МАТЕРИАЛЫ 5-ой МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ

СЕКЦИЯ ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ АВТОМАТИЗИРОВАННОЙ ПОДДЕРЖКИ НАУЧНЫХ ИССЛЕДОВАНИЙ

Руководитель секции
В.К. ФИНН

СЕКЦИЯ
ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ АВТОМАТИЗИРОВАННОЙ ПОДДЕРЖКИ НАУЧНЫХ ИССЛЕДОВАНИЙ

Руководитель секции
В.К. ФИНН

БЫСТРЫЕ АЛГОРИТМЫ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ ТИПА ДСМ

Д.В. Виноградов

FAST ALGORITHMS FOR INTELLIGENT SYSTEMS OF JSM-TYPE

D.V. Vinogradov

В настоящем докладе представлен обзор некоторых результатов, связанных с быстрым поиском аналогий с помощью цепей Маркова. Исследованы свойства полученных цепей, приведены примеры с равномерным и неравномерным стационарным распределением.

§1. Немного теории графов

Определение 1. Двудольный граф – это граф $G = \langle V, E\rangle$, в котором множество вершин разбито на два подмножества (называемых долями) $V = V_1 \cup V_2$, $V_1 \cap V_2 = \varnothing$, так, что $E \cap ([V_1]^2 \cup [V_2]^2) = \varnothing$. Иначе говоря, ребра соединяют вершины из различных долей ($E \subseteq [V_1 \times V_2]$).

Пример 1. Пример двудольного графа можно образовать из множества X^+ положительных примеров. Положим $V_1 = X^+$, $V_2 = \Sigma$ и в E добавим ребро $\{X, a\}$, где $X \in X^+$ и $a \in \Sigma$, если и только если пример X имеет признак a. Этот граф соответствует матрице инцидентности для множества примеров X^+ и признаков Σ.

Определение 2. Полным двудольным подграфом в двудольном графе $G = \langle V_1 \cup V_2, E\rangle$ называется такая пара подмножеств $W_1 \subseteq V_1$ и $W_2 \subseteq V_2$, что $[W_1 \times W_2] \subseteq E$. Иначе говоря, каждая вершина одного подмножества смежна каждой вершине другого.

Определение 3. Полный двудольный подграф $\langle W_1, W_2\rangle$ двудольного графа G называется максимальным (по включению), если никакой полный двудольный подграф графа G не содержит графа $\langle W_1 \cup W_2, [W_1 \times W_2]\rangle$. Другими словами, каждая вершина, не попавшая в W_1 (соответственно W_2), не является смежной некоторой вершине из W_2 (соответственно W_1).

Пример 2. Глобальные положительные сходства в смысле ДСМ–метода соответствуют максимальным полным двудольным подграфам двудольного графа, построенного по положительным примерам.

Определение 4. Двудольный граф $\langle V_1 \cup V_2, E\rangle$ определяет две антитонные функции замыкания $d: 2^{V_1} \to 2^{V_2}$ и $i: 2^{V_2} \to 2^{V_1}$ по правилам $d(W_1) = \{y \in V_2 \mid \forall x \in W_1 \{x, y\} \in E\}$ и $i(W_2) = \{x \in V_1 \mid \forall y \in W_2 \{x, y\} \in E\}$.

Определение 5. Замыканием максимального полного двудольного подграфа $\langle W_1, W_2\rangle$ в двудольном графе $\langle V_1 \cup V_2, E\rangle$ относительно элемента $\alpha \in (V_1 \setminus W_1)$ (соответственно относительно $\alpha \in (V_2 \setminus W_2)$) называется максимальный полный двудольный подграф $\langle i(d(\{\alpha\} \cup W_1)), d(\{\alpha\} \cup W_1)\rangle$ (соответственно $\langle i(\{\alpha\} \cup W_2), d(i(\{\alpha\} \cup W_2))\rangle$).

§2. Основной алгоритм

Для обеспечения хороших свойств цепи Маркова мы к гипотезам классического ДСМ–метода добавим дополнительно:

1. наибольшую гипотезу, имеющую все множество признаков и объявленную сходством пустого множества примеров, если никакой объект не содержит все признаки;

2. гипотезы, состоящие из единственного примера с соответствующим множеством признаков, если последнее множество не включается во множество признаков никакого другого примера.

3. наименьшую гипотезу, являющуюся сходством всех положительных примеров и имеющую (возможно пустое!) множество общих признаков.

Зададим переходы из состояния $\langle W_1, W_2\rangle$ следующим правилом:

1. с вероятностью S состояние не изменяется;
2. в противном случае, из множества $(X^+ \setminus W_1) \cup (\Sigma \setminus W_2)$ выберем равномерно распределенный случайный элемент α.
3. если $\alpha \in (\Sigma \setminus W_2)$, то перейдем на замыкание $\langle W_1, W_2\rangle$ относительно α.
4. если $\alpha \in (X^+ \setminus W_1)$, то перейдем на замыкание $\langle W_1, W_2\rangle$ относительно α.

На шаге (4) мы уменьшаем гипотезу, вычисляя сходство имеющейся гипотезы с некоторым новым примером. Он соответствует шагу известного алгоритма “замыкай по одному”. На шаге (3) мы, наоборот, увеличиваем гипотезу, добавляя несколько признаков.

§3. Анализ алгоритма

Мы предполагаем знание основ конечных цепей. Мы рассматриваем только однородные цепи Маркова, вероятности перехода
$p_{ij}(n) = P(X_{n+1} = j | X_n = i)$ которой не зависят от n.

Пронумеруем множество состояний конечной цепи Маркова: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. В дальнейшем элементы Ω отождествим с их номерами 1, 2, …, n, т.е. положим $\Omega = \{1, 2, \ldots, n\}$.

Тогда вероятностная мера μ на Ω может быть отождествлена с таким набором чисел $(\mu_1, \ldots, \mu_n)$, что $\forall j\ 0 \le \mu_j \le 1$ и $\Sigma_j \mu_j = 1$.

Определение 1. n×n матрица $P = (p_{ij})$ называется стохастической, если $\forall i\ \forall j\ 0 \le p_{ij} \le 1$ и $\forall i\ \Sigma_j p_{ij} = 1$.

Каждой цепи Маркова (X_k) соответствует стохастическая матрица $p_{ij} = P(X_1 = j | X_0 = i)$. Наоборот, по каждой стохастической матрице $P = (p_{ij})$ можно построить однородную цепь Маркова, вероятности переходов которой равны $P(X_{n+1} = j | X_n = i) = p_{ij}$.

Определение 1. Вероятностная мера $\mu = (\mu_1, \ldots, \mu_n)$ на множестве Ω называется инвариантной мерой цепи Маркова, если $\mu P = \mu$.

Утверждение 1. Для любой конечной цепи Маркова всегда существует инвариантная мера.

Доказательство. Запишем условие инвариантности меры в виде $\mu(P - E) = \mu P - \mu E = \mu P - \mu = 0$, где E - единичная матрица. Другими словами, μ – собственный вектор матрицы с собственным значением 1. Осталось проверить, что матрица P - E вырождена. Но это следует из того, что столбцы матрицы P - E линейно зависимы, так как их сумма равна 0, так как $\forall i\ \Sigma_j (p_{ij} - \delta_{ij}) = \Sigma_j p_{ij} - 1 = 0$.

Приведенное доказательство особенно просто, но применимо только к конечным цепям Маркова.

Из основного свойства цепей Маркова легко следует, что вероятность перехода за k шагов $P(X_{n+k} = j | X_n = i)$ равна $p_{ij}^{(k)}$, где $p_{ij}^{(k)}$ – (i, j)-й элемент стохастической матрицы $P^k = P \times P \times \ldots \times P$ (k-кратное произведение матриц).

Определение 2. Состояние i называется несущественным, если существуют такие m и j, что $p_{ij}^{(m)} > 0$, но для всех t $p_{ji}^{(t)} = 0$. В противном случае состояние называется существенным.

Множество существенных состояний обладает, очевидно, тем свойством, что, попав в него, цепь Маркова никогда уже его не покинет.

Можно показать, что на несущественных состояниях любая инвариантная мера равна 0.

Утверждение 2. Цепь Маркова, соответствующая алгоритму §2, не содержит несущественных состояний.

Доказательство. Пусть количество (+)-примеров равно n, а количество признаков – k. Тогда никакое состояние i не может быть несущественным, так как из любого состояния j можно достичь i за самое большее n+k шагов. Сначала добавляем признаки по одному до тех пор, пока не достигнем наибольшей гипотезы (1) из §2. Затем начинаем добавлять примеры до тех пор, пока не окажемся в i. Здесь используется тот факт, что по (2) из предыдущего параграфа каждый пример является гипотезой. Ясно, что $p_{ji}^{(n+k)} > 1/2^{n+k}(n+k)^{n+k}$.

Определение 3. Цепь Маркова называется приводимой, если пространство состояний Ω разбито на два подмножества A и B такие, что $p_{ij} = 0$ для всех $i \in A$, $j \in B$ и всех $j \in A$, $i \in B$. В противном случае, цепь называется неприводимой.

Если цепь Маркова X приводима, то существует много инвариантных мер. Достаточно рассмотреть все комбинации $a \cdot \mu_A + (1-a) \cdot \mu_B$, где $0 \le a \le 1$ и

μ_A, μ_B – инвариантные меры сужений $\{X_n\}$ на A и B соответственно, доопределенные 0 на B и A.

Утверждение 3. Цепь Маркова, соответствующая алгоритму §2, неприводима.

Доказательство. Аналогично доказательству утверждения 2.

Определение 4. Состояние $j \in \Omega$ цепи Маркова называется периодическим, если найдется такое целое число $m > 1$, что $p_{jj}^{(n)} = 0$, когда n не является кратным m. Наибольшее такое m называется периодом состояния j. Цепь называется апериодической, если она не имеет периодических состояний.

Утверждение 4. Цепь Маркова, соответствующая алгоритму §2, не содержит периодических состояний.

Доказательство. Согласно (1) шагу алгоритма $p_{jj}^{(1)} = S$. Но любой делитель 1 сам равен 1.

Теорема 1. (эргодическая теорема) Пусть цепь Маркова X_t неприводима. Определим

$N_\pi^i(m) = |\{t < m | X_t = i, P(X_0 = j) = \pi_j\}|$.

Тогда для любого начального распределения π почти всюду $m^{-1} \cdot N_\pi^i(m) \to \mu_i$ при $m \to \infty$,

где $\mu = (\mu_1, \ldots, \mu_n)$ – единственная инвариантная мера.

Теорема 2. (теорема сходимости) Пусть цепь Маркова X_t неприводима и апериодична. Тогда для любого начального распределения π для любого состояния i $P_\pi(X_t = i) \to \mu_i$ при $m \to \infty$,

где $\mu = (\mu_1, \ldots, \mu_n)$ – единственная инвариантная мера.

Каковы инвариантные меры для нашего алгоритма?

Пример 1. (равномерная мера). Рассмотрим n объектов $X^+ = \{1, 2, \ldots, n\}$ и n признаков $\Sigma = \{1, 2, \ldots, n\}$. Соединим каждый пример j со всеми признаками, кроме j. Тогда каждая гипотеза соответствует в точности одному подмножеству $W_1 \subseteq X^+$ (и в точности одному подмножеству $W_2 \subseteq \Sigma$). Соответствующее W_2 находится по правилу $W_2 = \Sigma \backslash W_1$. Таким образом, мы получаем Булеву алгебру всех гипотез. Очевидно, что полученная цепь Маркова имеет равномерную инвариантную меру.

Пример 2. (неравномерная мера). Рассмотрим 4 объекта $X^+ = \{1, 2, 3, 4\}$ и 3 признака $\Sigma = \{a, b, c\}$. Пусть примеры таковы, что $1 = \{a, b, c\}$, $2 = \{a, b\}$, $3 = \{a, c\}$ и $4 = \{a\}$. Матрица переходов P нашей цепи Маркова равна:

$$\begin{array}{c} \\ \langle\{1\},\{a,b,c\}\rangle \\ \langle\{1,2\},\{a,b\}\rangle \\ \langle\{1,3\},\{a,c\}\rangle \\ \langle\{1,2,3,4\},\{a\}\rangle \end{array} \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 0 & 2/3 \\ 1/3 & 0 & 0 & 2/3 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Вычисляя собственный вектор μ с собственным значением 1, получаем $\mu = (3/19, 9/38, 9/38, 7/19)$.

Пример 2 также показывает, что в общем случае цепь Маркова из предыдущего параграфа не может быть сведена к случайному блужданию на неориентированном графе, так как наименьшая гипотеза может быть достигнута из наибольшей гипотезы за один шаг, а обратное утверждение не верно.

Основной вопрос, требующий дальнейшего исследования, касается количества шагов, необходимых для достаточного перемешивания цепи, т.е. чтобы порождаемое алгоритмом распределение гипотез было достаточно близко к инвариантному.

Предложенный алгоритм, на взгляд автора, отражает реальные механизмы правдоподобных рассуждений. Общая схема абдукции, хотя и имеет фундаментальный характер, вряд ли может быть использована в системах реального времени (за исключением наиболее простых и/или наиболее понятых).

На практике, недостатком нового подхода является возможность пропуска интересных гипотез. Но главный недостаток – отсутствие механизма управления обучающей выборкой. В классическом ДСМ-методе таким механизмом является проверка критерия достаточного основания индуктивного вывода.

К достоинствам нового метода нужно отнести скорость вычислений.

О МЕТОДОЛОГИЧЕСКИХ ПРИНЦИПАХ ПОСТРОЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ ДЛЯ НАУК О СОЦИАЛЬНОМ ПОВЕДЕНИИ

В. К. Финн

ON METHODOLOGICAL PRINCIPLES OF CONSTRUCTION OF INTELLIGENT SYSTEMS FOR SOCIAL BEHAVIOR SCIENCE.

V.K. Finn

Строение интеллектуальных систем

Интеллектуальные системы являются инструментом решения задач в трудно формализуемых областях, требующих обзора массивов исходных данных большого объема и применения логических и вычислительных процедур значительной алгоритмической сложности.

Проблема формализации исходных данных и знаний в науках о социальном поведении весьма сложна из-за отсутствия точных языков исследования.

Создание же интеллектуальных систем (ИС) для наук о социальном поведении помимо практической пользы окажет стимулирующее влияние на развитие метатеоретических средств, адекватных и необходимых этим дисциплинам.

ИС состоят из Решателя задач, базы знаний (БЗ), Информационной Среды и интерфейса.

Решатель ИС образован Рассуждателем, Вычислителем и Синтезатором.

Рассуждатель реализует взаимодействие познавательных процедур, например, эмпирической индукции, аналогии, абдукции и дедукции (примером организации такого взаимодействия являются ИС типа ДСМ [1, 2]).

Отметим, что в [1, 2] абдукция осуществляет недедуктивное объяснение начального состояния данных посредством результатов, полученных интеллектуальной системой (в силу этого абдукция является критерием достаточного основания принятия гипотез [2]).

Синтезатор осуществляет управление взаимодействием познавательных процедур, выполняемых Рассуждателем, и процедур, выполняемых Вычислителем, реализующим необходимые численные методы (например, статистические), а также процедур, реализуемых семиотическим герменевтикозатором.

Семиотический герменевтикозатор есть подсистема, применяемая к Информационной Среде, где Информационная Среда (ИнфСр) есть подсистема, образованная массивом текстов (МТ) и базой данных (точнее, **состояниями** БД). Обозначим *n*-ое состояние БД посредством $БД_n$, семантический гер-

меневтикозатор посредством СГ, тогда
ИнфСр = МТ + БД, где

$$БД = \bigcup_{n=0}^{s} БД_n, СГ(МТ, БД_n) = БД_{n+1},$$

а *s* - заключительное состояние для процесса решения задачи посредством ИС.

Отметим, что наличие СГ и МТ является специфическими особенностями ИС для гуманитарных дисциплин. Следует обратить внимание на тот факт, что БЗ содержит реляционный тезаурус, в котором имеются уточнения употребляемых в МТ терминов, а также отношения между ними (включение, ассоциации, часть - целое, сходство, синонимия и т. д.).

Расширение *n*-ого состояния БД, т. е. $БД_n$, до $БД_{n+1}$ происходит за счет анализа МТ посредством СГ, применяемого к МТ и $БД_n$ (в том числе и с использованием **машинного обучения** [3, 4]).

Таким образом, СГ должен осуществлять, по крайней мере, следующие трудоемкие процедуры:

1. автоматическую классификацию текстов из МТ по темам [3, 4],

2. извлечение из МТ новых (относительно $БД_n$) сведений (в соответствии с атрибутами БД) для расширения $БД_n$ до $БД_{n+1}$ (подчеркнем при этом открытость МТ).

В [1, 2] содержатся определения квазиаксиоматических теорий (КАТ), являющихся средством представления знаний для открытых предметных областей. Выше мы охарактеризовали тезаурус, в котором имеются уточнения употребляемых в МТ терминов. Эти термины обозначают идеи, используемые в науках о социальном поведении, которые в уточненном виде задаются в тезаурусе. Их уточнения будем называть концептами (средства сопоставления идеям соответствующих концептов содержатся в [5]).

Результатом работы ИС типа ДСМ [1] является множество гипотез о причинно-следственных отношениях ((+)-причины, (-)-причины и (0)-причины, представляющие источники “фактических противоречий” или конфликты). В силу условия каузальной полноты в КАТ, являющегося критерием достаточного основания для принятия гипотез в ИС типа ДСМ [1, 2], в идеальном случае каждый (±)- или (0)-факт имеет соответствующую (±)- или (0)-причину. Знания о (±)-причинах и (0)-причинах могут быть использованы для установления причинно-следственных зависимостей между фактами (каузальной связности), организацией которых может быть каузальная сеть (эта сеть может содержать циклы, представляющие конфликты).

БЗ для наук о социальном поведении есть подсистема ИС, образованная КАТ, тезаурусом (система концептов и их зависимостей) и каузальной сетью.

Для обработки данных о конфликтах могут быть использованы специальные процедуры, расширяющие ДСМ-метод автоматического порождения гипотез [6].

Общие принципы построения интеллектуальных систем анализа социального поведения

I. Имеются три типа “миров” (онтологий) и соответствующих им способов оценивания знаний об этих “мирах” (эпистемологий) [6, 7]: “мир”-1 со случайными событиями и статистическими средствами их оценивания, “мир”-2 с детерминацией всех событий и логическими средствами оценивания знаний о них, “мир”-3, содержащий как детерминацию событий, так и случайные возмущения, влияющие на заключительные состояния “мира”-3. Знания этого “мира” оцениваются с помощью логических средств, учитывающих статистические соображения [8].

II. Язык для представления знаний в ИС должен обладать достаточной выразительной силой, чтобы осуществлять аргументацию и автоматизированные рассуждения для порождения гипотез о “мирах” 1-3.

III. В ИС должны содержаться средства для уточнения идей, преобразующих их в концепты [5], а также средства для изменения состояний $БД_n$ с использованием СГ.

IV. Процедурные средства ИС должны быть достаточны для применения принципа **каузального структурализма:** сходство о строении фактов должно извлекаться из БД как основание для гипотез о о (±)-причинах, (0)-причинах изучаемых эффектов (для “миров”-2, 3).

V. В связи с неполнотой информации в БД и БЗ должна использоваться **неклассическая** (**неаристотелевская**) теория истины (этот факт является современным расширением рационализма и по смыслу противоположен постмодернистскому релятивизму).

V*I*. **Семиотическая герменевтика** должна быть формализована как аппарат **понимания** текста посредством **аргументации** (в смысле [9])[1] , которая должна быть реализована в **семиотическом герменевтикозаторе**.

Постулаты анализа социологических данных как пример конкретизации данной методологии построения ИС.

В соответствии с изложенными принципами предлагаются следующие постулаты для анализа социологических данных (эти постулаты используются при формировании БЗ для интеллектуальных систем типа ДСМ, применимых в социологии [10]).

P1. Постулат поведения: пусть (SC) - множество характеристик, представляющих социальный характер субъекта, (*I*P) - множество характеристик, выражающих индивидуальные особенности личности, а (BD) - характеристики, представляющие биографические данные о субъекте.

Пусть далее $A=A_1\cup A_2\cup A_3$, $(A_1\subseteq(SC))$ & $(A_2\subseteq(IP))$& $(A_3\subseteq(BD))$, где некоторые из $A_i\neq\varnothing$, i=1,2,3, тогда A является **детерминантой** (или частью детерминанты) поведения субъекта [10].

В соответствии с классификацией М. Вебера [11] имеются четыре рода поведения - целерациональное (р), ценностнорациональное (ц), традиционное (т), аффективное (а). Пусть $\Delta = (SC)\cup(IP)\cup(BD)$, $\Omega = \{(p),(ц),(т),(а)\}$, рассмотрим множество отображений H вида $h: 2^{\Delta}\rightarrow 2^{\Omega}$, где $h\in H$. Следующая классификация детерминант поведения (в соответствии с P1) характеризует три сорта поведения - собственно социальное, слабо социальное и несоциальное (индивидуальное).

ПОВЕДЕНИЕ		
социальное		**не-социальное**
собственно социальное	**слабо социальное**	
$A_1\neq\varnothing$; $A_1\cup A_2$, $A_i\neq\varnothing$, i = 1, 2; $A_1\cup A_3$, $A_i\neq\varnothing$, i = 1, 3; $A_1\cup A_2\cup A_3$, $A_i\neq\varnothing$, i = 1, 2, 3;	$A_3\neq\varnothing$; $A_3\cup A_2$, $A_i\neq\varnothing$, i = 2, 3;	$A_2\neq\varnothing$ $(A_1 = A_3 =\varnothing)$

Пусть h(X)=Y, где $X\subseteq\Delta$, и

(1) $X = A_1$, или $X = A_1\cup A_2$, или $X = A_1\cup A_2\cup A_3$;

(2) $X = A_3$, или $X = A_3\cup A_2$;

(3) $X = A_2$,

Тогда, соответственно, будем говорить, что типы поведения, определяемые отображением h, являются собственно социальными, слабо социальными и не-социальными (индивидуальными).

Легко теперь подсчитать количество типов поведения в соответствии с определением отображения h: всего имеется 105 типов поведения ("таблица Менделеева" для типов поведения), из них типов собственно социального поведения имеется $4\cdot15=60$, типов слабо социального поведения - $2\cdot15=30$, а типов индивидуального поведения - 15.

В ИС массивы текстов (МТ) и БД должны быть проблемно-ориентированными для элементов приведенной выше типологии.

В связи со сказанным возникают следующие задачи:

Задача №1. Отнести текст из МТ к собственно социальному или индивидуальному поведению, конкретизировать это отнесение посредством распознавания вида поведения (в данном типе); задача решается посредством Семиотического герменевтикозатора (СГ);

Задача №2. Общую БД переструктурировать согласно типам (и видам в них) поведения.

Очевидно, что из четырех возможных веберовских типов поведения (р, ц, т, а) мы образовали 15 возможных комбинаций, отнесенных посредством h к возможным случаям характеризации субъекта согласно постулату поведения P1.

P2. Постулат индивидуализации социального отношения R между субъектами X и Z: если имеет место XRZ, то

$(X\cap(IP)\cap(SC))\cup(Z\cap(IP)\cap(SC))\cup(X\cap(BD)\cap(SC))\cup(Z\cap(BD)\cap(SC))\neq\varnothing$.

Таким образом, необходимым условием принадлежности пары <X,Z> социальному отношению R является $X\cap(IP)\neq\varnothing$, или $Z\cap(IP)\neq\varnothing$, или $X\cap(BD)\neq\varnothing$, или $Z\cap(BD)\neq\varnothing$ (этот постулат соответствует идеям М. Вебера о том, что цель социологии -выявить подлинные мотивы поведения индивида, а социальное действие есть действие индивида, зависящее от действий других людей [11]).

P3. Постулат ситуационизма [13] :

отсутствие действия социального субъекта (индивида или общности индивидов) при наличии потенциальной причины действия (в том числе установки) объясняется влиянием ситуации [12].

В соответствии с P3 может быть расширен постулат P1: детерминантой поведения может быть <A,S>, где S - представленная ситуация и (a) $A\neq\varnothing$, $S=\varnothing$ или (в) $A\neq\varnothing$, $S\neq\varnothing$, (c) $A=\varnothing$, $S\neq\varnothing$, или (d)$A\neq\varnothing$ или $S\neq\varnothing$.

В соответствии с видами детерминации (a) - (d) специфицируются и типы поведения, отображающие влияние ситуации S: число возможных типов поведения $\beta = 105\cdot4=420$.

Р4. Территориально очерченный фрагмент общества (поселок, город, район, регион, страна) является покрытием социальными общностями, а не разбиением непересекающихся общностей (социальных групп).

Р5. Постулат прогнозирования социального поведения, основанного на стратификационной модели:

Прогнозирование социального поведения (в том числе электорального) является социологически осмысленным, если оно использует представительные выборки, учитывающие стратификационные модели такие, что они сформированы посредством установленных детерминант поведения в соответствии с постулатами Р1 - Р4.

Таким образом, статистические методы как способ измерения должны использовать реляционные модели стратификации, основанные на обнаруженных детерминациях социального поведения.

Р6. Постулат концептуализации: социологическая концепция должна формулироваться как система концептов (в смысле [5]), являющихся уточнением соответствующих идей, имеющих эмпирическую экземплификацию и неопровергнутую аргументацию [9].

Опытным подтверждением осмысленности постулатов Р1 - Р6 являются результаты исследования солидарного поведения в [10]. Полезным применением Р1 - Р6 в ИС является возможность применения ДСМ-метода автоматического порождения гипотез для изучения электорального поведения с использованием машинного обучения.

Предполагаемая архитектура ИС для анализа и прогнозирования социального поведения согласуется с пониманием П.А. Сорокиным социокультурных образований как причинно-смысловых систем [14], изучение которых может осуществляться посредством порожденных детерминант и содержания БЗ. Таким образом, автоматизированное построение фрагмента БЗ - каузальной сети является важной проблемой, решаемой посредством ИС. Создание каузальной сети для исследуемого социума фактически является моделированием социальных отношений в соответствии с постулатами Р1 - Р6.[2]

Для каждого начального состояния БД и БЗ интеллектуальной системы порожденная каузальная сеть имитирует синхронную картину социума, но пополнение БД, БЗ и МТ (массива текстов) создает возможность посредством последовательных приближений имитировать и диахронную картину изучаемого социума (это обстоятельство может служить основанием для создания проблемно-ориентированного мониторинга социума).

Создание интеллектуальных систем для наук о социальном поведении - трудоемкая, сложная творческая задача, однако современное состояние точных методов и компьютерных средств и идейного багажа гуманитарных наук является основанием для практической реализации предлагаемой методологии, возможность применения которой в связи с этим не следует рассматривать как неопределенное пророчество в стиле Нострадамуса.

Работа выполнена при поддержке Российского гуманитарного научного фонда (проект № 99-03-19686а).

Литература

[1] **Финн В.К.** Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники, сер. Информатика. Т.15. – М.: ВИНИТИ, 1991. – с.54-101.

[2] **Финн В.К.** Синтез познавательных процедур и проблема индукции // НТИ. Сер.2. 1999. – № 1-2. – С.8-45.

[3] **Селисская М.А.** Использование машинного обучения при автоматической классификации текстов // НТИ. Сер.2. – 1999. – № 1-2. – С.28-33.

[4] **Царева П.В.** Алгоритмы для распознавания позитивных и негативных вхождений дескрипторов в текст и процедура автоматической классификации текстов // Там же –С.15-27.

[5] **Финн В.К.** Интеллектуальные системы и общество // НТИ. Сер.2. – 1999. – № 10. – С.6-20.

[6] **Финн В.К.** ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных сиситемах // НТИ. Сер.2. – 2000. – № 11 (в печати).

[7] **Аншаков О.М.** Каузальные модели предметных областей // НТИ. Сер.2. – № 3. – 2000. – С.3-17.

[8] **Григорьев П.А.** Об одном методе автоматического порождения гипотез, схожем с ДСМ-методом: применение статистических соображений // НТИ. Сер.2. – № 5-6. – 1996. – С.52-55.

[9] **Финн В.К.** Об одном варианте логики аргументации // Там же. – С.3-19.

[10] **Климова С.Г., Михеенкова М.А., Панкратов Д.В.** ДСМ-метод как метод выявления детерминант социального поведения // НТИ. Сер.2. – 1999. – № 12. – С.3-14.

[11] **Вебер М.** Основные понятия социологии // Избранные произведения, М.: Прогресс, 1990.

[12] **Финн В.К., Михеенкова М.А.** О ситуационном расширении ДСМ-метода автоматического порождения гипотез // НТИ. Сер.2. – 2000. – № 1.- С.20-30 (в печати).

[13] **Росс Л., Нисбетт Р.** Человек и ситуация (уроки социальной психологии). М.: АСПЕНТ ПРЕСС, 1999.

[14] **Сорокин П.А.** Социологические теории современности. М.: ИНИОН, 1992.

[1] Эту идею сообщил автору Ю. А. Шрейдер.

[2] Однако строение такой каузальной сети и процедуры ее порождения должны быть уточнены посредством специальных исследований.

О НОВЫХ ВАРИАНТАХ ДСМ-МЕТОДА АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ ГИПОТЕЗ И ИХ ПРИМЕНЕНИИ

С.М. Гусакова, М.А. Михеенкова, В.К. Финн

ON NEW VARIANTS OF JSM-METHOD OF AUTOMATIC HYPOTHESES GENERATION AND ITS APPLICATIONS

S.M. Gousakova, M.A. MIkheyenkova, V.K. FInn

ДСМ-метод автоматического порождения гипотез предоставляет возможности для проведения рассуждений, объединяющих логические и вычислительные средства, в открытых предметных областях. Как было показано в [1], метод представляет собой синтез познавательных процедур – эмпирической индукции, аналогии и абдукции. Последняя реализуется посредством специального критерия достаточного основания принятия гипотез – к.д.о.п.г. Средством формализации ДСМ-расуждения (представленного указанным синтезом) является специальный аппарат многозначных логик [2], а само рассуждение опирается на частично формализованные знания (как объективные, так и субъективные) о предметной области.

Ряд особенностей ДСМ-метода автоматического порождения гипотез позволяет считать его адекватным формальным аппаратом для качественного анализа данных в социологии. В рамках подхода возникает возможность отображения и исследования индивидуальных характеристик субъектов, потенциально детерминирующих тот или иной тип социального поведения. А это, в свою очередь, позволяет построить последующие исследования по типологии социальных общностей и построению моделей социальной стратификации.

Достаточно подробно результаты изучения отношения “субъект⇒поведение” средствами интеллектуальной системы типа ДСМ изложены в работе [3]. Существенное влияние контекста поведения (ситуации) на само поведение субъекта, выявленное в результате эмпирического исследования, в сочетании с принципом адекватности средств ДСМ-метода природе изучаемой задачи потребовало расширения самого метода. Существование внеличностных факторов, способствующих или препятствующих тому или иному поведению, послужило основой перехода от изучения отношения “субъект⇒поведение” к изучению отношения “⟨субъект, ситуация⟩⇒поведение” [4, 5]. При этом реализуется прямой ДСМ-метод, опирающийся на принцип “от сходства объектов – к сходству эффектов (свойств)” (сходство субъектов есть основа сходства их поведения). Одновременно было начато развитие обратного ДСМ-метода (реализующего ДСМ-рассуждения типа “от сходства эффектов – к сходству объектов”) в качестве процедурной модели анализа электорального поведения [6]. Обратный ДСМ-метод также является синтезом познавательных процедур. Порождаемые им отношения причинности, формирующие базу знаний, извлекаются из исходной базы фактов посредством специальных обратных ДСМ-предикатов, отличных от прямых.

1. Основные принципы ДСМ-анализа социологических данных

Предпочтение, которое, на наш взгляд, следует отдавать качественному формализованному анализу социологических данных, опирается на представление, что логическая систематизация данных и знаний (описание системы отношений, оценивание высказываний, выбор аргументации) должна предшествовать построению численной модели (предусматривающему применение численных методов). Онтологической особенностью социальной реальности является симметрия позитивных и негативных причин, соответственно, вынуждающих и тормозящих поведенческие акты. Наличие знаний о позитивных и негативных причинах рассматриваемого множества фактов и гипотез приводит к возможности построения каузальной конструктивной аргументации гипотез о моделируемой социологической реальности. Конструктивность принятия гипотез выражается в порождении их оценок посредством правил правдоподобного вывода для индукции и аналогии.

Применение ДСМ-метода основано на возможности реализации принципа структурализма в анализе данных: сходство на объектах есть фактор для распознавания детерминаций. Для структурного описания субъекта поведения используется три множества характеристик, представляющих социальный характер субъекта (SC), индивидуальные черты личности (IP) и биографические данные (BD). Поведение A субъекта определяется подмножеством дифференциальных признаков C таким, что $C=C_{SC}\cup C_{IP}\cup C_{BD}$, где $(C_{SC}\subseteq SC)\&(C_{IP}\subseteq IP)\&(C_{BD}\subseteq BD)$. В этом случае C является детерминантой поведения субъекта, представленного знанием A.

2. *Введение ситуаций*

Для формализации эмпирической индукции в ДСМ-методе используются два отношения – “обладание множеством свойств” и “причина” (применительно к социологическим исследованиям под

"причиной" понимается устойчивая совокупность факторов, вынуждающих некоторый эффект или препятствующих ему). Представляют эти отношения, соответственно, предикаты $X\Rightarrow_1 Y$ и $V\Rightarrow_2 W$. Здесь X – совокупность характеристик социального субъекта, Y – множество свойств (поведенческих активностей или готовностей), V – подмножество характеристик социального субъекта, вынуждающих или запрещающих появление множества свойств W. Очевидно, однако, что предикаты $X\Rightarrow_1 Y$ и $V\Rightarrow_2 W$, используемые в ДСМ-рассуждениях о социальном поведении (индивидов и их общностей), являются все же априорным способом представления знаний (несмотря на свою экспериментально установленную полезность для анализа социального поведения [3]). Необходимость формализации логики социальных наук как логики ситуационной обоснована К.Р. Поппером [7]. С другой стороны, опыт практического применения ДСМ-метода к анализу солидарного поведения – готовности к забастовкам на заводах в Санкт-Петербурге и Ельце – выявил заметные эффекты влияния ситуаций на детерминанты поведения. Таким образом, соображения как теоретического, так и эмпирического характера привели к необходимости введения параметра ситуации S.

В предлагаемой модели для представления исходных фактов взамен бинарного предиката $\Rightarrow_1$ используется тернарный предикат $P(X,Y,S)$, означающий "субъект X обладает (не обладает) множеством свойств Y в ситуации S". Причина наличия тех или иных свойств у объекта, в соответствии с высказанным выше предположением, может равно корениться как в самом объекте, так и в ситуации проявления свойств. Соответственно, предикат причинности по-прежнему бинарный, но причина включает два параметра: подобъект и фрагмент ситуации (или всю ситуацию целиком). Предикат $\Rightarrow_2$ заменяется предикатом $R_i(\langle V,S\rangle,W)$ – "подмножество характеристик V и фрагмент ситуации S′ есть причина наличия (отсутствия) множества свойств W". R_i (i=1, 2, 3, 4) характеризует структуру мира – насколько в ней существенны сам объект и ситуация проявления свойств.

$R=R_1\vee R_2\vee R_3$; $R_1\leftrightarrow R\&(V\neq\varnothing\&S=\varnothing)$; $R_2\leftrightarrow R\&(V\neq\varnothing\&S\neq\varnothing)$; $R_3\leftrightarrow R\&(V=\varnothing\&S\neq\varnothing)$; $R_4=R$.

Для ситуаций возможны различные структуры данных: булевская (когда задается множество параметров – характеристик – ситуации), кортежная, реляционная. Мы ограничимся пока что булевским представлением данных. Пусть даны конечные множества

$\mathbf{U}^{(i)}$, i=1, 2, 3, $\mathbf{U}^{(1)}=\{d_1,\ldots,d_{r_1}\}$, $\mathbf{U}^{(2)}=\{a_1,\ldots,a_{r_2}\}$, $\mathbf{U}^{(3)}=\{s_1,\ \ ,s_{r_3}\}$. Определим на них 3 булевых алгебры $\boldsymbol{B}_i=\{2^{\mathbf{U}^{(i)}},\ \text{—},\cap,\cup\}$. $\boldsymbol{B}_1$ – алгебра объектов, $\boldsymbol{B}_2$ – алгебра свойств, $\boldsymbol{B}_3$ – алгебра ситуаций (внешних обстоятельств). Назовем *объектом* V *в ситуации* S пару $\langle V,S\rangle\in\mathbf{U}^{(1)}\times\mathbf{U}^{(3)}$.

Операцию сходства k ($k\geq 2$) объектов и соответствующих им ситуаций определим следующим образом: $\rho^{(k)}(X_1,S_1,\ldots,X_k,S_k)=\langle V,S\rangle$, где $V=X_1\cap\ldots\cap X_k$, $S=S_1\cap\ldots\cap S_k$ ($\cap$ - стандартная операция пересечения для булевских данных, для других типов данных рассматривается алгебраически определенная операция сходства).

Мы будем рассматривать язык $\boldsymbol{L}$ с кванторами по кортежам и кортежными термами (см. [2]). Введем, в соответствии с представлением о ситуационном характере причинно-следственных зависимостей, переменные для сорта 3 (ситуации и подситуации, т.е. фрагменты ситуаций) S, $S_1,\ldots,S_n,\ldots$ из $2^{\mathbf{U}^{(3)}}$, соответствующие константы обозначим $\overline{S}$, $\overline{S}_1,\ldots,\overline{S}_n,\ldots$ (переменные и константы сортов 1 и 2 – объектов $X\in 2^{\mathbf{U}^{(1)}}$ и множеств свойств $Y\in 2^{\mathbf{U}^{(2)}}$, соответственно –, определяются стандартным образом, например, в [2]). Элементарным фактом будем называть формулы вида $J_{(v,n)}P(C,A,\overline{S})$, $J_{(v,n)}R(\langle C',\overline{S}\rangle,A,)$. Здесь C, A, $\overline{S}$ - константы сортов 1, 2 и 3, соответственно. Напомним, что для представления фактов в ДСМ-методе вводятся следующие типы внутренних истинностных значений : +1 - "фактическая истина", -1 - "фактическая ложь", 0 - "эмпирическое противоречие" ("конфликт"), τ - недоопределенность.

Внутренние истинностные значения имеют вид $\overline{\nu}=\langle\nu,n\rangle$ или (τ,n), где $\nu\in\{+1,-1,0\}$ - тип истинностного значения, приписанного на n-м шаге применения ППВ; $n\in N\cup\{\omega\}$, $N=\{0,1,2,\ldots,n,\ldots\}$ - множество натуральных чисел, ω - предельный "номер" и вводится из формальных соображений, для удобства определения кванторов в получаемой логике, $\omega\notin N$.

Типы внешних истинностных значений (для представления фактов с оценками и правил правдоподобного вывода ППВ): τ - логическая истина, f - логическая ложь. Пусть $J_\nu\Phi$ - оператор, $J_\nu\Phi=t$, если $\nu(\Phi)=\nu$, $J_\nu\Phi=f$, если $\nu(\Phi)\neq\nu$, где $\nu[\Phi]$ - функция

оценки, $J_{(\mathrm{v},\,\mathrm{n})}\,\Phi \stackrel{df}{=} \bigvee_{i=1}^{n} J_{\langle \mathrm{v},\,\mathrm{i}\rangle}\,\Phi$.

Сформулируем теперь предикат тернарного положительного сходства-2:

$$(\mathrm{V},\mathrm{W},\mathrm{S}_0,k)\quad \exists \mathrm{X}_1\exists \mathrm{Y}_1\exists \mathrm{S}_1\ldots\exists \mathrm{X}_k\exists \mathrm{Y}_k\exists \mathrm{S}_k$$

$$\begin{aligned}&((\mathop{\&}_{i=1}^{k}(J_{(1,\mathrm{n})}P(\mathrm{X}_i,\mathrm{Y}_i,\mathrm{S}_i)\&\\&\forall \mathrm{U}(J_{(1,\mathrm{n})}P(\mathrm{X}_i,\mathrm{U},\mathrm{S}_i)\rightarrow \mathrm{U}\subseteq \mathrm{Y}_i))\&\\&\rho^{(k)}(\mathrm{X}_1,\mathrm{S}_1,\ldots,\mathrm{X}_k,\mathrm{S}_k)=\langle \mathrm{V},\mathrm{S}_0\rangle\&\mathrm{V}\neq\varnothing\&\mathrm{S}_0\neq\varnothing\&\\&\forall i\forall j((\neg(i=j)\&1\le i,j\le k)\rightarrow\neg(\mathrm{X}_i=\mathrm{X}_j))\&\\&\forall \mathrm{X}\forall \mathrm{Y}\forall \mathrm{S}((J_{(1,\mathrm{n})}P(\mathrm{X},\mathrm{Y},\mathrm{S})\&\forall \mathrm{U}(J_{(1,\mathrm{n})}P(\mathrm{X},\mathrm{U},\mathrm{S})\rightarrow\\&\mathrm{U}\subseteq \mathrm{Y})\&(\mathrm{V}\subset \mathrm{X})\&(\mathrm{S}_0\subseteq \mathrm{S}))\rightarrow(\mathrm{W}\subseteq \mathrm{Y}\&\mathrm{W}\neq\varnothing\&\\&(\bigvee_{i=1}^{k}(\mathrm{X}=\mathrm{X}_i)))\&k\ge 2)\end{aligned}$$

Непараметрический предикат ${}_2^3M^{+}_{a,n}\,(\mathrm{V},\mathrm{W},\mathrm{S}_0)$

$\rightleftharpoons \exists k\,{}_2^3\widetilde{M}^{+}_{a,n}\,(\mathrm{V},\mathrm{W},\mathrm{S}_0,k)$.

Подформула

$\forall \mathrm{X}\forall \mathrm{Y}\forall \mathrm{S}((J_{(1,\mathrm{n})}P(\mathrm{X},\mathrm{Y},\mathrm{S})\&\forall \mathrm{U}(J_{(1,\mathrm{n})}P(\mathrm{X},\mathrm{U},\mathrm{S})\rightarrow \mathrm{U}\subseteq \mathrm{Y})\&(\mathrm{V}\subset \mathrm{X})\&(\mathrm{S}_0\subseteq \mathrm{S}))\rightarrow$

$(\mathrm{W}\subseteq \mathrm{Y}\&\mathrm{W}\neq\varnothing\&(\bigvee_{i=1}^{k}(\mathrm{X}=\mathrm{X}_i)))$ описывает эмпирическую зависимость (ЭмЗ), прогнозируемую как искомое причинно-следственное отношение ("подобъект V есть причина наличия свойств W в ситуации S_0"). Подформула

$$\mathop{\&}_{i=1}^{k}(J_{(1,\mathrm{n})}P(\mathrm{X}_i,\mathrm{Y}_i,\mathrm{S}_i)\&\forall \mathrm{U}(J_{(1,\mathrm{n})}P(\mathrm{X}_i,\mathrm{U},\mathrm{S}_i)\rightarrow \mathrm{U}\subseteq \mathrm{Y}_i))$$

описывает экзистенциальное условие (ЭкзУ), характеризующее рассматриваемое множество примеров. ЭмЗ и ЭкзУ являются непременными составляющими всех решающих ДСМ-предикатов.

Если заменить в формулировке предиката тернарного положительного сходства-2 входящее в него условие-2 $\mathrm{V}\neq\varnothing\&\mathrm{S}_0\neq\varnothing$ (относительно сходства *объекта в ситуации* $\langle \mathrm{V},\mathrm{S}_0\rangle$) условиями-1 или -3, приведенными выше, мы получим предикаты тернарного положительного сходства-1

${}_1^3\widetilde{M}^{+}_{a,n}\,(\mathrm{V},\mathrm{W},\mathrm{S}_0,k)$ и тернарного положительного сходства-3

$(\mathrm{V},\mathrm{W},\mathrm{S}_0,k)$, соответственно.

Правила правдоподобного вывода 1-го и 2-го рода (для индуктивного порождения гипотез о причинах и переноса найденных закономерностей по аналогии на объекты с неизвестными свойствами, соответственно, ППВ-*I* и ППВ-*II*) для ситуационного ДСМ-метода приведены в [4, 5].

Нельзя не отметить, что к числу онтологических особенностей социальной реальности следует отнести наличие в исходных данных утверждений с оценкой 0 – "эмпирическое противоречие", т.е. фактов вида $J_{\langle 0,0\rangle}(\mathrm{C}\Rightarrow_1\mathrm{A})$. Конструктивное порождение такой оценки позволяет осуществлять логический анализ модели конфликта. Для этого формулируется предикат тернарного конфликтного сходства [4, 5]. Рассматриваемому миру может отвечать несколько моделей конфликта – "сильная" (когда существуют причины (0)-гипотез), "симметрическая" (когда 0 – результат столкновения (+)- и (-)-причин) и смешанная (объединяющая оба случая). Соответственно, этим моделям отвечают различные варианты ППВ-*I* и ППВ-*II*, а также аксиом каузальной полноты – к.д.о.п.г. (см. [2]).

К предикатам , $i = 1, 2, 3$, $\sigma\in\{+,-\}$ могут быть добавлены эмпирические зависимости ЭмЗ, усиливающие ППВ (аналогично [2]): аналог запрета на контрпримеры, аналог единственности причины, аналог метода различия. Метод различия в ситуационном ДСМ-методе естественным образом допускает три модификации. В соответствии с рассматриваемой структурой мира, выраженной индексом $i = 1, 2, 3$ (который, в свою очередь характеризует, существенность объекта V и ситуации S), может быть рассмотрено как различие объектов, так и различие ситуаций.

Нельзя не остановиться на существенной особенности ситуационного ДСМ-метода: различные комбинации причин R_i и возможных покрытий свойств в P_j $(i, j = 1, 2, 3)$ открывают возможности для построения типологии причинно-следственных (детерминационных) зависимостей в исследуемой области [5]. Стратегия поиска эмпирических зависимостей средствами ДСМ-метода расширяется введением предварительных процедур тестирования. ДСМ-эвристики дополняются диагностическими тестами, устанавливающими типологию исследуемой модели мира $\mathrm{W}^{(\pm)}$.

3. *Обратный метод*

Использование ДСМ-метода для анализа социологических данных не ограничивается изучением детерминант поведения. Метод может оказаться полезным и для изучения опросов общественного мнения (первые попытки такого рода описаны в [8]), в частности, электорального поведения [6]. Следует подчеркнуть, что социологические опросы, включающие, в соответствии с "постулатом поведения", вопросы социального характера и биографические данные, являются непременным этапом подготовки данных и при решении первой задачи. На основании исходных данных, непосредственно полученных как ответы респондента на вопросы социологической анкеты и тестов, формируется множество дифференциальных признаков субъекта. При изучении опросов общественного мнения источником множества характеристик субъекта остается эта же процедура. Лишь после получения структурного описания субъекта мы можем обратиться собственно к изучению опросов, выявляя особенности субъектов, детерминирующие выбор тех или иных ответов на вопросы социологической анкеты (мнение субъекта). Это достигается использованием ДСМ-рассуждений, реализующих обратный ДСМ-метод.

Приведем здесь формулировку решающего предиката обратного положительного сходства из [6]:

$$\widetilde{\widetilde{M}}^{+}_{a,n}(\mathrm{V},\mathrm{W},k) \qquad \exists \mathrm{X}_1 \ldots \exists \mathrm{X}_k \exists \mathrm{Y}_1 \ldots \exists \mathrm{Y}_k$$

$$\Big(\Big(\&_{i=1}^{k}(J_{(1,n)}(\mathrm{X}_i \Rightarrow_1 \mathrm{Y}_i) \& \forall \mathrm{U}(J_{(1,n)}(\mathrm{X}_i \Rightarrow_1 \mathrm{U}) \rightarrow \mathrm{U} \subseteq \mathrm{Y}_i) \&$$

$$\Big(\quad \mathrm{Y}_h = \mathrm{W}) \& \mathrm{W} \neq \varnothing) \& \forall i \forall j((i \neq j \& 1 \leq i, j \leq k) \rightarrow \mathrm{X}_i \neq \mathrm{X}_j) \&$$

$$\forall \mathrm{X} \forall \mathrm{Y}((J_{(1,n)}(\mathrm{X} \Rightarrow_1 \mathrm{Y}) \& \forall \mathrm{U}(J_{(1,n)}(\mathrm{X} \Rightarrow_1 \mathrm{Y}) \rightarrow \mathrm{U} \subseteq \mathrm{Y}) \& \mathrm{W} \subseteq \mathrm{Y})$$

$$\rightarrow (\mathrm{V} \subset \mathrm{X} \& \mathrm{V} \neq \varnothing \& (\vee_{i=1}^{k}(\mathrm{X} = \mathrm{X}_i)))) \& k \geq 2),$$

$$(\mathrm{V}, \mathrm{W}) \qquad \exists k \ \widetilde{\widetilde{M}}^{+}_{a,n}(\mathrm{V},\mathrm{W},k).$$

Там же приводятся определения ППВ-*I* и ППВ-*II* для обратного ДСМ-метода. В результате применения ППВ-*I* ($^{(\sigma)}$), $\sigma \in \{+, -, 0, \tau\}$, порождаются гипотезы о причинах (тех или иных мнений респондентов), т.е. формулы вида $J_{\langle 1,n+1\rangle}(\mathrm{V}_3 \Leftarrow \mathrm{W})$, $J_{\langle -1,n+1\rangle}(\mathrm{V}_3 \Leftarrow \mathrm{W})$, $J_{\langle 0,n+1\rangle}(\mathrm{V}_3 \Leftarrow \mathrm{W})$ и $J_{(\tau,n+1)}(\mathrm{V}_3 \Leftarrow \mathrm{W})$. Соответственно, полученные гипотезы используются для порождения гипотез вида $J_{\langle \nu,n+1\rangle}(\mathrm{X} \Rightarrow_1 \mathrm{Y})$ и $J_{(\tau,n+1)}(\mathrm{X} \Rightarrow_1 \mathrm{Y})$, где $\nu \in \{+1, -1, 0\}$, посредством ППВ-*II*. Для предиката $\mathrm{V}_3 \Leftarrow \mathrm{W}$ формулируются аксиомы каузальной полноты (к.д.о.п.г. [2]) и основанная на них процедура абдукции.

4. *Изучение общественного мнения*

Опишем теперь, какой может быть модель изучения общественного мнения с помощью обратного ДСМ-метода.

Пусть задано множество высказываний $\mathbf{P}=\{\mathrm{p}_1, \ldots, \mathrm{p}_n\}$ (вопросов в социологической анкете), ν - тип истинностного значения (см. выше), $\nu \in \{+1, -1, 0, \tau\}$; пусть, далее, ? - оператор вопроса. Тогда выражение $?J_\nu \mathrm{p}_i$, где $\mathrm{p}_i \in \mathbf{P}$, есть терм, представляющий **вопрос**, p_i будем называть **корнем** вопроса. $?J_\nu \mathrm{p}_i$ интерпретируется как "Имеет ли высказывание p_i оценку ν?".

$?J_\nu \mathrm{p}_i$ является общим ли-вопросом (по терминологии [9], где развита общая теория вопросов), так как ν - переменная, а его примерами являются следующие ли-вопросы: $?J_1\mathrm{p}_i$, $?J_{-1}\mathrm{p}_i$, $?J_0\mathrm{p}_i$, $?J_\tau\mathrm{p}_i$. Соответственно, множество $\{J_1 \mathrm{p}_i, J_{-1} \mathrm{p}_i, J_0 \mathrm{p}_i, J_\tau \mathrm{p}_i\}$, следуя [9], будем называть субъектом ли-вопроса $?J_\nu \mathrm{p}_i$.

Рассмотрим $\mathbf{U}^{(2)}=\{J_\nu \mathrm{p}_i \mid \mathrm{p}_i \in \mathbf{P}, \nu \in \{+1, -1, 0, \tau\}\}$, пусть $\Rightarrow_1 *$, $*_3 \Leftarrow$ - отношения, соответствующие предикатам $\mathrm{X} \Rightarrow_1 \mathrm{Y}$, $\mathrm{Y}_3 \Leftarrow \mathrm{X}$. $\mathbf{U}^{(1)}$ – исходное множество дифференциальных признаков, используемых для задания субъектов и их сходств (возможных причин поведения в соответствии с приведенным выше "постулатом поведения").

Пусть $\mathbf{X}_1$ - множество субъектов, $\mathbf{X}_1 \subset$, а $\mathbf{X}_2$ - множество всех непустых подмножеств элементов из $\mathbf{X}_1$.

Пусть $\Omega = \{\mathrm{q}_j \mid \mathrm{q}_j = J_{\nu_1^{(i)}} \mathrm{p}_1 \& \ldots \& J_{\nu_n^{(i)}} \mathrm{p}_n, i = 1, \ldots, 4^n;\ \nu_1^{(i)}, \ldots, \nu_n^{(i)} \in \{+1, -1, 0, \tau\}\}$, $\Rightarrow_1 * \subseteq \mathbf{X}_1 \times \Omega$, $*_3 \Leftarrow \subseteq \mathbf{X}_2 \times \Omega$. Соответственно, будем рассматривать предикаты: $\mathrm{X} \Rightarrow_1 \mathrm{Y}$, $\mathrm{Y}_3 \Leftarrow \mathrm{X}$, имеющие указанную выше интерпретацию.

Введем предикат $\mathbf{Q}(z, \mathrm{X}, y)$: субъект X дает ответ z на вопрос y. Область определения y - множество ли-вопросов, соответствующее $\mathbf{U}^{(2)}$; область определения X - множество $\mathbf{X}_1$, область определения z - множество $\mathbf{U}^{(2)}$.

Обратный ДСМ-метод порождает гипотезы, позволяющие ответить на вопросы:

(1) "почему субъект X дает ответ $J_{\overline{\mu}}\ \mathrm{p}_i$ на вопрос $?J_\nu \mathrm{p}_i$", где $\overline{\mu} = \langle \mu, n\rangle$, $\mu \in \{+1, -1, 0, \tau\}$;

(2) “почему субъект X имеет мнение q_j”, где $q_j = p_1 \& \ldots \& J_{\nu_n}^{(j)} p_n, j = 1, \ldots, 4^n$, есть ответ на набор вопросов $\langle ? J_{\nu_1}^{(j)} p_1, \ldots ? J_{\nu_n}^{(j)} p_n \rangle$, называемый **мнением**.

Отметим следующий интересный факт: если выполняется к.д.о.п.г для (+)- и (-)-гипотез, то имеет место утверждение
$\forall X \exists V \exists n (\mathbf{Q}(J_\mu p_i, X, ? J_\nu p_i) \rightarrow (V \subset X \& V \neq \varnothing \& J_{\langle \mu, n \rangle}(\{J_\mu p_i\}_3 \Leftarrow V)))$, где $\mu = \pm 1$.

Это утверждение соответствует утверждению, что ДСМ-рассуждение есть конструктивная аргументация. Это означает, что в рамках рассматриваемой модели дается абдуктивное объяснение электорального поведения (ответ на вопрос “почему?”) и указываются соответствующие аргументы.

Можно надеяться, что объединение предложенных расширений ДСМ-метода – ситуационной модели и модели изучения общественного мнения – явится инструментом более глубокого и содержательного качественного анализа социологических данных.

Работа выполнена при поддержке Российского гуманитарного научного фонда (проект № 99-03-19686а).

$J_{\nu_1}^{(j)}$

Литература

[1] Финн В.К. Синтез познавательных процедур и проблема индукции // НТИ. Сер.2. 1999. – № 1-2. – С.8-45.

[2] Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники, сер. Информатика. Т.15. – М.: ВИНИТИ, 1991. – с.54-101.

[3] Климова С.Г., Михеенкова М.А., Панкратов Д.В. ДСМ-метод как метод выявления детерминант социального поведения // НТИ. Сер.2. – 1999. – № 12. – С.3-14.

[4] Михеенкова М.А., Финн В.К. Правдоподобные рассуждения с информацией о ситуации // VII Национальная конференция с международным участием “Искусственный интеллект-2000”, Переславль-Залесский, Октябрь 24-26, 2000, Труды конференции (в печати).

[5] Финн В.К., Михеенкова М.А. О ситуационном расширении ДСМ-метода автоматического порождения гипотез // НТИ. Сер.2. – 2000. – № 11 (в печати).

[6] Гусакова С.М., Финн В.К. Об одной модели электоральногоповедения // VII Национальная конференция с международным участием “Искусственный интеллект-2000”, Переславль-Залесский, Октябрь 24-26, 2000, Труды конференции (в печати).

[7] Поппер К. Логика социальных наук // Вопросы философии. – 1992. – № 10. – С. 65-75.

[8] Finn V.K., Mikheyenkova M.A. On the appiIcation of JSM-Method of automatic hypotheses Generation in sociological investigations // Artificial Intelligence News, special issue. – Moscow. – 1993. – Pp.91-98.

[9] Белнап Н., Стил Т. Логика вопросов и ответов // М.: Прогресс. – 1981.

JSM МЕТОД КАК СПОСОБ ИНТЕГРАЦИИ ДАННЫХ

Robert W. Burch
Department of Philosophy
Texas A&M University

THE JSM METHOD AS A DEVICE OF DATA FUSION

Robert W. Burch

Стандартная техника интеграции данных ограничена ее количественным характером. JSM метод, предложенный В.К.Финном, может быть рассмотрен как способ интеграции данных, который является качественным. Данная статья является кратким исследованием метода Финна.

Standard techniques for data fusion are limited by being quantitative in character. The JSM Method of V. K. Finn can be considered an approach to data fusion that is qualitative in character. This report is a brief survey of Finn’s approach.

1. Introduction

Well-known techniques for data fusion typically involve the application of methods—statistical analysis and inference, Bayesian methods, Kalman Filtering, and the like—that are quantitative. Not all information, however, is natively numerical. When information derives, for example, from reports given in natural language, either written or oral, it is often given in non-numerical terms.

Rather than containing or embodying quantities, such information involves only qualitative assessments. In order for standard quantitative methods to apply to such information, the information must first be converted into some sort of numerical representation. This approach gives at best a rough approximation; at worst it is sheer divination.

Most conversions of qualitative information into quantitative information presuppose answers to a variety of prior questions and often are little more than arbitrary assignments of numbers to qualities. An

example of a technique that lies midway between the best and worst cases is the familiar technique of assigning ratings based on questionnaires with multiple-choice entries such as "always," "usually," "sometimes," "usually not," and "never."

For approximately twenty years the research team of Professor Victor Konstantinovich Finn has been elaborating and implementing a novel approach to problems of automating non-deductive reasoning (also known as plausible inference) that has proved itself in practice to be extremely powerful. Moreover, Finn's approach is fundamentally a qualitative one that bases itself upon mathematical logic (also known as "semeiotic methods") rather than numerical methods. It also provides techniques for certain types of data fusion. Unfortunately, outside of Russia knowledge of Finn's approach is extremely rare. This report is an attempt to disseminate knowledge of Finn's work in the West.

2. The JSM Method

Finn's work depends upon a long logico-philosophical tradition in the theory of inquiry, experiment, and plausible reasoning. The major thinkers in this tradition to which Finn refers are Francis Bacon, John Stuart Mill, Charles Sanders Peirce, and Karl R. Popper. What Finn finds in these thinkers are patterns of non-deductive inference that can be used to generate hypotheses about complex sets of data. The resultant interred hypotheses are the fused data. Thus, in the present context data fusion is the process of reaching such hypotheses on the basis of raw input data. The American logician Charles S. Peirce (1839-1914) coined the word "semeiotic" to designate logico-linguistic topics. He distinguished at least three types of non-deductive inference. Induction was basically statistical inference, inference from sample to population (by means of the so-called "straight rule"). Abduction was inference to an explanatory theory or an explaining cause. Analogy was argument from one thing or group of things to another that is based on similarity between the two things or groups. Any general account of non-deductive inference must certainly provide an account of all three forms: induction, abduction, and analogy.

Finn's most recent work argues that his fundamental technique is a synthesis of all three forms of plausible inference. Finn calls his technique the "JSM Method of Automatic Generation of Hypotheses" in honor of the English thinker John Stuart Mill. Its basic process has two stages, a stage at which hypotheses are generated (brought forth, adduced) and a stage at which hypotheses are either accepted or not accepted. The context in which Finn's approach is applied is reasoning concerning causes and effects of various phenomena.

More exactly, the JSM Method operates after the fashion of ideas that were conceived initially by Francis Bacon in the sixteenth century and were developed extensively by John Stuart Mill in the nineteenth century. The methods of causal reasoning elaborated by Mill are known as "Mill's Methods of Causal Reasoning." These methods include the Method of Agreement (in three variations: the Direct Method of Agreement, the Inverse Method of Agreement, and the Double Method of Agreement), the Method of Difference, the Joint Method of Agreement and Difference, the Method of Residues, and the Method of Concomitant Variation.

Usually, only one of these "Mill's Methods" is relevant to the work Finn has completely implemented so far: the Direct Method of Agreement. In the Direct Method of Agreement, an effect *e* is identified and we are searching for a cause *c* of this effect. In order to find *c* we first list all the situations we know in which *e* is present and then look for the *maximal* factor *c* that is present in these situations. If such a maximal factor *c* is found, then the relation found so far between *e* and c is that ife then c. Hence, a candidate c for being a causal condition of a given effect e is found (in the sense of being a necessary condition for *e)*.

Finn's approach is to generalize this simple idea of Mill. At any given stage of constructing causal hypotheses, that is to say, the crucial task is to construct the *maximal* factor that is present whenever the effect *e* being investigated (which might be either the existence or the non-existence of some set of properties) is present. In order to explicate Finn's thinking, let us begin with the basic semeiotic idea of a "formal language." A formal language consists of a set of symbols (the "vocabulary" of the language), together with rules of formation for constructing the "well-formed formulas" (the wfs of the language). The wfs of a language represent the propositions or sentences of the language. A formal language may be a purely "prepositional" language. Prepositional languages are the simplest sorts of formal languages. Such languages are able to express only whole propositions and the ways in which whole propositions are connected with one another in order to produce other whole propositions. A formal language may be a "quantificational" or "predicational" language (also known as a "predicate" language). Such languages are able to express the ways that the parts of individual propositions are connected with each other in order to express claims that entities have properties and interrelationships.

Predicational languages may have very sophisticated

powers of expression. Indeed, complicated formal languages have been used to express the general structure of many subject domains. The languages that Finn uses to express the subject domains concerning which hypothesis generation is accomplished are predicational languages, not merely prepositional languages.

The signs and the formulas of formal languages are to be regarded as existing quite independently of any logical "meanings" or "values" whatsoever that these signs or sequences of signs might have or might be given. They are, as such, uninterpreted. The discussion of languages and their structures, independently of logical "meanings" or "values," is known as logical "syntax." Syntax is contrasted with logical "semantics," which does attend to logical "interpretations," "meanings," or "values."

Even though logic begins with a language in which inference is accomplished, there also needs to be a second language in which to refer to elements of the first language and in which perhaps to prove theorems related to the structure of this first language. Such a second language was called a "metalanguage" by the Polish logician Alfred Tarski, and theorems about the structure of the first language he called "metatheorems."

The Russian logician Dmitri Bochvar, who was Finn's teacher, discovered a way of making metastatements *about* a language *L* with statements *in* a larger language *L'* that itself contains *L.* The construction is a clever technical device. Bochvar referred to the first language as the "internal" language, and he referred to what serves as the metalanguage as the "external" language. This usage of Bochvar is retained by Finn. For those not overly concerned with technical details, the idea is simply that "internal language" is in effect the object language, and the "external language" is in effect the metalanguage for the object language.

When they are supplemented with what is called an "inferential structure," formal languages become what are called "formal theories," or simply "theories." An inferential structure is a set of what are called "axioms" and a set of what are called "rules of inference." Hence, a theory consists of a language, a set of axioms, and a set of rules of inference. Axioms are intended to express the most basic truths about the subject domain that a formal language is designed to represent. Rules of inference typically have the job of allowing transitions to be made from the basic truths of a theory to other truths, which are typically called "theorems" of the theory. Even after a suitable language is chosen for a subject domain, the subject domain might still be represented in various ways, depending on various choices of axioms and rules of inference for the language.

One of the most distinctive features of Finn's approach is that he uses rules of inference in certain theory-like constructions—which he calls "quasiaxiomatic theories"— to express non-deductive inferences. In particular, his rules express (certain forms of) inference to causes and (certain kinds of) inferences by analogy.

Let us now turn to the topic of formal "semantics," that is to say, the topic of assigning "interpretations," "meanings," or logical "valuations" to propositions of a logical language.

The most basic idea of a "meaning" in logic is the idea of what is called a "truth value." In two-valued logic, the truth values are the truth value T and the truth value F, or the "set of truth values {T, F}." Here we regard every proposition as having the one or the other (but not both) of these two truth values. Also, in two-valued logic we speak of a truth-determination of a proposition as being an assignment to that proposition of a member of the set { T, F}. Thus, we represent the fact that each proposition of a prepositional language has one of the two truth-values T and F by the existence of a function from the set of all propositions of the language to the set {T, F}. We may call such a function a "truth-determination function" or an "interpretation function" or a "valuation function."

It is, however, often desirable to have more than two truth values and a valuation function whose values are other than (merely) T and F. A logical semantics with such values is the basis for what is called a "many-valued logic." For example the set of truth values might be {T, F, I}, where "I" is the "indeterminate truth value." This set could also be represented as {1,-1, 0} or as {0, 1/2, 1} or in some other way. Valuations need not be understood as designating some or other degree of truth; they might, for example, be understood to represent degrees of confidence in (the truth of) a proposition. A similar usage may occur, for example, when probabilities are assigned to propositions. Such valuation functions as probabilities are infinite-valued; indeed, in this case the values are the range of real numbers between 0 and 1, so that the logic is uncountably infinite valued. In case there are some truth values other than T, some of them may represent rather close approximations to truth; we may speak of the union of the set of these and {T} as being the "distinguished (or: designated) truth values." The union of {F} and the truth values that do not represent

close approximations to truth we may call the "non-distinguished (or: non-designated) truth values." Finn has employed various multi-valued logical arrangements to represent certain factors in the process of constructing hypotheses. A typical arrangement he has used, for example, is a set of truth values, in which there two external truth values *T* and *F,* and a countably infinite set of internal truth values of four *types,* namely +1, -1, 0, and t. The value-type +1 is used to represent "factual truth," -1 is used to represent "factual falsity," 0 is used to represent "factual conflictedness," and t is used to represent indefiniteness. In Finn's theories the "external" truth values *T* and *F* are used strictly for what is in effect metalogical appraisals. The truth value types +1, -1, and 0 (and +1/2 and -1/2, if they are used) are said to be the *definite* truth value types or the truth value types of definiteness; t is said to be the *indefinite* truth value type or the truth value type of indefiniteness.

With such an arrangement Finn represents information *gaps.* The truth value type t is assigned to a proposition when there is no knowledge whether the proposition is true or false. And the truth value type 0 is assigned to a proposition for which there is both evidence that it is true and evidence that it is false.

Finn's basic procedure begins from a database with incomplete information, which, as we saw, is represented by assigning the truth value type t to a proposition. Then, evidence is search-for that may enable plausible hypotheses to be generated. Such evidence is used to fill in the missing information with "good guesses." If positive evidence is found and no negative evidence is found, then the truth value type 1 is assigned. If negative evidence is found and no positive evidence is found, then the truth value type -1 is assigned. If both positive and negative information is found, then the truth value type 0 is assigned. If neither positive nor negative evidence is found, then the truth value type is left as t. In general the process results in the diminution of cases in which the value t is assigned. Let us turn briefly to the notion of a "quasiaxiomatic theory" (QAT). A QAT is like an axiomatic theory in that it consists of language, axioms, and rules of inference. But, rather than consisting ofjust a single language, it consists at a minimum of an "internal language" and an "external language." Recall that the internal language is the language for expressing the structure of the subject domain and that the external language expresses the semantics for the internal language.

The axioms of a quasiaxiomatic theory are stratified. One subset of the axioms is called the "knowledge, base." It in turn consists of axioms of two sorts, "core axioms" and "specific axioms." Core axioms contain the basic axioms that apply to all quasiaxiomatic theories. Specific axioms encode the general structure of the subject domain under examination and will vary from subject domain to subject domain. Each quasiaxiomatic theory must be "fine-tuned" to the subject domain it is designed for. Moreover, the specific axioms do not contain any empirical data about any of the particular objects of the domain other than their most general structure. Rather, the empirical data about individual objects of the subject domain is contained in another subset of the axioms, which is called the "database."

The database contains empirical information about the particular objects of the subject domain. For any object and any basic property the database contains either the information that the object has the property, the information that the object does not have the property, or no information at all. In the first event the proposition that the object has the property is assigned the truth value type of +1; in the second event it is assigned the truth value type -1; and in the third event it is assigned the truth value type t. (In the initial database there are no assignments representing conflicting information.) The initial set of all +1 and –1 propositions is sometimes called the "training set" or "training sample" of the initial database, in accord with the vocabulary generally used in the theory of machine learning.

The rules of inference of a quasiaxiomatic theory divide into rules of reliable (that is, deductive) inference (RRI's) and rules of plausible (that is, non-deductive) inference (RPI's). Rules of reliable inference specify how complex properties relate to elementary ones. (A typical arrangement is that a "property" is understood simply as a subset of some given set of basic "attributes.") Rules of plausible inference are rules of two types, the first (RPI-1) specifying how causal propositions (causal hypotheses) are to be inferred from propositions about the properties of objects, and the second (RPI-2) specifying how predictions about previously-unknown properties of objects are to be inferred from causal propositions. These rules will be given only scant further elaboration here.

What is important to note about the RPI's is that they may be formulated in various ways, depending on the underlying criteria that are used for identifying causes, that is to say (in effect) which of the group of Mill's various methods are being formulated in the RPI's. In theory, any subset of the whole class of Mill's methods could be used. In practice, however, the use

of many of Mill's methods simultaneously quickly drives us into computational intractability, and the usual approach is to use only the direct method of agreement together with a condition called "prohibition of counterexamples."

Although the topic of quasiaxiomatic theories is important to Finn's thinking, no further discussion of them will be given here. In fact, as it turns out, there are ways of formulating the JSM Method that do not use quasiaxiomatic theories, as Finn's colleagues have shown. For this reason, the notion of a quasiaxiomatic theory (QAT) should not be conflated with the JSM-Melhod of Automatic Generation of Hypotheses" (the JSM-MAGH) itself. The two ideas are distinct and should be kept separate. A QAT is a logical structure for representing knowledge about a subject domain. The JSM-Method is a method for constructing hypotheses.

To employ the JSM Method, we must begin with some subject domain. This subject domain will contain entities of some more or less uniform sort. (Perhaps these are all chemical compounds, perhaps they are all persons, etc.) The subject domain will also include various properties that the entities may or may not have. We may call a specification of the entities of the subject domain and of the properties that are of interest to us an "ontology" for the subject domain. An ontology for the domain naturally involves a representation of the composition of the entities of the domain.

In addition to including the entities and the relevant properties ofthe subject domain, creating a suitable ontology may be accomplished in various ways. Indeed, the design of a suitable descriptive and identifying system of representation for the entities of a particular subject domain might involve a great deal of creativity. The goal of an ontology is a system of representation such that there is a one-to-one correspondence between descriptions and entities: every entity has a unique descriptive or specifying representation in the system. One way, for example, in which the composition of an entity might be specified is to stipulate that each entity ofthe subject domain is to be understood as a set of "elements" or elementary building blocks. If, ahead of time, we are given the entire set *E* of "elements," then an entity is understood as being a particular subset of *E*. Another way that the composition of an entity might be given is as a vector or "tuple" whose components are entries of +l (for a property known to be present), -1 (for a property known to be absent), or T (for lack of knowledge whether a property is present or absent). Ignoring the difference, therefore, between an entity and its unique representation, we can introduce the neutral word "object." As an object we may consider either an entity a s uniquely represented in the ontology, or else the representation itself of this entity. We can also introduce the word "subobject," which means either some proper part of an entity or else (if one prefers) a fragment ofthe total representation ofthe entity.

Thus, the entities of the subject domain are expressed in terms of an ontology of "objects" and their "subobjects." An example of objects might be chemical compounds, and an example of subobjects might be certain parts of chemical compounds, such as covalent bonds, benzine rings, (OH)- radicals, and the like. The language of representation might be some more or less standard form of representation of chemical compounds, such as three-dimensional chemical diagrams (graphs), the representations of Chemical Markup Language, or the like. A typical way Finn represents chemical compounds is as a set of "pharmacophors," which are standard structural chemical features that have been found by chemists and pharmacologists to be associated with varieties of pharmacological activity.

There is one feature of any ontology of a subject domain, however, that is absolutely crucial if the JSM Method is to be applied to the domain. Finn expresses this necessary feature by saying that the ontology must admit the definition of an "operation of similarity." This means that the ontology is expressible in terms of a semi-lattice. This crucial idea will not be explained in any detail here, but it is necessary in order to define the rules of plausible inference. Finn, in formulating this requirement, is implementing the idea that, in Mill's direct method of agreement, the candidate for being a cause is the *maximal* set of features that are common to all the cases in which a given effect is present. This maximal set of features in Mill's work is expressed in the idea of the *meet* operation of a downward semi-lattice.

As Finn uses the words, "a similarity" of a set of various objects means a subobject that all the objects of the set have in common. Finn's "operation of similarity" is an operation that, when applied to any finite collection of objects, produces the maximal subobject that all the objects of the set have in common. That is to say, the operation of similarity produces the maximal similarity of the objects in the collection. Finn has an additional distinction between a "local" and a "global" similarity. A (more exactly: the) local similarity of a set of objects is their maximal set of common

features. A global similarity is best defined by showing how one is found. We begin with a set of objects; we then find the local similarity of this set of objects (their maximal set of common features); then we find the maximal set of objects that have this maximal set of common features. At this point we have found a pair, which consists of a set of features and a set of objects. The set of features is the maximal set of features possessed by every object in the set of objects; and the set of objects is the maximal set of objects each of which possesses every one of the features in the set of features. Such a pair is called by Finn a "global similarity." By referring to the entire set of global similarities Finn is able to formulate rules of plausible inference that enable predictions to be made concerning which objects have which sets of properties.

The point of insisting on an operation of similarity is simple: in order to construct a candidate for the cause of a given property (more generally: of a given set of properties) in an object, we consider the collection of *all* the objects that do have this property (or: this set of properties), and we find their *maximal* similarity (that is, the maximal subobject that all these objects have). This maximal similarity is a good candidate for being a cause of the presence of the property in an object. For example, if in all chemical compounds (the objects) that are basic (the property under investigation), we find that there is an (OH)- radical (a similarity) and that there is nothing further that all these objects have in common (the similarity is maximal), then having an (OH)- radical is a good candidate for being a cause of being basic. (Finn's actual rules also require that having an (OH)- radical *not* be common to all the objects that are not basic in order that having an (OH)- radical be a good candidate for being a cause of being basic; but this point can be left ignored in the present context.)

To repeat what was said about the first stage of the JSM Method: the first, inductive, sort of step generates hypotheses about causes, while the second, analogical, sort of step generates hypotheses about possessed properties.

The rules for the two kinds of hypothesis generation are said to be rules of plausible inference of the first sort (RPI- 1) and rules of plausible inference of the second sort (RPI-2), respectively. Here we leave the rules of the first and second sorts unspecified. The fundamental idea is that at the completion of each step of application of the JSM Method to a subject domain we have a particular (perhaps new) state of knowledge about the domain. Each state of knowledge divides into knowledge about the properties possessed by objects—let us call it Γ knowledge—and knowledge about the causal relations of subobjects to sets of properties—let us call it D knowledge.

The fundamental idea of the first stage of the JSM Method is to use Γ knowledge at one step to hypothesize as to D knowledge at the next step, and then to use the D knowledge at this step to hypothesize as to Γ knowledge at the still next step. The process zigzags from Γ to D and back again to Γ, over and over until stabilization is achieved.

We begin the whole process with some initial database or base of facts about the subject domain. Each determinate fact in this initial database says that some particular object possesses some particular property or that some particular object does not possess some particular property. There will also be what we might call "indeterminate facts," or (more precisely) facts about the indeterminacy of facts. These "quasi-facts," as we might also call them, express the incomplete information with which the process begins. In the ontology for the subject domain, there will of course be many objects and many properties. In some cases we do not know whether a particular object possesses a particular property, so we have incomplete information. Let us consider every possible pair of the form (object, property). Then, to any such pair we assign the truth value type 1 if the object is known to have the property, -1 if the object is known not to have the property, and T if it is not known whether the object has or does not have the property. Thus, such an assignment constitutes our initial database. The initial database will contain no assignments of 0 (for contradictoriness, or conflictedness); so it contains no overt information about conflictedness of properties in objects, although it may covertly (that is, by implication) contain information which might lead us to hypothesize such at a later point.

Also, the initial database contains no overt information about the causes of the presence or absence of properties in objects, although by virtue of containing information about the properties of objects (and because our ontology contains information about the subobjects of objects), the initial database may covertly (that is, by implication) contain causal information. We may represent the fact that the initial database contains no overt causal information as follows. Let us consider every possible pair of the form (subobject, set of properties). Let us agree to the following: to any such pair we assign the truth value type 1 if the subobject is known to be a cause of the presence of the set of properties in an object, -1 if the subobject is known to be a cause ofthe absence of the

set of properties in an object, and T if it is not known whether the subobject is or is not a cause of the set of properties in an obiect. Since we are assuming that the initial database contains no overt causal information, *all* such pairs will end up being assigned the value type T in the initial database.

It is at this point that Finn's four *types* of (internal) truth values are related to the countably infinite set of (internal) truth values themselves. A truth value in Finn's sense is an ordered pair <v, *n*>, where v is one of the *types* of truth values: 1, -1, 0, and t, and where *n* is a non-negative integer. This non-negative integer is a representative of the step of the application of rules of plausible inference of sort I or sort 2 at which a type of truth value is assigned to a proposition. Hence, what it means for a proposition to have the truth value <*v*, *n*> is simply for that proposition to be assigned the truth value v at step *n* of application of the rules (of either sort). Propositions together with truth values whose second entry is 0 are simply the facts of the initial database S_0 (more specifically of Γ_0). Propositions together with truth values whose second entry is an integer greater than 0 are hypotheses that are generated asconclusions of the rules of plausible inference at some step.

Now that we have discussed the first stage (hypothesis-adducing) of the JSM Method, we are now ready to discuss its second stage (hypothesis-testing and acceptance or rejection). When stabilization is reached at the end of the first stage, a test is performed to check on the explanatory adequacy of the sets of formed hypotheses. This test is based on what Finn calls the "Axiom of Causal Completeness" (ACC).

The ACC test is simple. After stabilization we look at all the facts in the training sample of the initial database, that is at all the pairs of the form (object, property) such that they initially had assigned to them the truth value type +1 or the truth value type -1. After stabilization each such fact should be "explained" by a relevant hypothesis. What it means for (such) a fact to be explained is the following. If the fact has the truth value type +1, then there must be some hypothesis that causally connects one of the subobjects of the relevant object to the presence of the property. If the fact has the truth value type -1, then there must be some hypothesis that causally connects one of the subobjects of the relevant object to the absence of the property. The ACC test is passed if and only if every fact in the initial database is thus explained. The test can be made more exacting by requiring not only that the facts of the initial database be explained by a relevant causal hypothesis but also that the generated predictions about the properties of objects be so explained. This more demanding form of ACC seems to be favored in Finn's recent work.

If the ACC test is passed, then the hypotheses are accepted for sufficient reason; if the ACC test is not passed, then the hypotheses are either rejected or else they tire accepted for insufficient reason. If the hypotheses are rejected, then this fact may signal that the "training sample" of facts in the original database was too small or was otherwise insufficient. In this case, the original database needs to be either expanded or replaced before the JSM Method can yield useful resultant hypotheses.

It has been demonstrated repeatedly, however, that when the initial database is reasonably fertile, even if it be small, the JSM Method can yield powerful and useful results that can be "accepted for sufficient reason."

3 Conclusion

The JSM Method represents genuine progress. It has, just to take one example, a facility for handling data that is qualitative rather than numerical in a way different from, and often superior to, statistical methods. It seems to be quite superior to statistical methods in connection with . qualitative data that are not easily or naturally put into numerical form, and it can produce useful results on databases that contain only small amounts of information. In its techniques for classifying objects, and by implication for forming concepts, it is a useful supplement to statistical methods of cluster analysis, which require the invention of a metric for data that are already numerical in form.

Moreover, the JSM Method of Automatic Generation of Hypotheses is a straightforward, well-motivated set of techniques whose rationale is obvious and whose range of applicability is wide. Additionally, it can be readily formulated in terms other than those of a quasiaxiomatic theory. Moreover, its basic ideas are not very difficult to implement computationally.

The JSM Method points to an avenue of approach to data fusion that can lead along many different particular paths, all united by the fact that methods of mathematical logic are crucial in them. It is a method that has been continually ongoing and changing in the hands of Finn and his colleagues.

Perhaps the most difficult thing to accomplish in applying and computerizing the JSM Method is to find a suitable ontology for the subject domain in such a way that the relevant rules of non-deductive inference can be formulated for it. In general, this task must be done "by hand tailoring" or "fine-tuning" the JSM Method to each separate subject domain.

Finn's research team at VINITI in Moscow maintains

an active research program for developing new JSM systems of application of the fundamental ideas and techniques of the JSM Method. Research domains include not only chemistry and pharmacology but also sociology and social psychology. Such a wide range of topics as these topics display indicates the extremely broad area of potential applications of this system of automatic plausible reasoning.

ВИРТУАЛЬНЫЕ ОТДЕЛЕНИЯ В ГОСПИТАЛЯХ

Tamas Gergely
Лаборатория прикладной логики

VIRTUAL DEPARTMENTS IN HOSPITALS

Tamas Gergely
Applied Logic Laboratory

Виртуальные отделения в госпиталях подразумевают компьютеризацию отделения и снабжение его сотрудников (врачей, специалистов, медицинских сестер, пациентов и т.д.) всеми необходимыми данными, информацией и знаниями. Виртуальные отделения могут быть использованы для таких целей как обучение и тренировка, поддержка совместных работ медицинского персонала в данной области, обеспечение медицинского ухода уникальным образом.

В работе описана одна из виртуальных моделей медицинской организации. Для обоснования использования этой модели, различные потенциальные типы виртуальных моделей будут проанализированы. Одновременно показано, как должен быть организован виртуальный отдел и какие преимущества имеют различные типы виртуальности.

Специальное внимание будет посвящено анализу типов возможных информационных систем, которые служат в качестве информационной базы для виртуального клинического отдела.

Рассмотрены возможные способы реализации создания виртуального отдела.

A virtual hospital department means the computer representation of a department and its actors (physicians, specialists, nurses, patients etc.) together with all the necessary data, information and knowledge. A virtual department may be used for different purposes such as education and training, support of the collaborative work of medical staff in a given discipline, provision of medical care in a unique way.

In order to explain the types of virtuality a model of a healthcare organisation will be described. By the use of this model the potential types of virtuality will be investigated. At the same time it will be shown how a virtual department can be organised and how it can advance from one type of virtuality to another.

Special attention will be devoted to the analysis of the types of possible information systems that serve as an informatics basis of a virtual clinical department.

The realisation of a virtual department means the way it can be used. Different possibilities of realization will also be considered.

1. Introduction

Informatics as one of the main active forces of today provides efficient facilities to satisfy new requirements in healthcare. The requirements appear on the one hand due to the development of clinical medicine and due to health policy on the other. Some of these requirements are as follows:

- To organise the collaborative work of a medical staff in a certain discipline without establishing the corresponding department,
- to introduce a new approach and the corresponding discipline into a hospital without organising a separate department for it;
- to organise hospitals in a more economical way by using informatic facilities as much as possible even for organising and making the departments and the whole organisation function [1];
- to organise distributed health service on the basis of a given discipline by using efficient methods of telemedicine [2].

To satisfy all these requirements an efficient information handling and processing method is required that permit to work with the information imprinting of real clinical situations. Thus appropriate tools are needed to substitute some real clinical situations with their information models, at least partially.

The quick progress of the technical facilities for networks and of the technology for knowledge-based systems allows the partial substitution of real situations by their information models in more and more areas. Under certain conditions these models provide virtual situations for the actors of the corresponding areas. The conditions of using virtuality appear gradually even in health care. However, due to the specificities of health care virtuality can be introduced only after thorough preparation.

In order to understand how virtuality may appear a model of medical organisations will be described in the form of two interacting spaces: the activity space and the information one. By the use of this model we can also answer the questions:

- What are the conditions of the appearance of virtuality?

- What types of virtuality may appear?

This model also helps to explain how virtuality works in the case of health care. It supports the consideration of both aspects the information representation and the realization of virtuality, respectively.

At the end it will be shown how a virtual department can be organised and how it can advance from one type of virtuality to another.

2. The space model

A clinical department has its two characteristic spaces; the information and the activity. The work of a department takes place in these spaces. The functioning of a department takes place in its activity space. The latter consists of the staff and those physical objects and tools that are used in the course of treatment and other processes that take place in the organization. At the same time the elements of the structure of an activity space and the events are fixed in the form of information. All the information about the activities is collected and processed in the information space. Hence the prescriptions and protocols of the activities also belong to this space.

However, the information space contains information and knowledge related to

- Medical professional activities, such as

- relevant biomedical disciplines, that contain theories, models, facts and cases;
- clinical theories, methods and cases with appropriate analysis;
- data processing methods;
- reasoning and decision making methods for clinical use;

- Organization activities, such as

- activity organisation,
- report generation,
- management;

- Infrastructure.

The activity and information spaces are strongly interrelated to each other. While the functioning of a department takes place in its activity space, it is still under the influence of its information space. In other words, the results of this functioning appear in the activity space and at the same time are represented in the information space.

In order to standardise the work, i.e. the activities in the activity space, appropriate prescriptions and protocols can be given in the information space for controlling the activities. Moreover, the methods of analysis and monitoring can be also fixed in the form of guidelines. The unification of the staff's activity in a department makes the latter more efficient. Now if the entire activity space is condensed in the information space we will have an informational copy of the department in question.

Communication necessarily takes place in the information space itself. So information is generated and transmitted as the element of information space. However, only such information is considered that is recorded and available repeatedly for several people. In principle we could also put in the information space information existing only in the mind of the actors of the activity space, or the one sounded without getting fixed, but this would contradict to the criterion of availability. Therefore information sounded inside a communication space but not recorded in proper use does not belong to the information space.

Gathering, storing, processing, and transmitting information takes place in the information space. Connections among the departments and the actors are realized through these. A major part of the economic and financial connections is realized this way as well.

Recording of information always generates a mapping between the activity space and the information space even if this mapping is not complete and is not unfailingly exact. Structures of the elements (data ensembles) of the information space correspond to the elements of activity space. Series of the elements of the information space, that is time series correspond to the processes. There is also a correspondence between the structure of activity space and the information space, e.g. access authority defined in the information space reflects the competency lattice of the activity space to a certain extent. Recording data creates mapping and fineness of the mapping is determined by what is recorded. Mapping need not be complete even in case of proper use. Not every event is relevant to the functioning of the organization, to the goals of the mapping.

Participants of the activity space play a double role concerning the information space: they are information producers and/or information consumers. Both roles are determined by activities done in the activity space. The source of information is the measures and observations necessary to the operation of the activity space or conclusions and decisions of the participants (actors) of the information space. Information generation acts from the activity space towards information space. At the same time the information space influences the activity space, since the actors of the activity space use in their work the information recorded in the information space. However the information space also generates processes that run in the information space. Their execution creates roles that are connected to the activity space only through

the information space (e.g. the whole accounting is such). The processes that take place exclusively in the information space *transform* information.

Activity space can be organized through the information space. The knowledge ensuring the operation of an organization becomes a part of the information space through being put into official documentation (e.g. technological descriptions, Structural Operational Regulations, job descriptions, etc.). These documents describe details of the activity space normatively, that is they say what type of certain components and what processes should take place and how and with which results they should flow. By this they characterize the activity space, and influence it. Therefore we say that knowledge assuring the operation of the organization put into official documents constitute the normative segment of the information space. Normative segment of the information space is guiding the operation of the activity space, and the comparison of the normative specification with the maps of the reflected activities is the control.

The connections of activity space and information space are as follows:

Mapping: The elements of the information space correspond to some components of the activity space. Namely information generation, that is information gathering and recording form such a mapping that acts from the activity space toward the information space;

Information consumption: the actors of the activity space use in the course of their activity information from the information space that is recorded as abstract expressions. This relation is the inverse of mapping;

Influencing: normative prescriptions recorded in the information space are guiding the processes of the activity space.

Processes taking place inside the information space are of two types:

Information processing: Information producing processes creating new information from the information that is realized in the information space. (Sometimes it is difficult to separate the processes of activity space and the processes of the information space. Sometimes a process is realized in both at the same time.)

Control: comparing data to the normative regulation.

Information space is not formed spontaneously from the activity space. As the information space is created only from the appropriately recorded information and is available regularly in a controlled way the information space is to be planned, realized and maintained. Therefore during the planning of the information space of an organization the following demands are to be kept in mind:

Information-consumption necessary for the operation of the information space must be satisfied.

All the processes taking place inside the information space must be met.

These demands are even more important when we intend to concentrate the activity space in an information unit in order to develop a virtual space of activities.

3. Virtual department

The evolution of the use of IT in medicine has reached the current situation where both the activity and information spaces can be represented in an information unit. This unit is considered as a *virtual clinical department (VCD)*.

The representation of an information space may be in the forms of the documents necessary for recording data, e.g. about the patients.

The information space represents the organisation of a hospital together with the department structure. The patients of a hospital may be grouped according to this structure. However, the classification may be done in a different way too, since information may be grouped according to different aspects and principles. Thus a group of patients can also be defined that belong only to a virtual department.

Recent developments in IT allow making a normative segment of the information space independent. Guidelines, represented knowledge and reasoning methods map a huge part of the care.

The development of a virtual department may start with the definition of a given group of patients together with all the necessary administrative data handling methods. Then an independent and so portable normative segment may be added to complete the development of a virtual department. Thus a virtual department will be an actively operating unit of the hospital.

A virtual department contains data, information, knowledge and reasoning methods. Data are the results of measuring or observation. By interpreting data in the corresponding elements of knowledge we obtain information. Facts and cases are examples of information, while models; theories, methods and algorithms are examples of knowledge. Reasoning consists of all the processes of data and information manipulation like relationship extraction, inference, argumentation etc. The four constituents may deal with the organization or patients or medical science and practice.

Depending on the level of presence of the above

constituents we distinguish the following types of VCD:

Administrative VCD, where reasoning methods are mainly oriented on data and information manipulation with respect to the organization and to a certain degree with respect to patient's data. That is, a VCD of this type, virtuality concerns common data and information handling for an organization that exists only virtually. This VCD is administrative in the sense that it supports the administrative functions of the department. Here reasoning is connected to data management.

Professional VCD, where medical professional aspect appears in the information system implicitly. Algorithms, structure of questioners and data models and other elements of the system may embody pieces of medical knowledge. Reasoning represented by data processing algorithms that deal mainly with patient data. However, the system can receive, provide and process data in an interpreted way, so it also executes information management.

Intelligent VCD, where medical knowledge is represented explicitly and appropriate reasoning tools support professional work. Therefore here virtuality is also armed with knowledge management.

Therefore, in VCD, we distinguish *data, information and knowledge management* that are strongly related to the corresponding data and knowledge bases.

Note that there is also a special educational function for VCD. A specially developed VCD can be used for training and education and here the patient records can be handled as virtual patients. If we put into an information unit all the information we know about the mechanism of a disease and all the relevant knowledge about the process evaluation of certain parameters we can define a virtual patient of given parameters. A virtual patient is a model of a patient either with respect to a given disease or with respect to a certain status of the organism of the patient. A virtual patient may be used as an experimental entity or a modelling object.

4. The informatic support of VCD

Since VCDs are realized as special information systems let us see a classification of these systems in order to make the above classification of VCDs clear. We distinguish four types of information systems depending on the functions they may provide.

The *administrative system* provides the following functions:

- Data collection,
- Data storage
- Data processing
- Report generation.

In this type of information systems data management is dominant.

The *professional system* supports the following functions additionally to the aforesaid:

- Definition of professionally important data. The data in this case will be discipline oriented.
- Collection of professionally determined data in the form of questionnaires.
- Information organization in the form of the description of medical cases
- Information handling.

In this type of information systems data and information management is dominant but elements of knowledge management also appear through the cases accumulated during the practice. Moreover, professional knowledge is embodied during the system development when the questionnaire is determined and the structure of the cases is fixed.

The *partner system* supports the following functions additionally to the aforesaid:

- Advising different members of the staff on guidelines and protocols.
- Decision making support with respect to guidelines and protocols.
- Controlling the order of the professional actions given in guidelines and protocols

In this type of information systems knowledge management becomes as important as data and information management and support of reasoning processes appears in the decision-making. Professional knowledge is embodied into the system as a knowledge base that contains the rules that help to realise guidelines and protocols in all possible situations.

The *intelligent partner system* supports the following functions additionally to the aforesaid:

- Information and knowledge acquisition.
- Support of the reasoning processes with various methods such as deductive, inductive and abductive.

Knowledge management is dominant in this type of information systems. These information systems perform professional functions as well as support reasoning and decision-making.

Depending on the information system that serves as the informatic basis for a VCD we can distinguish the following types of VCDs:

- Data oriented administrative VCD (with an administrative information system).
- Information oriented professional VCD (with a professional information system).
- Weakly knowledge oriented VCD (with partner information system).
- Knowledge oriented VCD (with intelligent partner information system).

5. Realisation of VCD

The VCDs are also differing in the way they may be realised. The use of a VCD presupposes its embedding into real clinical situations. This embedding means that the constituents (representations of the information and activity spaces) of a VCD should be mapped into the existing objects and actors of an actual real situation.

5.1 Projection of the constituents of a VCD.

The realisation of a VCD is the projection of its constituents.

The realisation of the virtual information space should be done in the following order:

1. It will be projected into the information space of the real clinical situation,
2. It will be made compatible with the real information space by defining appropriate interfaces,
3. Appropriate communication that involves the virtual information space into the functioning of the clinic in question will be defined.

To realise the virtual activity space the following steps should be taken:

1. The organization should be adequately defined.
2. Adequate actors that can realise the given prescriptions should be selected.
3. The information projection of the activities should be defined.
4. An appropriate infrastructure should be organised for the implementation of the actions.
5. Communication should be appropriately organized.

5.2 Realisation of different types of VCD

The realisation of different type of VCD will be different. The common aspects of the realisation are as follows:

- The organization represented by VCD should be projected onto the real organization and if it is necessary a compromise should be defined.
- The infrastructure should be projected onto the infrastructure of the realising place that may be e.g. one or more clinical departments.
- All the data and information that will be generated during the realized functioning should be collected in the database of the VCD. Of course, this database may be distributed among the databases of the realizing departments, but in this case it should be virtually handled by the VCD.
- The represented knowledge should be transferred to the medical staff and interpreted by it. I.e. specialists of various levels of the given discipline of the VCD should interpret it.

In our approach a VCD can be realized in three different ways:

1. Directly at the *patient's home*. In this case the actors may be in the surroundings of the patient. In this case all the components of a given VCD may be further placed in a distance from the home in question and the constituents will be realized when they are needed. The connection may be occasional or permanent through the Internet.

2. Through *real medical departments*. E.g. in such a realization all the patient records may be controlled by the given VCD. Physically, records may be in the database of the departments but as a whole it belongs to the VCD. Moreover, VCD realizes case analysis and consultations in teleconferencing among the professionals of different departments. Therefore in the real departments the VCD under realization provides a common background for

- Handling patient's data in a uniform way;
- Establishing common case demonstration and discussion;
- Using a given tool kit for data analysis and processing.
- Developing a unique database.

Therefore this type of realization of VCD may take place in various departments, such that the patients physically belong to the realizing department, but all the actions (diagnostics and treatment) will be done in combination with the prescriptions and protocols of the VCD. Members of the staff of the involved departments may have two different kinds of role depending on the VCD to be realized. The first one is related to the staff with members who are professional in the discipline of the VCD. In this case these members of different departments will realize the necessary actions. This part of the staff can be organised into a unique discussion group. The realization of the VDC is nothing but a distributed department with a unique information space. All data processing takes place either according to the needs of the staff or by the prescriptions of VCD.

The second kind of role is when such type of VCD is to be realised that provides reasoning and data processing tools, too. The realization does not require professionals in the discipline in question. It is enough to have in the realising departments professionals who understand the discipline in question and can realize the prescriptions and guidelines with understanding and responsibility. Here the consultation may be realized even with the participation of the intelligent models of the VCD. All the data information are collected and processed by the VCD. Among others this processing results in knowledge maintenance and acquisition on the basis of the accumulated cases.

Therefore the realization of a VCD depends on the type of system that represents it. This means that the simplest VCD is the one where the computer supports the patient records and all the knowledge in the corresponding discipline belongs to the staff. The next step is the professional VCD, where the virtual system supports the specialist with appropriate questionnaires and guidelines. However the latter are only in virtual form. The computer support of the guidelines permits to control the staff behaviour and/or advices the appropriate steps.

3. In a *virtual hospital* form, where a VCD is realized when it is required by the state of a patient. In this case all the constituents should be adequately realized starting from the realisation of the organisation. This can be done in a hotel-based structure using matrix principles [1]. In this case in the same structure different virtual disciplines given as VCD can be realized.

6. Virtual immunological department

In clinical practice a new situation has appeared. The traditional nosological approach in more and more cases is inadequate. Due to the growth of the external and internal pollution people's organism became overactive i.e. allergic. This often changes the morbidity and so the course of a disease will be different from case to case. Therefore treatment requires an individual approach. The solution in this situation is to add a special discipline to enforce and support the nosological approach. The required approach exists in immunology with appropriate methods to give more accurate diagnosis and to provide individually optimal treatment that could be used as adjuvant toolset for most of the clinical disciplines (see e.g. [3]). Thus the organisation of a virtual immunological department can improve the efficiency of a clinic. Such a virtual immunological department is under development in the Military Hospital, Budapest.

References

[1] Gergely, T., Mikola, I., Patient centered healthcare institution, Modern Hospital, Kecskemйt, 1998, pp.76-86.(in Hungarian)

[2]Viegas, S.F., Dunn K. (Eds.), Telemedicine, Lippincott-Raven, Philadelphia, 1998.

[3]Gergely, T., Seniouk, O., Immunological Diagnosis and Treatment Optimisation, Naukova Dumka, Kiev, 1993. (in Russian)

СРАВНЕНИЕ АЛГОРИТМОВ ПОСТРОЕНИЯ МНОЖЕСТВА ПОНЯТИЙ И ЕГО ДИАГРАММЫ ХАССЕ

С. А. Объедков

ON THEORETICAL AND EXPERIMENTAL COMPARISON OF ALGORITHMS GENERATING THE SET OF ALL CONCEPTS AND ITS LINE DIAGRAM

S. A. Ob"edkov

Several algorithms that generate the set of all formal concepts and line (Hasse) diagrams of concept lattices are revIewed. Algorithmic complexity of the algorithms was studied both theoretically (in the worst case) and experimentally. The results of this study are presented. Conditions of preferable use of some algorithms are given in terms of density/sparsity of underlying formal contexts.

1. Основные определения

В формальном анализе понятий (Formal Concept Analysis, FCA) [Ganter et al., 1998] формальный контекст определяется как тройка $<G, M, I>$, где G - множество объектов, M - множество признаков, $I \subseteq G \times M$.

Для и :

;

$$B' = \left\{ g \in G \mid \forall m \in B \left(\langle g, m \rangle \in I \right) \right\}.$$

Формальное понятие формального контекста $<G, M, I>$ есть пара (A, B), где $A \subseteq G$, , $A' = B$ и $B' = A$. A называется объемом, а B - содержанием понятия (A, B).

Задача ДСМ-метода автоматического порождения гипотез [Финн, 1991] может быть сформулирована в терминологии FCA следующим образом. Рассматриваются положительный $<G^+, M, I^+>$ и отрицательный $<G^-, M, I^+>$ контексты, являющиеся таковыми относительно некоторого целевого атрибута: объекты положительного контекста обладают свойством, обозначаемым этим атрибутом, а объекты отрицательного контекста - не обладают. Гипотеза ДСМ-метода о причине данного свойства формулируется как содержание понятия положительного контекста, не входящее ни в какое содержание отрицательного понятия (гипотеза с запретом на контр-пример). Построение ДСМ-гипотез подразумевает использование алгоритма порождения всех (или, по крайней мере, некоторых) понятий.

На множестве понятий определен частичный порядок:

, если $A \subseteq C$, что эквивалентно

(и соответствующий строгий порядок <); в этом случае понятие (A, B) называется менее общим, чем понятие (C, D). Говорят, что понятие X является соседом понятия Y снизу ( ), если X < Y и не существует понятия Z, такого что X < Z < Y. Множество понятий формального контекста образует решетку. Диаграмма Хассе - стандартный способ представления решетки. Диаграммой Хассе для контекста (G, M, I) называется пара $\langle L(G,M,I), \prec \rangle$, где L($G$, M, I) - множество понятий контекста (G, M, I), а $\prec$ - определенное выше отношение.

2. Обзор алгоритмов порождения понятий

В литературе уделяется большое внимание проблеме построения множества понятий и диаграммы Хассе данного контекста. Известно, что число понятий может быть экспоненциальным по отношению к размеру исходного контекста, а вычисление этого числа является #P-полной задачей. Поэтому, с точки зрения сложности в худшем случае, оптимальным можно считать алгоритм, порождающий все понятия и/или строящий диаграмму понятий за время линейное от размера входного контекста. С другой стороны, "плотные" контексты, где число понятий экспоненциально, не столь часто встречаются на практике. Нами было проведено теоретическое и экспериментальное сравнение алгоритмов, позволившее выявить обстоятельства, в которых те или иные алгоритмы работают быстрее прочих.

В большинстве случаев оказалось возможным предложить существенные улучшения и дополнения к оригинальным версиям рассматриваемых алгоритмов. Только два алгоритма в оригинальной формулировке строят диаграмму Хассе; оказалось возможным, не нарушая сложностных оценок, модифицировать прочие алгоритмы с тем, чтобы их также можно было использовать для построения диаграммы.

Алгоритмы делятся на две категории: пошаговые алгоритмы [Carpineto et al., 1996], [Godin et al., 1995], [Norris, 1978], на I-ом шаге строящие множество понятий или диаграмму Хассе для множества, состоящего из первых I объектов исходного контекста, и пакетные алгоритмы [Bordat, 1986], [Chein, 1969], [Ganter, 1984], [Lindig, 1999], [Забежайло и др., 1987], [Кузнецов, 1993], строящие множество понятий и его диаграмму сразу для всего множества исходных объектов. Кроме того, данные задачи обычно решаются в рамках одной из стратегий: сверху вниз (от максимального объема к минимальному) или снизу вверх (от минимального объема к максимальному).

К пакетным алгоритмам относятся алгоритм Борда [Bordat, 1986] и его модификации, работающие в стратегии сверху вниз: сначала порождается понятие с максимальным объемом, затем для каждого порожденного понятия строятся его соседи снизу. Эти алгоритмы различаются способом порождения соседей снизу и механизмами, позволяющими предотвратить повторное порождение некоторых понятий. Пакетные алгоритмы Гантера [Ganter, 1984] и "Замыкай по одному" [Кузнецов, 1993] работают в стратегии снизу вверх и используют понятие каноничности порождения, основанное на лексикографическом порядке, что позволяет эффективно распознавать случаи повторного порождения. Эти алгоритмы имеют временную сложность, линейную от числа понятий.

Среди пошаговых алгоритмов особого внимания заслуживают алгоритмы Норриса [Norris, 1978] и Годана [Godin et al., 1995]. Первый из них также линеен от числа понятий и по сути является пошаговым вариантом алгоритма "Замыкай по одному". Сложность алгоритма Годана в худшем случае квадратична относительно числа понятий, что делает нецелесообразным применение этого алгоритма для построения диаграммы больших и "плотных" контекстов (контекстов, где $|I|$ не слишком мало по сравнению с $|G \times M|$). Однако в случае разреженных контекстов алгоритм Годана работает лучше прочих алгоритмов.

3. Результаты экспериментального сравнения

Было проведено тщательное экспериментальное сравнение вышеназванных и некоторых других алгоритмов на контекстах различных размеров и плотности, которое показало, что выбор алгоритма для построения множества понятий и диаграммы Хассе имеет смысл соотносить со свойствами исходных данных. Общий принцип таков: алгоритм Годана стоит применять на малых разреженных контекстах; в случае плотных контекстов разумно использовать алгоритмы, основанные на линейной по времени от числа исходных объектов проверке каноничности: "Замыкай по одному", Норриса. Наша реализация алгоритма Борда демонстрирует хорошие результаты на контекстах средней плотности, особенно при построении диаграммы Хассе.

Результаты теоретического и экспериментального сравнения алгоритмов изложены более подробно в [Kuznetsov et al., 2000].

Литература

[Забежайло и др., 1987] Забежайло М.И., Ивашко

В.Г., Кузнецов С.О., Михеенкова М.А., Хазановский К.П., Аншаков О.М. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез. // НТИ, Сер. 2. 1987. № 10.

[Кузнецов, 1993] Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из конечной полурешетки. // НТИ, Сер.2. 1993. № 1.

[Финн, 1991] Финн, В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ. // Итоги науки и техники, Сер. Информатика. 1991. Т. 15 (Интеллектуальные информационные системы).

[Bordat, 1986] Bordat J.P. Calcul pratique du treillis de Galois d'une correspondance. // Mathematiques et Sciences humaines. 1986. P. 96.

[Carpineto et al., 1996] Carpineto C. and Giovanni R. A Lattice Conceptual Clustering System and its Application to Browsing Retrieval. // Machine Learning. 1996. P. 24.

[Chein, 1969] Chein M. Algorithme de recherche des sous-Matrices premieres d'une Matrice. // Bull. Math. Soc. Sci. Math. R.S. Roumanie. 1969. P. 13.

[Ganter, 1984] Ganter B. Two Basic Algorithms In Concept Analysis, FB4-Preprint No. 831 - TH Darmstadt, 1984.

[Ganter et al., 1998] Ganter B. and Wille R. Formal Concept Analysis. Mathematical Foundations - Springer, 1998.

[Godin, 1995] Godin R., Missaoui R., and Alaoui H. Incremental Concept Formation Algorithms Based on Galois Lattices. // Computation Intelligence. 1995.

[Kuznetsov et al., 2000] Kuznetsov, S.O. and Objedkov, S.A., Algorithms for the Construction of the Set of All Concepts and Their Line Diagram, Preprint Technische UniversItat -Dresden, MATH-AL-05-2000, June 2000.

[Lindig, 1999] Lindig C., Algorithmen zur Begriffsanalyse und Ihre Anwendung bei Softwarebibliotheken, (Dr.-InG.) Dissertation - Techn. Univ. Braunschweig, 1999.

[Norris, 1978] Norris E.M. An Algorithm for Computing the Maximal Rectangles in a Binary Relation. // Revue Roumaine de Mathematiques Pures et Appliquees. 1978. № 23(2).

К ПРОБЛЕМЕ ПРЕДСТАВЛЕНИЯ ИНТЕГРИРОВАННЫХ ХИМИЧЕСКИХ И БИОЛОГИЧЕСКИХ ЗНАНИЙ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ ТИПА ДСМ

Е.С.Панкратова
ВИНИТИ

ON PROBLEM OF REPRESENTATION OF HYBRID BIOCHEMISTRY KNOWLEDGE FOR INTELLIGENT SYSTEMS OF JSM-TYPE

E.S. Pankratova

Интеллектуальная система (ИНТС) ДСМ [1] представляет собой интегрированную интерактивную систему, в которой на базе развитого логико-математического обеспечения, реализующего ДСМ-метод автоматического порождения гипотез (АПГ), осуществляется интеллектуальная обработка информации, представленной в БД с неполной информацией (БДНИ) и в БЗ. ИНТС ДСМ применяется для прогнозирования свойств структурированных объектов в БДНИ для задачи фармакологии, технической диагностики и социологии.

ИНТС ДСМ отличается открытостью решающих средств, т.е. возможностью настраивать ее на предметную область. ИНТС ДСМ применялась ранее для решения задачи прогнозирования зависимости “структура химического соединения - множество биологических активностей” и успешно использовалась экспертами - фармакологами.

В настоящее время исследуется возможность применения ИНТС ДСМ в новой предметной области.

Цитологические и цитохимические исследования [2] могут дать достаточное представление об эффективности противоопухолевой химиотерапии как в ходе лечения, так и после него. В то же время очень важно определять эффективность антибластических препаратов перед лечением, чтобы выбрать наиболее подходящий из них в каждом отдельном случае. Но ввиду многообразия морфологических признаков дистрофии клетки возникают трудности по их объективной оценке. Некоторые исследователи ввиду невозможности одновременного охвата всех морфологических признаков, идут по пути минимизации количества значимых (по их мнению) признаков. При этом, естественно, не все объективные данные учитываются при принятии окончательного решения.

Все морфологические признаки, характеризующие повреждение клетки цитостатиками, можно разделить на групповые, отражающие расположение опухолевых клеток в исследуемом препарате, и на признаки, характеризующие отдельную клетку и ее структурные элементы: ядро, цитоплазму, ядрышки, т.е. признаки клетки.[1] Таким образом, возможно структурированное описание опухоли, - как мно-

жества морфологических признаков.

Очень важно определять эффективность антибластических препаратов перед лечением, чтобы выбрать наиболее подходящий из них в каждом отдельном случае.

С целью более точного выявления цитологических признаков, характеризующих индивидуальную чувствительность, больные разделяются на три группы соответственно данным радиоиндикации:

1) с чувствительными опухолями, т.е. с выраженной ингибицией синтеза ДНК и/или РНК под влиянием нескольких химиопрепаратов и со слабо выраженной чувствительностью к нескольким химиопрепаратам;

2) с опухолями, нечувствительными ни к одному из проверенных химиопрепаратов;

3) с опухолями, малочувствительными к одному или нескольким химиопрепаратам или более чувствительными только к одному из них.

Учитывая перечисленные сведения о предметной области, мы предлагаем следующую возможную постановку ДСМ-задачи.

Объектом в терминологии ДСМ является множество морфологических признаков, характеризующих исходную цитологическую картину. Возможен также вариант двухкомпонентного объекта, где, кроме цитологической картины, учитывается антибластический препарат (существует несколько форм представления его химической структуры).

Свойство - наличие или отсутствие терапевтического эффекта (или, другими словами, определение индивидуальной чувствительности опухоли).

Гипотеза 1-го рода в терминологии ДСМ представляет собой следующее утверждение: “совокупность морфологических признаков опухоли является причиной положительного/отрицательного терапевтического эффекта от введения некоторого химического вещества ”.

В результате работы системы предлагается прогноз: ”будет ли терапевтический эффект от воздействия конкретного химического препарата на конкретную разновидность опухоли, описанную в виде множества морфологических признаков[2] ”.

Возможность проведения подобной работы зависит от предоставления конкретных данных из цитологических лабораторий, которых мы в настоящее время еще не имеем.

[1] В описании цитологической картины участвует более 80 морфологических признаков.

[2] Более полной эта задача была бы с привлечением данных об истории болезни конкретного больного.

Литература

1.Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ //Итоги науки и техн. Сер.Информатика.-М.:ВИНИТИ,1991.-Т.15.С. 54-101.

2.Зитаре И.Я. Патоморфоз при химиотерапии опухолей //Рига,”ЗИНАТНЕ”,1984.

НОВАЯ ВЕРСИЯ ФКСП И ЕЕ ПРИМЕНЕНИЕ

В.Г. Блинова, Д.А. Добрынин

A NEW VERSION OF FCSS-NOTATION AND ITS APPLICATIONS

V.G. Blinova, D.A. Dobrynin

Язык ФКСП (фрагментарный код суперпозиций подструктур) используется для описания структуры химических соединений и представления их в форме, пригодной для компьютерной обработки. Этот язык был предложен В.В.Авидоном в начале 70-х годов. Он предназначен для дискретного описания химического соединения в виде набора всех имеющихся в нем подструктур, представляющих собой центры локализации пи-электронов, как-то: гетероатомы или ароматические циклические системы, комплексы кратных углеродных связей, соединенные между собой цепью атомов углерода. Эти подструктуры, несущие элестростатический заряд, являются активными центрами, которые потенциально могут взаимодействовать с рецепторами ферментов, отвечающих за химические реакции внутри живых организмов. Таким образом, при кодировании с помощью ФКСП молекулярная структура химического вещества преобразуется во множество символьных дескрипторов, которые несут полезную информацию о его биологических свойствах.

В настоящее время реализована новая версия кодировщика ФКСП, которая имеет следующие отличительные особенности от предыдущей версии, созданной Лейбовым А. :

- для преобразования в набор дескрипторов ФКСП химическое соединение должно быть представлено файлом в MDL-формате. Данный формат является международным стандартом и поддерживается многими химическими редакторами.
- набор активных центров, которые используются при построении дескрипторов ФКСП кода, можно изменять.
- кодируются полициклические структуры путем разбиения на цепочки из циклов различной длины. Возможность сравнения между собой полициклических структур сильно расширяет область применения ФКСП кода.
- при кодировании химического соединения сохраняются и могут быть использованы в дальнейшем привязки ФКСП дескрипторов к атомам и связям, которые образуют данный дескриптор.

Новая версия ФКСП кодировщика реализована в виде независимого модуля на языке C++. Это позволяет легко встраивать данный ФКСП кодировщик в различные системы, использующие ФКСП код в качестве языка представления химических соединений.

В настощий момент ФКСП кодировщик используется в новой системе анализа структура-активность с помощью ДСМ-метода порождения гипотез с визуализацией исходных химических соединений и полученных гипотез в виде пространственных 3D структур. В качестве исходных данных используется набор файлов MOL формата. Одно соединение представляется одним файлом. Для правильного кодирования соединение, представленное MOL файлом, должно включать в себя атомы водорода, указанные в явном виде. Кроме того, для корректного представления результатов эксперимента в виде пространственных 3D структур, атомы молекулы должны иметь истинные или близкие к ним трехмерные координаты. Оба эти требования выполняются, если подготовку исходных данных проводить с помощью трехмерного химического редактора. Авторами использовался широко распространенный редактор HiperChem, который позволяет строить и оптимизировать трехмерные молекулы химических соединений.

После подготовки исходных файлов соединений, они загружаются в базу данных эксперимента. При загрузке незаметно для пользователя осуществляется преобразование исходного соединения во множество дескрипторов ФКСП кода. Реальная скорость преобразования в ФКСП код так высока, что практически не заметна для пользователя.

Свойства новой реализации ФКСП кодировщика позволяют использовать его для системы исследования биотрансформаций. В отличие от первого варианта на вход кодировщика подается MOL файл, в котором присутствуют дополнительные параметры - признаки помеченных атомов. Алгоритм кодирования отличается тем, что для тех дескрипторов ФКСП, которые содержат помеченные атомы, создаются копии с модифицированным именем дескриптора. Например, в начало строки такого дескриптора добавляется какой-нибудь символ. Такой двойной набор дескрипторов позволяет в результате дальнейшего эксперимента отследить окружение выделенных атомов молекулы. При исследовании реакций биотрансформации нужно определить, при каком окружении, т.е. наличии каких дополнительных центров идет реакция с конкретным выбранным центром молекулы. Двойное ФКСП кодирование позволяет эффективно решить эту задачу.

ПОИСК СТРУКТУРЫ В ТЕКСТЕ ОГЛАВЛЕНИЯ

С.В. Израйлит
специализация "Искусственный интеллект", ФПМЭ, МФТИ.

TEXT STRUCTURE ANALYSIS FOR TABLES OF CONTENTS

S.V. Israilit

1. Введение - ценность данной работы

Как известно, систематизация информации сегодня является одной из важнейших задач развития. В этом свете становится важным классифицировать тексты и их фрагменты, находить и узнавать структурные образования в тексте.

Существует ряд стереотипов, связанных с оформлением текстов, которые содержат набор различных по своему смысловому содержанию элементов. Человек, стремясь сделать текст более наглядным, несознательно или намеренно выделяет разные элементы, выравнивая по-разному, шрифтом, специальными символами, акцентируя внимание жирным шрифтом или курсивом и т.д. А раз так, то подобные закономерности могут быть выявлены, опираясь на эти стереотипы, в частности методом правдоподобных рассуждений В.К. Финна [1].

2. Постановка задачи и необходимый формализм

Для применения ДСМ-метода правдоподобных рассуждений мы должны сформулировать некоторое поле свойств текстового фрагмента, чтобы затем реализовать процедуру поиска и выявления сходства по этому кругу свойств. Этот вопрос является ключевым в данной работе, поскольку текст в чистом виде содержит большое количество информации, организация поиска свойств по которой, вероятно, бессмысленна. Таким образом, нам следует абстрагироваться от отдельных шрифтов, форм, или позиций фрагментов (а тем более от отдельных букв), и оперировать только избранным набором простых характеристик каждого фрагмента текста.

В данной работе наименьшим фрагментом текста считается слово. На вход алгоритма разбора подаётся слово и минимальный набор признаков, в частности, координаты его абстрактного расположения на листе документа, номер шрифта, цвета и размера по (некоторому фиксированному набору). Точное описание входного формата представлено в приложении 1.

Самую явно выраженную структуру имеют различные отчёты, списки, оглавления. Поэтому будем рассматривать здесь обработку одного из них, например, оглавления.

Итак, в нашем исходном файле мы имеем оглавление. Основной задачей распознавания следует отметить выделение статей, и разнесения их содержимого по заданному набору полей. Также отличительной чертой статей в оглавлении является присутствие обязательного элемента - страниц.

Таким образом, конечная постановка задачи будет следующая - есть файл, содержащий слова и некоторый набор их признаков, получить базу данных статей, где каждая статья будет представлена в виде совокупности полей.

3. Первый этап алгоритма - построение формального представления

Изначально необходимо преобразовать набор слов в расширенный, где служебные и грамматические фрагменты текста будут отделены от, собственно, слов. Для этого в программе реализован набор инструментов, позволяющих в автоматическом режиме выделить эти куски (разделители), а также внутренняя система связи разделителей с текстом. Далее производится разбор каждого слова по таблице вхождения в него 4 основных типов символов - маленьких и больших букв, цифр и прочих символов. Выделяются степень заполнения и наличие крайних вхождений. Аналогичной процедуре подвергается разделители.

Далее всё множество слов делится на классы эквивалентности по совокупности входных и полученных характеристик. Сами наборы характеристик выделяются в виде стилистических векторов.

Следующим шагом производятся поиск порядка по вертикальной координате и выделение строк. Далее всё множество слов упорядочивается по строкам и по своему положению в строке. Аналогичная процедура проводится по горизонтали (по началу и концу слова). Только в результате мы получаем не набор строк, а набор вертикальных выравниваний. Все выравнивания делятся на две группы - вовлекающие большое удельное количество элементов и "случайные" совпадения, которые могли получиться из-за совпадений тех или иных координат. Следует особенно отметить, что в реальных поставщиках текстов (например, сканирование с бумажного носителя) возможен некоторый шум по каждой из координат, поэтому и строки и столбцы формируются не по равенству соответственных координат, а по попаданию в некоторый диапазон значений относительно друг друга.

Далее, выделяются те наборы элементов, которые находятся в начале и конце строки, и отмечаются соответствующим образом. И, наконец, выделяются группы, которые образованы сходными стилистически словами, и имеют одно и то же выравнивание.

Совокупность всех перечисленных подразделений и представляет исходный формализм, который предстоит обрабатывать.

4. Второй этап алгоритма - построение гипотез интерпретаций

Далее нам необходимо породить гипотезы того, каковы могут быть поля. Это делается на основании введения нескольких размеченных примеров. Этот блок находится сейчас в стадии разработки, а в текущей реализации программы используется небольшой набор гипотез, которые были занесены в ручную.

Здесь следует отметить, что из-за по-прежнему большого набора информации (который, не смотря на свою громоздкость всё же гораздо более адекватен поставленной задаче) предполагается использовать модификацию ДСМ-метода. Суть модификации в том, что гипотеза включает в себя ровно тот набор характеристик положения слова и его окружения, который является достаточным, чтобы опровергнуть все отрицательные примеры. Если количество отрицательных примеров расширяется, то расширяется и описательная мощность гипотезы (т.е. сужается поле её деятельности). Если за счёт этого некоторые примеры выпадают из области уверенной интерпретации всего набора гипотез, то на их множестве порождаются новые и т.д.

Такой метод обработки информации является попыткой застраховаться от комбинаторного взрыва, неизбежного в таких случаях. Однако подробные исследования устойчивости метода ещё не проводились, так что это является основой дальнейших работ.

5. Третий этап алгоритма - использование гипотез интерпретации

Когда гипотезы доступны, начинается анализ набора слов и их характеристик. С учётом специфики предмета (обязательный элемент - номер страницы) анализ начинается с выяснения расположения номеров страниц. Здесь предполагается, что порядок полей в статье всегда одинаковый, и номера страниц представлены в ней один раз.

На основании этого свойства и гипотез, касающихся номеров страниц, находятся страницы, и фиксируется сходство в их расположении относительно других элементов. Далее по этому признаку образуются статьи, и проводится анализ остальных элементов. Найденные элементы разносятся по полям и записываются в файл требуемого формата.

Приложение 1 - Формат входного файла.

mainstr = string[65];

```
{Такие слова используются в данной версии}
minwrd = record
{Считываемый тип слова}
      l  :byte;
{Длина слова}
 x1,x2,y,y2  :word;
{Координаты коробки слова}
 s,r,c :byte;
{Тип написания, Размер, Цвет}
  ins  :mainstr;
end;
f    :file of minwrd;
```

{Файл слов - служебная информация каждого слова расположена в порядке объявления, т.е. в начале 1 байт - длинна слова, потом 2 байта - меньшая координата X1, 2 байта - большая - X2, 2 байта - большая (основная) У, 2 байта -меньшая У2 и т.д., на каждую букву - 1 байт, нумерация - в пикселях от верхнего левого края}

Данный файл открывается по стандартным правилам языка программирования Pascal 7.0 for DOS. На сегодняшний момент автоматическая перекодировка файлов другого типа в требуемый находится в стадии разработки. Принципиальный алгоритм заключается в следующем:

1) Открывается файл вида .rtf или сходный с ним по структуре;

2) По порядку читаются блоки данного формата, в специальной среде выставляются текущие свойства блока;

3) Блок разделяется на слова, потом с помощью встроенных функций определения размеров текстового сегмента каждому слову приписывается текущий координатный набор, соответствующий оформлению его блока и его местоположению, текущий шрифт, размер и т.д.;

4) Берётся следующее слово, текущий координатный набор изменяется в соответствии с размерами и положением предыдущего слова, потом повторяется 3), 4);

5) Набор слов сохраняется в файл требуемого формата. Какая-либо сортировка не требуется, программа-приёмник сама всё сделает.

Литература

[1] Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники, сер. Информатика. Т.15. - М.: ВИНИТИ, 1991. - с.54-101.

СЕГМЕНТАЦИЯ РУССКОГО ПРЕДЛОЖЕНИЯ (ПОВЕРХНОСТНО-СИНТАКСИЧЕСКИЙ АНАЛИЗ КАК САМОСТОЯТЕЛЬНЫЙ МОДУЛЬ АНАЛИЗА ТЕКСТА)

Т.Ю. Кобзарева, Д.Г. Лахути, И.М. Ножов

RUSSIAN SENTENCES SEGMENTATION: SURFACE-SYNTACTIC ANALYSIS AS INDEPENDENT MODULE OF TEXT ANALYSIS

T.Yu. Kobzareva, D.G. Lakhutl, I.M. Nozhov

Работа, о которой идет речь в предлагаемом докладе, является фрагментом системы поверхностно-синтаксического анализа предложений на русском языке, разработка которой была начата в 1971 году в Информэлектро в отделе Д.Г.Лахути группой Г.А.Лесскиса, участником которой была Т.Ю.Кобзарева. В настоящем докладе будут рассмотрены, во-первых, некоторые теоретические основы принятого в этой работе подхода, и во-вторых, состояние работ по ее программной реализации, осуществляемой И.М. Ножовым.

Об уровне поверхностно-синтаксическом (о синтаксисе линейной структуры текста - на первом этапе - предложения) можно говорить с двух точек зрения: правомерно ли рассматривать явления этого уровня отдельно, т.е. о лингвистической обоснованности отдельного его рассмотрения и о прагматическом интересе вычленения его в отдельную задачу.

При анализе связей в предложении “Ясным солнечным днем он любовался дорогой“ к построению двух способов "понимания" (он любовался (чем?) дорогой (когда?) ясным солнечным днем и он любовался (чем?) ясным солнечным днем (когда?) дорогой) приводит омонимия слова дорогой.

Факт существования синтаксических неоднозначностей (неоднозначностей при разбиении предложения на сегменты, их интерпретации и связей), разрешение которых часто определяется экстралингвистическим контекстом (не структурой компонент), относится к проблеме денотативной структуры предложения (S), однако уже на нашем уровне анализа требует исчисления синтаксической омонимии.

Предложение “Рядом с домом, на берегу, под горой, росли кусты” можно понять как 1. рядом с домом и на берегу, а именно под горой, где уточняющий оборот под горой можно отнести как а) ко всей сочиненной группе, так и б) только к последнему ее члену 2. рядом с домом, а именно на берегу, а именно под горой, т.е. как включающее два уточняющих оборота. Т.о. возникают 3 варианта членения и интерпретации сегментов и их связей.

“Он видел отца, работающего в саду соседа, старика и Ваню”. В этом примере неоднозначны функции обеих запятых: первая может быть левой границей обособленного причастного оборота или же сочинять слова отца и соседа. Вторая запятая мо-

жет, соответственно, быть правой границей причастного оборота или же сочинять слова соседа и старика. Конечно же, отца и соседа случайно совпали по роду, и при замене на работающую в саду мать неоднозначность исчезает. Но при замене на работающего в саду соседа брата возникает дополнительно (внутри причастного оборота) возможность неоднозначной интерпретации связей: то ли "видел отца и брата, работающего в саду соседа", то ли "видел отца и соседа брата, работающего в саду ".

Очевидно, что максимально десемантикализованный анализ текста, облегчающий пополнение словарей и увеличивающий быстродействие систем, не просто сохраняет свой прикладной интерес. Если задачу автоматического морфологического анализа в принципе (после появления словаря А.А. Зализняка) можно считать решенной, так как существует множество его реализаций, проблемой анализа следующего - поверхностно-синтаксического уровня чаще занимаются ad hoc в рамках разного рода глубинно-синтаксических и семантических исследований, предполагающих компьютерную реализацию (ср. Леонтьева 1996).

Компьютеризация культуры в целом порождает необходимость искать оптимальных решений прикладных лингвистических задач, в частности задачи сегментации предложения, являющейся необходимым этапом поверхностно-синтаксического анализа естественного текста, необходимым при любом его дальнейшем использовании. Задача описания линейной структуры предложения (и далее - текста), лежащая на стыке морфологии и синтаксиса, остается практически нетронутой (если не считать работ 20-30-летней давности, наметивших возникающие здесь проблемы, как, например, проблему синтаксических неоднозначностей - Иорданская 1967 или проблему проективности - Шрейдер 1971, Падучева 1971).

Предлагаемая работа рассматривает эту задачу как отдельную область исследования и ставит целью понять, сколь далеко можно продвинуться в поверхностно-синтаксическом анализе без обращения к семантике (по крайней мере, к лексической), т.е. используя для анализа морфологические и линейно-комбинаторные характеристики текста

Имплицитно любое описание языка, в том числе и на уровне фонологическом, не обходится без использования семантики. При этом степень семантикализации описания определяется необходимостью эксплицитно вводить семантические параметры. Степень грамматичности модели (алгоритмов) "обратно пропорциональна" степени индивидуализации семантической информации. Помимо морфологических характеристик в нашей системе используется простейшая грамматическая модель управления (не лексикализованная, задающая управление с точностью до падежа или части речи) и некоторые простейшие семантические классы слов.

Эта работа представляет собой развитие "древнего", возникшего на заре прикладной лингвистики, направления работ, где ad hok строились алгоритмы анализа предложения. Наша система, изначально в некотором смысле такая же "адхоковая", предполагает возможность работы с естественными текстами на русском языке и имплицитно содержит элементы грамматики комбинаторных возможностей поверхностно-синтаксических структур, существенные для решения проблем этого уровня анализа.

Текст, порождаемый в процессе речевой деятельности, протекающей во времени и пространстве, физически одномерен. Предложение - линейно-упорядоченная одномерная структура - является отображением некоторой сложно организованной подструктуры ментального пространства. Базисные "простые компоненты", сосуществующие в сознании и находящиеся в сложных контекстуальных отношениях, манифестируются строго упорядоченной последовательностью сегментов (фрагментов) предложения.

При этом в цепочку простых предложений, служащих костяком линейной структуры (β-сегментов), вставляются - в форме придаточных предложений, деепричастных оборотов, обособленных согласованных определений, выраженных причастиями, прилагательными, существительными, и др. - зависимые сегменты (α-сегменты). В них в свою очередь могут быть вставлены другие подчиненные им сегменты и т.п. Возникают разрывы, цепочки соподчиненных как β-, так и α-сегментов, цепочки соподчиненных или последовательно связанных а-сегментов.

Предлагаемая система является попыткой строить анализ предложения, минимизируя словарные средства и опираясь на высокую информативность грамматического контекста - порядка слов предложения на русском языке в их конкретном морфологическом представлении. Эту модель анализа можно рассматривать как синтаксис линейной структуры предложения.

Собственно сегментации предшествует блок двух алгоритмов, чрезвычайно значимых в ходе анализа (в частности, при анализе сочинения). Эти алгоритмы строят синтагматические связи, определяющие

проективные, но не вычленяемые знаками препинания фрагменты предложения: построение предложных групп с любой структурой вложений между предлогом и существительным-слугою и анализ необособленных согласованных определений в препозиции к хозяину с любыми вставлениями любой глубины (явления вставлений на уровне внутрисегментных структур отмечены Хомский 1972). Например, над широко раскинувшейся перед возвышающимся на негусто заросшей лесом горе замком поляной... , ... проработавший неделю каникул инженером старик... и т.п.

Первая процедура собственно сегментации - определение α-запятых - левых границ α-сегментов (отрезков S, в которых есть "сегменто-образующие" слова : подчинительные союзы, деепричастия, обособляемые предлоги и т.п.), наиболее надежных операторов членения (с обращением как к подпрограммам к алгоритму определения левой границы обособленного согласованного определения с вершиной - причастием или прилагательным и к алгоритму анализа сочинения). На следующем этапе они будут служить нам отправными точками при поиске правых границ α-сегментов.

Второй этап анализа вложенных сегментов - алгоритм поиска правой границы α-сегмента с обращением к алгоритму анализа сочинения. Двигаясь по α-сегментам в S справа налево (т.е. от самых "глубоких" вложений), максимально удлинняем каждый очередной α-сегмент: двигаясь от его левой границы (α-запятой) направо, анализируем функцию каждой очередной запятой, обращаясь при определенных условиях к алгоритму анализа сочинения (..он шел, вглядываясь в окружающие его деревья, краснеющие кусты, странные облака, кружево листвы, прямо и прямо.).

После определения границ α-сегментов мы получаем возможность, исключив их из рассмотрения, перейти - с учетом разрывов - к анализу β-сегментов.

В основе анализа цепочки "сочиненных" простых предложений, составляющих анализируемое сложное, лежит анализ сочинения предикатов, в ходе которого определяются сказуемые и подлежащие (в их "школьном" понимании).

Далее проводится дополнительная сегментация по союзам соподчиненных α-сегментов одного типа, затем - внутрисегментный анализ сочинения на уровне союзного внутрисегментного сочинения.

После алгоритмов сегментации должны работать алгоритмы построения внутрисегментных синтагматических связей, поиск которых весьма облегчается тем, что во многих случаях помимо границ сегментов уже найдены их "главные члены" и все сочиненные группы, тоже вычленяющие проективные отрезки (это свойство сочинения отмечено в Падучева 1971).

В настоящее время существует программная реализация значительной части алгоритмов системы, демонстрирующая как лингвистическую, так и прикладную эффективность найденной стратегии.

Линейная структура предложения S естественного языка состоит из множества словоформ

$S = \{W_1, W_2, \ldots, W_n\}$, где каждая словоформа является множеством морфологических омонимов

$W_i = \{H_1, H2, \ldots, H_m\}$. Таким образом, предложение можно представить как упорядоченную цепочку элементов $S' = \{E_{11}, E_{12}, \ldots, E_{1m}, \ldots, E_{np}\}$, где первый индекс элемента соответствует номеру словоформы в предложении, а второй - номеру морфологического омонима словоформы. Первоначальный этап синтаксической сегментации начинает работать с линейным представлением S. При построении синтагм и поиске предикатов происходит активизация омонимов, в результате чего возникают смешанные цепочки типа $S'' = \{W_1, E_{2j}, W_3, \ldots, E_{np}\}$. Существует динамически пополняемый список $L = \{S''_1, S''_2, \ldots, S''_k\}$, активизация нового омонима является событием, которое вызывает пополнение списка. Каждый S''_i содержит минимальное число синтагм, необходимых для дальнейшей сегментации. Синтаксическая сегментация проводится для каждого элемента списка L , собирая разорванные вложениями α и β сегменты. Предложение представляется в виде графа, в узлах которого находятся сегменты, а ребра являются связями между сегментами, такой граф можно представить как множество узлов $ST = \{Segm_1, Segm_2, \ldots, Segm_n\}$, где $Segm_i \subset S''$. Каждый S''_i из списка L преобразуется в множество графов $S'' \Rightarrow \{ST_1, ST_2, \ldots, ST_m\}$; неоднозначность интерпретаций S'' обусловлена возникновением синтаксической омонимии.

После того, как проанализированы все элементы списка L, мы получаем множество всех возможных графов сегментов данного предложения $M = \{ST_1, ST_2, \ldots, ST_q\}$; путем сравнения и оценки элементов M должны выбираться лучшие структуры. Множественность синтаксических интерпретаций зачастую определяется естественной смысловой омонимией в предложении, как это было показано на вышеприведенных примерах. Уже на этой точке выбора отсекается часть активизированных морфологических омонимов. После завершения сегментации возможно проведение полного синтаксического анализа внутри простых синтаксических единиц, каковыми

являются α и β сегменты.

Метод активизации омонимов и алгоритм синтаксической сегментации позволяют сократить число омонимичных структур предложения и избежать декартового произведения омонимов. Полученная структура сегментов - помимо того, что она существенно облегчает задачу дальнейшего синтаксического анализа - может, даже без полного построения синтаксических связей, быть использована в различных задачах автоматического анализа текста (поиск, аннотирование и др.).

Библиография

Иорданская 1967: Л.Н.Иорданская. Синтаксическая омонимия в РЯ (с точки зрения автоматического анализа и синтеза). НТИ, 1967, N5.

Леонтьева 1996: Н.Н.Леонтьева. О предмете "прикладная лингвистика". Московский лингвистический альманах "Спорное в лингвистике". 1996. Вып.1.

Падучева 1971: Е.В.Падучева. О порядке слов в предложениях с сочинением: сочинительная проективность. НТИ сер.2, 1971, N3.

Хомский 1972: Н.Хомский. Аспекты теории синтаксиса. 1972.

Шрейдер 1964: Ю.А.Шрейдер. Свойство проективности языка. НТИ, 1964, N8.

СИСТЕМА РУССКО-АНГЛИЙСКОГО И АНГЛО-РУССКОГО МАШИННОГО ПЕРЕВОДА RETRANS В 2000 Г.

Г.Г. Белоногов, В.С. Егоров, Ю.Г. Зеленков, А.П. Новоселов, Ал-др А. Хорошилов, Ал-сей А. Хорошилов, А.Н. Шогин
ВИНИТИ

RETRANS MACHINE TRANSLATION SYSTEM FROM RUSSIAN INTO ENGLISH AND FROM ENGLISH INTO RUSSIAN IN 2000

G.G.Belonogov,V.S.Egorov, Y.G.Zelenkov,A.P.Novoselov, Alex.A.Khoroshilov, Al.A.Khoroshilov, A.N.Shogin

В ВИНИТИ РАН разработана система автоматического (машинного) перевода политематических текстов с русского языка на английский и с английского языка на русский (система RETRANS). В основу построения этой системы положена перспективная концепция фразеологического машинного перевода, сформулированная профессором Г.Г. Белоноговым в 1975 году. Согласно этой концепции, основными единицами языка и речи считаются не отдельные слова, а фразеологические единицы, выражающие понятия и отношения между понятиями. Такой подход позволяет более точно передавать смысл переводимого текста.

Машинный словарь системы содержит более 3,4 миллиона лексических единиц. Среди них 2,8 миллиона являются словосочетаниями длиной от двух до 17-ти слов. Это самый большой русско-английский (англо-русский) словарь в мире! С помощью системы RETRANS можно переводить тексты по всем естественным и техническим наукам, а также по экономике, политике и бизнесу.

Система RETRANS может работать в среде операционных систем MS DOS, WINDOWS 95, 98, NT и UNIX (сетевая версия системы RETRANS). Скорость перевода зависит от мощности используемой ЭВМ. На ПЭВМ типа PENTIUM при работе в автоматическом режиме она превосходит 100 слов/сек. (90 авторских листов в час). Система поставляется на CD ROM и после загрузки в ЭВМ занимает 150 МБ памяти на жестком диске.

При работе с системой RETRANS ее можно использовать в двух режимах:

1) в автоматическом режиме;

2) в диалоговом режиме. В последнем случае есть возможность выбирать различные варианты перевода отдельных слов и словосочетаний.

В течение ряда лет система RETRANS используется в государственных учреждениях Англии, США, Франции и России (ВИНИТИ, ВНТИЦентр, Минауки и Минобороны) и в частных фирмах. В ВИНИТИ с ее помощью переведено с русского языка на английский более 40 книг.

At the VINITI of the Russian Academy of Sciences a system of automatic (computer aided) translation of polythematic texts from Russian into English and from English into Russian has been developed. The basic idea of this system is the conception of the phraseological machine translation, formulated by professor G.G. Belonogov in 1975. According to this conception, as the main language and speech units not isolated words but phraseological units, signifying concepts and relationships between concepts, are considered. Such conception makes it possible to convey more strictly the meaning of texts to be translated.

The machine dictionary of the system includes more than 3.4 million lexical units. 2.8 million of them are word combinations, consisting of two up to seventeen words. It is the most powerful Russian-English (English-Russian) dictionary in the world! By means of the RETRANS system one can translate different texts in all natural and technical sciences as well as on economics, politics and business.

The RETRANS system can operate within such

operation systems as MS DOS, WINDOWS 95, 98 and UNIX (in case of the network version of the system). The translation rate depends on the capacity of the computer used. When using a PENTIUM type computer in automatic translation mode, the translation rate exceeds 100 words/sec. (more than 2000 sheets per hour). The system is supplied on CD ROM and, after loading into the computer, it occupies 150 MB on the hard disk.

The RETRANS system can be used in two translation modes: in an automatic mode and in an interactive one. In the latter case there is a possibility of selecting different variants of translating equivalents of separate words and word combinations.

For a number of years the RETRAS system has been used in the state establishments of Great Britain, the USA, France and Russia (the VINITI, the Ministry of Science and Technology, the Ministry of Defense) and in the private firms. At the VINITI with its help more than 40 books have been translated from Russian into English.

БОЛЬШОЙ ПОЛИТЕМАТИЧЕСКИЙ АНГЛО-РУССКИЙ (РУССКО-АНГЛИЙСКИЙ) МАШИННЫЙ СЛОВАРЬ ПО ЕСТЕСТВЕННЫМ И ТЕХНИЧЕСКИМ НАУКАМ, ЭКОНОМИКЕ И ПОЛИТИКЕ

Г.Г. Белоногов, Ал-др А. Хорошилов, Ал-сей А.Хорошилов, И.Л. Ефременко, Е.Ю. Рыжова, Л.Ю. Гуськова
ВИНИТИ

LARGE POLYTHEMATIC ENGLISH-RUSSIAN (RUSSIAN-ENGLISH) MACHINE DICTIONARY ON NATURAL AND TECHNICAL SCIENCES, ECONOMICS AND POLITICS (DEVELOPMENT HISTORY AND CURRENT STATE)

G.G.Belonogov, Alex.Khoroshilov, Al.Khoroshilov, I.L.Efremenko, E.Y.Ryzhkova, L.Y.Guskova

В Отделе лингвистических исследований ВИНИТИ РАН в течение двенадцати лет создавался политематический англо-русский (русско-английский) машинный словарь по естественным и техническим наукам, экономике и политике. Составление словаря началось со статистической обработки массива поисковых образов документов, извлеченных из баз данных ВИНИТИ. В результате были созданы частотные словари ключевых слов и словосочетаний по тематическим областям, представленным в этих базах данных. Далее словники частотных словарей были переданы в отраслевые отделы ВИНИТИ и там для каждого русского термина (слова или словосочетания) были указаны их английские переводные эквиваленты. Таким образом был составлен двуязычный словарь объемом более 140 тыс. словарных статей.

Другим исходным материалом для составления словаря явились заголовки англоязычных документов и их переводы на русский язык, сделанные специалистами по соответствующим отраслям знаний. Такие заголовки широко представлены в базах данных ВИНИТИ. Авторы доклада извлекли из баз данных более одного миллиона пар заголовков на русском и английском языках и с помощью ЭВМ составили по ним англо-русские (русско-английские) словари слов и словосочетаний.

Параллельно с составлением двуязычного политематического словаря велась разработка систем русско-английского (RETRANS) и англо-русского (ERTRANS) машинного перевода, и на определенном этапе пополнение этого словаря стало вестись с опорой на системы перевода. При этом использовались различные методы. Например, заголовки англоязычных документов переводились с помощью системы ERTRANS на русский язык и их переводы сопоставлялись с переводами, ранее выполненными экспертами-референтами. Если при этом переводы слов и словосочетаний, сделанные референтами, оказывались точнее, то они включались в словарь. В словарь включались также новые переводные соответствия между словами и словосочетаниями, которые отсутствовали в словаре, но содержались в переводах референтов.

Для пополнения двуязычного политематического словаря новыми переводными соответствиями использовались также готовые словари, составленные другими авторами. Но основным и наиболее надежным источником его пополнения служили переводные соответствия, получаемые в процессе перевода оригинальных русских и английских текстов с помощью систем RETRANS и ERTRANS. Это оказалось возможным благодаря созданным авторами доклада специальным программным средствам.

В результате к концу 2000 года англо-русский (русско-английский) словарь ВИНИТИ вырос до объема 1,7 миллиона словарных статей. Наряду с авторами статьи в его составлении принимали участие многие лица. Среди них наибольший вклад внесли Е.Г. Дружинина, Е.Б. Дудин, Б.А. Кузнецов, В.И. Макаров, И.П. Рыбакова, Е.А. Хорошилова и С.Б. Черноног.

ИНФОРМАЦИОННО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Е.Ф. Глебова, Л.С. Ломакина, Д.В. Ломакин, А.З. Панкратова
Нижнегородский государственный университет

INFORMATION-STATISTICAL MODEL FOR

LINGUISTIC RESEARCH

E.F.Glebova, L.S.Lomakina, D.V.Lomakin, A.Z.Pankratova

The algorithm of statistical processing of the text is offered which allows to construct original "portrait" of its structure and to carry out identification and classification of appropriate "portraits" by criterion of the consent.

Язык в процессе речевой деятельности выступает в его коммуникативно-обусловленной системности, в единстве формы, содержания и функционального назначения текстов. Построение такой модели языка помогает понять, что правила грамматики существуют как средства управления коммуникативным процессом, как средства понимания и порождения различных текстов, помогает осмыслить систему языка не только в статике, но и в динамике.

Модель строится на основе анализа текста, который является продуктом языковой деятельности. Каждый текст отражает структуру своего источника, является типичным для него и поэтому по тексту можно оценить структуру языка.

Используется самая общая модель системы, состоящая из непроизводных элементов (морфем, слов), которые образуют пространство элементарных событий. Множество событий называется пространством элементарных событий, если все события несовместны, т.е. наступление одного из них исключает возможность наступления другого, и образуют полную группу событий, т.е. сумма вероятностей всех событий равна единице. Любое событие, которое может быть зафиксировано в тексте, определяется как соответствующее подмножество элементарных событий. Структура системы определяется как совокупность статистических зависимостей между событиями.

Предлагается алгоритм статистической обработки текста, который позволяет построить своеобразный "портрет" его структуры и осуществить идентификацию и классификацию соответствующих "портретов" по критерию согласия.

Алгоритм инвариантен по отношению к родовому и национальному происхождению языка, породившему данный текст и тем более к особенностям стиля автора, и, поэтому, индивидуальные особенности языка обнаруживаются в виде особенностей соответствующего "портрета" текста.

В отличие от известных методов статистической обработки текста [1] в работе предлагается описывать статистическую зависимость между словами, находящимися в тексте на данном расстоянии, не только посредством условных вероятностей, но и с помощью взаимной информации, что способствует более эффективной классификации слов.

Алгоритм классификации основан на следующей гипотезе: чем меньше взаимная информация между соседними словами, тем больше вероятность того, что данные слова принадлежат одному и тому же классу.

Информационный "портрет" структуры текста строится на множестве всех возможных комбинаций из двух слов, находящихся на заданном расстоянии друг от друга. Каждую комбинацию из двух слов можно изобразить точкой в декартовой системе координат. Слова по каждой из координат располагаются в порядке убывания их вероятностей. Каждой комбинации из двух слов ставится в соответствие количественная мера взаимной информации между этими словами:

где $p(x_i,y_j)$ - вероятность появления пары слов xi и yj, i и j - порядковые номера слов на координатных осях;

$p(x_i)$ и $p(y_j)$ - безусловные вероятности появления слов в тексте.

В качестве оценки вероятностей берется относительная частота появления в тексте соответствующих событий.

Слова, образующие данную комбинацию, выбираются из одного и того же предложения, поскольку предложение представляет собой априорную известную структурную единицу языка, которую не следует разрушать при анализе текста.

Каждому расстоянию между словами соответствует свой "портрет" структуры текста, причем синтаксическая (информационная) связь между словами не определяется однозначно расстоянием между словами и может увеличиваться с его увеличением.

В результате обработки "портретов", соответствующих разным расстояниям между словами, можно получить более подробное описание структуры текста с помощью графа.

Таким образом, предлагаемый алгоритм статистической обработки текста позволяет оценить структуру языка, породившего данный текст без заранее заданной грамматики и тем самым идентифицировать текст и язык.

Литература:

1. Сухотин Б В. Оптимизационные методы иссле-

дования языка. - М.: Наука, 1976. - 170 с.

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ УПАКОВКИ И ГЕНЕРАЦИИ ПОДГРАФОВ МЕТОДОМ СУБСТИТУЦИЙ

А.Ф. Горшков
Институт конструкторско-технологической информатики РАН

INFORMATION TECHNOLOGY OF PACKING AND GENERATION SUBGRAPHS BY A SUBSTITUTIONS METHOD

A.F. Gorshkov

The information technology of packing-unpacking of the data about structure graphs is considered. This process is reduced to construction of a tree not isomorphic graphs. Area of applications: CAD-CAM, chemistry, molecular biology, crime consultation, genetic etc.

Во многих областях знаний используются структурные объекты, представленные моделями в виде графов или подграфов. К таким областям знаний можно отнести экономику, машиностроение, биологию, генетику, химию, криминалистику и т. д. Методы теории графов дают исследователям и практикам мощный и доступный инструмент построения моделей и решения задач структурного упорядочения объектов. В связи с развитием современных телекоммуникаций возникают актуальные проблемы хранения в компьютерных архивах и передачи больших объемов информации по сетям Интернет, а также проблемы идентификации и каталогизации структурированных объектов различного происхождения, находящихся в электронных архивах.

Постановка проблемы. В данной работе рассматривается новая информационная технология решения указанных выше проблем для объектов, различного происхождения, предварительно преобразованных в связные неориентированные графы, мультиграфы или подграфы [1]. (В дальнейшем для краткости - графы.)

В зависимости от физического смысла той или иной модели, формируемой с помощью графов, исследователь (или прикладник) использует либо граф с помеченными вершинами, либо непомеченный граф. В первом случае для хранения и передачи информации достаточно использовать матрицу смежности или матрицу весов графа. При этом идентификация обеспечивается с точностью до изоморфизма. Однако параллельно произвести автоматическую каталогизацию в этом случае невозможно или, по крайней мере, весьма затруднительно. Во втором случае (непомеченный граф) возникает задача отыскания всех неизоморфных графов с постоянными заранее заданными числами вершин и ребер. Известны асимптотические формулы для числа неизоморфных деревьев с n вершинами [3] и для числа неизоморфных графов без петель и кратных ребер с n вершинами.

Информационная технология. С помощью предложенной в данной работе информационной технологии, решается задача отыскания неизоморфных графов (ОНГ). При этом графы могут иметь петли. Кратность ребер k≥1. Обозначим, наперед заданные, число вершин - n и число ребер - m. Для решения такой задачи аналитических формул не существует.

Информационная технология решения задачи ОНГ сводится к построению так называемого дерева формирования неизоморфных графов (ДФНГ). ДФНГ представляет собой многоярусное дерево, растущее из корня. Корневой узел ДФНГ находится на нулевом ярусе и содержит информацию об объекте в виде непомеченного графа, информацию о котором требуется преобразовать с целью получения следующих результатов:

- компактное представление объекта, удобное для хранения или пересылки;
- полный набор сгенерированных неизоморфных графов, то есть объектов, порожденных объектом, находящимся в корневом узле;
- структуру данных в виде вектора, обеспечивающего поиск интересующего нас объекта, а также его идентификацию с точностью до изоморфизма.

Следующий ярус ДФНГ имеет порядковый номер 1 и называется фундаментальным ярусом. Фундаментальность первого яруса заключается в том, что он должен присутствовать в любом ДФНГ, независимо от области знаний исследуемого объекта и его происхождения. Каждый узел первого яруса представляет собой вектор подвижных степеней [4] тех графов, которые будут формироваться на последующих ярусах. Рассмотрим вектор подвижных степеней

$$D = (\delta_1, \delta_2, \delta_3, \ldots, \delta_{n+1})$$

где

δ_i - компонента вектора подвижных степеней, равная количеству вершин в искомом подграфе со степенью, равной i.

Последующие ярусы, начиная со второго и кончая самым нижним, называются прикладными ярусами. Прикладные ярусы формируются с помощью алгоритмов, реализующих физический смысл соответствующих объектов, зависящих от области его

приложения. Фундаментальный ярус является инвариантным относительно происхождения исследуемых объектов. Рассмотрим процесс формирования первого яруса ДФНГ более подробно.

Формирование фундаментального яруса. Теоретической основой вычислительной схемы построения первого яруса ДФНГ является следующая теорема [4].

Теорема. Пусть G(V,E) - неориентированный граф; тогда параметры n,m искомого подграфа и компоненты вектора подвижных степеней (1) связаны между собой системой диофантовых уравнений

$$\sum_{i=1}^{n+1} i\delta_i = 2m.$$

Данная теорема дает ключ к формированию узлов первого яруса ДФНГ.

Решая систему диофантовых уравнений, получаем некоторое множество допустимых решений, каждое из которых становится узлом фундаментального яруса, являясь вектором (1). При этом каждый узел первого яруса становится узлом-предком для прикладных ярусов ДФНГ.

Однако весьма актуальным становится вопрос о существовании метода или алгоритма решения описанной в теореме системы диофантовых уравнений, поскольку существует проблема разрешимости однородных диофантовых уравнений в целых числах [5, 6, 7]. Следует заметить, что диофантовы уравнения теоремы не относятся к виду ax+by=1 или ax2+bxy+cy2+dx+ey+f=0, поэтому число решений не может быть бесконечным [8]. Решение диофантовых уравнений теоремы выполняется методом простого перебора и имеет трудоемкость, не превышающую O(n3), где n - число слагаемых в уравнении. В экспериментальной программе используется почленное вычитание переменных на первом шаге с последующим перебором допустимых значений переменных. Реализованный в программе алгоритм имеет вложенный цикл второго порядка, в каждом из которых осуществляется перебор допустимых значений с отсечением результатов, не являющихся целыми. Кроме того, необходимо иметь ввиду, что отсутствие петель и кратных ребер в исследуемых графах позволяет снизить число компонент в векторе подвижных степеней до величины n-1.

Порядок функционирования информационной технологии. Допустим, что некий объект представлен двумя параметрами графа: числом вершин n и числом ребер m. На первом этапе формируется фундаментальный ярус ДФНГ в виде набора векторов подвижных степеней. С помощью алгоритмов, реализующих физический смысл конкретного приложения, формируется число поддеревьев, равное числу различных векторов подвижных степеней. В каждом висячем узле всех поддеревьев ДФНГ сформирован конкретный неизоморфный граф, исчерпывающую информацию о котором необходимо преобразовать в такое компактное представление (упаковать), которое занимает существенно меньший объем памяти. В упакованном виде информация должна обеспечивать возможность автоматической каталогизации и идентификации объекта, интересующего пользователя. Одновременно каталогом, идентификатором и упакованным форматом является вектор, компонентами которого являются номера узлов-сыновей, порождаемых узлами-отцами в порядке следования поколений, начиная от корневого узла и кончая висячим. По существу полученный вектор представляет собой путь (pathname) доступа к висячей вершине ДФНГ и содержит число компонент, равное числу ярусов дерева. Например, для графа с n=2000 и m=8000 ДФНГ, содержащее пять ярусов обладает коэффициентом упаковки равным 1/2000. Что означает экономию памяти при хранении и снижение занятости каналов связи в 2000 раз. Примером подобного вектора (упакованного ключа) может быть такой P=(10, 7, 14, 8, 23), где порядковый номер компоненты равен номеру яруса ДФНГ, а числовое значение компоненты соответствует порядковому номеру узла на данном ярусе. Разумеется, за такое "сжатие" информации приходится платить дополнительным расходом машинного времени при упаковке до ее пересылки и распаковке после ее получения адресатом.

Распаковка "сжатой" информации выполняется путем генерации интересующего пользователя объекта с помощью вектора P, который был получен при упаковке информации. При этом используется программа-распаковщик (генератор), аналогичная программе-упаковщику. Ядром программ упаковщика и распаковщика являются алгоритмы, основанные на методе субституций (замещений) [4,8,9].

Некоторые приложения информационной технологии.

При разработке проектно-конструкторской документации в области машиностроения, конструкторские бюро неизбежно сталкиваются с проблемой хранения больших объемов информации в электронных библиотеках, а также при пересылке проектной документации заводам изготовителям. Существенную долю объема такой документации составляют чертежи деталей и размерных цепей [10], разрабатываемые в среде систем автоматизированно-

го проектирования (САПР). Как известно, чертеж детали с размерной цепью формата А4 (картинка) требует затрат электронной памяти, в лучшем случае, порядка нескольких килобайт. В то же время, с помощью предлагаемой информационной технологии, некоторые детали и размерные цепи могут быть преобразованы в графы, каталогизированы и упакованы с целью хранения и последующей передачи заводам изготовителям или фирмам - заказчикам.

Примером еще одного приложения новой информационной технологии в области органической химии, может служить разработка экспериментального электронного каталогизатора изомеров предельных углеводородов (алканы), общая химическая формула которых CnH2n+2. В основу каталогизатора изомеров положены правила Международного союза теоретической и прикладной химии (ИЮПАК). ДФНГ каталогизатора изомеров содержит один фундаментальный ярус и четыре прикладных. Задачей перечисления структурных изомеров занимались многие математики, начиная с Кэли [3] и до Пойа и Харари [11].

Ниже приводятся сравнительные результаты начальной части гомологического ряда алканов, полученные Кэли с одной стороны и ИКТИ РАН - с другой. Каждая цифра соответствующего ряда показывает число изомеров, а индекс - число атомов углеродного скелета.

11, 12, 13, 24, 35, 56, 97, 188, 359, 7510…(Кэли, Англия, 1875).

11, 12, 13, 24, 35, 56, 97, 208, 419, 8210…(Горшков, Россия, 2000).

Сравнительный анализ двух моделей показывает, что расхождение в результатах начинается с молекулярной формулы C8H18. Новая информационная технология упаковки и генерации графов позволила получить, хотя и побочный, но интересный научный результат.

Перспективы использования информационной технологии.

В связи с развитием систем электронных коммуникаций, систем автоматизированного проектирования, созданием идентифицирующих баз данных с автоматической каталогизацией, систем искусственного интеллекта и т.п. актуальность применения информационных технологий для упаковки релевантных данных будет возрастать. Проблема размещения структуры данных в памяти ЭВМ или при передаче по линиям связи с минимальными затратами ресурсов эффективно разрешима с помощью описанной выше информационной технологии. При этом необходимо, чтобы исходные объекты предварительно были преобразованы в графы. Во многих областях знаний существуют объекты, удобные для преобразования в графы. Достаточно упомянуть такие области знаний, как органическую химию, молекулярную биологию, генетику и т.д. Поскольку, предложенная информационная технология использует ДФНГ, содержащее фундаментальный ярус, то в будущем, при создании специализированных ЭВМ-упаковщиков, возможна аппаратная реализация решателя диофантовых уравнений, что может оказаться экономичней использования супер-ЭВМ. Прикладные уровни ДФНГ целесообразно реализовать программно, так как множество приложений по упаковке, каталогизации и идентификации будет постоянно возрастать. Примером области знаний по прикладной задаче идентификации может служить криминалистика. Эта проблема в настоящее время обсуждается в зарубежной прессе, где предлагается использовать базу данных образцов ДНК для лиц, состоящих на криминалистическом учете.

Как известно, развитие методов чтения ДНК привело к возникновению новой дисциплины - компьютерной генетики [12, 13]. В настоящее время суммарный объем банка нуклеотидных последовательностей (GenBank) возрастает с поразительной скоростью [14]. Есть все основания предположить, что автоматическая "прогонка" (при использовании данной информационной технологии) отдельных разделов GenBank, в порядке вычислительного эксперимента, позволит выполнить формализованную каталогизацию и идентификацию некоторых фрагментов ДНК, выбрать перспективные способы упаковки и каталогизации генных карт, генных ансамблей и других объектов, содержащих генетическую информацию. При этом GenBank и упакованные ключи будут связаны между собой информационно.

Содержание