номенклатуры соединений. Для этого необходимо было построить способ однозначной записи структурных формул в виде линейных последовательностей знаков, т. е. в виде так называемых линейных шифров. За последнее десятилетие было предложено много различных систем шифровки структурных формул [25], среди которых наибольшей известностью пользуются системы Дайсона [26], Виссвессера [27], а также Гордона, Кендалла и Дейвисона [28].

На рис. 14 показана структурная формула 2-диметиламино-4-ацетокситолуола I и соответствующие ей линейные шифры по этим системам.

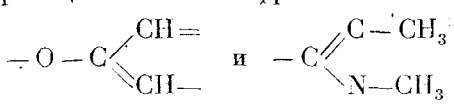
Хотя эти системы шифровки были разработаны первоначально в расчете на использование в обычных указателях, до настоящего времени они не были применены на практике для этих целей.

Если подобные шифры неудобны с точки зрения человека, то для механизации информационных поисков их использование представляется

необходимым, поскольку как в случае механически сортируемых перфокарт, так и в случае использования электронных машин существующих типов, записываемая и обрабатываемая информация должна представлять собой линейную последовательность знаков.

При помощи любой системы шифровки на перфокарты можно нанести коды соединений и наряду с этим соответствующую информацию о значениях некоторых физических констант, например, о положении максимумов в спектре поглощения соединения, а также сведения о наличии у него тех или иных видов физиологической активности и т. д. После этого путем механической сортировки можно найти все соединения, обладающие определенными заданными значениями физических констант или заданным набором физиологических свойств. Значения физических констант могут быть указаны в поисковом предписании как вполне точно, так и в виде некоторых допустимых интервалов. Проведение подобных многомерных поисков, невыполнимых при помощи обычных одномерных указателей, может интересовать, кроме химиков, также весь упомянутый уже круг исследователей, имеющих дело со свойствами веществ. Наиболее важными из решаемых этим путем задач представляются задачи установления корреляции между двумя или несколькими свойствами соединений. В целом же проблематика решения всех описанных здесь задач полностью соответствует тому, что было сказано в предыдущем разделе о методах документации при помощи перфокарт в других отраслях науки.

Однако в химии приходится иметь дело с одним важнейшим видом признаков особого рода, используемых для характеристики химических соединений, проведение поисков по которым в общем случае не представляется возможным при помощи перфокарт. Речь идет о признаке присутствия в молекуле соединения определенного элемента строения или точнее некоторого заданного фрагмента структуры. Под фрагментом структуры подразумевается некоторая связная подструктура, содержащаяся в структурной формуле соединения. Примерами структурных фрагментов являются: фрагмент $>C=O$, содержащийся в $(CH_3)_2C=O$ и в вышеприведенной структуре **I** (рис. 14), содержащиеся также в **I** фрагменты



или, наконец, изображенный на рис. 15 фрагмент *а*), входящий в состав приведенной там же структуры **II**.

Как легко убедиться сопоставлением шифра фрагмента *а*) с шифром содержащей его структуры **II**, присутствие фрагмента не может быть установлено путем простого «сканирования» шифра структуры, поскольку шифр фрагмента не содержится в нем в качестве линейного вхождения. Это показывает, что для проведения поиска соединений по признаку заданного фрагмента структуры шифры соединений должны подвергаться более сложным видам логической обработки, осуществление которых возможно только при помощи электронных цифровых машин с программным управлением.

Необходимо еще отметить, что множество всевозможных структурных фрагментов, также как и множество соединений, бесконечны, причем в поисковых предписаниях могут фигурировать произвольные фрагменты структуры.

Важность задачи поиска классов соединений по фрагментам структуры объясняется тем, что один из основных приемов констатации химических закономерностей состоит в приписывании некоторым фрагментам структуры роли носителей определенных специфических видов свойств,

в частности физических свойств (например, спектральных), химических свойств, т. е. реакционной способности, или биологических свойств (например, физиологической активности *) соединений. Поэтому в процессе установления подобных закономерностей часто приходится проводить поиски классов соединений, задаваемых различными структурными признаками, для того чтобы получить подтверждение или опровержение некоторой выдвигаемой гипотезы о существовании корреляции между определенными свойствами и структурными фрагментами.

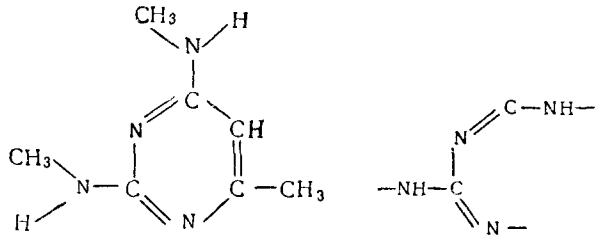


Рис. 15. Пример записи фрагмента *а*), содержащегося в структуре **II**. Приведены линейные шифры по системе ГКД.

Поскольку эти задачи имеют большое практическое значение, для их частичного решения при помощи перфокарт выбирают некоторые, считающиеся наиболее важными с точки зрения данной частной задачи, структурные фрагменты обычно сравнительно небольшой протяженности, рассматриваемые в дальнейшем в качестве рядовых признаков, присутствие или отсутствие которых кодируется на перфокарте обычным путем. Так поступают, например, в случае подробно описанных в [15] перфокарт, предназначенных для документации молекулярных спектров. Для различных других целей было предложено несколько вариантов кодировки структур соединений на перфокартах в виде линейной совокупности определенных структурных элементов [30,31,31a]. Первая из этих систем была широко испытана на практике в деятельности центра Chemical-Biological Coordination Center (США) и дала положительные результаты при установлении корреляций между биологической активностью и строением; вторая, разработанная под руководством Пича для неорганических соединений, используется в работе Gmelin Institut (ФРГ). Необходимо отметить, что коды соединений, построенные по изложенному принципу, нельзя рассматривать в качестве однозначных линейных шифров структурных формул, поскольку нескольким различным структурам (например, построенным из одних и тех же «основных» структурных единиц, но по-разному скрепленных между собой) будет соответствовать один и тот же код.

Не отрицая пользы, которую дает подобное применение перфокарт для поиска по структурным признакам, все же легко показать, хотя бы на том же примере документации молекулярных спектров, недостаточность данного подхода к решению задачи. Известно, например, что частоты колебаний карбонильной группы в разных соединениях изменяются в довольно широких пределах (от $\sim$1975 $см^{-1}$ до $\sim$1550 $см^{-1}$), поэтому между структурным фрагментом $>C=O$ и положением соответствующего

*) Например, изображенный на рис. 15 структурный фрагмент *а*) считается носителем антималярийной активности [29].

максимума может быть установлена лишь очень грубая корреляция. Корреляционные связи могут быть существенно уточнены, если рассмотреть такие фрагменты структуры, как —C(=O)CH_2C(=O)— (β-дикетоны), —C(=O)CCl< (α-галогенкетоны) или фрагменты еще большей протяженности, характеризующие такие классы соединений, как, например, «циклические кетоны 5-членные» или «β-лактамы» и т. д. Для частот колебаний C=O групп, входящих в состав каждой из этих группировок, получаются намного более точные значения [32]. Как известно, частота полос инфракрасных спектров поглощения связана с определенным взаимным расположением и взаимодействием одних структурных элементов с другими, поэтому вполне естественно, что для эффективного изучения зависимости этих спектров от строения соединений нужно уметь производить поиски по самым различным видам структурных фрагментов. С этой точки зрения несомненно большой практический интерес представляет применение быстродействующих электронных устройств для решения в общем виде задачи поисков классов соединений по структурным признакам.

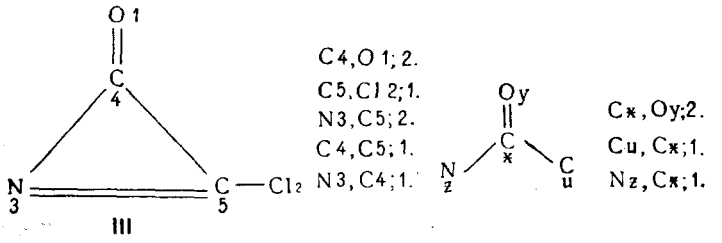


Рис. 16. Примеры записей структуры и структурного фрагмента в виде таблиц связей.

Наиболее приспособленными для работы на таких машинах оказались так называемые топологические системы шифровки структур соединений, отличительной особенностью которых служит то, что они рассматривают структурную формулу в качестве набора занумерованных атомов (или некоторых более крупных структурных блоков) и связей между ними, причем шифр включает в принципе только указания о парах связанных между собой атомов.

На рис. 16 изображена простейшая структура **III** и показан один из принципиально наиболее простых способов ее топологической записи в виде «таблицы связей», т. е. в виде пар номеров связанных между собой атомов, с указанием вида связей между ними (1—простая связь, 2 — двойная связь и т. д.). Номера приписаны атомам произвольно, при единственном условии, что все атомы структуры должны иметь разные номера. Следует отметить, что такая запись структуры, которая представлена здесь в удобной для обозрения табличной форме, без труда может быть развернута в линейную последовательность знаков. Легко убедиться, что на основе такой записи структура может быть восстановлена однозначным образом и что преобразования записи, состоящие в перемене мест двух занумерованных атомных символов, фигурирующих в одной и той же строке таблицы или в перемене мест любых двух строк таблицы не меняют структуру, получаемую на основе записи. На рис. 16 показан также структурный фрагмент, содержащийся в структуре **III**, и его запись в виде таблицы связей, причем вместо произвольных конкретных номеров атомам фрагмента приписаны буквенные обозначения x, y, z и т. д. Для того чтобы установить присутствие в некоторой структуре определенного фрагмента, заданного в поисковом предписании, достаточно найти такое преобразование записи структуры и такой набор конкретных значений x, y, z и т. д.

для которых запись фрагмента обнаруживается в качестве линейного вхождения в записи структуры. Если после выполнения всех допустимых преобразований записи структуры и приписывания всевозможных значений номерам атомов фрагмента такое побуквенное совпадение части записи структуры с записью фрагмента не выявляется, можно сделать вывод об отсутствии искомого фрагмента в рассматриваемой структуре.

Поатомный вариант топологического шифра, близкий к вышеописанному, используется в опытах Рей и Кирша[33] в Патентном Бюро США, а поблочный топологический вариант — в работах Оплера и Нортона [34, 34a], выполненных в фирме Dow Chemical Company. В обоих вариантах допускается произвольная нумерация атомов или соответственно блоков составляющих структуру.

Для каждой конкретной системы записи структур разработаны программы поиска, основывающиеся на выполнении преобразований описанного типа, но с таким расчетом, чтобы свести к минимуму число необходимых операций по перебору всех возможностей.

Оплер и Нортон записывали на магнитную ленту, используемую в качестве внешней машинной «памяти», шифры структурных формул соединений, имеющие вид последовательностей цифр вместе с порядковыми номерами, брутто-формулами и названиями соединений. Магнитная лента присоединялась к цифровой вычислительной машине IBM-704 (или 701), способной выполнять около 2,5 миллиона элементарных операций (сложений, вычитаний, сравнений) в минуту. Поисковое предписание, в виде зашифрованного таким же образом структурного фрагмента, кодировалось на перфокарте, которая также вводилась в машину, вместе со стандартной программой поисков. После нажатия пусковой кнопки машина начинала просматривать последовательно все записи на магнитной ленте и сравнивать шифр соединений с шифром структурного фрагмента, заданного в вопросе. Путем ряда преобразований, включающих в среднем для каждой просматриваемой записи около 1200 элементарных логических операций, машина устанавливала наличие или отсутствие искомого фрагмента в данном соединении и, в конечном счете, печатала порядковые номера или названия соединений, удовлетворяющих поисковому предписанию. Вся операция от момента ввода вопроса до получения напечатанного ответа в случае поиска среди 10 000 соединений длилась около одной минуты.

В аналогичной работе, проводимой в фирме Monsanto Chemical Company Уальдо и де Беккером [35,35a], машина IBM-701 используется не только для проведения операции поиска, но также и для целей частичной автоматизации шифровки структурных формул. После того, как химик записал структурную формулу на листке клетчатой бумаги, в соответствии с некоторыми стандартами, все дальнейшие операции по переводу этой записи на машинный язык и по проверке полученного кода производятся при участии только технических сотрудников — нехимиков, с существенным использованием самой машины. В этой работе на магнитных лентах накапливались сведения, взятые непосредственно из лабораторных журналов различных отделов исследовательской лаборатории. Из одного отдела поступали сведения о строении и физико-химических свойствах вновь синтезированных соединений, из другого отдела сведения об испытаниях биологической активности этих соединений. Основываясь на этом запасе сведений, машина автоматически печатает «обзорные отчеты» по такого рода вопросам как например: на какие виды физиологической активности было испытано определенное индивидуальное соединение, какие соединения дали положительные результаты при испытании на один вид физиологической активности, но вместе с тем отрицательные результаты по другому виду активности, какой вид биологической активности является общим

для всех соединений, содержащих определенный заданный фрагмент структуры и т. д. В этом случае машина печатала не только названия или порядковые номера соединений, но и их структурные формулы, правда, в несколько непривычном «квадратизированном» виде. На рис. 17 в качестве примера показаны напечатанные машиной структурные формулы фенола и ацетата калия. В этих формулах атомы водорода опущены, а цифры между атомами означают вид связи.

Первые работы по применению быстродействующих электронных вычислительных машин для поиска химической информации показали большие возможности этих устройств как в смысле высоких скоростей обработки значительных массивов информации, так и в смысле разнообразия выполняемых поисковых операций и операций по преобразованию информации. Следует иметь в виду, что одна и та же вычислительная машина может быть использована не только для самых различных видов поисковых операций,

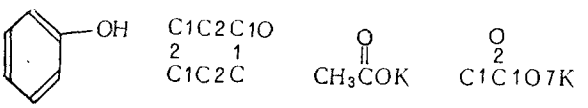


Рис. 17. Структурные формулы фенола и ацетата калия в обычной форме и в том виде, в котором они печатались машиной в опытах Уальдо и Де Беккера:
1—обозначение простой связи, *2*—двойной связи, *7*—ионной связи [35а].

но вместе с тем и для решения чисто вычислительных задач. По данным [36] около 100 больших электронных вычислительных машин эксплуатируется на предприятиях химической промышленности США; они используются для коммерческих расчетов и вместе с тем примерно на каждой второй машине решаются различные задачи научного значения, в том числе и задачи информационного поиска.

Если для экспериментальных работ в этой области вполне целесообразно применение вычислительных машин универсального типа, то для достижения результатов серьезного практического значения нужно иметь специализированные информационные машины. Главная отличительная особенность этих машин должна заключаться в наличии у них больших объемов долговременной и быстродействующей машинной памяти. Магнитные ленты, применяемые в качестве запоминающих устройств современных машин, обладают весьма существенным недостатком, обусловленным необходимостью механического движения (перемотка ленты) при считывании информации, что в значительной мере ограничивает скорость считывания. Поэтому для создания информационных машин исключительное значение имеет вопрос о разработке таких видов машинной памяти, которые были бы способны к долговременному хранению больших объемов сведений и работали бы без механического движения, на основе чисто электрических принципов считывания. Такой вид машинной памяти разработан под руководством проф. Л. И. Гутенмахера в Лаборатории электромоделирования ВИНИТИ АН СССР [37].

Наличие этой памяти наряду с современными техническими возможностями создания специализированных устройств, способных к выполнению самых различных логических операций, дает возможность ставить вопрос о создании большой информационно-логической машины для сведений о химических соединениях и химических реакциях. Кроме технических предпосылок, при этом существенное значение имеет то обстоятельство, что химические структурные формулы, легко переводимые в линейные шифры, представляют собой весьма удобный и почти универсаль-

ный язык для описания свойств и поведения химических соединений. В связи с этим представляется возможным моделирование при помощи электронных информационно-логических машин некоторых сторон «химического мышления», в частности решение такой задачи, как нахождение на основе записанных в машинной памяти «химических аналогий» правдоподобных путей синтеза еще не полученного соединения[38].

§ 5. ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТАЛЬНЫХ РАБОТ В ФОРМЕ, УДОБНОЙ ДЛЯ КОДИРОВАНИЯ

В химии, экспериментальной физике и ее технических приложениях рассмотренные выше методы документации могут быть эффективны только в том случае, если результаты измерений будут представляться в компактной форме, удобной для кодирования. Свертка информации может производиться методами современной математической статистики. Результаты измерений обычно можно интерпретировать как множество значений случайной величины; тогда громоздкие таблицы могут быть представлены компактным образом с помощью параметров распределения. Для многомерных случайных величин линейные графики могут быть заменены коэффициентами регрессии и корреляции. В сложных опытах при изучении действия и взаимодействия многих факторов результаты исследования могут быть обработаны с помощью дисперсионного анализа, и тогда громоздкий цифровой материал будет представлен с помощью дисперсий и величин, оценивающих их статистическую значимость.

Среди многих экспериментаторов до сих пор распространено неверное мнение о том, что математическая статистика применима только к очень большому цифровому материалу. Современная математическая статистика может применяться и для обработки малых выборок, состоящих всего из трех, а иногда даже из двух измерений. При этом, естественно, увеличивается элемент неопределенности в оценках измеряемых величин, который опять-таки учитывается методами статистического анализа и выражается в форме, удобной для кодирования. Возможность применять статистические методы для свертки информации в выборках любого размера позволяет представлять экспериментальный материал независимо от его объема строго стандартным образом.

В лабораторной работе часто приходится сталкиваться с тем обстоятельством, что на случайные процессы накладываются некоторые, не существенные с точки зрения экспериментатора, закономерные явления. С подобной ситуацией приходится иметь дело и при исследованиях в технике, сельском хозяйстве и пр. Эти несущественные для экспериментатора закономерности всегда могут быть нарушены при помощи специального приема—рандомизации *), позволяющего планировать эксперименты так, чтобы обеспечить случайный характер изменения всех неконтролируемых факторов.

Опыт показывает, что применение математической статистики для свертки информации становится эффективным только в том случае, если статистические методы используются не только для окончательной обработки материала, но также и для планирования эксперимента. (Выбор схемы проведения эксперимента, расположение материала, рандомизация условий эксперимента, выбор числа параллельных измерений и т. д.).

Математическая статистика начинает находить применение также для компактного представления информации, содержащейся в тех громадных архивах, которые накопились в научно-исследовательских институтах

*) От английского слова random—случайный.

и заводских лабораториях. До сих пор эти материалы, как правило, оставались неиспользованными, так как в необработанном виде в силу своей громоздкости они не могли передаваться через обычные каналы научно-технической информации.

Идея применения математической статистики для свертки информации принадлежит Р. Фишеру [39]. Известно, что основатели англо-американской статистической школы стояли на махистских позициях и, исходя из этих позиций, обосновывали задачи статистики. Теперь в связи с развитием идей кибернетики к проблеме свертки информации можно подойти с принципиально иных позиций. Свертку информации нужно рассматривать не как самоцель, а как некоторый промежуточный этап познавательного процесса. Во многих областях исследований и, в частности, в технической физике часто приходится иметь дело со столь сложной ситуацией, когда одна работа еще не позволяет проникнуть в природу изучаемого явления. Если результат исследования представить компактным образом и оценить элемент неопределенности, связанный со сложностью экспериментальных условий, то каждое исследование можно будет рассматривать как часть некоторой коллективной работы. Представленную таким образом информацию можно будет заносить в долговременную память машины и сопоставлять с результатами других работ. Такое сопоставление большого числа работ даст возможность глубже проникать в физическую сущность явлений. Эту идею можно иллюстрировать следующим примером: в эмиссионном спектральном анализе вопрос о влиянии третьих элементов рассматривается в сотнях работ. Ни одна из этих работ, взятая в отдельности, не позволяет проникнуть в сущность этого явления. Сделать какие-либо обобщения из сопоставления этих работ не представляется возможным, так как результаты исследований представлены нестандартным образом и не оценен элемент неопределенности, связанный с результатами исследований. При таком положении дел применение информационных машин не может принести существенной пользы. В память машины нельзя заносить результаты, полученные в этих работах. Единственно, что может здесь сделать машина,— выбрать по заданной тематике несопоставимые между собой работы.

В настоящее время имеется большое число руководств, посвященных статистическим методам планирования эксперимента и обработки его результатов [40,41,42,43,44,45,46]. Возникает необходимость в строгой стандартизации способов представления результатов эксперимента в соответствии с особенностями той или иной области исследования.

Одновременно в ряде случаев возникает необходимость и в стандартизации способов проведения экспериментов. Например, документация молекулярных спектров может быть эффективной только в том случае, если применяемая аппаратура, способы измерений и представления спектров, критерии для оценки чистоты веществ и пр.— достаточно строго стандартизированы. Служба информации оказывается вынужденной предъявлять свои требования экспериментаторам.

§ 6. ДЕСКРИПТОРНОЕ ИНДЕКСИРОВАНИЕ

Приведенные в разделах 3 и 4 примеры информационных поисковых систем представляют собой лишь частные случаи многочисленных различных систем документации, разработанных в основном за последние 10—15 лет. В течение этого периода как исследовательские работы по созданию новых информационных поисковых систем, так и различные формы их практического приложения приобрели весьма широкий размах. Возникла новая область теоретической и практической деятельности, которая стала

известной под названием документалистики. Целью документалистики является разработка методов наиболее эффективного использования сведений, хранимых в различных коллекциях документов. Под последними в документалистике имеются в виду всевозможные виды записей сведений, в частности, журнальные статьи, книги, отчеты, записи в лабораторных журналах и т. п. В течение последних лет документалистика все в большей мере приобретает характер самостоятельной научной дисциплины, стремящейся к разработке общей теории информационных поисковых систем с использованием математического аппарата и аппарата математической логики. Ныне издается целый ряд научных журналов, посвященных вопросам документалистики, среди которых наибольшей известностью пользуется издающийся с 1950 г. журнал American Documentation.

В настоящее время под влиянием кибернетических концепций рождается новая научная дисциплина, преследующая более широкую цель: передачу кибернетическим машинам различных видов интеллектуального труда, поддающихся алгоритмическому описанию, в частности, процессов поиска научной информации, нужной для решения определенной задачи, а также различных процессов преобразований научной информации (перевод, свертывание, выработка решений и т. д.). Необходимость в этой дисциплине диктуется, как это было показано в первом разделе статьи, экспоненциальным характером развития науки.

Поскольку для развития этой новой, не получившей еще названия дисциплины, достижения документалистики имеют существенное значение, а также учитывая самостоятельный практический интерес ее результатов, представляется целесообразным рассмотреть основные направления современной документалистики. Среди них в первую очередь следует упомянуть о так называемом дескрипторном индексировании.

Это направление документалистики, занимающееся в основном проблемой индексирования, представлено группой М. Таубе и его сотрудников [48,49,50]. Они предлагают принять так называемую координирующую систему индексации, при которой для составления каталога документов используется набор в некотором смысле произвольно выбранных единичных терминов (унитермов или дескрипторов). Дескрипторы — это понятия, каждое из которых необходимо, а все вместе достаточны для идентификации данного документа (в рамках определенной коллекции документов). Например, для доклада на тему: «Равновесный состав и термодинамические свойства горючих газов» дескрипторами будут следующие понятия: горючие вещества, газы, вычисление, топливо, импульс, давление, температура, энтропия, теплосодержание, адиабатичность[51]. Как только документ поступает, ему приписывается определенный номер, который наносится на все карточки дескрипторов, действительно необходимых для полного распознавания данного документа. Карточка для каждого дескриптора (имеющая собственный отличительный номер) делится на 10 колонок, нумеруемых цифрами от 0 до 9 включительно. На эту карточку наносятся номера всех тех документов, в которых фигурирует данный дескриптор, причем номер документа попадает в ту колонку, номер которой совпадает с его последней цифрой. Внутри каждой колонки дескрипторной карточки номера документов располагаются столбцом в порядке возрастающих номеров. Грубая схема поиска документа, идентифицированного с помощью n дескрипторов, сводится к сравнению чисел на n дескрипторных карточках для того, чтобы выявить числа, общие всем этим n карточкам. Процедура такого сравнения существенно облегчается упорядочением чисел номеров документов на дескрипторных карточках. Пример дескрипторной карточки показан в таблице II.

Таблица II

Дескрипторная карточка, содержащая номера тех работ, в которых имеется информация о спектрах

Спектры									
0	1	2	3	4	5	6	7	8	9
120 190	1951 2451 2471 6121 6421	1432 2567 4502 5662 6262	5713 6263	1064 2894 4524 5714	2885 6025	4576 6026 6106 6506 6636	1277 2877 4107 4897 5707 6247 6457 6517	1278 5728	1279 4489 5069

В настоящее время фирма Information for Industry в Вашингтоне издает периодически две серии указателей, основанных на принципе дескрипторного индексирования и известных по названиям Uniterm index, выпускаемые отдельно для патентов США в области химии и электроники. Эти указатели представляют собой собранные в виде книги (блокнотного формата) дескрипторные карточки, расположенные в алфавитном порядке унитермов (т. е. дескрипторов). Для того чтобы облегчить сравнение чисел, фигурирующих под двумя различными унитермами, в одном переплете помещены рядом два идентичных экземпляра указателя таким образом, чтобы можно было бы рассмотреть одновременно два любых разных унитерма.

Проведена работа по выяснению логических оснований метода координирующей системы индексирования. Итогом этого исследования явился вывод о том, что логикой координирующего индексирования является исчисление классов современной математической логики, которое исторически базируется на булевой алгебре свойств*). Каждый индексируемый термин порождает класс всех тех предметов, описание которых требует использования заданного термина. В таком случае отношения между терминами при координирующей индексации могут описываться соответ-

*) Алгебра Буля или, как ее иногда называют иначе, алгебра логики, является приложением математических методов к формальной логике. Алгебра логики является наукой, широко используемой сейчас в математике (в теории вероятностей), в технике (в теории электрических релейно-контактных схем слабого тока), в логике (в так называемом исчислении высказываний, где с помощью методов алгебры логики решается, например, проблема выведения всех следствий из данной системы посылок, или же проблема подыскивания гипотез, из которых могут быть получены данные высказывания). Элементарное введение в алгебру логики изложено в [47], а в советской литературе—в комментариях проф. С. А. Яновской к книге [67]. Алгебра Буля названа так по имени ее создателя—известного английского математика Джорджа Буля (1815—1864). Частично она являстся обобщением аристотелевой силлогистики, теории логических умозаключений.

Основными операциями у Буля являются:

1. Сложение, обозначавшееся знаком «+»; в вычислении классов булевской формуле $x+y$ соответствует объединение классов x и y с исключением их общей части; в исчислении высказываний—так называемая строгая дизъюнкция.

2. Умножение, обозначавшееся знаком «·»; в исчислении классов этой операции соответствует пересечение, в исчислении высказываний—конъюнкция.

3. Дополнение до единицы, обозначавшееся записью $1-x$; в исчислении классов формула $1-x$ означает дополнение к классу x; в исчислении высказываний—отрицание x.

ствующими теоремами из области исчисления классов в терминах операций объединения, пересечения и взятия дополнения.

Проведен целый ряд экспериментальных исследований по приложению идей и методов координирующей индексации к процедуре библиотечного поиска информации. Любопытные результаты в попытках такого рода описаны и обоснованы, например, в работе[51]. В ней рассматривалась задача упрощения процедуры для библиотечного поиска технической информации под углом зрения метода координирующей индексации. Для поиска информации использовалась вычислительная машина IBM-701, установленная на американской экспериментальной станции морской артиллерии в Калифорнии. Объектами поиска являлись доклады, относящиеся к проблемам развития морской артиллерии и издающиеся рядом соответствующих агентств США.

Лицо, желающее получить информацию по определенному вопросу, предлагает список ключевых слов (дескрипторов), по которым и будет вестись поиск. Цель программирования поисков на машине IBM-701 состоит в механизации процедуры поисков и в разработке ежедневной программы библиотечных поисков. Дескрипторные карточки, заготовляемые в виде перфокарт, несущих собственный номер дескриптора вместе с номерами соответствующих ему докладов, переписывались на магнитной ленте в порядке роста собственных номеров дескрипторов. Предусмотрен ввод и стирание новых дескрипторов или новых номеров докладов для каждого дескриптора. Ежедневная программа, включающая до 75 независимых поисков, предусматривает, что каждое заинтересованное лицо представляет не более 8 дескрипторов, определяющих нужный ему доклад. Эта группа дескрипторов определяет независимый поиск. Дескрипторы поисков, входящих в ежедневную программу (исключая повторения), переносятся с общей ленты на рабочую ленту. При составлении поисковой программы авторы принимали во внимание средние объемы карточек с учетом их естественного прироста в течение нескольких ближайших лет. Для записей на ленте использовалось около 9600 дескрипторов и 14 000 номеров докладов. На $^2/_3$ катушки ленты хранилось до 120 000 единиц информации. Ежемесячно присоединялось около 300 новых докладов. Число новых дескрипторов, присоединяемых каждый месяц, было достаточно малым. Библиотечная поисковая программа подразделяется на три следующих друг за другом независимых периода: 1) ввод, 2) поиск и 3) вывод. Фаза (1) требует для своего осуществления около 6 минут. Время периода (2) варьирует в зависимости от числа дескрипторов в каждом поиске; средняя продолжительность его не превышает четырех минут. Фаза (3) продолжается всего одну минуту.

§ 7. ОПЫТ ПОСТРОЕНИЯ ИСКУССТВЕННОГО ИНФОРМАЦИОННОГО ЯЗЫКА ДЛЯ МАШИННЫХ ПОИСКОВ ИНФОРМАЦИИ

Естественные языки не пригодны для машинной информации в силу своей чрезвычайной гибкости. Большое количество омонимов, синонимов и полусинонимов, зависимость смысла слов от так называемых семантических полей, которые создаются в силу сложного взаимодействия слов, зависящего от их расположения во фразе и от конструкции фразы,— все это делает машинные поиски информации на базе естественных языков невозможными или малоэффективными. Поэтому механизация информационных поисков существенно зависит от построения некоторого искусственного языка, который мы в дальнейшем будем называть информационным.

Информационный язык является средством для компактной, строго однозначной записи содержания документа в форме, приемлемой для

машинного поиска информации. Один из способов перевода с естественного языка на информационный будет рассмотрен ниже. Необходимо подчеркнуть разницу между понятиями «информационного языка» и «формализованного языка» (к последнему близко по содержанию понятие формальной системы). Хотя всякий информационный язык включает в себя элементы формализации, однако он не строится вполне формально, т. е. в нем нет аксиоматики и строго фиксированных правил формально-логического вывода. Первые призывы к созданию формализованного языка встречаются в работах Г. В. Лейбница и Р. Декарта*). Первые намеки на информационный язык восходят к 1884 г. и содержатся в трудах Г. Голлерита [54,55], в которых впервые высказана также идея о возможности применения математической логики к проблемам, связанным с информационными поисками.

Большая работа как по теоретическим основам, так и по практическому созданию и приложениям информационного языка была проведена в США Д. Перри, М. Берри, А. Кентом и их коллегами [56,57]. В [56] прежде всего ставится задача уточнения списка терминов, которая рассматривается как необходимое предварительное условие для построения информационного языка. Необходимость специального анализа терминологии диктуется, в частности, тем, что поисковые машины, основанные на электронной технике, способны отождествлять некоторые последовательности символов, но не могут проводить интерпретацию значения слов. Этапы проведения этого анализа таковы: 1) выясняется прежде всего, какой тип объекта или понятия характеризует данный термин, 2) к какой области науки он относится, 3) наконец, выполняется предварительная классификация терминов. Возможный вариант такой классификации применительно к химии мог бы, например, выглядеть так: а) машины (озонизатор, спектрограф и др.), b) процессы (абсорбция, нейтрализация и др.), c) материалы (кислота, спирт и др.), атрибуты (магнитный, уксуснокислый и др.), e) понятия (свободная энергия, энтропия и др.).

Основное требование, предъявляемое к информационному языку, сводится к тому, чтобы он давал возможность выражать существенные стороны содержания индексируемого и кодируемого документа в форме, пригодной для поиска при помощи машины. При кодировании содержания документа важно провести эту операцию так, чтобы при поисках информации наиболее существенные сведения, содержащиеся в документе, находились легче, чем второстепенные. В [56] предлагается оригинальное решение этой задачи, основанное на так называемом организованном обозначении аспектов содержания документа. Вводится компактный, основанный на мнемонике, код. Используется понятие семантических единиц (соответствующих признакам кодируемого термина). Например, семантическими единицами термина «термометр» будут следующие понятия: «предназначен для измерения», «включается в класс приборов», «подвержен действию температуры» и др. Между кодируемым термином и его семантическими единицами существуют разнообразные отношения. В [57] показано, что список таких отношений не только конечен, но и может быть сделан поразительно малым. С помощью очень небольшого числа отношений и несколько большего, но вполне обозримого числа семантических единиц в [57] представлено на информационном языке около 30 000 научных и технических терминов. Отношения между терминами и их семантическими единицами подразделяются на два вида: 1) аналитические и 2) синтетические. Под аналитиче-

*) В новейшее время пионерами построения формализованных языков в конкретных областях естественных наук выступали Р. Карнап [52] и Вуджер [53]. В [52] содержатся попытки аксиоматического изложения теории родства (в смысле юриспруденции и биологии), а в [53] представлен формализованный язык для ряда разделов биологии, широко использующий достижения математической логики и теории множеств.

ским отношением понимается отношение, существующее между терминами лишь в силу их определения или области применения. Например, в высказывании «флот состоит из кораблей» выражено аналитическое отношение («состоять из») между термином «флот» и понятием «корабли». Отношение, не являющееся аналитическим, считается синтетическим. Примеры синтетических отношений: отношение между конечными и исходными материалами какой-либо реакции, отношение кодируемого объекта к реагенту, способствующему протеканию некоторого процесса, и др. В формализации синтетических отношений заключается, по мнению авторов[56], одна из основных трудностей, возникающих на пути к созданию удовлетворительной системы информационного языка. Для символического представления семантических единиц употребляются следующие 13 заглавных букв латинского алфавита:

$$B,\ C,\ D,\ F,\ G,\ H,\ L,\ M,\ N,\ P,\ R,\ S,\ T, \tag{1}$$

а для аналитических отношений — следующие 10 букв этого же алфавита:

$$A,\ E,\ I,\ O,\ U,\ Q,\ M,\ X,\ Y,\ Z. \tag{2}$$

Семантические единицы имеют вид:

$$\mu \square \lambda\theta,$$

где μ, λ, θ—какие-нибудь буквы из ряда (1), причем две из них не обязательно различны, пустое место «□» заполняется каким-либо аналитическим отношением. Были предложены следующие аналитические отношения:

A—отношение включения элемента в класс, или класса в более широкий по объему класс. Пример: *M—TL* — сокращенное обозначение для понятия «металл»; выражение *MATL* указывает, что тот или иной элемент включается в класс металлов.

E—отношение кодируемого объекта к тому материалу, из которого он состоит. Пример: выражение *METL* указывает, что тот или иной объект состоит из металла.

I—отношение части к целому. Пример: *M—CH* сокращенное обозначение для понятия «машина»; семантическая единица *MICH* указывает на то, что кодируемый объект является частью машины.

O—отношение целого к своим частям. Пример: *S—HP* — сокращенное обозначение понятия «корабль»; семантическая единица *SOHP* используется при кодировании понятия «флот» («состоит из кораблей»).

U—предназначено для. Пример: *M—SR* — сокращенное обозначение для понятия «измерение»; семантическая единица *MUSR* указывает на то, что тот или иной кодируемый объект является измерительным прибором.

W—подвержено действию. Пример: *L—CT* — сокращенное обозначение для понятия «электричество»; семантическая единица *LWCT* указывает на то, что на кодируемый объект оказывает действие электричество. Эта семантическая единица будет использована при кодировании понятия «амперметр».

X—отсутствие какого-либо качества или характеристики. Пример: *H—DR* — сокращенное обозначение понятия «вода»; семантическая единица *HXDR* используется для кодирования понятия «ангидрид» («лишенный воды»).

Q—используется. Пример: *LQCT* — часть кода для «гальваностегии».

Y—положительная характеристика, не совпадающая ни с одним из предыдущих аналитических отношений.

Приведем пример кодирования. Код для понятия «термометра» выглядит следующим образом:

MACH MUSR RWHT 4X 002.

В этом выражении символ «4*X*» означает, что предшествующий его появлению код удовлетворяет не только содержанию понятия «термометр», но и содержанию некоторых других понятий, в частности содержанию понятия «пирометр» (код последнего будет отличаться от кода термометра отсутствием символа «4*X*»). Числовое окончание «002» служит для дальнейшей дифференциации семантического кода.

Рассмотрим еще ряд примеров кодирования по методу, изложенному в [56,57].

1. Абсорбция — *BASB* 001.

B—SB — семантическая единица, обозначающая понятие абсорбции. Символ *A* указывает на то, что здесь семантическая единица употреблена для обозначения свойства, являющегося частью этого широкого понятия.

2. Абсорбционная полоса *BWSB GARP MYPR* 98*X* 001.

B—SB, как уже указывалось, обозначает абсорбцию. Символ *W* указывает на то, что полоса возникает в результате абсорбции. *G—RP* служит для обозначения понятия «коллекция». Символ *A* указывает на то, что здесь речь идет об объекте, являющемся частью коллекции (полоса поглощения — часть спектра поглощения). *M—PR* служит для обозначения свойства вещества. Символ *Y* указывает на то, что полоса поглощения является характеристикой свойств вещества.

3. Абсорбционная башня *BUSB MACH* 005.

B—SB, как и прежде, обозначает абсорбцию, символ *U* указывает на то, что здесь абсорбция рассматривается как процесс, для протекания которого предназначен кодируемый объект. *M—CH* обозначает машину или устройство. Символ *A* указывает на то, что здесь речь идет о некотором объекте, который является частью широкого понятия — машин. Числовой суффикс «005» позволяет выделить понятие «абсорбционная башня» из широкого класса абсорбционных устройств.

Занесение информации в долговременную память машины производится так: для каждой статьи пишется реферат в так называемом телеграфном стиле, его содержание переводится затем на информационный язык при помощи семантических единиц и некоторых индексов, служащих для обозначения синтетических отношений, отображающих телеграфную грамматику. Если, например, имеется химическая работа, то важно закодировать не только участвующие в реакции вещества, но также отметить исходные и конечные продукты реакции и вещества, способствующие протеканию реакции. Для этой цели пользуются индексами: *KAY* — исходные материалы, *KWY* — полученные вещества, *KQY* — реагенты, способствующие протеканию некоторого процесса, и т. д.

Поиски могут проводиться по одному аспекту (по коду термина, коду для одной семантической единицы или по комбинации какой-либо семантической единицы с аналитическим отношением) или по ряду аспектов содержания требуемого документа. В последнем случае поиски обычно идут по логическому произведению семантических единиц, их логической сумме, логической разности, по комбинации логических сумм, произведений и разностей. Допустим, например, что нам нужно найти вещества класса *A*, обладающие одновременно свойствами *) *B* и *C* (заштрихованная область на рис. 18). В этом случае программа поисков может быть за-

*) Здесь используется алгебра логики (см. примечание на стр. 40) и ее топологическая модель (окружности на рис. 18).

несена в виде логического произведения (пересечения) $A \cdot B \cdot C$. Если же нам нужно найти вещества класса A, обладающие свойствами B и не обладающие свойством C (эта область на рис. 18 заполнена точками), то эта задача записывается в виде логической разности $A \cdot B - C$. Наконец, допустим, что мы ищем все работы, в которых для некоторого класса веществ X измерена хотя бы одна из характеристик A, B или C. Эта задача может быть записана в виде логической суммы $X \cdot A + X \cdot B + X \cdot C$. Возможна и более сложная постановка задач. Роль логико-математических взаимоотношений в многоаспектных поисках будет подробно рассмотрена ниже, в § 8.

В США в течение ряда лет ведутся широкие экспериментальные работы по разработке и изучению информационных систем, основанных на применении описанного выше языка. Работы проводились на медленно действующей машине с электромагнитными реле в качестве бинарных элементов. С начала 1959 г. возглавляемый Перри исследовательский центр Center for Documentation and Communication Research при университете Western Reserve University в Кливленде в сотрудничестве с Американским металлургическим обществом начал выдавать машинную информацию по вопросам металлургии. В долговременную память машины занесено содержание 25 000 статей по металлургии, появившихся в печати после ноября 1955 г. В последнее время начато построение быстродействующей машины марки GEL-250, предназначенной для просмотра информации, записанной на магнитной ленте со скоростью 100 000 рефератов в час. В 1958 г. Перри и Кент выпустили обширную монографию [57], которая представляет собой развернутую рабочую инструкцию по применению машинного языка в информационной службе. Первые примеры машинных поисков информации в области металлургии приведены в [58].

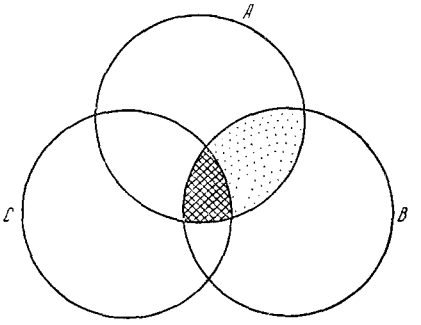


Рис. 18. Диаграмма отношений между признаками [53].

Рассмотрим вопрос о возможности применения информационных машин такого типа в творческой работе. При планировании экспериментов исследователь, на основании знакомства с литературными данными, выдвигает вероятностную (правдоподобную) гипотезу. Успех здесь определяется, с одной стороны, способностью работника к ассоциированию идей, а с другой стороны, его эрудицией и умением перерабатывать большое количество информации. В США иногда практикуется такая организация работы: руководитель исследовательскими работами собирает специалистов самого разнообразного профиля и просит их высказать свои соображения по интересующей его проблеме. Каждый из присутствующих может предлагать свое объяснение, если даже оно носит интуитивный характер и не может быть строго обосновано; единственное условие—не выдвигать какие-либо возражения или сомнения против любой высказанной идеи. Все выступления стенографируются и затем руководитель работ выбирает те гипотезы, которые кажутся ему наиболее правдоподобными [59]. Вероятностные оценки правдоподобности гипотез никогда не доводятся до количественных оценок. Механизм творческого процесса, приводящий к принятию тех или иных гипотез при планировании эксперимента, нам не известен—этот процесс, по крайней мере на сегодняшнем уровне формализации научных теорий, не может быть описан формально, хотя ясно, что

информационные машины могут здесь оказать существенную помощь исследователю. Допустим, например, что мы планируем эксперименты по изучению влияния третьих элементов на результаты спектрального анализа. Машина должна будет нам дать сведения о всех тех работах, в которых изучалось влияние третьих элементов на коэффициент диффузии в твердом теле и газовом облаке, а также работах, в которых рассматривалось влияние третьих элементов на кинетику испарения и т. д. При этом машина, конечно, выдаст информацию и не представляющую интерес с точки зрения рассматриваемой задачи, например, информацию о проблеме трех тел в механике. Экспериментатор, очевидно, без труда сможет отобрать из этой информации все интересующие его концепции и использовать их для формулировки гипотезы при планировании экспериментов, направленных на изучение влияния третьих элементов при спектральном анализе. Машина заменит высказывания группы специалистов широкого профиля.

При построении информационной системы, подобно описанной здесь, самым важным и ответственным моментом является выбор уровня информационных поисков, который определяется, с одной стороны, разветвленностью информационного языка, а с другой стороны, полнотой передачи информации в кодируемых рефератах. Трудно предложить четкий критерий, который бы позволил выбрать уровень информационных поисков, не приводящий к излишнему усложнению и удорожанию этого процесса и в то же время не дающий поверхностных результатов. Здесь приходится иметь дело с ситуацией, аналогичной той, которая описывается в математической теории игр: лицо, проектирующее информационную систему, ведет игру с лицом, заинтересованным в получении информации, пытаясь предвидеть стратегию поисков, в условиях, когда не достаточно ясны пути развития соответствующих научных дисциплин, а следовательно, и не достаточно четко известна сфера интересов исследователей, которые будут обращаться за поисками информации.

§ 8. ОПЫТ ПОСТРОЕНИЯ ЛОГИКО-МАТЕМАТИЧЕСКИХ ТЕОРИЙ ИНФОРМАЦИОННЫХ ПОИСКОВЫХ СИСТЕМ

Довольно значительный практический опыт, накопленный в связи с функционированием разнообразных информационных поисковых систем, дал возможность сформулировать ряд существенных теоретических обобщений, использующих результаты современной математической логики и алгебры. В § 7 настоящей статьи был приведен пример определения стратегии информационного поиска при помощи булевой алгебры. В настоящем параграфе будет изложена общая теория использования математической логики и современной алгебры для формализации различных систем информационного поиска. В этом направлении получены интересные результаты в работе Муэрса [60]. Автор ее различает три различных абстрактных поисковых системы: (1) дескрипторная (сокращенно, ИСД), (2) иерархического типа (т. е. такая, в которой элементы располагаются в порядке подчинения и соподчинения; сокращенно системы типа (2) будем обозначать через ИСИ), (3) логического типа (т. е. такая, элементы которой упорядочены отношениями логического характера: дизъюнкцией, конъюнкцией, импликацией и т. д.; сокращенно системы типа (3) будем обозначать через ИСЛ).

В [60] разрабатывается общая математическая модель произвольных информационных систем, которая применяется затем к описанию ИСД, ИСИ и ИСЛ, выявляет сходство и различие между ними. В этой общей теории употребляются следующие основные понятия.

(1) пространство P всех возможных поисковых предписаний;

(2) пространство L всех возможных подмножеств документов, выбираемых из данной коллекции документов;

(3) поисковые преобразования T, устанавливающие связь между P и L.

Пространство L описывается с помощью булевой алгебры. Каждой точке из P соответствуют какие-нибудь точки из L, но не каждой точке из L соответствует какая-либо точка из P. Соответствие между P и L не является, таким образом, взаимно однозначным.

Строение P сложнее строения L. Поэтому в целях удобства исследования желательно разложить P на компоненты. Пространство P рассматривается как некоторое произведение всех элементов, входящих в данное хранилище R. Хранилище R—это четвертое основное понятие общей теории информационных поисковых систем. В нем элементы системы (дескрипторы, объекты иерархического типа, или же объекты, упорядоченные логическими отношениями) представляются в виде некоторых частично-упорядоченных систем. При описании каждого из R, P, L и T последовательно в каждой из ИСД, ИСИ и ИСЛ в [60] существенно используются понятия из обобщенной теории современной алгебры (частично-упорядоченной системы, кардинального произведения и пр. *).

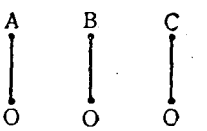


Рис. 19. Частично-упорядоченная система для дескрипторов [60].

Остановимся теперь на том, как в терминах введенных понятий описывается ИСД. Дескрипторы будем обозначать большими латинскими буквами (с индексами или без них). Элементами $R_{\text{ИСД}}$ здесь будут следующие графически представляемые выражения, показанные на рис. 19. Каждое из них — частично-упорядоченная система. Если взять все множество документов и все множество дескрипторов, относящихся к определенной более или менее узкой области науки, то отношение между произвольным дескриптором и произвольным документом может быть описано следующим образом: либо взятый дескриптор необходим для идентификации данного документа (является, так сказать, ключевым для него), либо не необходим для такой идентификации. Эта неизбежная альтернатива и зафиксирована в графическом изображении элементов из R, где буквы A, B, C (стоящие в верхнем конце «спички» вида $\overset{\mathfrak{A}}{\underset{O}{|}}$) соответствуют тому случаю, когда дескриптор необходим, а O—той ситуации, когда он не является необходимым для идентификации документа. Рассматривая частично-упорядоченные системы вида $\overset{\mathfrak{A}}{\underset{O}{|}}$, видим, что в них $\mathfrak{A}$ играет роль единицы, а O—нуля. Следовательно, такие частично-упорядоченные системы подчиня-

*) Дадим краткое определение упомянутых понятий, имея в виду читателя, не изучавшего специально математическую теорию структур. Частично-упорядоченная система есть множество элементов вместе со специфическим упорядочивающим отношением таким, что для любых двух элементов x и y из этого множества можно однозначно решить, x предшествует ли y или же наоборот, либо между x и y вообще нельзя определить отношение предшествования. Понятие кардинального произведения определяется следующим образом. Пусть даны две частично-упорядоченные системы X и Y, типические элементы которых обозначены посредством x и y; в таком случае кардинальное произведение $X \cdot Y$ также является частично-упорядоченной системой. $X \cdot Y$ складывается из всех пар формы (x, y), которые упорядочены с помощью того правила, что $(x, y) \leqslant (x', y')$, где $x \leqslant x'$ в X, а $y \leqslant y'$ в Y (отношение $\leqslant$ понимается в смысле *отношения* предшествования). Как частично-упорядоченная система, так и процедура образования кардинального произведения могут быть проиллюстрированы на диаграмме.

ются законам алгебры логики, а на языке теории структур—законам дистрибутивной структуры с дополнением. Логическим аппаратом ИСД является поэтому алгебра логики. [См. примечание на стр. 40.]

$P_{\text{ИСД}}$ образуется как кардинальное произведение элементов из $R_{\text{ИСД}}$. Пусть, например, число элементов в $R_{\text{ИСД}}$ равно 2. Образуем их кардинальное произведение на диаграмме так, как это показано на рис. 20 для двух дескрипторов A и B. Последняя из фигур на этом рисунке может быть истолкована следующим образом: документы, идентифицированные пересечением дескрипторов A и B (символически: $A \cdot B$), ищутся после того, как были найдены все документы, идентифицируемые дескриптором A и дескриптором B. Каждый документ просматривается, и решается вопрос о том, идентифицируется ли он или не идентифицируется при помощи каждого из дескрипторов A и B. Здесь мы имеем как бы намек на программу поиска документов, удовлетворяющих условию $A \cdot B$ для дескрипторной системы. Каждое из выражений $\overset{\mathfrak{A}}{\underset{0}{|}}$ и $\overset{\mathfrak{B}}{\underset{0}{|}}$ есть булева структура. Опишем теперь $T_{1\text{ИСД}}$ и $T_{2\text{ИСД}}$. Определение $T_{1\text{ИСД}}$ выглядит так: можно связать каждую точку из $P_{\text{ИСД}}$ с некоторыми точками из $L_{\text{ИСД}}$, хотя и не каждая точка в $L_{\text{ИСД}}$ может быть связана с какой-либо точкой из $P_{\text{ИСД}}$. Определение $T_{2\text{ИСД}}$ может быть сформулировано следующим образом. При задании любой точки x в P тем самым выделяется большое семейство X других точек P, таких, для которых выполняется соотношение $x \leqslant X$ (на графике все такие точки семейства X лежат выше точки x; отношение «$\leqslant$» соответствует отношению включения). Существует много документов в документной коллекции, связанное с которыми подмножество дескрипторов является одной из точек семейства X. Рассмотрим наибольший класс документов, такой, что каждый документ в этом классе имеет подмножество дескрипторов, представленных некоторыми точками из X. Пусть этот класс документов представлен точкой x^* в L. Преобразование $T_{2\text{ИСД}}$ переводит точку x из $P_{\text{ИСД}}$ в точку x^* из $L_{\text{ИСД}}$. Обозначим отношение перевода x в x^* через «$\in\!\!\rightarrow$». Пусть J и O соответственно наибольший и наименьший элементы в P и L. Возникающие тогда следующие два соотношения очевидны *):

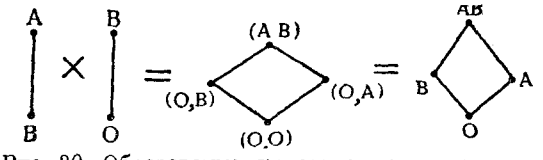


Рис. 20. Образование кардинального произведения для двух частично-упорядоченных дескрипторных систем [60].

(a) $\quad 0(P) \in\!\!\rightarrow J(L),$

(b) $\quad J(P) \in\!\!\rightarrow 0(L).$

$R_{\text{ИСД}}$—структура, поэтому для любых x и y в P выполняется соотношение: $x \cup y = z$, где «$\cup$» есть операция объединения **). При $T_{2\text{ИСД}}$

*) Содержательно соотношения (a) и (b) могут быть «расшифрованы» так: соотношение (a) выражает ту мысль, что отсутствие дескрипторов в поисковом предписании равносильно требованию выбрать все документы коллекции; соотношение (b) утверждает, что ни один документ в коллекции не может идентифицироваться всем множеством дескрипторов.

**) Операцию «$\cup$» определим следующим образом. Пусть $x \leqslant z$ и $y \leqslant z$, причем z является самым маленьким из элементов, для которых выполняются два вышенаписанных соотношения; в таком случае z называется о б ъ е д и н е н и е м x и y и обозна-

$x \in \longrightarrow x^*$, $y \in \longrightarrow y^*$, $z \in \longrightarrow z^*$ в L^* имеем (L^* есть результат поиска из L): $x^* \bigcap y^* = z^*$ (где «$\bigcap$» есть операция пересечения). Однако это отношение не выполняется в L. Соотношения, содержащие «$\bigcup$» и «$\bigcap$» в P, выполняются и в L^* (при условии взаимного перемещения знаков «$\bigcup$» и «$\bigcap$» один на место другого). Аналогичное изучение проводится и для ИСИ и ИСЛ. Остановимся вкратце на системах типа ИСИ. Элементы ИСИ будем изображать малыми латинскими буквами или комбинациями таких букв, заключаемыми в круглые скобки. Для двух элементов (a) и (b) частично-упорядоченная система (ИСИ) выглядит так, как это показано на рис. 21 ($b = ab$, т. е. $b \leqslant a$). Движение от (0) к (ab) идет по линии увеличения содержания и уменьшения объема понятия (т. е. по линии конкретизации употребляемых понятий). При построении $R_{\text{ИСИ}}$ выражение вида (a, a) редуцируется к a. Будем различать два вида иерархии: сильную и слабую. При с и л ь н о й и е р а р х и и каждый элемент не может иметь более одного непосредственно предшествующего ему элемента. Если же это условие не выполняется, то мы имеем дело со с л а б о й и е р а р х и е й. Пример сильной иерархии: иерархия в десятичной системе, пример слабой: классификация патентов в американском патентном бюро. Сравнивая пригодность слабой и сильной иерархии к машинному поиску, К. Муэрс отдает предпочтение системам со слабой иерархией, как более гибким. На наш взгляд, это объясняется тем, что в системах лишь с сильной иерархией можно всегда ввести операцию $\bigcap$, но не операцию $\bigcup$, так как если бы можно было в них определить $\bigcup$, то нашелся бы такой элемент Φ, который имел бы два непосредственно предшествующих члена Φ' и Φ'', а это нарушило бы условие, конституирующее само понятие сильной иерархии. Рассмотрим примеры образования кардинального произведения частично-упорядоченных систем с иерархиями различного типа. (Для сильной иерархии на диаграмме отношение непосредственного предшествования интерпретируется так: лежать ровно на одну точку ниже.) Пример образования кардинального произведения показан на рис. 22. Аналогичный пример для слабой иерархии приведен на рис. 23. Читатель легко может дать истолкование последних фигур на каждом из этих рисунков по аналогии с тем, как мы это делали для последней фигуры на рис. 20 в случае дескрипторной информационной системы. Сравним теперь систему сильной иерархии с иерархией слабой. В сильной иерархической системе каждый элемент в последовательности может иметь лишь один непосредственно предшествующий элемент (лежащий на одну точку ниже на соответствующей диаграмме). В слабой же иерархической системе каждый элемент в последовательности может иметь один или более непосредственно предшествующих элементов (включая нулевой элемент). Сильная иерархия графически изображается в виде некоторого дерева, слабая иерархия напоминает сеть. Возможны классификационные системы, которые включают

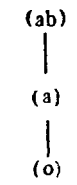


Рис. 21. Пример частично-упорядоченной системы в иерархических классификациях [60].

чается через $x \bigcup y$. Операцию «$\bigcup$» иногда называют условно «чашкой» (от английского слова «cup» (чашка)). Используемую в дальнейшем операцию «$\bigcap$» определим так. Пусть $z \leqslant x$ и $z \leqslant y$, причем z является самым большим из элементов, для которых выполняются два вышенаписанных соотношения; в таком случае z называется п е р е с е ч е н и е м x и y и обозначается через $x \bigcap y$. Операцию «$\bigcap$» иногда условно называют «крышкой» (от английского слова «cap» (крышка)). Учитывая определения, приведенные в настоящем примечании, а также в примечании на предыдущей странице, легко теперь дать и общее определение понятия структуры. Именно с т р у к т у р а есть частично-упорядоченная система, в которой для любой пары элементов можно определить как операцию «$\bigcap$», так и операцию «$\bigcup$». Выражение $x \bigcap y$ в литературе иногда записывают так: $x \cdot y$, а $x \bigcup y$ в виде $x + y$.

в себя как элементы сильной, так и элементы слабой перархии. Мы не будем останавливаться на них, а отсылаем читателя, который заинтересуется ими, а также общими вопросами применения математических методов к теории классификационных систем, к работе [61]. $R_{\text{ИСИ}}$, а также $P_{\text{ИСИ}}$,

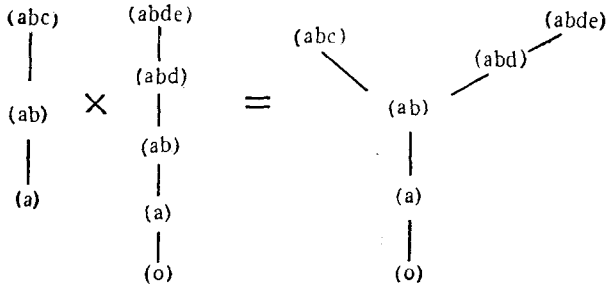


Рис. 22. Пример образования кардинального произведения при сильной перархии [60].

$L_{\text{ИСИ}}$, $T_{1\text{ИСИ}}$ и $T_{2\text{ИСИ}}$ определяются по аналогии с теми же понятиями в случае ИСД.

Муэрс рассматривает также информационные поисковые системы, элементы которых упорядочены отношениями логического типа: в качестве последних фигурируют логические константы, соответствующие союзам «и», «или» (не разделительному) и частице «не». Каждое поисковое предписание формулируется тогда в виде некоторого логического многочлена, построенного на операциях «и», «или» и «не». Если иметь в виду соответствие описанных в [60] абстрактных информационных систем ИСД, ИСИ

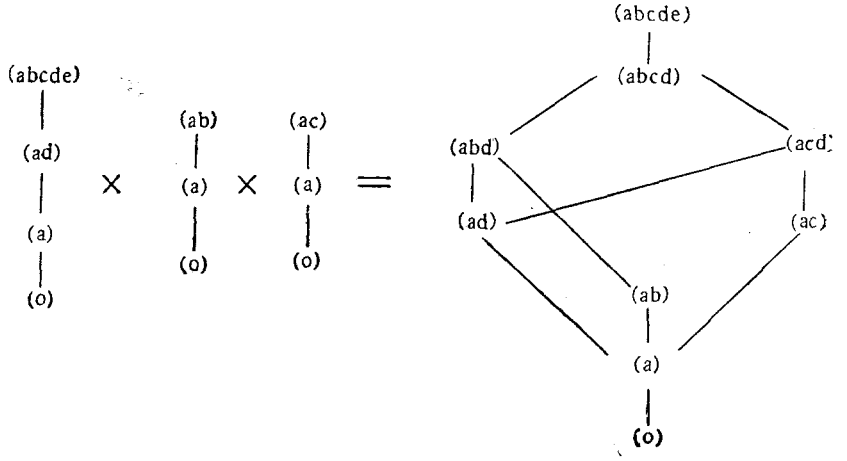


Рис. 23. Пример образования кардинального произведения при слабой перархии [60].

и ИСЛ реально действующим поисковым системам машинного (или полумашинного) характера, то система ИСД соответствует методам координирующей индексации группы М. Таубе и его сотрудников, а комбинация систем ИСД и ИСЛ соответствует приблизительно методам Д. Перри, М. Берри и А. Кента (мы говорим «приблизительно», имея в виду то обстоятельство, что в системах ИСД и ИСЛ не ставится проблема формализации некоторых синтетических (дескриптивных) отношений, что, как известно, входит в качестве составной части в методы, предлагаемые в [56]). Система же ИСИ

может рассматриваться в качестве формализации существующих систем библиотечной классификации, а также различных вариантов усовершенствования и видоизменения этих систем.

Интересно поставлена проблема в работе Ф. Джонкера [62], постулирующего обобщенную теорию индексирования, которая рассматривает все системы индексирования как относительный континуум (спектр). Отличительным параметром этого континуума признается средняя длина отдельной записи или используемого при индексации заголовка. Этот параметр, который определяет позицию системы индексирования в континууме, исчисляется просто как среднее число букв в индексационных терминах. На одном конце континуума или спектра—индексирование с помощью ключевых слов (дескрипторов); индексирование предметными рубриками находится посередине, в то время как иерархические классификации находятся на другом краю (спектра). Схематический описательный континуум изображается так, как это показано на рис. 24. Линия *ab* символизирует позицию дескрипторного индексирования («короткотерминный» конец спектра), *bc*—позицию индексирования предметными рубриками, *cd*—индексирования с помощью систем иерархических классификаций («длиннотерминный» конец спектра). Каждая документная коллекция должна индексироваться с помощью некоторой оптимальной индексационной системы, т. е. такой системы, которая обеспечивала бы достаточную быстроту поиска и надежность хранения информации при допустимых затратах.

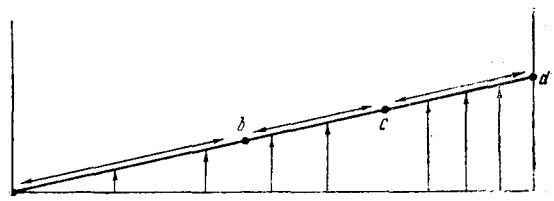


Рис. 24. Описательный континуум [62].

В [62] утверждается, что все другие параметры континуума ведут себя как функции основного параметра—средней длины заголовка при индексировании. В качестве производных параметров рассматриваются:

(1) потенциальная глубина индексирования (пропорциональна числу использованных критериев индексирования),

(2) перестановочность критериев индексации,

(3) степень иерархического подчинения индексирования,

(4) степень возможности справиться с семантической неопределенностью,

(5) поисковый шум (выдача излишней информации),

(6) потенциальная потребность в координирующем механизме, и некоторые другие параметры.

Несмотря на интересную постановку вопросов, в [62] имеется, однако, ряд слабых мест. Сформулированный там «основной закон описательного континуума» (степень иерархического подчинения × глубина индексирования = const) имеет скорее качественный, чем количественный характер. Во-вторых, отсутствует количественная обработка производных параметров. Заметна и неравноценность выделенных Джонкером производных параметров: так, перестановочность индексационного критерия есть параметр, производный от производного же—потенциальной глубины индексирования. И, главное, неясна функциональная зависимость (вплоть до того, существует ли она вообще) некоторых производных параметров от основного параметра—средней длины индексационного заголовка. Бросается в глаза также несколько преувеличенная оценка дескрипторной системы по сравнению с другими индексационными системами. И тем не менее нужно подчеркнуть общеметодологическую ценность основной

идеи[62], постулирующей необходимость создания обобщенной теории индексирования документов, в рамках которой можно было бы описывать как машинные, так и «ручные» методы обработки информации.

Естественно возникает вопрос: чем могут быть полезны новейшие теории документалистики для облегчения современных трудоемких процедур библиотечного поиска? Уже при современном состоянии библиотековедения и документалистики можно привести ряд примеров, иллюстрирующих практическую применимость логико-математических представлений информационных поисков. Как известно, в математике подвергают специальному исследованию свойства так называемых лабиринтных систем. Под лабиринтом понимается некоторая сеть, состоящая из ключевых точек и линий между ними. Библиотечный предметный каталог может рассматриваться как некоторый лабиринт. Лабиринтные свойства такого каталога обнаруживаются в силу того обстоятельства, что в нем существует система так называемых перекрестных ссылок, выявляющая взаимосвязь предметных рубрик и заголовков. Эта общая идея и вытекающие из нее следствия обсуждаются в [63].

Мур в [64] поставил элементарную проблему, касающуюся теории лабиринтов: заданы исходная и конечная точки в лабиринте; найти путь, соединяющий эти точки и содержащий минимальное число линий (звеньев). Для решения этой проблемы Муром был предложен специальный алгоритм (этот алгоритм Эстрин применяет для упрощения процедур оперирования с предметным каталогом и не требует отказа от десятичной системы классификации документов). Алгоритм в [64] для определения кратчайшего пути в лабиринте состоит из нескольких шагов и может быть описан следующим образом:

Начальный шаг. Фиксируются начальная и конечная точки. Начальной точке приписывается индекс 0. Возможность распознавания конечной точки также должна быть обеспечена вне зависимости от того, где именно эта конечная точка выбрана.

Шаги i. Сначала (на первом шаге типа i) индексом 1 помечаются все точки, смежные с начальной. Далее, на втором шаге, индексом 2 отмечаются точки, смежные с точками, получившими уже отметку 1. Процесс продолжается вплоть до того момента, пока не будет достигнута конечная точка. Этой конечной точке приписывается индекс k. Значения, пробегаемые i, таковы:

$$i=1,\ 2,\ 3,\ \ldots,\ k.$$

Число k в точности равно числу звеньев в искомом кратчайшем пути.

Шаги j ($j=k+1,\ k+2,\ k+3,\ldots,\ 2k$). Если шаги i дают информацию о числе звеньев в искомом кратчайшем пути, то следующие за ними шаги j явно определяют сам этот путь. Двигаемся вспять от конечной точки к некоторым точкам, с нею смежным. На каждом из шагов выбирается такая смежная точка, которая имеет наименьший индекс i из всех тех индексов i, которые имеют остальные точки, смежные с данной. Приписываем соответствующей точке с наименьшим индексом i новый дополнительный индекс $i'=i$. Каждая из таким образом вновь проиндексированных точек берется как базис для осуществления следующего шага. Этот процесс продолжаем вплоть до начальной точки, индекс которой равен $i'=0$. Все точки, помеченные индексом i', лежат на искомом пути.

На наш взгляд, алгоритм Мура, очевидно, не является наилучшим среди других возможных алгоритмов по нахождению кратчайшего пути в лабиринте. Более удобным, во всяком случае, представляется такой алгоритм, который начинает не только с исходной, но и одновременно с конечной точки. В этом последнем случае необходимость перебора всех

без исключения точек лабиринта (что, как мы видели, имеет место в алгоритме Мура) уже отпадает. В самом деле, представим себе лабиринт, от каждой точки которого отходит по два звена. Если двигаться по этому лабиринту от точки 0 до точки n, пользуясь алгоритмом Мура, то нужно будет осуществить 2^{n+1} переборов точек, в то время как при другом возможном подходе придется выполнить не более $2^{\frac{n+3}{2}}$ переборов точек.

Однако и оригинальный алгоритм Мура, по-видимому, является достаточно простым и пригодным для механизации с помощью существующей вычислительной техники. Весьма важной особенностью лабиринта является возможность его переориентации в зависимости от выбора начальной точки; при осуществлении переориентации (предполагающей предварительное упорядочивание точек по известным уровням) вид лабиринта может резко меняться. В информационных поисковых системах применимость алгоритма Мура почти очевидна и связана с решением такой задачи: пусть информационная система охарактеризована графическим способом и выглядит как некоторая лабиринтная сеть; построить алгоритм Мура (или модифицировать его) на сети в согласии с заданными поисковыми предписаниями. В терминах библиотековедения ситуация допускает такую формулировку: представим себе, что точкам лабиринта соответствуют те библиотечные рубрики предметного каталога, благодаря которым обнаруживается система перекрестных ссылок. Выберем начальную точку лабиринта, соответствующую исходной предметной рубрике, относящейся к поисковому предписанию. В таком случае кратчайший путь между заданной начальной и конечной точками (рубриками) будет описывать наилучшую стратегию библиотечного поиска. Заметим, что здесь необходимо также составить схему лабиринта каталога относительно входной точки, т. е. произвести переориентацию лабиринта (или его фрагмента—локальной области каталога) относительно начального предметного заголовка.

Итак, алгоритм Мура для нахождения кратчайшего пути в лабиринте может найти себе эффективное применение для облегчения процедуры библиотечных поисков, и он имеет также ту отличительную особенность (важную хотя бы с точки зрения нынешних условий немеханизированной библиотечной работы), что не связан с обязательным требованием отказа от десятичной системы классификации документов (что совершенно неизбежно при использовании, например, информационных поисковых систем дескрипторного типа).

Среди других попыток построения логико-математических моделей систем поиска информации должны быть кратко отмечены также две статьи Б. Виккери [65, 66]. Использование результатов современной математической логики носит в них, однако, гораздо менее конструктивный характер, чем в тех работах, которые мы рассмотрели выше. Основными понятиями в [65] являются: (1) понятие терма как единицы смысла и (2) понятие документа как единицы записи информации. Вводятся понятия семантических связей между термами и между документами, а также рассматриваются различные способы записи информации. На уровне предметного анализа исследуют одни только единицы информации, на уровне структурного анализа переходят к представлению информационной системы в виде некоторой сети единиц информации. Большое внимание уделяется анализу отношений включения и соподчинения для термов. В [65] предлагаются затем правила перевода информационной сети в двумерную матрицу, в номерах столбцов которой ставятся соответствующие номера документов, а в номерах строк которой помещаются соответствующие термы. На пересечении i-й строки и k-го столбца может быть зафиксировано вхождение (или его отсутствие) данного терма в данный документ. Далее

производится перевод двумерной матрицы в одномерную запись. В [66] исследуется, в частности, понятие модулянтов—категорий семантических единиц в том смысле, как это рассматривается в [56]. Общим недостатком подхода Виккери к решению проблемы разработки единой теории информационных систем следует считать его отказ от выделения синтаксических (в некотором смысле синтетических) связей между термами.

ЗАКЛЮЧЕНИЕ

Поиски новых путей накопления, передачи и обработки научной и технической информации привели к созданию научной дисциплины (еще не получившей своего названия), которую можно рассматривать как одну из ветвей кибернетики. Эта дисциплина возникла на пересечении нескольких наук: математической логики, теории вероятностей и математической статистики, документалистики, лингвистики, психологии мышления, электроники, вычислительной техники. В дальнейшем (в связи с поисками новых технических средств для свертывания информации) она, по-видимому, будет широко использовать результаты биофизических исследований—поскольку наиболее компактное свертывание информации имеет место в нуклеиновых кислотах хромозом, в которых закодирована структура всего организма [68, 69].

Успех развития этой новой дисциплины в значительной степени будет зависеть от того, как будет решен основной вопрос кибернетики: о взаимоотношении возможностей вычислительной машины и мышления [70].

Уже сейчас очевидно, что развитие этой дисциплины приведет к новым формам организации науки и прежде всего будет способствовать ее дальнейшей математизации. Для разработки этой дисциплины потребуются коллективные усилия специалистов всех упомянутых выше областей науки.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

1. D. J. Price, Archives Internationales d'Histoire des Sciences **30**, № 14, 85 (1953).
2. D. J. Price, Discovery **17**, № 6, 240 (1956).
3. E. Pietsch, Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen **38**, 34 (1954) (Westdeutscher Verlag, Köln und Opladen).
4. Biological Abstracts **31**, № 7, X (1957).
5. R. H. Ewell, Chemical and Engineering News **33**, № 29, 2980 (1955).
6. Д. Томсон, Предвидимое будущее, 1958, ИЛ, Москва.
7. M. H. Halbert, R. L. Ackoff, Preprints of papers for the International conference on scientific information, Washington, D. C., November 16—21, Area 1, стр. 87 (1958).
8. F. Liebesny—там же, Area 2, стр. 161.
9. D. A. Bruning—там же, Area 3, стр. 7.
10. H. C. Lehman, Scientific Monthly **78**, № 5, 321 (1954).
11. W. Shockley, Proceedings of the IRE **45**, № 3, 279 (1957).
12. H. P. Luhn, IBM Journal of research and development **2**, № 2, 159 (1958).
13. S. Ullmann, Journal de Psychologie normale et pathologique **55**, № 3, 338 (1955).
14. P. B. Baxendale, IBM Journal of research and development **2**, № 2, 159 (1958).
15. В. В. Налимов, Б. С. Непорент, Успехи физических наук **65**, № 3, 521 (1958).
16. T. E. Beukelman, Analytical Chemistry **29**, № 9, 1269 (1957).
17. I. Wachtel, American Documentation **3**, № 1, 56 (1952).
18. L. M. Addison, E. H. Spencer, E. M. Charlet, Analytical Chemistry **30**, № 5, 885 (1958).
19. R. S. Casey, J. W. Perry, A. Kent, M. Berry (Editors), Punched cards. Their application to Science and Industry. 2 Ed. Reinhold Publishiny Corporation. New York, Chapman and Hall. Ltd., London, стр. 697 (1958).
20. H. D. Ashthorpe, ASLIB Proceedings **4**, № 2, 101 (1952).
21. H. E. Stiles, American Documentation **9**, № 1, 42 (1958).
22. J. D. Bernal, см. [7], Area 1, стр. 67.

23. F. Heumann, E. Dale, Progress report in chemical literature retrieval, Interscience Publishers, стр. 201 (1957).
24. Beilsteins Handbuch der organischen Chemie, B. 29, 1 Teil General-Formelregister für das Hauptwerk und die Ergänzungswerke I und II, Berlin, Springer Verlag, 1956.
25. Chemical and Engineering News **33**, 2838 (1955).
26. G. M. Dyson, A New Notation and Enumeration System for Organic Compounds, 2 Ed., Longmans, Green & Co, London and New York, 1949.
27. W. J. Wisswesser, A Line-Formula Chemical Notation, N. Y. Growell Co., New York, 1955.
28. M. Gordon, C. E. Kendall, W. H. T. Davison, Chemical Ciphering: A Universal Code as an Aid to Chemical Systematics, Royal Institute of Chemistry of Great Britain and Ireland, 1948.
29. F. H. S. Curd, G. F. L. Rose, J. Chem. Soc., 729 (1946).
30. Chemical Codification Panel, National Research Council, A Method of Coding Chemicals for Correlation and Classification, Washington, 1950.
31. E. Pietsch, Die IBM-Lochkarte in der Gmelin-Dokumentation, FID Manual on Document Reproduction and Selection, Part II, 1956.
31a. W. Steidle, Pharmazeutische Industrie **19**, № 3, 88 (1957).
32. Л. Беллами, Инфракрасные спектры молекул, ИЛ, Москва, 1957.
33. L. C. Ray, R. A. Kirsch, Science **126**, 814 (1957).
34. A. Opler, T. R. Norton, Chemical and Engineering News **34**, 2812 (1956).
34a. A. Opler, Chemical and Engineering News **35**, № 33, 92 (1957).
35. W. H. Waldo, M. De Backer, см. [7], Area 4, стр. 49.
35a. W. H. Waldo, R. S. Gordon, J. D. Porter, American Documentation 9, 28 (1958).
36. J. Sherman, Industrial and Engineering, Chemistry **50**, № 11, 2441 (1958).
37. Л. Н. Гутенмахер, Вестник АН СССР **88**, № 10 (1957).
38. Г. Э. Влэдуц. Некоторые вопросы научной информации в области химии. I. О путях усовершенствования химических указателей. Изд. ВИНИТИ АН СССР, Москва, 1958.
39. Р. А. Фишер, Статистические методы для исследователей. Госстатиздат, М., 268 стр., 1958.
40. W. J. Youden, Statistical methods for chemists. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London, стр. 127, 1951.
41. W. T. Federer, Experimental design, theory and application. The Macmillan Comp. New York, 544 стр. (1955).
42. O. Kempthorne, The design and analysis of experiments. John Wiley and Sons, Inc., New York., Chapman and Hall, Ltd., London, стр. 631 (1952).
43. C. A. Bennett, N. L. Franklin, Statistical analysis in chemistry and the chemical industry. John Wiley and Sons, Inc, New York, Chapman and Hall, Ltd., London., 724 стр., 1954.
44. O. L. Davies (Editor), Design and analysis of industrial experiments. Imperial Chemical Industries Ltd., Oliver and Boyl, London, Edinburgh, 636 стр., 1956.
45. O. L. Davies (Editor), Statistical methods in research and production. Imperial Chemical Industries Ltd., Oliver and Boyl, London, Edinburgh, 396 стр., 1957.
46. D. R. Cox, Planning of experiments. John Wiley and Sons, Inc., New York, Chapman and Hall, Ltd., London, стр. 308, 1958.
47. Кутюра Луи, Алгебра логики, Одесса, изд. Mathesis, 1909.
48. M. Taube, C. D. Gull, Irma S. Wachtel, American Documentation **3**, № 4, стр. 213–218 (1952).
49. M. Taube, Irma S. Wachtel, American Documentation, april, стр. 67–69 (1953).
50. M. Taube, American Documentation **6**, № 4 (1955).
51. R. Brakken, G. Tillitt, Journal Assoc. Comput. Machinery **4**, № 2, стр. 131–136 (1957).
52. R. Carnap, Der logische Aufbau der Welt, Springer, Berlin, 1928.
53. J. H. Woodger, The axiomatic method in biology, Cambridge, University press, 1937.
54. H. Hollerith, Art of compiling statistics, U. S. Patent № 395782, Januar, 1889.
55. H. Hollerith, Apparaturus for compiling statistics, U. S. Patent № 395787, September 23, 1884.
56. J. W. Perry, A. Kent, M. Berry, Machine literature searching. Interscience Publishers Inc, New York, стр. 162, 1956.
57. J. W. Perry, A. Kent, Tools for machine literature searching. New York, Interscience Publishers Inc., стр. 992.
58. J. Rees, A. Kent, American Documentation 9, № 4, 277 (1958).
59. В. Плэтт, Информационная работа в стратегической разведке, ИЛ, М., стр. 341, 1958.

60. C. N. Mooers, см. [7], Area VI, стр. 57—94.
61. R. A. Fairthore, The Mathematics of Classification, Proc. British Society of International Bibliography 9 (4) (1947).
62. Frederick Jonker, см. [7], Area VI, стр. 22—41.
63. Gerald Estrin, там же, 113—123.
64. Moore E. F., International Symposium on Switching Theory, Harvard University, April 1957 (в печати).
65. B. C. Vickery, см. [7], Area VI, стр. 5—20.
66. B. C. Vickery, там же, Area V, стр. 41—52.
67. Д. Гильберт и В. Аккерман, Основы теоретической логики, ИЛ, М., 1947.
68. Э. Шредингер, Что такое жизнь с точки зрения физики?, ИЛ, 1948.
69. Г. Гамов, А. Рич, М. Икас, статья в сборнике «Вопросы биофизики», ИЛ, М., 1957.
70. А. А. Ляпунов, статья в сборнике «Проблемы кибернетики», Физматгиз, М., 1958, вып. 1, стр. 5.