

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 7

Москва 2015

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.89

Д.Л. Ольшанский

Подбор алгоритма для параллельной реализации метода сходства в интеллектуальных ДСМ-системах

Описывается принцип выбора подходящих алгоритмов для параллельной реализации метода сходства в решателе ДСМ-системы. Выделяется семейство алгоритмов наиболее перспективное для параллельных модификаций, рассматриваются его различные представители. Создается экспериментальная реализация ряда алгоритмов и их параллельных версий на языке C++, при общем уровне оптимизации. По результатам оценивается метод распараллеливания, приводятся дальнейшие усовершенствования и направления для исследования.

Ключевые слова: анализ формальных понятий, ДСМ-решатель, параллельные вычисления

ВВЕДЕНИЕ

Обнаружение связи между структурой объекта и его поведением (свойствами) является одной из важных задач в области машинного обучения. ДСМ-метод автоматического порождения гипотез хорошо приспособлен для этих целей. Связь между структурой и целевыми свойствами интерпретируется как

причинно-следственная связь. Как правило, характерные черты структуры объекта считаются причинами особенностей его поведения. Детальное описание интеллектуальных ДСМ-систем можно найти в работах [1-3]. Самой трудоемкой процедурой ДСМ-метода является шаг сходства (индукция) – процесс порождения гипотез по множествам положительных и отрицательных примеров, для этого необходимо

получить все возможные комбинации пересечений признаков объектов, что является экспоненциально сложной задачей.

Связь ДСМ-метода и анализа формальных понятий (АФП) установил С.О. Кузнецов [4, 5], что позволило применять алгоритмы порождения формальных понятий в ДСМ-системах. Теоретические результаты АФП также применяются в различных областях, в том числе в машинном обучении, поиске информации, инженерии (открытие) знаний и другие. Задача генерации множества всех понятий и решетки понятий детально изучена во множестве опубликованных работ [5-10]. Также было показано, что в общем случае задача относится к классу #P-полных задач [5].

Большинство текущих интеллектуальных ДСМ-систем построены на последовательных алгоритмах, как правило используются алгоритм Норриса или «замыкай под одному» [4]. Вместе с тем, темпы развития компьютерной индустрии таковы, что структура вычислительных комплексов радикально изменилась. Так, примерно с 2005 г. многоядерная архитектура (симметричная мультипроцессорность) и алгоритмы параллельной обработки стали ключевым механизмом роста производительности компьютерных систем [11]. Как следствие, эффективное использование ресурсов требует создания новых параллельных версий алгоритмов и компьютерных систем на их основе. Это позволит расширить область применения интеллектуальных систем к большим массивам данных.

В настоящей статье кратко рассматривается ряд алгоритмов порождения формальных понятий с точки зрения возможности их распараллеливания. Описываются несколько алгоритмов на основе «замыкай по одному» как наиболее подходящие для параллельной реализации. При этом нас не будут интересовать построение связей между полученными понятиями.

Взяв за основу одну из удачных работ по разработке параллельного алгоритма [12, 13], впервые были созданы параллельные модификации для алгоритмов InClose-2/InClose-3 [14]. Для получения равноценных условий для сравнения все алгоритмы были реализованы на языке C++, с аналогичными оптимизациями и уровнем абстракции.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Введем необходимые определения из теории формального анализа понятий.

Определение 1. Формальный контекст описывается тройкой $(X; Y; I)$, где X и Y – конечные множества, а I – отношение между X и Y , $I \subseteq X \times Y$. Элементы множества X называются объектами, а элементы множества Y – атрибутами. Записью xIy будем обозначать, что $(x, y) \in I$, т.е. объект x обладает атрибутом y .

Таким образом, формальный контекст имеет очевидное представление в виде прямоугольной таблицы, где колонкам соответствуют атрибуты, а строкам – объекты. Отношение I обозначается крестом в позициях, где объекту принадлежит соответствующий атрибут (табл. 1).

Формальный контекст

	a	b	c	d
1	X			X
2	X	X	X	
3			X	X
4		X		X
5	X	X		X

Определение 2. Для любого подмножества A из X объектов контекста $(X; Y; I)$ определим оператор $\downarrow$, порождающий множество $A^\downarrow$ общих атрибутов для каждого объекта из A :

$$A^\downarrow = \{ y \in Y \mid xIy, \forall x \in A \}.$$

Аналогично, для любого множества атрибутов B из Y , введем оператор $\uparrow$ порождающий множество $B^\uparrow$ объектов имеющих все атрибуты из B .

$$B^\uparrow = \{ x \in X \mid xIy \forall y \in B \}.$$

В совокупности эти операторы являются связью Галуа в $(X; Y; I)$ и используются, чтобы дать определение формального понятия.

Пусть B – подмножество атрибутов, тогда $B^\uparrow$ – множество объектов, и следовательно $B^{\uparrow\downarrow}$ – множество атрибутов. Получаем оператор: $B \subseteq Y \Rightarrow B^{\uparrow\downarrow} \subseteq Y$. Аналогично для объектов, и подмножества A , получим $A \subseteq X \Rightarrow A^{\downarrow\uparrow} \subseteq X$.

Определение 3. Формальное понятие в контексте $(X; Y; I)$ – это пара (A, B) , где $A \subseteq X$, $B \subseteq Y$, и $A = B^\uparrow$, $B = A^\downarrow$. A называется экстенсионалом (объемом) понятия, B называется интенсиналом (содержанием).

В приведенном ранее примере понятиями будут: $(12345, \emptyset)$, $(125, a)$, $(245, b)$, $(23, c)$, $(1345, d)$, $(25, ab)$, $(45, bd)$, $(3, cd)$, $(2, abc)$, $(\emptyset, abcd)$.

Определение 4. Если для понятий $C1 = (A1, B1)$ и $C2 = (A2, B2)$, $A1 \subseteq A2$ (или аналогично $B2 \subseteq B1$), то понятие X является менее общим или равным $C2$ ($C1 \leq C2$).

Таким образом, для понятий $(C1, C2)$ в $(X; Y; I)$ имеется частичный порядок. Все понятия формального контекста с построенными связями порядка между ними образуют решетку понятий.

Определение 5. Упорядоченным контекстом называется формальный контекст, если его атрибуты упорядочены от меньшего к большему по числу объектов, имеющих данный атрибут.

Упорядоченный контекст имеет ту же решетку понятий, и соответственно то же множество понятий. Поскольку множества X , Y и I не изменились от перестановки элементов, то не изменится и решетка понятий.

ОПИСАНИЯ АЛГОРИТМОВ

Следуя классификации, предложенной С.О. Кузнецовым [5], можно разделить алгоритмы по порядку генерации понятий на четыре основных группы: сверху-вниз, снизу-вверх, упорядоченные и инкрементные.

Сверху-вниз. Такие алгоритмы как у J.P. Bordat [9], начинают построение с максимального понятия, а затем из каждого вычисленного понятия порождаются все его нижние соседи. Затем процесс повторяется для каждого полученного понятия. J.P. Bordat не только генерирует все понятия, но и строит связи между узлами решетки, что увеличивает затраты по времени. Также существенно возрастают затраты по памяти.

Снизу-вверх. Такие методы генерируют понятия снизу, последовательно замыкая операцией пересечения все большее число объектов. Примером такого алгоритма является Chein [8] и его модификация МСА [10].

Для каждого объекта x_i , $(x_i, x_i^\uparrow)$ считается первым слоем L_1 . L_k – это множество понятий на уровне k , произвольный элемент $L_k - (X_i, X_i^\uparrow)$. Из уровня L_k строится уровень L_{k+1} . Для каждой пары из $L_k (X_i, X_i^\uparrow)$ и $(X_j, X_j^\uparrow)$, если $X_i^\uparrow \cap X_j^\uparrow \notin L_{k+1}$, то $(X_i \cup X_j, X_i^\uparrow \cap X_j^\uparrow)$ добавляется к L_{k+1} . В противном случае объединение всех пар с тем же $X_i^\uparrow \cap X_j^\uparrow$ является элементом L_{k+1} . После мы удаляем из L_k все элементы с тем же множествами атрибутов, которые есть в L_{k+1} .

Упорядоченные. Алгоритмы обходят решетку в некотором заданном порядке. Например, NextClosure [7] и Close by One [5] используют лексикографический порядок, чтобы определить порождалось ли данное понятие в первый раз. Полученное понятие считается каноничным, если оно не предшествовало текущему по порядку. Алгоритмы семейства Close by One мы рассмотрим далее.

Инкрементные – такие как Norris [6], достраивают решетку посредством постепенного добавления объектов и пересечения с имеющимися понятиями. В противовес этому принципу «пакетные» алгоритмы выполняют все построение в один проход. Принцип работы алгоритма Norris – мы добавляем объекты x_k по одному, создавая множество понятий L_{k+1} из L_k . При этом множество L_1 содержит лишь $(x_1, x_1^\uparrow)$. Добавляя один объект x_{k+1} к L_k , мы получаем L_{k+1} следующим образом: $\forall (X_i, X_i^\uparrow) \in L_k$, если $X_i^\uparrow \subseteq x_{k+1}^\uparrow$, то $(X_i \cup x_{k+1}, X_i^\uparrow) \in L_{k+1}$. Иначе, $(X_i, X_i^\uparrow) \in L_{k+1}$, а также в L_{k+1} добавляются $(X_i \cup x_{k+1}, X_i^\uparrow \cap x_{k+1}^\uparrow)$, если $(X_i, X_i^\uparrow \cap x_{k+1}^\uparrow)$ является максимальным.

Наиболее доступными для параллельной реализации являются алгоритмы, в которых присутствует наименьшее число точек синхронизации. Если говорить об алгоритмах семейства Chein, то очевидной точкой синхронизации является поисковая структура, соответствующая каждому уровню уже созданных понятий. Аналогичные проблемы можно обнаружить у алгоритма Bordat, а инкрементные алгоритмы, по определению, хранят полное множество понятий в структуре, которую необходимо обновлять на каждом шаге алгоритма.

Таким образом, алгоритмы, основанные на порядке, имеют преимущество с позиции сокращения числа точек синхронизации – за счет отсутствия дополнительных структур данных. Единственным состоянием, которое необходимо сохранить, является текущее понятие и множество его атрибутов. Каждая ветвь рекурсии таких алгоритмов фактически не зависит от других и может быть вычислена в изоляции, в том числе на другом процессоре.

Рассмотрим более подробно различные модификации алгоритма Close by One (CbO), главного представителя упорядоченных алгоритмов. Его версии общепризнанно являются одними из самых быстрых для порождения всех понятий формального контекста [13]. Четыре модификации именно этих алгоритмов мы реализуем в настоящей работе, а также три параллельные версии.

Алгоритм CbO [5, 12] использует следующий тест на каноничность – интенционал D , полученный из B , является каноническим, если B и D согласуются по всем атрибутам до текущего атрибута j . Если же в D есть атрибут, идущий до j , которого нет в B , то понятие считается неканоническим. То есть понятие каноническое, если:

$$B \cap Y_j = D \cap Y_j,$$

где Y_j – множество атрибутов до j , не включая j :

$$Y_j = \{y \in Y \mid y < j\}.$$

Алгоритм также передает множество атрибутов родительского понятия на следующий уровень рекурсии, таким образом, входящие в него атрибуты можно будет пропустить. Это достаточно простой и эффективный тест, чтобы избежать повторной генерации родительского понятия.

Схема работы алгоритма CbO (рис.1):

Строка 2. Передача понятия (A, B) в процедуру вывода Output, которая обрабатывает его необходимым образом (например, записывая в файл).

Строка 3. Выбор следующего атрибута из контекста.

Строка 4. Проверка встречается ли следующий атрибут в текущем экстенционале B , если да, то пропустить его, избегая генерации того же самого понятия.

Строка 5. Получаем новый экстенционал, пересекая текущий A с объектами, имеющими атрибут j .

Строка 6. Замыкаем экстенционал и получаем интенционал D . Таким образом, понятие (C, D) полностью формируется до проведения теста на каноничность.

Строка 7. Выполняем тест на каноничность проверяя, что все атрибуты в B и D согласуются до текущего атрибута. Если да, то (C, D) каноническое понятие.

Строка 8. В этом случае из (C, D) рекурсивно вычисляются новые понятия, начиная со следующего атрибута.

Основным недостатком алгоритма CbO являются необходимость вычисления полного замыкания до его проверки на каноничность.

Как альтернативу выполнению полного замыкания, автор алгоритма InClose предлагает выполнять

замыкание интенционала инкрементно до текущего атрибута j [15]. Заметим, что для теста на каноничность достаточно просмотреть атрибуты до j , а атрибуты после j используются лишь в том случае, если понятие оказывается каноничным. Соответственно, вводится оператор частичного замыкания $\uparrow_j$:

$$A\uparrow_j = \{y \in Y_j \mid \forall x \in :xIy\}.$$

И соответственно, тест на каноничность проверяет сходятся ли атрибуты интенционала B вплоть до j с частичным замыканием C до j :

$$B \cap Y_j = C\uparrow_j.$$

При этом большая эффективность может достигаться за счет того, что частичное замыкание является менее дорогостоящей операцией. Число пересечений, требуемых для полного замыкания, пропорционально числу атрибутов n , в то время как для частичного замыкания это число всегда меньше n .

Тем не менее, после того как тест на каноничность выполнен, понятие еще предстоит полностью замкнуть, что производится уже на следующем уровне

рекурсии. Отсюда и название алгоритма Incremental Closure.

Схема работы алгоритма InClose (рис. 2):

Строка 2. Выбор следующего атрибута.

Строка 3. Получаем новый экстенционал, пересекая текущий A с объектами имеющими атрибут j .

Строка 4. Если экстенционал C равен текущему A , то...

Строка 5. ... продолжаем замыкание экстенционала B , добавив в него текущий атрибут j

Строка 6. В противном случае применяется новый частичный тест каноничности. Дополнительное упрощение связано с тем, что на данный момент B не содержит атрибутов $\geq j$, т.е. $B = B \cap Y_j$, и тест сводится к простому равенству.

Строка 7. Создается новый интенционал D , содержащий все атрибуты B и текущий атрибут j .

Строка 8. (C,D) передается на следующий уровень рекурсии с атрибутом $j+1$, чтобы продолжить замыкание D .

Строка 9. Уже полностью замкнутое к концу процедуры понятие (A,B) передается процедуре вывода Output.

```

1 CbO((A,B), y):
2   Output(A, B)
3   for j=y to n-1:
4     if j ∉ B:
5       C = A ∩ {j} ↓
6       D = C ↑
7       if B ∩ Yj == D ∩ Yj:
8         CbO((C,D), j+1)

```

Старт рекурсии - CbO(X, X↑)

Рис. 1. Псевдокод алгоритма CbO («Замыкай по одному»).

```

1 InClose((A,B), y):
2   for j=y to n-1:
3     C = A ∩ {j} ↓
4     if A == C:
5       B = B ∪ {j}
6     else if B == C ↑j
7       D = B ∪ {j}
8       InClose((C,D), j+1)
9   Output(A,B)

```

Старт рекурсии - InClose(X, ∅)

Рис.2. Псевдокод алгоритма InClose

Можно заметить, что по сравнению с CbO отсутствует тест на принадлежность атрибута родительскому понятию, это связано с тем, что на данный момент атрибуты экстенционала В не содержат атрибутов после j. В связи с этим отсутствует возможность исключить повторную генерацию понятия на основе унаследованных атрибутов.

Другим существенным улучшением алгоритма CbO стала версия FCbO [16], где результаты неудачного теста на каноничность используются в последующих уровнях рекурсии. Более того, это позволяет избежать вычисления полного замыкания.

Этот результат был получен из наблюдения, что, если первоначальный тест на каноничность $B \cap Y_j = D \cap Y_j$ не выполняется для атрибута j, то тест также не выполнится для любого $B' \subseteq B$, если $(D \setminus B) \cap Y_j$ содержит атрибут, которого нет в B' . При этом интенционал D и атрибут j, для которого не выполнялся тест, передаются на следующий уровень рекурсии как записи об ошибке – Nj. Если для определенного атрибута j существует атрибут в Nj до j, которого нет в B, то нет необходимости выполнять полный тест на каноничность.

Для того чтобы получить все интенционалы, провалившие тест на каноничность, перед переходом на следующий уровень рекурсии, используется комбинированный подход с обходом в ширину и в глубину. Вместо непосредственной передачи новых понятий в рекурсивный вызов, теперь в основном цикле они помещаются в очередь для дальнейшей обработки. И после того как все понятия из данного уровня получены, очередь обрабатывается рекурсивным вызовом алгоритма уже с полным набором записей об ошибке.

Схема работы алгоритма FCbO (рис. 3):

- Строка 2. Передача понятия (A,B) в процедуру вывода Output.
- Строка 3. Выбор следующего атрибута.
- Строка 4. Наследуем запись об ошибке из предыдущего уровня Nj в Mj, записи об ошибке для следующего уровня.
- Строка 5. Пропускаем атрибуты, которые есть в B или те, для которых есть запись ошибки.
- Строка 6. Получаем новый экстенционал, пересекая текущий A с объектами, имеющими атрибут j.
- Строка 7. Замыкаем экстенционал и получаем интенционал D.
- Строка 8. Выполняем изначальный тест на каноничность.
- Строка 9. Если понятие оказывается каноничным, помещаем его в очередь вместе с атрибутом, для которого оно было получено.
- Строка 11. Иначе записываем интенционал D в текущую запись об ошибке Mj.
- Строка 12. Извлекаем понятия и атрибуты из очереди и переходим к рекурсивной части.
- Строка 13. Передаем новое понятие на следующий уровень рекурсии вместе со стартовым атрибутом для следующего уровня и записями об ошибках M, полученными на текущем уровне.

Алгоритм InClose2 создан как объединяющий свойства частичного замыкания и полного наследования атрибутов родительских понятий, присущего алгоритмам (F)CbO [14]. Для этого принимается комбинированный подход с обходом понятий в ширину и в глубину, как в FCbO. Теперь основной цикл завершается до рекурсивных вызовов на следующие уровни, что позволяет передать полный набор родительских атрибутов в дочерние. В основном цикле новые экстенционалы помещаются в очередь подобно алгоритму FCbO.

Схема алгоритма InClose2 (рис. 4):

- Строка 2. Выбор следующего атрибута.
- Строка 3. Пропускаем атрибуты, которые уже есть в B. Этот тест стал возможен, поскольку интенционалы теперь наследуют родительские атрибуты.
- Строка 4. Получаем новый экстенционал, пересекая текущий A с объектами, имеющими атрибут j.
- Строка 5. Если полученный экстенционал C равен текущему A, чей интенционал замыкается в данный момент, то ...
- Строка 6. Добавляем атрибут j к текущему интенционалу B.
- Строка 7. Иначе выполняем тест на каноничность. Теперь необходимо использовать $B \cap Y_j$, поскольку B может содержать унаследованные атрибуты после j.
- Строка 8. В случае успешного теста на каноничность? помещаем новый экстенционал и атрибут, для которого он был получен в очередь.
- Строка 9. Передача понятия (A,B) в процедуру вывода Output.
- Строка 10. Обработка очереди, извлечение новых экстенционалов и атрибутов.
- Строка 11. Каждый новый интенционал D включает все атрибуты из B, добавляя j, атрибут для которого был сформирован экстенционал.
- Строка 12. Рекурсивно вызываем процедуру для получения всех дочерних понятий, начиная с j + 1 и завершения замыкания D.

Алгоритм InClose3 – это дальнейшее развитие идей объединения особенностей FCbO с алгоритмом InClose [14]. Следующий логический шаг – включение записей об ошибках в качестве дополнительного теста до частичного замыкания.

Схема алгоритма InClose3 (рис. 5):

- Строка 2. Выбор следующего атрибута.
- Строка 3. Наследуем запись об ошибке из предыдущего уровня Nj в Mj, записи об ошибке для следующего уровня.
- Строка 4. Пропускаем атрибуты, которые уже есть в B и те? для которых есть запись об ошибке.
- Строка 5. В противном случае формируем новый экстенционал, пересекая текущий A с объектами, имеющими атрибут j.
- Строка 6. Если полученный экстенционал C равен текущему A, чей интенционал замыкается в данный момент, то ...
- Строка 7. Добавляем атрибут j к текущему интенционалу B.

Строка 8. Иначе выполняем тест на каноничность.
 Строка 9. В случае успешного теста на каноничность, новый экстенционал и атрибут, для которого он был получен, помещаются в очередь.
 Строка 11. Иначе записываем частично замкнутый интенционал D в текущую запись об ошибке Mj.
 Строка 12. Передача понятия (A,B) в процедуру вывода Output.

Строка 13. Извлекаем экстенционалы и атрибуты из очереди и переходим к рекурсивной части.
 Строка 14. Каждый новый интенционал D включает все атрибуты из B, добавляя j, атрибут, для которого был сформирован экстенционал.
 Строка 15. Рекурсивно вызываем процедуру для завершения замыкания D и получения всех дочерних понятий, начиная с j + 1. Также теперь в вызов передаются все записи об ошибках Mj.

```

1 FcBO((A,B), y, {Ny | y ∈ Y}):
2   Output(A,B)
3   for j=y to n-1:
4     Mj = Nj
5     if j ∉ B and Nj ∩ Yj ⊆ B ∩ Yj:
6       C = C = A ∩ {j} ↓
7       D = C ↑
8       if B ∩ Yj == D ∩ Yj:
9         PutInQueue((C,D), j)
10      else
11        Mj = D
12  while GetFromQueue((C,D), j) :
13    FcBO((C,D), j+1, {My | y ∈ Y})

```

Старт рекурсии - FcBO(X, X ↑, 0, {})

Рис. 3. Псевдокод алгоритма FcBO

```

1 InClose2((A,B), y):
2   for j=y to n-1:
3     if j ∉ B
4       C = A ∩ {j} ↓
5       if A == C:
6         B = B ∪ {j}
7       else if B ∩ Yj == C ↑ y:
8         PutInQueue(C, j)
9   Output(A, B)
10  while GetQueue((C, j)):
11    D = B ∪ {j}
12    InClose2((C,D), j+1)

```

Старт рекурсии – InClose2(X, ∅, 0)

Рис. 4. Псевдокод алгоритма InClose2

```

1 InClose3((A,B), y, {Ny | y ∈ Y):
2   for j=y to n-1:
3     Mj = Nj
4     if j ∉ B and Nj ∩ Yj ⊆ B ∩ Yj:
5       C = A ∩ {j} ↓
6       if A == C:
7         B = B ∪ {j}
8       else if B ∩ Yj == C ↑ y:
9         PutInQueue(C, j)
10      else:
11        Mj = C ↑ y
12  Output(A, B)
13  while GetQueue((C, j)):
14    D = B ∪ {j}
15    InClose3((C,D), j+1, { My | y ∈ Y })

```

Старт рекурсии - InClose3(X, ∅, 0, {∅})

Рис. 5. Псевдокод алгоритма InClose3

Параллельные версии получены из исходных алгоритмов по схеме, предложенной в статье [12, 13] для алгоритма CbO, вводится некоторый порог рекурсии L , до которого выполняется прежний последовательный алгоритм. На уровне же L вместо вызовов рекурсивной процедуры следует вызов процедуры `distribute`, помещающей указанные параметры в одну из t очередей. Параметр t определяет число параллельных потоков выполнения, которые будут запущены по завершению последовательной фазы. В типичном случае количество потоков соответствует числу ядер процессоров в системе. Механизм распределения по очередям может быть различным. Эффективность работы будет во многом зависеть от сбалансированности нагрузки, среди методов можно рассмотреть простой перебор по кругу, использование некоторой хеш-функции от понятия, и генератор случайных чисел.

МОДИФИКАЦИИ И ОСОБЕННОСТИ РЕАЛИЗАЦИИ

Большинство из рассмотренных в прошлом разделе алгоритмов были реализованы на языке C++11 при соблюдении сходного уровня оптимизации. В рамках исследования были созданы CbO, InClose2, InClose3 и FCbO, а также их параллельные варианты PInClose2, PInClose3, PFCbO.

Контекст был реализован как массив битовых строк, аналогичным образом были реализованы все множества, фигурирующие в описаниях алгоритмов. Тем не менее, за счет механизма шаблонов языка C++, битовые строки в процедурах могут быть легко заменены любым другим типом, поддерживающим нужный набор операций. Выбор оптимальных структур данных является важным вопросом для исследования, и может оказать радикальное влияние на производительность алгоритмов.

Вторая оптимизация, общепринятая среди реализаций CbO, это сортировка атрибутов по их весу, т.е. числу объектов имеющих данный атрибут. Несложно заметить, что, переставляя наиболее редкие атрибуты в начало контекста, мы снижаем вероятность получения неканонического интенционала (объекты с меньшим номером атрибутов более редки) после замыкания.

Для алгоритмов CbO и FCbO также была реализована оптимизация, объединяющая шаги формирования экстенционала C (отсев объектов, не имеющих атрибут j) и его интенционала D [16]. При этом интенционал D получается пересечениями объектов из C по мере проверки наличия атрибута j в очередном объекте (рис. 6). Подобный способ также выгоден тем, что требует лишь операции пересечения, которая эффективно реализуется на битовых строках.

```

1 ComputeConcept((A,B), j):
2   C = ∅
3   D = Y
4   for x in A:
5       if j ∈ {x}↑:
6           C = C ∪ {x}
7           D = D ∩ {x}↑
8   return (C,D)

```

Рис. 6. Псевдокод вычисления экстенционала C одновременно с его интенционалом

Предложенная оптимизация формирования понятия не подходит для алгоритмов InClose2/InClose3, поскольку экстенционал C требуется до выполнения его частичного замыкания. За счет необходимости повторной итерации по всем объектам из C, при вычислении частичного замыкания эффективность этих алгоритмов вызывает сомнения.

Большой сложностью в реализации оказался вопрос выделения памяти для множеств, в особенности в версиях алгоритмов, применяющих наследование записей об ошибках. Для упрощения этой задачи был разработан механизм, позволяющий выделять и освобождать объекты из заранее выделенной области, или пула. На каждом шаге рекурсии выделялся пул для множеств, используемых в пределах этого рекурсивного вызова. Для записей об ошибках не выделяется дополнительная память, вместо этого они указывают на уже вычисленные интенционалы.

Для эффективной реализации записей об ошибках требуется механизм быстрой передачи массива множеств на следующий уровень рекурсии. При этом явное копирование таких объемов данных нежелательно, поскольку накладные расходы могут легко превысить преимущество предварительного теста на каноничность. Это проблема была решена с помощью фиксированного стека указателей. При этом заранее выделенный массив из $N*(N+1)$ – указателей применяется как стек из «окон» по N позиций для записей об ошибках. При рекурсивном вызове текущее окно смещается на N позиций вниз, и возвращается назад по завершении. Таким образом процедура передачи сводится к увеличению/уменьшению указателя на соответствующую величину.

ЭКСПЕРИМЕНТЫ

Все эксперименты с данной реализацией проводились на системе с 4-х ядерным процессором Intel i5-4670 3.4ГГц с 8Гб оперативной памятью, ни в одном из испытаний пределы доступной памяти не исчерпывались. Операции вывода понятий были отключены, и измерялось лишь время, затраченное непосредственно этапом генерации, исключая ввод данных и предобработку, которые в большинстве случаев и так занимали незначительное время.

В качестве первого блока испытаний тестировались сгенерированные случайным образом наборы данных, с применением генератора Mersenne Twister, из стандарта языка C++11. При этом параллельные модификации запускались с числом потоков, равным числу вычислительных ядер. Основная цель данного блока тестов – оценить зависимость эффективности параллельных алгоритмов от ключевых параметров контекста.

Рассматривались ряды данных со следующими свойствами:

- Фиксированные число объектов и плотность, число атрибутов изменялось от 200 до 800 (см. рис. 7).
- Фиксированные число атрибутов и плотность, число объектов изменялось от 2000 до 80000 (см. рис. 8).
- Фиксированные число объектов и атрибутов, плотность контекста менялась от 2% до 10% (см. рис. 9).

По результатам первого блока данных можно увидеть, что преимущество параллельных версий проявляется лишь при достаточном объеме вычислений, вне зависимости от вида контекстов. Также версии алгоритмов с наследованием записей об ошибках (FCbO, InCose3) имеют существенное преимущество по сравнению с версиями без этого теста (CbO, InClose2).

Далее рассмотрим работу параллельных версий алгоритма на реальных данных из репозитория для машинного обучения UCI [UCI]. Использовались выборки данных mushroom и adult, предварительно все атрибуты заданные как перечисление были переведены в двоичные. В тестах были задействованы от одного до четырех ядер процессора, чтобы оценить масштабируемость предложенных модификаций. Результаты тестов представлены в табл. 2, графики для обратной величины времени (скорость выполнения)

представлены на рис. 10 и 11. Как и ожидалось из построения, все алгоритмы семейства СбО дают практически линейное масштабирование скорости от числа задействованных ядер. Незначительное снижение эффективности с ростом числа потоков является типичным для параллельных алгоритмов, этот эффект может быть связан с накладными расходами на создание потоков или конкуренцией потоков за общий ресурс, например – распределитель памяти операционной системы.

Таблица 2

Время выполнения на реальных выборках (сек) в зависимости от числа потоков

Adult	1	2	3	4
pfcbo	8.731	5.33	4.121	3.619
pinclose2	47.673	25.48	17.668	14.323
pinclose3	12.679	7.334	5.496	4.644
Mushroom	1	2	3	4
pfcbo	1.044	0.584	0.435	0.369
pinclose2	3.83	2.016	1.451	1.13
pinclose3	1.606	0.873	0.64	0.528

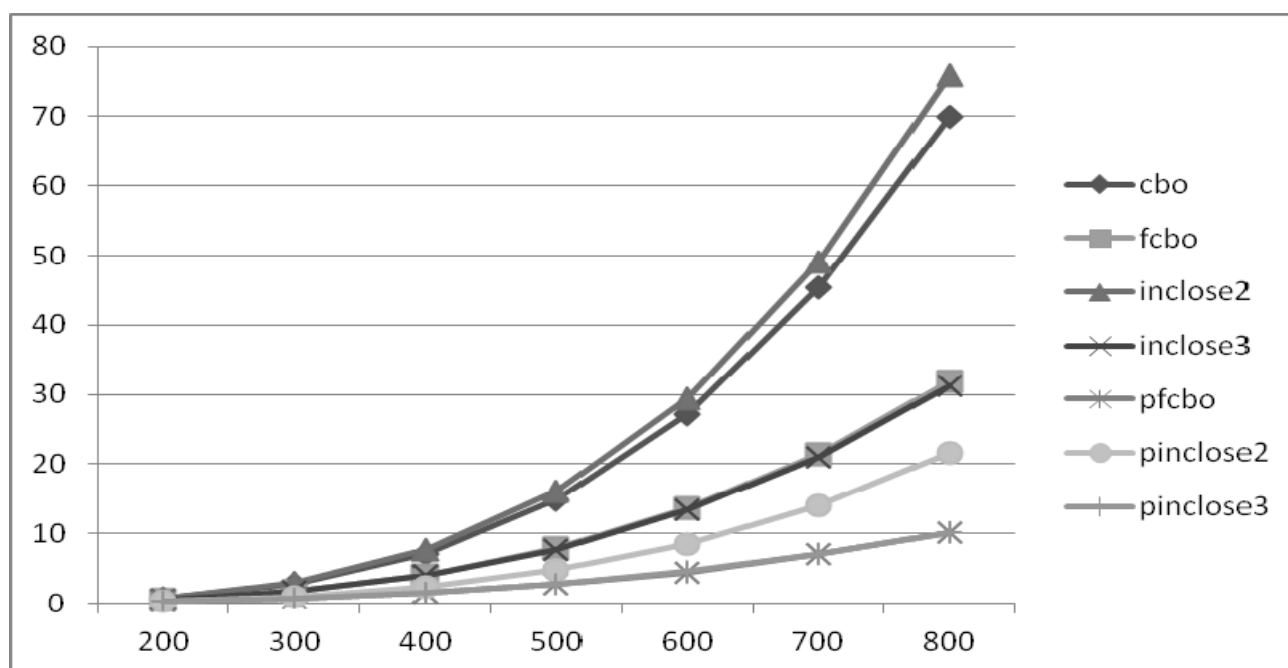


Рис. 7. Синтетические контексты, число объектов 2500, плотность 5%.
Ось Y – время в секундах, X – число атрибутов

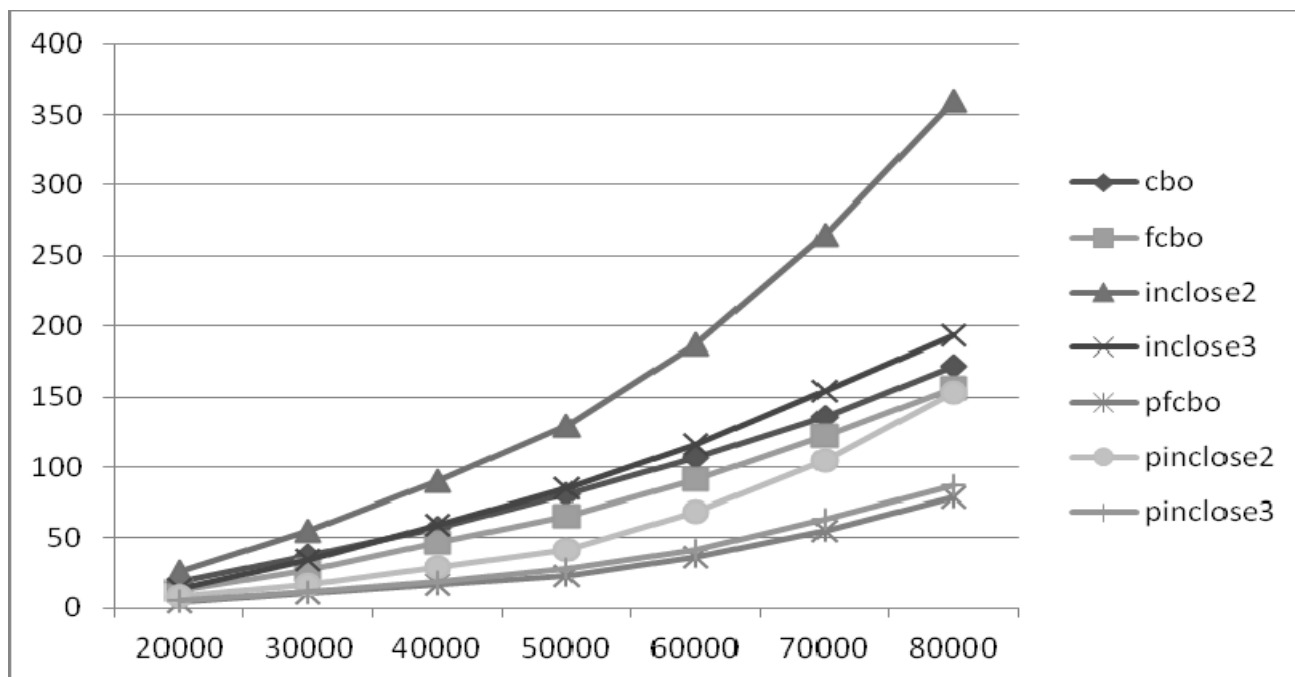


Рис. 8. Синтетические контексты, число атрибутов 200, плотность 5%.
Ось Y – время в секундах, X – число атрибутов

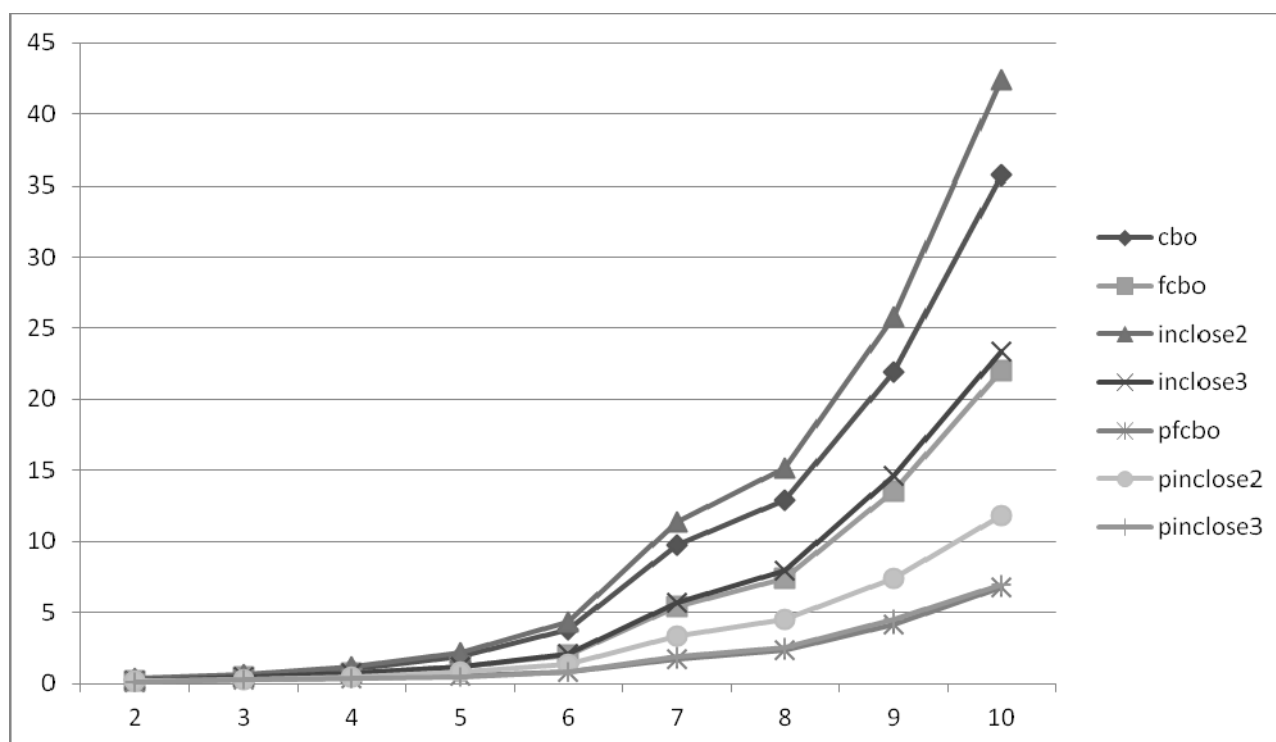


Рис. 9. Синтетические контексты, число объектов 5000, атрибутов 200.
Ось Y – время в секундах, X – плотность контекста в %

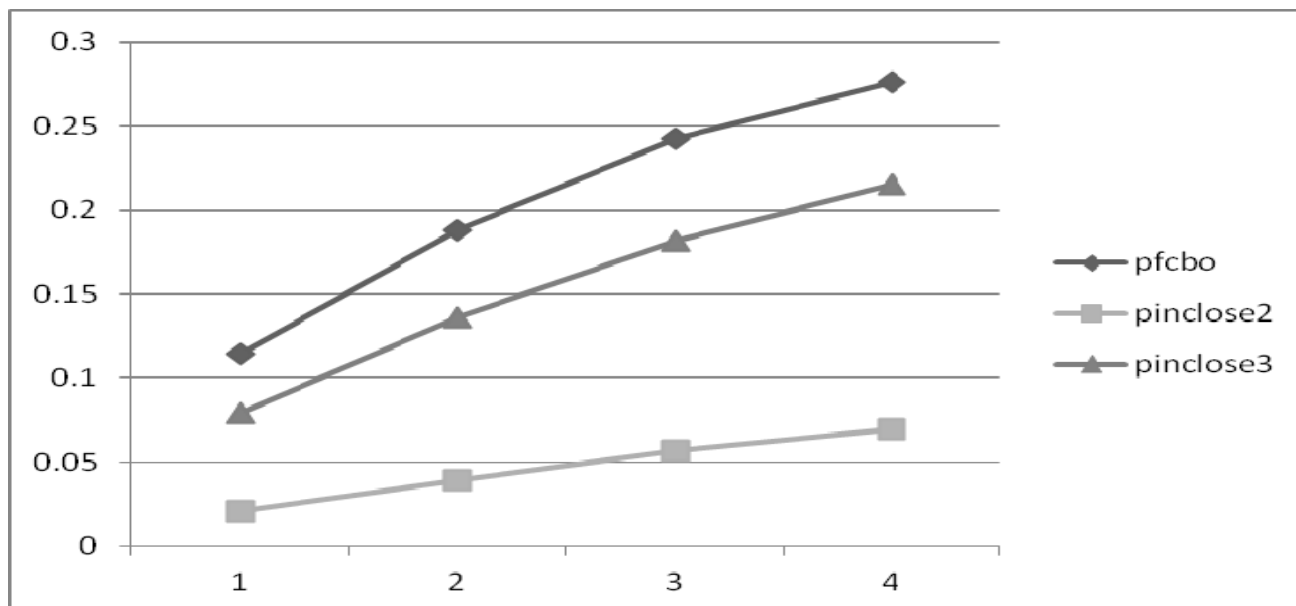


Рис. 10. Выборка данных adult.
Ось Y – обратное время на выполнение 1/с, ось X – число потоков

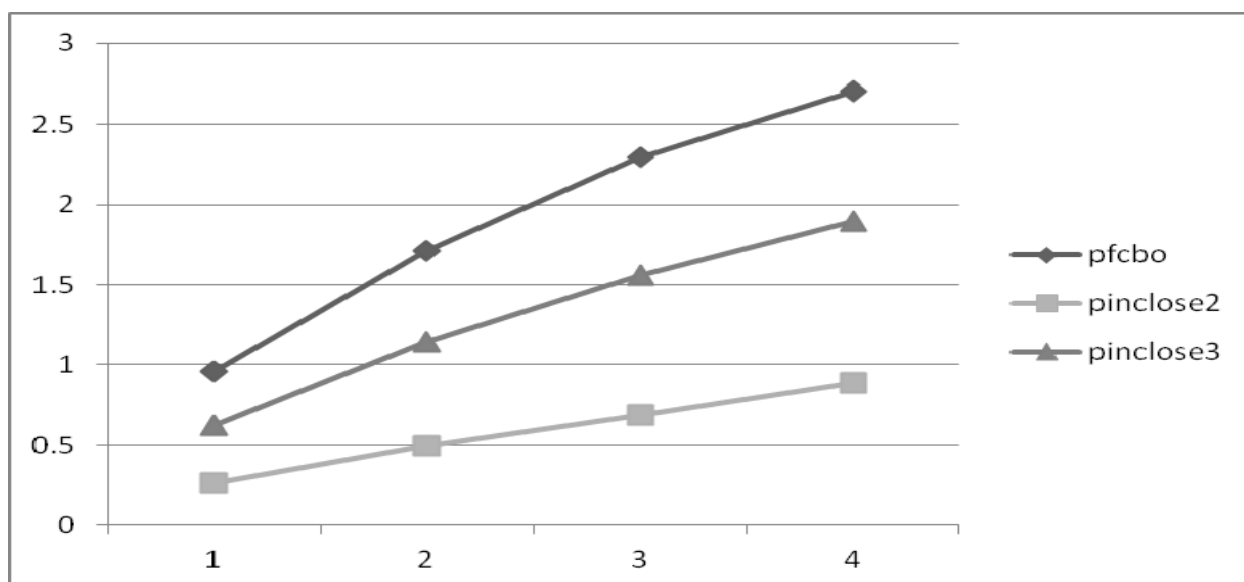


Рис. 11. Выборка данных mushroom.
Ось Y - обратное время на выполнение 1/с, ось X – число потоков

ЗАКЛЮЧЕНИЕ

В качестве дальнейшего направления исследования помимо общепринятых битовых строк, важно рассмотреть применение других структур данных. Рассматривая псевдокод текущих алгоритмов, можно заметить, что набор требуемых операций для интенционала и экстенционала – различен. Поэтому наибольший интерес представляют неоднородные комбинации, где интенционал и экстенционал имеют разное представление.

Серьезной помехой для исследований на больших масштабах параллельности может оказаться механизм управления памятью, поскольку представляет собой неявную точку синхронизации, в

том числе на уровне системных вызовов к операционной системе. Следующее усовершенствование, которое можно рассмотреть, – это применение для каждого потока локальных распределителей памяти и, соответственно, пулов. Подобные техники могут существенно уменьшить неявное влияние потоков друг на друга.

Из рассмотренных алгоритмов очевидным преимуществом оказывается механизм наследования ошибок из FCbO, а также в его параллельных версиях. Вместе с тем, FCbO имеет недостаток в виде дополнительных затрат по объему памяти, в том числе неявных – в виде необходимости сохранять интенционалы в памяти дольше, чем в CbO или InClose2.

Некоторые аспекты исследованной реализации еще предстоит оптимизировать для полномасштабного применения, в частности, решить задачу буферизованного ввода/вывода для многих потоков. Тем не менее, полученные результаты показывают, что предложенный механизм параллельного порождения гипотез для ДСМ-решателя дает существенное ускорение на реальных выборках. При этом наблюдается близкое к линейному масштабирование, вне зависимости от конкретной модификации базового алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений. – 2010. – №3. – С. 3–21.
2. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть II // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 14–40.
3. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ-метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 41–48
4. Аншаков О.М. ДСМ-метод : теоретико-множественное объяснение // Научно-техническая информация. Сер. 2. – 2012. – № 9. – С. 1-19.
5. Kuznetsov S.O., Obiedkov S.A. Comparing Performance of Algorithms for Generating Concept Lattices // Journal of Experimental and Theoretical Artificial Intelligence. – 2002. – Vol. 14, № 2–3. – P. 189–216.
6. Norris E.M. An Algorithm for Computing the Maximal Rectangles in a Binary Relation // Revue Roumaine de Mathématiques Pures et Appliquées. – 1978. – № 23(2). – P. 243–250.
7. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. – Berlin: Springer-Verlag, 1999.
8. Chein M. Algorithme de recherche des sous-matrices premières d'une matrice // Bull. Math. Soc.Sci. Math. R.S. Roumanie. – 1969. – Vol.13. – P. 21–25.
9. Bordat J. P. Calcul pratique du treillis de Galois d'une correspondance // Math. Sci. Hum. – 1986. – Vol. 96. – P. 31–47.
10. Troy A.D. Faster Concept Analysis // 15th International Conference on Conceptual Structures, ICCS — 2007.
11. Sutter H. The Free Lunch Is Over - A Fundamental Turn Toward Concurrency in Software // Dr. Dobbs's Journal. – 2005. – Vol.30.
12. Krajca P., Outrata J., Vychodil V. Parallel Recursive Algorithm for FCA. // Proc. CLA CEUR WS, 433 — 2008
13. Krajca P., Outrata J., Vychodil V. Advances in Algorithms Based on CbO // Proc. CLA. – 2010. – P. 325-337.
14. Andrews S. A 'Best-of-Breed' approach for designing a fast algorithm for computing fixpoints of Galois Connections // Information Sciences. – 2015. – P. 633-649.
15. Andrews S. In-Close, a Fast Algorithm for Computing Formal Concepts. // Supplementary Proceedings of ICCS'09 CEUR WS 483 — 2009
16. Outrata J., Vychodil V. Fast Algorithm for Computing Fixpoints of Galois Connections Induced by Object-Attribute Relational Data. //Information Sciences. – 2012 – P. 114–127

Материал поступил в редакцию 28.04.15.

Сведения об авторах

ОЛЬШАНСКИЙ Дмитрий Леонидович – аспирант Всероссийского института научной и технической информации Российской академии наук, Москва
e-mail: dmitry.olsh@gmail.com

Семантический подход к оценке качества научных публикаций

Для улучшения разведки знаний и поиска релевантной информации в Интернете целесообразно учесть компетентность авторов статей, материалов, а также оценить надежность источников информации.

Предлагается мера оценки компетентности авторов, которая кроме формальных библиографических ссылок учитывает также упоминание в тексте авторов идей и отдельные фразы, представляющие релевантные идеи. Предлагаемая мера также учитывает независимость ссылающихся авторов и хронологические характеристики опубликованных статей.

Ключевые слова: автоматическая обработка текста, Большие Данные, извлечение знаний, индекс цитирования, качество поиска, компетентность авторов, релевантные идеи, лексико-семантический анализ, онтология, семантика, сетевой опыт

ПРОБЛЕМЫ ОЦЕНКИ КАЧЕСТВА НАУЧНОЙ РАБОТЫ

Особое значение проблема оценки качества научных работ приобретает на современном этапе, особенно в связи необходимостью рыночной ориентации научной деятельности, потребностью повышения добавленной стоимости продукции за счет инноваций. В этом контексте постоянно растущий поток научной информации требует разработки автоматизированных методов и средств решения этой проблемы.

Оценка качества научных публикаций – основная проблема наукометрии. Специалистами в этой области давно признано, что качество научной работы – настолько тонкая и сложная вещь, что более или менее объективная его оценка может быть произведена только коллегами – специалистами в данной предметной области. В 60-х годах прошлого столетия Юджином Гарфилдом был введен научный индекс цитируемости (Scientific Citation Index - SCI) [1], ставший первым формальным показателем качества научной работы. Основным показателем SCI – количество явных ссылок на данную работу. Позже были введены другие показатели, в частности индекс Хирша для оценки научной продуктивности авторов. Безусловно, явные (т.е. указанные в списках литературы или сносках) ссылки на работы являются достаточно универсальным и информативным показателем их научной значимости. Помимо оценки качества, явные ссылки позволяют решать и другие задачи, например, осуществлять автоматическую классификацию работ.

Тем не менее видны и недостатки использования явных ссылок. Прежде всего, далеко небезразлична значимость ссылки: она может идти от источников с

разной авторитетностью. Для решения этой проблемы журналы, в которых публикуются статьи, разбивают на классы значимости, образуя различные «базы цитирования»: Web of Science, Scopus и проч. Однако этот способ зачастую не достигает цели, поскольку он сильно зависит от субъективного фактора и многие требующие оценивания работы печатаются в изданиях, которые в эти базы не попадают. Требовать же от авторов публиковаться в заданном списке журналов представляется не всегда реальным, особенно для закрытых и узкоспециализированных работ.

Еще более существенный недостаток методов, основанных на явных ссылках, выявляется в связи с проблемой независимости ссылок. Проблема эта имеет несколько аспектов. Прежде всего, в силу различного рода плагиата, контрафакта и других схем информационного копирования, ссылки на определенную работу могут необоснованно дублироваться, повышая индекс работы. В то же время, вполне возможно, существует и заимствование и присваивание научных идей, плагиат без цитирования, который занижает подлинную значимость научной работы. Другой аспект связан с определенной структурой научного сообщества [2]. Жизненный опыт показывает, что невозможно отдельных научных работников считать полностью самостоятельными и независимыми. Существуют научные коллективы с соотношениями начальник–подчиненный, существуют иерархические связи типа руководитель–аспирант, преподаватель–студент и т.д., в результате чего цитирование оказывается зависимым. Наконец, существуют определенные неформальные объединения ученых, принимающих какую-либо парадигму научного исследования или,

наоборот, ее отвергающих, в результате чего ссылки становятся зависимыми: если некоторый автор сослался на работу, то очень вероятно, что другой автор тоже на нее сошлется. Разумеется, такое зависимое цитирование (или, наоборот, нарочитое нецитирование) искажает рейтинг работы и ее авторов.

Схему формирования индекса цитирования, по-видимому, можно усовершенствовать, применив методы разведки знаний или методику подсчета рейтингов экспертов, обычно используемую в схемах принятия коллективных решений. Привлечение подобных методов наиболее естественно, поскольку цитирующие авторы фактически и являются экспертами, принимающими решение о значимости работы.

РАЗВЕДКА НАДЕЖНЫХ ЗНАНИЙ

Разведка надежных знаний – это автоматизированное обнаружение текстов, содержащих искомые, новые и скрытые знания, релевантные выделенной научной предметной области и поисковым предписаниям.

Авторы настоящей работы исходят из того, что семантику выделенной предметной области знаний (тему) можно интерпретировать экстенционально через ее отождествление с множеством текстов/статей по теме, над которым определена *метаструктура категорий*, описываемая названиями категорий, ассоциированными с текстами и их взаимосвязями. При этом категория понимается в более широком смысле [3], чем это трактует классический подход к категоризации через объединение объектов на основе общих свойств. Предполагается также, что множество текстов и его метаструктура реализованы в вычислительной среде как набор хранилищ и баз данных с развитым поисковым аппаратом.

Назовем выделенную предметную область знаний темой, а множество метаструктурированных текстов этой области – базой знаний по теме. Множество текстов темы содержит определенное множество входящих в них терминов, которые встречаются чаще, чем в текстах других тем. Причем это множество включает термины, частота которых, как правило, превышает некоторый порог или находится в зоне частот, определяемой специальным образом [4, 5]. Наиболее характерные для отражения тематики предметной области термины назовем ключевыми терминами. В отдельных случаях ключевые термины для повышения качества (прежде всего полноты) поиска могут быть структурированы в виде тезауруса [6]. Набор ключевых терминов также определяет тему, так как с помощью полнотекстового поиска по заданному набору ключевых терминов с их статистическими характеристиками можно с определенной полнотой и точностью найти тексты по этой теме. Таким образом, тему можно задать как непосредственно через набор текстов, так и через набор ключевых терминов.

Кроме метаструктуры категорий выделенная предметная область обладает естественной динамически ориентированной структурой – структурой гипертекста, реализуемой библиографическими ссылками из текста на другие тексты из этой же предметной области. Динамика гипертекста – это появление новых тек-

стов, содержащих новые или дополнительные знания, релевантные выделенной предметной области, их связывание с существующими категориями и погружение в базу данных предметной области.

Релевантность, а скорее, пертинентность нового текста в выделенной предметной области не всегда достаточно точно может быть определена автоматически, поэтому традиционно определяется экспертами-индексаторами после ознакомления с текстом. Индексаторы определяют потенциальное место текста в структуре категорий. В настоящее время существуют методы автоматической классификации (рубрицирования) статей по категориям и рубрикам. Методы автоматической классификации используют алгоритмы машинного обучения, которые составляют терминологические портреты рубрик. Однако, как известно, качество (полнота и точность) автоматического рубрицирования полнотекстовых документов даже при использовании тезаурусов для представления их поисковых образов не всегда может удовлетворять пользователей. Поэтому традиционные подходы к улучшению ручных процедур экспертов-индексаторов остаются актуальными. Для этого могут быть использованы продвинутое технологии сетевой экспертизы [7], помогающие группе экспертов достаточно быстро приходить к согласию относительно экспертных заключений, даже работая в сетевой среде.

Между тем категория нового текста уже в какой-то степени неявно определена авторами нового текста, которые дают текстам названия, формируют аннотации и ключевые слова, дают ссылки на использованные знания из других текстов. Такие ссылки при подготовке текстов для ведущих мировых издательств выделяются в специальный раздел и записываются по международным стандартам. Таким образом, возникает возможность автоматического определения категории нового текста путем приписывания новому тексту категорий цитируемых им текстов.

В работе [8] показано, что по описанию темы в виде набора формальных библиографических ссылок можно с достаточно высокой точностью найти новые релевантные тексты по теме и определить их категорию. Проблема остается в том, что культура цитирования часто оставляет желать лучшего, и не всегда авторы дают все необходимые ссылки на источники заимствованных идей и понятий.

НЕФОРМАЛЬНЫЕ ССЫЛКИ И ВИЗУАЛИЗАЦИЯ

Кроме формальной (явной) структуры библиографических ссылок выделенная предметная область обладает также неформальной (неявной, латентной, скрытой) структурой связей, представленных в текстах статей через упоминания идей, понятий и людей. Идеи могут выражаться одной или несколькими фразами. Тексты статей, описывающие общие идеи, можно считать связанными, а направление этой связи определяется хронологическим порядком статей. В простейшем случае с ненулевой степенью полноты и точности можно выявлять общие идеи через наличие общих одинаковых фраз в рассматриваемых текстах. Более сложный случай выявления общих идей – это

наличие похожих фраз, расширенных с применением тезауруса или словаря синонимов. В общем случае идею можно представить как постоянно пополняемое множество фраз, по-разному выражающих эту идею. Учет неявных связей при поиске наряду с обычными библиографическими связями позволит поднять показатель полноты поиска релевантных текстов.

Обозримое множество статей со ссылками между ними может быть представлено на экране компьютера визуально, например, в виде графа. В таком графе вершинам соответствуют статьи, а дугам – ссылки между статьями. Визуальное представление помогает анализировать информацию с помощью инструментов визуальной аналитики. Формальные и неформальные ссылки можно отобразить по-разному, например, цветом или толщиной (фактурой). Направление ссылок может определяться по хронологическому порядку статей. В таком графе можно выделить сильно связанные кластеры. Информацию можно также кластеризовать по потенциальным заказчикам: странам, университетам, авторам и группам соавторов, издательствам, журналам, корпорациям, и т.д.

Перечисленные аспекты семантической интерпретации текстовых материалов, включая хронологическое ранжирование на графах и визуализация представления документов, может помочь повысить качество оценки цитируемости авторов, сделать его более независимым от субъективных соображений и политической ангажированности и тем самым повысить качество оценки научной деятельности, строить более инвариантные индексы цитирования, рейтинги авторов и источников.

ФАКТОРЫ КАЧЕСТВА В ОЦЕНКЕ НАУЧНОЙ ДЕЯТЕЛЬНОСТИ

К факторам, которые повышают надежность и значимость источника информации можно отнести:

- рыночную адекватность – соответствие потребностям практики, которая выявляется маркетинговыми методами, анализом больших данных;
- дату публикации – чем раньше был сформулирован факт, тем больше этот факт дает веса автору или источнику;
- простоту описания факта – основу взаимопонимания, и, следовательно, уровня истинности в групповой оценке компетентности авторов.
- независимость подтверждающих источников (авторов);
- наличие методов концептуального и численного моделирования для построения оптимальных решающих правил агрегирования и категоризации;
- репрезентативность семантической интерпретации темы набором статей, ключевых терминов или фраз или ссылок;
- соответствие метаструктуры категорий принципам конвергентности.

Остановимся более подробно на сравнительно новом, последнем из перечисленных факторов. Известна семантическая закономерность – чем более сложен синтаксис языка описания проблемы, тем, с одной стороны, сложнее интенсионал предиката (метаструктуры) обозначения темы, а с другой – меньше экстенсионал (объем класса интерпретации) [9]. При-

веденная закономерность может быть использована для повышения целостности семантической интерпретации темы, а именно – из имеющихся у исследователя вариантов интерпретации стоит попробовать выбрать более обобщенный. Например, если в проекционном пространстве введена метрика, то стоит для начала от нее отказаться. Если доверие некоторой группы экспертов к модели представления темы, сделанной с применением многозначной логики, необходимо повысить, то стоит попробовать перейти к более концептуальному представлению модели, например с применением бинарной логики.

С учетом этой закономерности в работе [10] показано, что для ускорения достижения согласованного группового решения, обеспечения устойчивой сходимости процесса принятия решений к заданным целям целесообразно создавать необходимые структурные условия, в частности: использовать конечные и одновременно интуитивно обозримые множества понятий, факторов и их взаимовлияний, значений делений на лингвистических шкалах, числа экспертов и рубрик в классификаторах, кластеров понятий и пр.

Таким образом, научная публикация содержит данные об авторах, категориях, ссылки и фразы, описывающие идеи. При этом тему можно задать одной или несколькими публикациями. При таких исходных данных рейтинг автора может быть определен через: число ссылок на идеи автора (количество упоминаний его оригинальных идей, число его оригинальных идей и ссылок на новые идеи); количество популярных/признанных идей, а также величину доли признанных идей среди всех фраз.

Для выявления наиболее достоверных факторов при разведке знаний по теме могут быть использованы различные подходы, начиная с применения процедур формирования оптимального решающего правила и оптимальных весов у авторов, экспертов, источников [11, 12] и заканчивая использованием процедур формирования ранговых распределений [13] и квантовой семантики для выявления наиболее значимых терминов [14]. Эти процедуры позволяют выявлять более достоверные факты при разведке знаний по теме.

Следует, однако, обратить внимание, что отдельные методики хорошо работают только для случаев с независимыми экспертами, т.е. необходимо условие, которое в контексте массовой оценки научных работ строго не выполняется. Кроме того, зачастую в методиках предполагается, что все эксперты оценивают все альтернативы; в случае схемы явного цитирования это не так: каждый автор ссылается только на малую часть других авторов. Все это делает проблематичным непосредственное применение отдельных методик в схеме явного индексирования. Вследствие указанных недостатков метода индекса цитируемости значительную актуальность приобретает разработка новых подходов к оценке научных работ.

МЕТОДЫ ОЦЕНКИ ВЕСОВ АВТОРОВ, ЭКСПЕРТОВ, ИДЕЙ, ФРАЗ

При написании статьи на определенную тему каждый автор принимает решение включать или нет ту или иную библиографическую ссылку в статью. Другими словами, авторы голосуют за ту или иную

ссылку. Каждая тема имеет свои характерные и популярные библиографические ссылки и на этом основании создаются методы автоматической классификации, например [8, 13]. Авторы, пишущие статьи на определенную тему, выполняют коллективную работу по определению характерных библиографических ссылок по этой теме. Это может быть использовано для оптимального объединения индивидуальных решений авторов, согласно чему автору может быть присвоен некоторый вес, который тем больше, чем выше его компетенция.

Так, каждая тема интерпретируется сильно связанным кластером (подграфом) в общем графе связанности всех статей. Такая кластеризация позволяет рейтинговать релевантности статей и компетентности авторов по теме. Алгоритм оптимального объединения решений авторов, предложенный в работе [15], включает следующие шаги.

- выберите некоторые начальные весовые коэффициенты (акции) для индивидуальных решений экспертов;
- для каждого эксперта, используя историю его экспертных решений, рассчитайте вероятность совпадения индивидуального решения с коллективным решением с учетом акций;
- рассчитайте новые весовые коэффициенты (акции) каждого эксперта по рассчитанной на предыдущем шаге вероятности;
- повторяйте три предыдущих шага до тех пор, пока новые весовые коэффициенты не совпадут с рассчитанными на предыдущем шаге.

Авторы этого алгоритма показали, что он работает лучше, чем любое другое правило объединения решений, в частности, лучше, чем правило принятия коллективного решения по принципу простого большинства.

Приведенный алгоритм представляет собой линейное правило объединения решений экспертов. Главным результатом этой работы было утверждение, что если вероятность правильного решения каждого эксперта известна, то линейное правило объединения их решений является оптимальным, и максимальная вероятность правильного коллективного решения достигается, когда весовые коэффициенты (акции) экспертов рассчитываются по формуле $W_i = \log(P_i/1-P_i)$, где P_i – это вероятность правильного решения эксперта с номером i .

Подход к нахождению оптимального коллективного решения может быть и иным. Например, одним из авторов настоящей работы разработан подход к организации сетевого экспертного мозгового штурма на основе логики генетического алгоритма и принципов конвергентного управления. Он может быть использован для групповой оценки событий, авторов, статей и пр.

Таким образом, существуют методы и алгоритмы, интегрирующие индивидуальные решения экспертов в оптимальное коллективное решение. В соответствии с этим алгоритмом могут приниматься оптимальные коллективные решения относительно включения той или иной библиографической ссылки в поисковый образ.

Вместе с тем, для работы описанных алгоритмов необходим полнотекстовый доступ к базам научных статей. В настоящее время российские исследователи должны платить деньги за доступ к статьям в зарубежных базах, а достаточно полной российской базы научных текстов в свободном доступе для российских ученых нет. Поэтому для обеспечения инновационных процессов в России необходимо создать собственную Российскую базу научных текстов, в которой имелись бы как отечественные, так и зарубежные статьи и материалы. Важным звеном в развитии этой работы является наполнение информацией Единой системы учёта результатов научно-исследовательских, опытно-конструкторских работ специального, двойного и гражданского назначения, выполняемых за счёт средств федерального бюджета.

ПРАКТИЧЕСКИЙ ПРИМЕР ИСПОЛЬЗОВАНИЯ НЕЯВНЫХ ССЫЛОК

В качестве примера использования неявных ссылок для оценки качества научных работ авторы предлагают следующий, прошедший апробацию метод [16]. Сфера его применения – оценка информативности работ из Интернета в определенной предметной области. Поэтому он включает в себя и метод описания предметной области.

Первоначально предметная область описывается множеством задаваемых пользователем ключевых терминов, из которых формируется поисковый запрос в Интернет. Алгоритм формирования поискового запроса заключается в переборе всех комбинаций ключевых терминов (единичные термины, пары терминов, тройки и т.д.) так, чтобы длина запроса не превышала определенной границы n . Это делается с тем расчетом, чтобы в результате поиска было найдено «среднее» количество (порядка нескольких сотен) релевантных документов. Разумеется, заранее предсказать количество найденных документов затруднительно, и потому алгоритмы объединения имеют эмпирический характер – значение границы n определяется опытным путем.

Затем с помощью известных поисковых систем (Google, Яндекс и пр.) запросы обрабатываются, и в результате получается множество M_1 текстовых документов. Это множество просматривается на предмет выделения URL-ссылок, и по ним формируется расширенное множество документов M_2 . В этом процессе необходимо подключение специальных механизмов отсеивания дубликатов, а также дублей статей. Это множество, в свою очередь, содержит URL-ссылки, и потому процесс расширения множества документов M_2 продолжается и далее, несколькими итерациями, подключая документы, на которые указывают ссылки. При этом необходимо проверять вновь найденные документы на наличие в них первоначальных ключевых терминов – в этом случае и последующие итерации вносят существенный вклад в пополнение множества искомых документов.

Далее, документы из множества M_2 делятся на предложения или на фрагменты, близкие по длине к

обычным предложениям. В результате составляется база данных с записями вида:

<текст фрагмента>—<URL, указывающее на этот фрагмент>.

Такая база данных имеет, как правило, значительный объем – до нескольких терабайт. Для хранения такой базы данных и осуществления поиска целесообразно использовать методы работы с Большими данными («Big Data»).

По полученной базе данных производятся статистические подсчеты: сколько различных URL ссылается на одно и то же предложение. При таком подсчете сравнение URL имеет смысл производить с точностью до доменов первого или второго уровней. В большинстве случаев можно ограничиться первым уровнем. Однако есть значительное количество хостингов, содержащих большое количество сайтов второго уровня, для них подсчеты следует выполнять с точностью до второго уровня. При проведении расчетов возможно возникновение проблемы независимости сайтов. Дело в том, что зачастую заводятся сайты-двойники, контент которых полностью или почти полностью дублирует контент основного сайта. При подсчетах такого рода сайты следует исключить. Выявить сайты-двойники можно по накопленной базе данных, поскольку она фактически хранит значимый контент всех сайтов.

Таким образом, каждый фрагмент получает определенный рейтинг, равный количеству его вхождений в независимые сайты. Далее, если для каждого документа подсчитать суммарный вес как сумму рейтингов всех входящих в него фрагментов, то такую сумму можно рассматривать как оценку информативности этого документа.

Предлагаемый подход может быть использован не только для оценки качества научных статей, но и для составления рейтинга сайтов, а также опробован на нескольких предметных областях как научного, так и прикладного характера.

* * *

Использование индексов цитирования в качестве меры компетентности авторов не является достаточно достоверной мерой оценки рейтинга авторов работ. В данной работе предлагается иная – по утверждению авторов, более продвинутая – мера оценки компетентности авторов. Кроме библиографических ссылок она учитывает также упоминание в тексте авторов идей и фразы, описывающие важные идеи. Предлагаемая мера также учитывает независимость ссылающихся авторов и временную динамику формирования идей, а также сетевые экспертные оценки.

СПИСОК ЛИТЕРАТУРЫ

1. Garfield E., Sher I.H., Torpic R.J. The Use of Citation Data in Writing the History of Science. – Philadelphia: Institute of Scientific Information, 1964. – 76 p.
2. Кун Т. Структура научных революций. С ввд-ной статьей и дополнениями 1969 г. – М.: Прогресс, 1977. – 300 с.
3. Язык и интеллект. Сб. / Пер. с англ. и нем. / Сост. и вступит. статья В.В. Петрова. – М.: Издательская группа «Прогресс», 1995. – 416 с.
4. Райков А.Н. Условия предельной сходимости диалогового поиска документов // Научно-техническая информация. Сер. 2. – 1990. – № 3. – С. 15–17.
5. Raikov A.N. Holistic Discourse in the Network Cognitive Modeling // Journal of Mathematics and System Science. – 2013. – № 3. – P. 519–530.
6. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во МГУ, 2011. – 512 с.
7. Gubanov D., Korgin N., Novikov D., Raikov A. E-Expertise: Modern Collective Intelligence // Studies in Computational Intelligence. – 2014. – Vol. 558, XVIII. – 112 p.
8. Ежела В.В., Бунаков В.Э., Луговский С.Б., Луговский В.С., Луговский К.С. Цитирование как средство обнаружения дополнительных знаний и их автоматического индексирования. // Труды Третьей Всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции" (RCDL-2001). – Петрозаводск, 2001. – URL: <http://rcdl.ru/doc/2001/ezhela.pdf>.
9. Райков А.Н. Алгебраическая семантика булевого поиска документов // Научно-техническая информация. Сер. 2. – 1990. – № 5. – С. 27–30.
10. Райков А.Н. Конвергентное управление и поддержка решений. – М.: Изд-во ИКАР, 2009. – 245 с.
11. Ruth Ben-Yashara, Samir Khullerb, Sarit Krausc. Optimal collective dichotomous choice under partial order constraints // Mathematical Social Sciences. – 2001. – Vol. 41. – P. 349–364.
12. Nitzan S., Paroush J. Optimal decision rules in uncertain dichotomous choice situations // International Economic Review. – 1982. – Vol. 23. – P. 289–297.
13. Райков А.Н. Принципы естественности в методах автоматического классифицирования документов // Научно-техническая информация. Сер. 2. – 1981. – № 6. – С. 25–27.
14. Райков А.Н. Когнитивное программирование // Экономические стратегии. – 2014. – Т. 16, № 4. – С. 108–113.
15. Baharad E., Goldberger J., Koppel M., Nitzan S. "Beyond Condorcet: Optimal aggregation rules using voting records", 2010, Working Papers, Bar-Ilan University, Department of Economics, No. 2010-20. – 40 p.
16. Charnine M., Somin N., Nikolaev V. Conceptual text generation based on key phrases // Proceedings of the 2014 International Conference on Ar-

tificial Intelligence (ICAI 2014), WORLDCOMP'14, July 21–24, 2014, Las Vegas, Nevada, USA, CSREA Press.

Материал поступил в редакцию 23.02.15.

Сведения об авторах

ЕЖЕЛА Владимир Владимирович – кандидат физико-математических наук, ведущий научный сотрудник, Институт физико-технической информатики, Протвино.

e-mail: vladimir.ezhela@ihep.ru

КЛИМЕНКО Станислав Владимирович – доктор физико-математических наук, профессор, Генераль-

ный директор Института физико-технической информатики, Протвино.

e-mail: stanislav.klimenko@gmail.com

РАЙКОВ Александр Николаевич – доктор технических наук, профессор, лауреат премии Правительства Российской Федерации в области науки и техники, ведущий научный сотрудник Института проблем управления РАН, Генеральный директор ООО «НСА», Москва.

e-mail: alexander.n.raikov@gmail.com

ШАРНИН Михаил Михайлович – кандидат технических наук, старший научный сотрудник, Институт проблем информатики РАН, Москва.

e-mail: mc@keywen.com

Обработка информации с использованием естественных и искусственных нейронных сетей: сравнительный анализ

Проанализированы способы обработки информации биологических и искусственных нейронных систем. Рассмотрены модель биологического нейрона и упрощенные способы преобразования информации в отдельных модулях нервной системы; модели искусственного нейрона и нейронной сети, их сходство и различие. Приведена краткая история развития искусственного интеллекта применительно к роботам и рассмотрены направления исследований.

Ключевые слова: искусственный интеллект, компьютеры, механизмы обработки информации, нейронные сети, биологические нейроны, искусственные нейроны, робототехника

ВВЕДЕНИЕ

При обработке информации часто говорят об аналогии между мозгом живого организма и ЭВМ. Реальное или воображаемое сходство между ними подчеркивают такие ходовые выражения, как «машинное мышление» и «электронный мозг». Никто не отрицает, что многие конечные результаты работы ЭВМ и мозга живого организма сходны. Имеется много примеров сходства работы мозговых и нервных механизмов, аналогичных по своим функциональным характеристикам техническим устройствам. И что различия велики – это тоже очевидно. Сравнивая внешний вид мозга и ЭВМ, вряд ли можно усмотреть «фамильное сходство» между ними. Однако сходство функциональных результатов и структурные различия могут оказаться чисто внешними. Следует подчеркнуть, что физические особенности существующих ЭВМ в большей степени определяются состоянием техники. В большинстве ЭВМ используются электронные элементы, но только потому, что никто еще не изобрел способ изготовления органического нейронного материала, общие функциональные свойства которого отвечали бы требованиям теории ЭВМ. Если невозможно найти таких сходств между ЭВМ и мозгом живого организма, которые оказались бы существенными при сравнении общей структуры этих систем или принципов устройства их элементов, то можно перенести поиски с анатомического уровня на физиологический уровень. И далее исследовать глубинные механизмы передачи информации и выяснить, имеется ли близкое сходство между функциями нейронов в мозгу и функциями отдельных электронных элементов в ЭВМ.

Фундаментальная особенность всего класса ЭВМ состоит в том, что ЭВМ достигает своих поразительных результатов путем выполнения огромного числа очень простых операций и что все сложные вычислительные и логические операции можно разбить на элементарные действия, с которыми могут справиться

очень простые устройства. В вычислительной технике существует теория, каким образом можно разбить искусственные интеллектуальные процессы на простые операции. В этой связи разрабатываются машинные программы, проводятся эксперименты на роботах и интерпретируются полученные результаты.

Дальнейшие исследования направлены на создание теоретических подходов, призванных объяснить фундаментальную задачу: каким образом элементарные операции, выполняемые нейронами в головном мозгу, могут в совокупности порождать функциональные атрибуты разумной деятельности. Применение знаний по механизмам обработки информации нейронами, почерпнутых из биологии, могут привести в дальнейшем к разработке моделей интеллектуальных процессов и к расширению набора функциональных элементов при создании «мыслящих» ЭВМ и интеллектуальных роботов.

В 1977 г. опубликование работ по исследованию структуры и архитектуры коры головного мозга и умственных процессов, протекающих в нем [1, 2], положило начало новому направлению в нейрофизиологии по исследованию мозга. В то время в рамках специальной исследовательской программы США – Neurosciences research Program (NRP) – была организована двухнедельная конференция – Программы Интенсивных Исследований. На этой конференции В. Маунт-кастл и Дж. Эдельман сделали два основополагающих доклада о кортикальной организации и селекции нейронных групп из теории высших функций головного мозга. Эти доклады по решению конференции были опубликованы отдельным изданием.

В этой связи Дж. Эдельман [1] пишет, что невозможно понять природу умственных процессов без знания структуры, функций и эволюции головного мозга. Вероятность отгадать, каким образом работает мозг без изучения его структуры чрезвычайно мала. Очевидно также, если согласиться с тем, что психи-

ческое состояние является продуктом эволюции мозга, то мы должны, по крайней мере, изучить процесс его эволюции, особенно его структуры.

В. Маунткастл предположил [2], что хотя генетически более древние части мозга могут играть значительную роль, основой для любых плодотворных теорий высших мозговых функций и интеллектуального поведения должны быть единственные в своем роде структура и свойства коры больших полушарий. В. Маунткастл и Дж. Эдельман выдвинули ряд идей, которые в сумме представляют организующий принцип функций мозга, построенный на сведениях, полученных в опытах, основанных на эволюционной теории и на иерархическом принципе, и включает в себя понятия о мозге как о динамической машине, обрабатывающей информацию, и элементарном модуле в распределенной системе. Главная задача, которую требуется разрешить, чтобы понять функции коры головного мозга, была определена как раскрытие внутренней структурной и функциональной организации модулей коры головного мозга и их функционирования как единого целого в умственных процессах.

Примерно в то же время Г.С. Пospelовым были определены основные проблемы в области искусственного интеллекта, которые имеют большое значение и сейчас при решении вопросов, связанных с исследованиями в области общей организации мозга и искусственного интеллекта [3, 4].

Исследование поколений постепенно усложняющихся моделей мозга и искусственных нейронных систем описано в работах [5-7]. Дальнейшее развитие это направление получило при создании моделей, обеспечивающих обработку информации и поведение интеллектуальных роботов.

ВОЗНИКНОВЕНИЕ НЕРВНЫХ СИГНАЛОВ: ЗАКОН «ВСЕ ИЛИ НИЧЕГО»

Предпосылки моделирования сетей биологических нейронов заложены в фундаментальных работах физиологов. Появился взгляд на функционирование мозга как электрической цепи, состоящей из генераторов импульса, усилителей и передаточных звеньев. Были описаны виды клеток, измерены электрические характеристики в разных условиях функционирования (сон, бодрствование и др.). Стало приблизительно известно число клеток мозга, способ их соединения, диапазоны электрических характеристик (сила тока, сопротивление, амплитуда, время затухания сигнала, передаваемого между клетками).

Работы проводились на земноводных организмах, потому что их ткани обладают необходимым размером (например, толстый нерв) и долгое время остаются живыми вне организма. Эти научные сведения позволили установить связь между сокращением мышц и одним из самых важных факторов науки о жизни — электрической основой действия нервов. Ученым удалось установить электрическую природу действия как двигательных нервов, по которым головной мозг посылает мышцам сигналы, заставляющие их сокращаться, так и чувствительных нервов, по которым мышцы посылают в мозг информацию о своем растяжении или сокращении. Как было выяснено, двигательные нервы всегда передают свои сигналы с помощью электричества, независимо от того, о

чем они должны сообщать — о растяжении мышц, прикосновении, боли, тепле или холоде, звуке, запахе, химическом составе и т.п. Точно также и все сигналы, посылаемые по чувствительным нервам к исполнительным механизмам, оказались электрическими, независимо от их функции — механической, как у мышц, или химической, как у желез. Однако растяжение мышц может быть зарегистрировано нервом лишь в том случае, если оно превысит некоторый нижний предел или порог.

Представим себе, что мы собрали электронную схему, позволяющую точно фиксировать любые электрические эффекты, которые возникают в нерве на определенном расстоянии от места его соединения с мышцей при воздействии на нее. Когда мышца находится в расслабленном состоянии, никаких электрических явлений в нерве не обнаруживается. Затем начинаем растягивать мышцу. Небольшое растяжение никак не сказывается на нашей схеме, даже если это растянутое состояние поддерживать неопределенно долго. Когда растяжение достигает порогового уровня, схема начинает обнаруживать в нерве электрический ток, который состоит из ряда импульсов. Начинаясь с нулевого уровня, за несколько десятитысячных секунды импульс возрастает до максимальной величины, а затем, немного медленнее, вновь снижается до нуля; весь этот процесс занимает около 0,001 секунды. Спустя какую-то долю секунды через точку, в которой производится измерение, проходит другой такой же заряд электричества. За вторым импульсом следует третий, затем четвертый и т.д. Это продолжается до тех пор, пока мышца остается растянутой. Таков электрический сигнал, вызываемый растяжением мышц, лишь слегка превышающим пороговую величину.

Что произойдет, если растянуть мышцу еще больше? Результат чрезвычайно интересен: отдельные импульсы останутся такими же, как и при слабом растяжении; единственным отличием будет их большая частота. По мере дальнейшего растяжения мышцы частота возникновения импульсов все возрастает, пока не достигнем максимума, являющегося «сигналом насыщения» для исследуемого нерва; дальнейшего повышения частоты добавочным растяжением мышцы вызвать нельзя. Эти электрические свойства сигнала, передаваемого чувствительным нервом, оказываются общими для всех чувствительных нервов всех животных. О чем бы ни сигнализировал тот или иной нерв, он передает свою информацию рядом последовательных электрических импульсов примерно одинаковой величины и длительности независимо от силы раздражителя, и только частота этих импульсов указывает, слабым или сильным было раздражение. Порог и насыщение также присущи всем чувствительным нервам. Такими же свойствами обладают и двигательные нервы — те, которые проводят сигналы к мышцам и железам.

Таким образом, одна из важных особенностей всей нервной системы состоит в том, что она передает информацию с помощью электрических сигналов и что используемый в ней род электрической передачи относится к типу «все или ничего». Все нервные импульсы кодируются сигналами «1» и «0» и имеют одинаковую форму (рис.1).

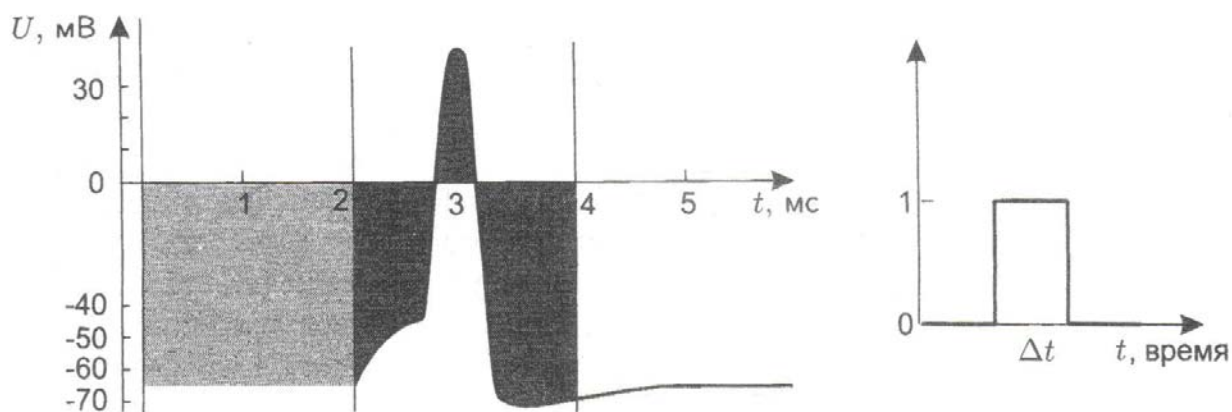


Рис. 1. Нервный импульс и его «модель» – импульсный единичный сигнал [8].

Почему можно ограничиваться только регистрацией самого воздействия и его частоты? Из теории управления известно [8], что для управления подавляющим большинством динамических процессов этих данных достаточно.

Итак, можно представить центральную нервную систему, частью которой является головной мозг, как некоторый черный ящик с входом и выходом. Входные сигналы оказывают воздействие на особые нервные клетки, называемые рецепторами; эти клетки реагируют на «внешнюю (сенсорную) информацию». Рецептор, по существу, представляет собой небольшую распознающую систему, для которой определен класс распознаваемых ею объектов (воздействий). К этой сенсорной информации относится свет, механическая деформация, химические вещества и т.п. Во всех этих случаях возникает электрический сигнал. На другом конце нервной системы имеем выход – двигательные нейроны, оканчивающиеся не на других нейронах, а на клетках мышц и желез. Все выходные сигналы вызывают мышечные сокращения или реакции клеток желез.

МОЗГ И НЕРВНАЯ СИСТЕМА ЖИВОГО ОРГАНИЗМА

Мозг живого организма – очень сложный орган. Он содержит 10^{12} клеток. Веский аргумент в пользу сложности мозга можно усмотреть в соединениях между его клетками. Обычная нервная клетка (нейрон) получает здесь информацию от сотен и тысяч других клеток и в свою очередь передает информацию сотням или тысячам нейронов. Общее число соединений в мозгу, таким образом, должно составить приблизительно $10^{14} - 10^{15}$ элементов.

Анатомическая сложность определяется не только числом элементов, но и характером организации, который трудно оценить числами. Можно провести аналогию между мозгом человека и большим компьютером, где аналогия заключается главным образом в наглядном представлении о множестве малых частей, смонтированных в определенном точном порядке, функции которых по отдельности уловить трудно.

Ниже рассмотрена обработка информации, представленная в изложении работ [9-11].

Детально разработанная область нейрoанатомии – это нечто вроде географии структур, функции которых все еще остаются загадкой или не вполне ясны. Нервная система состоит из следующих основных областей: лобная доля, височная доля, теменная доля, затылочная доля, стриарная кора, мозжечок, ствол мозга, спинной мозг. В этих областях локализованы следующие главные функции головного мозга: моторная область, сенсорная область, визуальная область, область слуха, область речи.

Наиболее важной для поведения живого организма является кора (cortex) головного мозга, содержащаяся в двух больших полушариях. Все доли коры головного мозга являются парными. Каждая пара состоит из связанных между собой структур, расположенных в левом и правом полушарии. Наличие двух полушарий является фундаментальным принципом организации мозга: правое полушарие обеспечивает, главным образом, организацию мозговых процессов, связанных с переработкой и хранением новой информации, а левое полушарие – с ранее известной информацией. При такой организации мозга оба полушария совместно участвуют в большинстве процессов, связанных с интеллектуальной деятельностью. Ранее неизвестные данные обрабатываются в правом полушарии и после такой переработки перемещаются в левое полушарие. Таким образом, в левом полушарии постепенно накапливаются знания, повышающие компетентность и способности нервной системы.

Кору можно определить как «ассоциативную» область мозга, т. е. область, в которой все ее подобласти действуют совместно. Поэтому, когда говорят о функциях отдельных частей мозга, подразумевают те процедуры, которые выполняют эти части во время их совместной работы по обеспечению интеллектуального поведения.

МОДЕЛЬ БИОЛОГИЧЕСКОГО НЕЙРОНА

Дадим упрощенное представление о том, что такое нервная система живого организма – как она построена и как работает. Основные строительные бло-

ки в мозгу – это нейроны или нервные клетки. Это не единственные клетки в нервной системе. В перечень структурных элементов мозга включены глиальные клетки, которые скрепляют нейроны и, вероятно, помогают питать их и удалять ненужные продукты обмена веществ; кровеносные сосуды и составляющие их клетки; различные покрывающие мозг оболочки.

В функциональном отношении нервные клетки делятся на три основных класса: 1) чувствительные (сенсорные), или рецепторные, нейроны; 2) двигательные (моторные) нейроны и 3) промежуточные нейроны. Эти три класса можно рассматривать как входные устройства, выходные устройства и все, что находится между ними.

Посмотрим теперь, как работает нейрон. Различные его части выполняют разные функции.

Тело нейрона обычно имеет шаровидную форму, свойственную многим клеткам (рис. 2), и содержит ядро. От тела клетки отходит отросток в виде цилиндрической нити – нервное волокно, передающее сигнал и называемое аксоном. Аксон представляет собой именно то, чем он кажется, – это путь проведения нервного импульса, возникшего в теле клетки, в какое-нибудь другое место, обычно к другому нейрону. Аксон – природный аналог соединительного проводника электрической цепи. Кроме аксона от тела отходит множество других ветвящихся и суживающихся к концу волосков, их называют дендритами. Вся нервная клетка – ее тело, аксон и дендриты – одета клеточной мембраной. Тело нейрона и дендритов получают информацию от других нейронов; аксон передает информацию от данного нейрона другим нейронам. Вблизи своего окончания аксон разделяется на многочисленные веточки, концевые участки которых очень близко подходят к телам или дендритам других клеток, но не соприкасаются с ними вплотную. Каждое соединение аксона одного нейрона с дендритом или телом другого называется

синапсом. В этой области информация передается от одной нервной клетки – пресинаптической, к следующей – постсинаптической.

Представляет интерес как первоначально возникают импульсы, что происходит на дальнем конце аксона, когда импульс прибывает туда и что происходит, когда информация передается от одной клетки к другой через синапс?

Сигналы в нерве возникают в точке аксона, близкой к месту его соединения с телом клетки; они передаются вдоль аксона, удаляясь от тела клетки, и доходят до области концевых разветвлений. Из окончаний аксона информация передается через синапсы следующей клетке или клеткам – здесь происходит химическая передача, которая имеет фундаментальное значение для нервных механизмов.

То, что проведение нервного возбуждения подчиняется закону «все или ничего», связано с общим свойством живых клеток. Протоплазма любой клетки отделена от окружающей среды очень тонкой прозрачной мембраной. Так как эта мембрана неодинаково проницаема для различных ионов, вещество клетки заряжено отрицательно по отношению к внеклеточной жидкости. В нервной клетке животного этот внутренний отрицательный потенциал составляет примерно 70 милливольт (см. рис.1). Подобным свойством обладают все клетки. Каждая клетка имеет мембрану, между двумя поверхностями которой существует разность потенциалов. Это интересный пример глубокого сходства между живыми организмами. Проведение нервного возбуждения становится возможным в результате специализации электрических свойств мембран живых клеток. Сигналы, поступающие через входные участки нейрона, вызывают эффекты, понижающие электрический потенциал протоплазмы в теле нейрона по сравнению с его нормальной величиной в 70 милливольт.

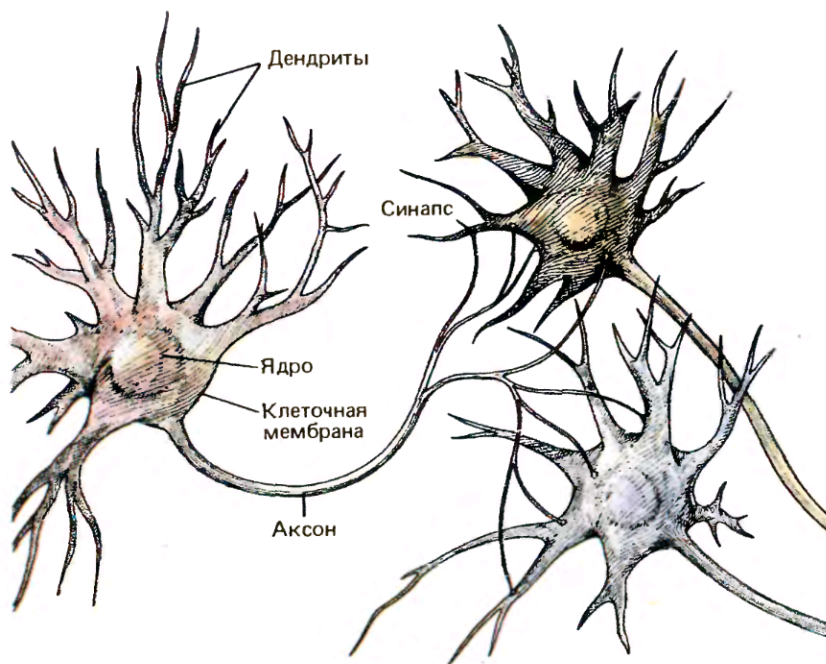


Рис. 2. Главные части нервной клетки-нейрона [9].

Это снижение потенциала распространяется на ближайший участок основания аксона. Если это снижение потенциала, или деполяризация, достигает достаточной величины, то аксон проявляет интересную, лишь ему свойственную особенность: происходит электрический «пробой» его оболочки. Это ведет к внезапному изменению проницаемости мембраны, отделяющей протоплазму аксона от окружающей жидкости. В результате этого изменения наружные ионы натрия, которые ранее не могли пройти через мембрану, устремляются внутрь аксона, тогда как избыток внутренних ионов калия выходит наружу. Суммарный электрический эффект этого перемещения ионов состоит в резком изменении внутреннего потенциала в том участке аксона, где происходит «пробой»: нормальная величина (-70 милливольт) сменяется слегка положительной величиной по отношению к окружающей жидкости. Возникающая при этом разность потенциалов между участком «пробоя» и соседней протоплазмой вызывает перемещение ионов, которое быстро приводит к деполяризации соседнего участка аксона, достаточной для того, чтобы вызвать «пробой» и в этом участке; при этом возникает ток, деполяризующий следующий участок аксона, и т.д. Таким образом, импульс деполяризации, или потенциал действия, распространяется вдоль аксона. Это и есть тот нервный импульс типа «все или ничего», который со времени его открытия составлял загадку для биологов.

Важной особенностью импульса деполяризации является скорость распространения. Она находится в сложной зависимости от химических и электрических свойств аксона и окружающей жидкости, а также от толщины аксона. В организме человека их скорость составляет от 1 до 100 м/сек в зависимости от толщины волокон. Другая особенность передачи импульса по аксону состоит в том, что он несколько не ослабевает при своем движении по нервному волокну. Когда это свойство было впервые открыто, казалось весьма удивительным, что нервный импульс, проходя расстояния в тысячу раз большие, чем диаметр проводящего волокна, прибывает к месту назначения таким же, каким он покинул тело породившего его нейрона. Сейчас видят в этом естественное следствие феномена распространения электрического «пробоя», при котором импульс эффективно возобновляется на каждом этапе своего продвижения. Энергия, необходимая для такой последовательной регенерации, первоначально образуется за счет различий в концентрации ионов натрия и калия внутри и вне аксона; эти различия, в свою очередь, являются результатом химических процессов, постоянно идущих на всем протяжении аксона, которые восстанавливают и поддерживают электрохимические условия в клетке, находящейся в состоянии покоя.

Возвращение протоплазмы аксона к нормальному состоянию равновесия после прохождения потенциала действия имеет очень большое значение для нервных механизмов. Аномальная проницаемость мембраны, лежащая в основе прохождения импульса по аксону, — явление весьма кратковременное; нормальные характеристики мембранной проницаемости и

обычная величина электрической поляризации восстанавливаются всего лишь за одну или две тысячные секунды. К тому моменту, когда участок аксона вновь приобретает способность к возбуждению, потенциал действия уже проходит расстояние, во много раз превосходящее диаметр аксона, и находится слишком далеко, чтобы вызвать повторный разряд в восстановившей свою возбудимость цитоплазме. В результате нервный импульс всегда распространяется в одном направлении — он удаляется от тела нейрона.

Когда протоплазма у основания аксона восстанавливает свои первоначальные свойства, в ней может быть вызван новый нервный импульс. Это происходит как только нервные импульсы, поступающие в нейрон через его входные участки, вновь создают в теле нейрона степень деполяризации, достаточную для возникновения нового разряда у основания аксона. Если приходящие сигналы сильны, потенциалы действия следуют друг за другом с короткими интервалами, и нервный сигнал может достичь частоты нескольких сотен импульсов в секунду. Если же сигналы слабы, то для повторной пороговой деполяризации тела нейрона после разряжающего действия каждого импульса может потребоваться сравнительно долгое время; в таком случае сигнал, распространяющийся по аксону, имеет частоту всего несколько импульсов в секунду.

Когда потенциал действия достигает синапса, происходят события совершенно иного рода. В синапсе приходящий электрический сигнал оказывается блокированным. Электронно-микроскопические исследования показали, в чем здесь дело: аксон передающего нейрона не соприкасается непосредственно с дендритом или телом нейрона, воспринимающего импульс. Всегда существует промежуток величиной около примерно 200 ангстрем. Путем экспериментов было установлено, что нервный импульс преодолевает этот промежуток с помощью химических передатчиков. Каждый приходящий импульс вызывает освобождение ничтожного количества вещества-передатчика, которое диффундирует через жидкость синаптической щели на другую ее сторону. Под влиянием этого вещества проницаемость мембраны воспринимающего нейрона изменяется, и это изменение приводит к перераспределению ионов, ведущему к изменению электрического заряда в протоплазме тела этого нейрона. За короткое время «импульс» электрического заряда распределяется по всему телу клетки и тем самым изменяет степень поляризации, которая и определяет, произойдет ли разряд у основания аксона. Важно отметить, что явления, происходящие на входных участках нейрона, протекают не по закону «все или ничего». Действительно, с точки зрения вычислительной техники они больше похожи на перевод данных из цифровой формы в аналоговую. После каждого пришедшего по аксону импульса на теле следующего нейрона накапливается некоторый заряд, медленно рассеивающийся. Хотя все приходящие импульсы имеют одну и ту же величину, поступление сигнала более высокой частоты вызывает в синапсе пропорционально большей эффект деполяризации. Последний в свою очередь может вызвать

импульсный разряд сравнительно высокой частоты в аксоне, отходящем от второго нейрона.

Описанная картина слишком упрощена, поскольку в редких случаях воспринимающий нейрон приводится в действие входным сигналом от одного единственного нейрона. В естественных условиях нейроны никогда не выступают поодиночке. Тело воспринимающего нейрона действует как своего рода «сумматор» для поляризующих эффектов, вызываемых различными входными сигналами. И чтобы входные сигналы могли суммироваться и произвести совместное действие, они должны приблизительно совпадать по времени. Обычно сенсорные стимулы приходят по большому числу смежных нервных волокон. Число промежуточных нейронов, вовлекаемых в процесс, также велико, и каждый из них имеет входные соединения с аксонами нескольких сенсорных нейронов; точно так же и любой выходной сигнал, т. е. результат обработки информации в нервной системе, передается на выход по нескольким аксонам, активирующим многочисленные мышечные волокна, участвующие в реакции. Таким образом, в нервной системе существуют параллельные каналы связи, что и объясняет ее относительно малую чувствительность к повреждению небольшого числа нейронов.

Интегрирующие свойства нейрона не сводятся к суммированию эффектов деполяризации в теле нейрона. Имеются механизмы, с помощью которых часть приходящих сигналов может ослаблять, а не усиливать общий эффект. Некоторые типы сигналов являются «тормозными». Химическая организация их такова, что в местах соединения их аксонов с дендритом или телом другого нейрона, приходящий нервный импульс вызывает выделение химического вещества, влияющего на проницаемость мембраны воспринимающего нейрона таким образом, что возникновение разряда в его аксоне не облегчается, а затрудняется. Этот электрический эффект называют тормозным постсинаптическим потенциалом в отличие от возбуждающего постсинаптического потенциала, вызываемого прибытием нервных импульсов через синапсы деполяризующего типа. Типичный промежуточный нейрон получает входные сигналы как от нескольких возбуждающих, так и от нескольких тормозных нейронов. Импульсы в его аксоне возникают или не возникают в каждый данный момент в зависимости от того, приведут ли интегрированные эффекты всех более или менее одновременных входных сигналов к надпороговой деполяризации внутренней стороны клеточной мембраны.

Перечисленные свойства нейрона придают ему большое сходство с элементом ЭВМ. При определенных условиях он, несомненно, обладает особенностями поведения, напоминающими свойства электронного устройства, применяемого в ЭВМ. Если бы многочисленные дендритные входы были сгруппированы параллельно и соединены только с двумя или тремя источниками сигналов и если бы использовались импульсные входные сигналы определенной частоты и амплитуды (как это обычно бывает в цепях цифровых вычислительных машин), то активное или неактивное состояние нейрона определялось бы ком-

бинацией входных сигналов. Таким путем можно было бы получить эквиваленты осуществляемых машинами логических операций типа «и», «или», «не»; а создание обратных связей позволило бы получить переключательные элементы типа триггера.

Однако нейрону присущи и другие свойства. Повышение выходной частоты с увеличением амплитуды входного сигнала, способность складывать и вычитать входные величины, наличие временного суммирования, изменчивость порога и другие свойства, – все это указывает на то, что нейрон – значительно более сложное устройство, чем электронное устройство в ЭВМ.

Связи между нейронами или группами нейронов мозга обычно не очевидны, для того чтобы выявить наиболее важные проводящие пути потребовалось столетие. Для изучения каждого пучка нейронов в отдельности нужны специальные методы. Ученые придумали эффективные способы выявления как отдельных клеток в какой-либо одиночной структуре, так и связей между разными структурами. Нейроны с одинаковыми или близкими функциями часто бывают связанными между собой. Тесно взаимосвязанные клетки в нервной системе группируются вместе по причине большей эффективности коротких аксонов: такие аксоны «дешевле изготовлять», они занимают меньше места и быстрее доставляют сообщения своим адресатам. Мозг поэтому содержит сотни клеточных скоплений, которые имеют форму шаров или параллельных слоев, шаровидные или пластинчатые структуры которых соединяются последовательно в проводящие пути. Примером такой цепи последовательных структур служит зрительный путь сетчатки двух глаз, содержащей светочувствительные рецепторные клетки. Две сетчатки глаз посылают свои выходные сигналы двум клеточным скоплениям, расположенным в глубине мозга. От этих структур идут волокна в зрительную область коры головного мозга.

Таким образом, за последнее время биологами исследованы разделы о том, как работают отдельные нейроны и как информация передается от клетки к клетке через синапсы, были изучены физико-химические механизмы проведения нервных импульсов и синаптической передачи. Задача нервной клетки состоит в том, чтобы принимать информацию от клеток, которые передают ее, суммировать, или интегрировать, эту информацию и доставлять интегрированную информацию другим клеткам. Информация обычно передается в форме кратковременных процессов, называемых нервными импульсами. В любой момент частота импульсов, посылаемых нейроном, определяется сигналами, только что полученными им от передающих клеток, и передает информацию клеткам, по отношению к которым этот нейрон является передающим. Частота импульсов варьирует от одного в каждые несколько секунд или еще ниже до максимума около тысячи в секунду. Что происходит, когда информация передается от одной клетки к другой через синапс? В первой – пресинаптической – клетке около основания аксона возникает электрический сигнал, или импульс. Импульс перемещается по аксону к его окончанию. Из каждого

окончания в результате этого импульса в узкий (0,02 мкм) заполненный жидкостью промежутки, отделяющий одну клетку от другой, – синаптическую щель – высвобождается химическое вещество, которое диффундирует ко второй – постсинаптической – клетке. Оно влияет на мембрану этой второй клетки таким образом, что вероятность возникновения в ней импульсов либо уменьшается, либо возрастает.

Каким образом происходит процесс распространения импульса? Перенос заряда по длине волокна требует времени. Активный участок возникает в одном месте и перемещается по волокну со скоростью ~ 10 м/сек. В любой момент времени существует один активный участок с реверсированным потенциалом, и эта область реверсии передвигается, удаляясь от тела нейрона. Впереди нее находится участок с еще не открытыми каналами, а сзади – участок, где каналы снова закрылись и временно неспособны к повторному открытию. Это и есть процесс распространения импульса. Он не похож, как видно, на физический импульс, т.е. на прохождение тока по медной проволоке. По всей длине нерва электрические заряды, ионы и другие материальные частицы не перемещаются. Здесь автором работы [9] приведена аналогия, сравнивающая смыкание лезвий ножниц, когда ничто не перемещается от соединительного винта до их кончиков. (Ионы образуют лишь местные токи, переходя внутрь и наружу, подобно тому, как лезвия ножниц движутся вверх и вниз). Перемещается некоторое событие или процесс – переключение лезвий ножниц или импульс в нерве. Характерная частота, с которой клетка или аксон способны генерировать импульсы, составляет 100 – 200 импульсов в секунду. Важной особенностью нервного импульса является то, что он возникает, как упоминалось выше, по принципу «все или ничего». Величина потенциала (т.е. импульса), распространяющегося по нерву, определяется самим нервом, а не степенью начальной деполяризации, которая привела к его возникновению. Уместно привести аналогию с любым взрывообразным процессом, когда скорость полета пули никак не связана с тем, насколько резко нажали на спусковой крючок.

МОДЕЛЬ БИОЛОГИЧЕСКОЙ НЕЙРОННОЙ СЕТИ

Всю деятельность головного мозга можно определить как некоторую переработку информации. Рецепторные нейроны доставляют сведения о состоянии внешнего мира и внутренних органов. Эти сведения дополняются записанной в памяти информацией о прошлом опыте, а затем эти комбинированные данные обрабатываются нейронными сетями и получается тот или иной конечный результат. У человека очень часто, а у низших животных почти всегда, этот конечный результат переработки информации представляет собой определенное сочетание двигательных команд, побуждающих мышцы или железы реагировать в соответствии с действующими раздражителями. Иногда – у человека и, вероятно, у

некоторых высших животных – конечный результат этой переработки носит «умственный» характер: это мысль или представление, которые не всегда ведут к видимой физической активности, так как для этого может потребоваться сочетание их с другими мыслями или представлениями. По существу, целью ученых является прийти к пониманию того, каким образом происходит переработка информации в головном мозгу. Поскольку у большинства животных значительная доля информации о внешнем мире воспринимается зрением, именно в зрительной системе больше всего возможностей и необходимости в переработке и реорганизации входных данных перед введением их в головной мозг.

Центральная нервная система, лежащая между входными и выходными нейронами, является тем аппаратом, который позволяет воспринимать, реагировать и помнить. Он же, в конечном счете, должен быть ответственен за наше сознание. Одна из главных целей биологов состоит в выяснении того, что происходит на этом пути – какая информация, приходящая к некоторой группе нейронов, преобразуется и затем передается дальше, а также какой смысл имеют эти преобразования для успешной адаптации живого организма к среде. Хотя схемы связей для многих отделов центральной нервной системы значительно варьируют в деталях, в большинстве случаев в основе их лежит сравнительно простая общая схема, показанная на рис.3, взятая из работы [9].

Прежде всего, на входе имеем не одну, а много сенсорных систем – зрение, осязание, вкус, обоняние и слух, где рецепторы преобразуют такие внешние воздействия, как свет, тепло или звук. Выходом могут быть сокращения мышц или реакции железистых клеток. И каждая система имеет свой собственный набор уровней в мозгу. Когда и где в мозгу различные наборы уровней сводятся вместе, если они действительно сводятся, все еще остается неясным. Рассмотрим эксперименты, давшие интересные сведения о переработке зрительной информации животными.

За последнее время проведено множество наблюдений за ползающими и летающими насекомыми, представляющими интерес для разработки роботов-насекомых. Особый интерес вызывает эксперимент над земноводными, ставший классическим. Глаз земноводных в основных чертах сходен с человеческим глазом. Изображение внешнего мира фокусируется на палочки и колбочки сетчатки. Эти первичные фоторецепторы воспроизводят в форме электрических сигналов распределение света и тени. Как показывают анатомические исследования, в глазу земноводных, точно так же как у человека, аксоны первичных фоторецепторов не соединены непосредственно со зрительным нервом, идущим к мозгу, а оканчиваются в нескольких слоях промежуточных нейронов, лежащих за слоем палочек и колбочек и соединенных между собой многочисленными перекрестными связями. Исследователи выяснили, какого рода «картина» передается по зрительному нерву и насколько она сходна с изображением на сетчатке.

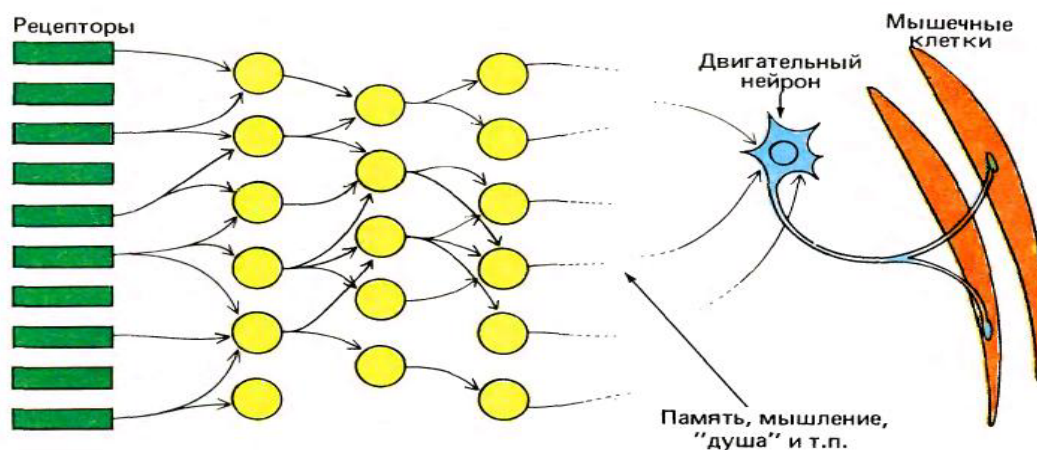


Рис.3. Упрощенная модель биологической нейронной сети [9].

В результате было выделено четыре различных изображения всего поля зрения, которые передаются одновременно по четырем группам волокон зрительного нерва. Эту картину можно сравнить с цветным телевидением, в котором отдельные изображения, соответствующие нескольким цветовым компонентам, одновременно передаются по каналу связи и затем воссоединяются в приемнике, создавая желаемый эффект. Далее оказалось, что в мозгу в месте приема, «изображения» накладываются друг на друга; зрительные волокна, несущие четыре различных вида информации о событиях в определенном участке сетчатки, оканчиваются в точности друг над другом в четырех последовательных слоях мозговой ткани. Однако свойства четырех зрительных «изображений» оказались различными. Исследователи охарактеризовали эти изображения по типам зрительной информации: детекторы зрительного контраста, детекторы движущихся границ, детекторы общего применения и детекторы кривизны. При проведении опытов биологов интересовало, как осуществляется переработка первичной зрительной информации от палочек и колбочек для получения окончательного «изображения» и как осуществляется детальная организация нейронных сетей, лежащая в основе этой переработки зрительных данных в глазу земноводных. Результаты экспериментов, проведенных на земноводных, побудили ученых искать аналогичные явления и у других животных.

Описанная выше работа была проведена на позвоночных Д. Хьюбелем и Т. Визелем [9]. Результаты указывают как на различие между земноводными и позвоночными (и, вероятно, человеком), так и на некоторые черты сходства. Авторы поставили опыты, в которых они вводили микроэлектрод в корковый нейрон и определяли участки сетчатки, которые необходимо было освещать, чтобы вызвать изменение частоты потенциала действия в этом нейроне. Оказалось, что рецептивное поле кортикального нейрона (та область сетчатки, в пределах которой изменение освещения вызывает эффект), никогда не имело форму круга, как для волокон зрительного нерва. В гео-

метрии рецептивных полей корковых нейронов преобладающими оказались не круги, а прямые линии.

Для отдельного нейрона наиболее эффективный световой раздражитель часто имел форму «щели», длина которой во много раз превышала ширину. Иногда это была линия резкого контраста, но всегда прямая, а не дугообразная. Ориентация этой щели или прямой границы на сетчатке обычно имела решающее значение: изменение в $5 - 10^0$ существенно влияло на ответы исследуемого коркового нейрона. Когда микроэлектрод вводили в зрительную кору точно под прямым углом к ее поверхности, получаемые результаты говорили о том, что «изображения», представленные электрической активностью нейронов в последовательных слоях коры, с большой точностью накладываются друг на друга. Кроме того, рецептивные поля нейронов одного «столбика» перекрывали на сетчатке друг друга, хотя и не были тождественными. Если микроэлектрод вынимали и снова вводили в близлежащую точку поверхности коры, то нейроны в этом новом «столбике» опять имели общую ось рецептивного поля, ориентация которого, однако, могла быть отличной от ориентации осей, характерной для других, соседних параллельных «столбиков» корковой ткани. Авторы заключили, что нет преимущественной ориентации осей рецептивных полей на сетчатке: число корковых столбиков, реагирующих на горизонтальные световые линии, на вертикальные линии или на линии, занимающие промежуточные положения, было примерно одинаковым. Было предположено, что преобразование зрительной информации из радиально-симметричной формы в линейно-симметричную происходит там, где она впервые вступает в кору, а не в последующих слоях нейронов, осуществляющих ее переработку. Такое преобразование, по-видимому, должно быть связано с особой схемой нейронных соединений: все волокна от различных рецепторов, лежащих на участке сетчатки, должны служить источниками входных сигналов для одного и того же кортикального нейрона. Они пришли к выводу, что значительная часть реорганизации первичных зрительных данных

в зрительной коре позвоночных осуществляется еще до того, как сигналы поступают в более глубокие слои коры.

Отметим, что Д. Хьюбелем и Т. Визелем выявлены особенности зрительного восприятия глаза. Психологов всегда изумляла необычная способность нашего глаза замечать прямые линии света и тени во всем, на что смотрим. Если организация зрительной системы у человека во многом сходна с ее организацией у позвоночных, то мы имеем реальное физиологическое объяснение этой особенности восприятия: мы так легко замечаем прямые линии потому, что каждый нейрон в нашей зрительной коре получает входные сигналы от линейных групп рецепторов сетчатки. Другой особенностью является реакция нейронов на освещение сетчатки. Так, например, в некоторых случаях отдельный корковый нейрон реагировал на световую «щель» или прямую контрастную границу независимо от положения ее на сетчатке. Однако определенная угловая ориентация этой щели или линии контраста должна была в точности сохраняться; незначительный поворот приводил к полному исчезновению реакции коркового нейрона.

Возможно, что именно такие нейронные элементы, индифферентные по отношению к локализации прямой линии в поле зрения, но чувствительные к ее угловой ориентации, лежат в основе нервных механизмов. Эти механизмы позволяют человеку и другим высшим позвоночным распознавать треугольники, квадраты и иные прямолинейные геометрические фигуры, независимо от их расстояния от глаза и положения в поле зрения. В настоящее время ведутся исследования, необходимые для полного понимания различий между изображением внешнего мира на сетчатке глаза и субъективным ощущением, возникающим в мозгу.

Подытожим сказанное. Наблюдаемая биологами морфологическая картина клетки-нейрона имеет несколько входов и только один выход, который может разветвляться на сеть отростков. Обычно один нейрон имеет от одного до нескольких сотен, возможно, около тысячи входов. Известно, что отдельный нейрон соединен не со всеми, а только с некоторыми из рецепторов-датчиков. Рецепторов, например, у человека несколько миллионов. Поэтому очевидно, что один нейрон не может быть соединен со всеми рецепторными клетками или всеми нейронами предыдущего слоя нервной сети, как это принято в современных традиционных искусственных нейронных сетях, где каждый нейрон соединяется со всеми датчиками или всеми нейронами предыдущего слоя, что явно не соответствует биологическим данным.

МОДЕЛЬ ИСКУССТВЕННОГО НЕЙРОНА

Основное понимание в годы появления искусственной нейронной сети было сосредоточено на попытках понять принцип действия нервной клетки и сетей из таких клеток, но не принцип действия всей нервной системы и мозга. Первая модель нейрона, построенная по законам математической логики, была предложена У. Маккалоком и У. Питсом [13]. Эта модель была, по-видимому, более других адекватна свойствам биологического нейрона и сетей из нейронов.

В дальнейшем схема этого нейрона развивалась фон Нейманом [14], Ф. Розенблатом [15] и др. В результате этих усилий была получена схема формального нейрона, которая демонстрировала очень полезное свойство – она могла быть обучена распознавать некоторые образы. С этого момента формальный нейрон уже существенно отличался от модели У. Маккалока и У. Питса, но был очень удобной моделью для программирования.

Формальный нейрон в современном варианте представляет собой элемент [15], на входах которого имеются возбуждающие и тормозящие синапсы. В нейроне определяется взвешенная сумма (с учетом весов синапсов) входных сигналов, при превышении этой суммой порога нейрона вырабатывается выходной сигнал. Работа нейрона (рис.4) может быть описана уравнениями:

$$Y_i = F(net_i - K_j)j$$

$$net_i = \sum_j w_{ji} X_j$$

где j – номер нейрона в сети, X_i – входные сигналы, Y_i – выходной сигнал нейрона, w_{ji} – веса синапсов, net_i – суммарное входное воздействие на нейрон, K_j – порог нейрона, $F(.)$ – активационная функция.

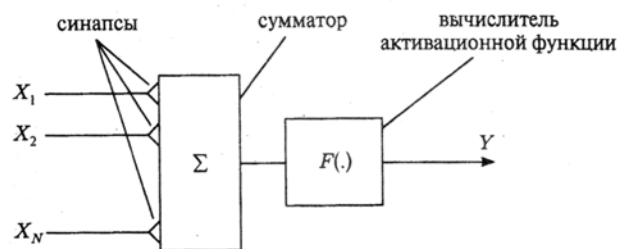


Рис.4. Схема формального нейрона:

X_i – входные сигналы,

Y_i – выходной сигнал нейрона [15].

Активационная функция характеризует реакцию нейрона на входное воздействие net_j , она может быть пороговой:

$$F(a) = \begin{cases} 0 & \text{при } a \leq 0, \\ 1 & \text{при } a > 0 \end{cases}$$

или некоторой непрерывной, например, линейной:

$$F(a) = ka$$

или логистической:

$$F(a) = 1/[1 + \exp(-a)]$$

В зависимости от реализуемого алгоритма на допустимые значения входов и выходов нейрона накладываются определенные ограничения. Значения X_i и Y_i могут быть бинарными (т. е. равными 0 или 1), бинарными биполярными (+1 или -1), действительными и принадлежащими интервалу (0, 1) и произвольными действительными числами. Аналогичные ограничения накладываются на веса синапсов нейронов w_{ji} .

Очень упрощая картину, работу биологического нейрона можно описать следующим образом. Дендриты получают сигналы от других клеток через синапсы, эти сигналы поступают в тело клетки, где они суммируются с другими такими же сигналами. Если суммарный сигнал в течение короткого промежутка времени является достаточно большим, то клетка возбуждается, вырабатывая в аксоне импульс, который передается на следующие клетки. Не вдаваясь в подробности, отметим, что искусственные нейроны только очень грубо отражали работу биологических нервных клеток. Из-за несовершенства модели произошло падение интереса к разработке искусственной нейронной сети. Однако естественное желание усовершенствовать искусственную модель нейрона и внести в нее возможно большее подобие биологическому нейрону отразилось в том, что продолжались разрабатываться более совершенные модели, в которых, однако, за большее биологическое правдоподобие приходилось расплачиваться большой вычислительной сложностью. Эти модели расширили арсенал разработчика прикладных систем и дали ему возможность выбирать модели нейронов, адекватные конструируемым прикладным системам.

На рис. 5 представлена традиционная нейронная сеть, которая состоит из сети нейронов, рецепторных клеток и вспомогательных структур. Рассмотрим нейронные сети, которые должны выполнять функции системы распознавания. Обратимся к наблюдаемой биологами морфологической картине. Здесь на множестве выходов рецепторных клеток, по которым передаются импульсы, начинается сеть нервных клеток, передающих информацию преимущественно по направлению от рецепторов вглубь нервной системы. Наиболее значимое направление исследований искусственных нейронных сетей принадлежит Ф. Розенблатту, который создал электромеханическое устройство, названное «перцептрон, модель Mark 1» [15]. Его элементарный перцептрон функционирует следующим образом (рис. 6).

Перцептрон состоит из элементов трех типов: S-элементов (рецепторы), A-элементов (слой ассоциативных нейронов) и R-элемента (выходного сумматора). На сетчатку (глаз, ухо, матрицу тактильных датчиков) проецируется некоторый образ (изображение, звук, карта давления).

Чувствительные элементы (сенсоры), освещенные сильнее некоторого порогового значения, «возбуждаются» и передают электрический сигнал (+1) суммирующим A-элементам. Не возбужденные чувствительные элементы передают сигнал «0». Далее в каждом A-элементе происходит суммирование сигналов от произвольного количества входящих связей.

Сигналы могут быть тормозящими (-1), нулевыми (нет сигнала) и возбуждающими (+1). Каждому входу может быть поставлен «вес». Ф. Розенблатт полагал, что номера индексов и само их количество у S-элементов следует подбирать случайным образом. Он считал, что, например, глаза живых организмов «случайно вычисляют» такие комбинации S, которые важны для распознавания. Такой вид обучения он назвал S-управляемым обучением.

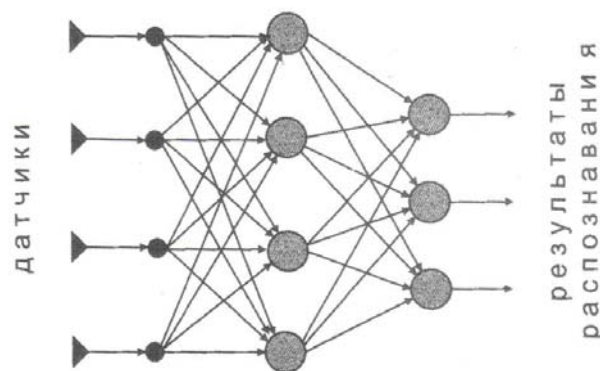


Рис. 5. Структура традиционной искусственной нейронной сети [8].

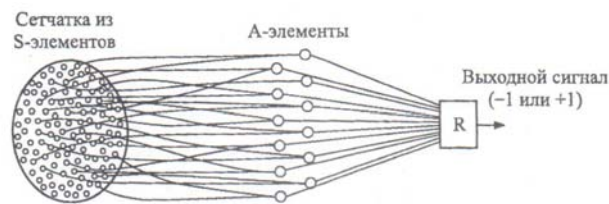


Рис. 6. Элементарный перцептрон [15].

Сигналы от всех A-элементов умножаются на веса (это уже другие веса), суммируются в R-элементе и сравниваются с пороговым значением [16]. В зависимости от того, превышен порог или нет, выдается сигнал -1 или +1. В процессе обучения перцептрона подбирают как раз эти веса и порог (рис. 7).

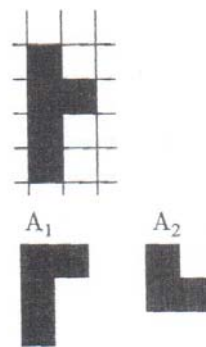


Рис. 7. Исходная сетчатка S и два элемента A, случайно настроившиеся на ее восприятие [16].

Исследования персептронов показали, что они способны обучаться, однако, способности довольно ограничены. Исследования также показали, что слабые стороны персептрона (в частности большое время обучения) в значительной степени зависят от случайных связей между его элементами. К этому времени исследователями был разработан ряд искусственных нейронных сетей, и в течение некоторого времени казалось, что ключ к интеллекту найден, а воспроизведение биологического мозга является лишь вопросом конструирования достаточно большой искусственной нейронной сети.

Однако возможности персептрона оказались довольно ограниченными. Задачи, которые, в принципе, могут быть решены персептроном, могут потребовать нереально больших времен и нереально большой памяти. Поэтому возник спад активности исследований искусственных нейронных сетей. Тем не менее, эксперименты, поставленные на модели «Марк 1», для начинающих экспериментаторов по-прежнему интересны и актуальны [16]. Для начального знакомства с моделированием нейронных сетей в Интернете имеются наборы программ, например, программа – Robot Execution Failures.

Современный вид модели нейронных сетей приняли под влиянием работ Д. Хопфилда [17], Т. Кохонена [18], К. Фукусимы [19], А.А. Фролова [20], Дж. Эдельмана [21] и др.

В эволюционном контексте интересен цикл работ Института нейронаук (США), в котором под руководством Дж. Эдельмана проводятся исследования поколений постепенно усложняющихся моделей мозга, строятся модели, обеспечивающие поведение робота NOMAD (Neurally Organized Mobile Adaptive Device) [21]. Здесь же разработана модель сокращенного описания нейрона [22], которая позволяет адекватно, но вместе с тем просто (с помощью системы обыкновенных дифференциальных уравнений) моделировать динамику нейронных сетей. Разработка этой модели позволит реализовать модель искусственного мозга с помощью компьютеров. Такой уровень интеграции соответствует уровню мозга человека. То есть, хотя принципы работы мозга во многом неизвестны, уже готовится инструмент для его моделирования. Работы Института нейронаук ориентированы на исследование формирования поведения робота NOMAD.

Разработка моделей адаптивного поведения для целей робототехники ведется в проекте «Мозг Анимата», который нацелен на разработку широкого спектра моделей адаптивного поведения. Основной подход этого направления – конструирование и исследование искусственных (в виде компьютерной программы или реального робота) организмов, способных приспосабливаться к внешней среде. Эти организмы называются «аниматами» (animal + robot = animat) [23–25].

Поведение аниматов имитирует поведение животных. Для описания живых организмов используются феноменологический (на уровне поведения) и микроскопический (на уровне нейронов и молекул) подходы. При феноменологическом подходе предполагается, что существуют формальные правила адаптивного поведения, и эти правила не обязательно связаны с конкретными микроскопическими нейрон-

ными или молекулярными структурами, которые есть у живых организмов. При этом считают, что для исследования систем управления адаптивным поведением феноменологический подход должен быть более эффективен (по крайней мере, на начальном этапе работ), так как очень трудно сформировать целостную картину поведения на основе анализа всего сложного многообразия функционирования нейронов, синапсов, молекул.

Исследователи направления «Адаптивное поведение» стараются строить именно такие модели, которые применимы к описанию поведения, как реального живого организма, так и искусственного анимата. Программа-минимум направления «Адаптивное поведение» предназначена исследовать архитектуры и принципы функционирования, которые позволяют животным или роботам жить и действовать в переменной внешней среде. Программа максимум предназначена для попыток проанализировать эволюцию когнитивных способностей животных и эволюционное происхождение человеческого интеллекта. Схемы управления адаптивным поведением тестируются на модельных (реализованных в компьютерных программах) или реальных роботах, что придает направлению исследований реальность и надежность.

ЗАКЛЮЧЕНИЕ

Анализ ряда работ в области теории высших функций головного мозга, искусственного интеллекта и организации компьютерных систем, работающих в режиме реального времени, показал наличие глубоких, более прогрессирующих связей между их идеями. Если ранее эта связь просматривалась на уровне общих концепций, то в настоящее время она приобретает глубинный характер. Независимо друг от друга, специалисты в области интеллекта биологических организмов начали описывать, а конструкторы, проектирующие компьютерные системы, начали создавать системы с адекватной структурой.

Дальнейшие исследования направлены на создание теоретических подходов, призванных объяснить, каким образом элементарные операции, выполняемые нейронами в головном мозгу, могут в совокупности порождать функциональные атрибуты разумной деятельности. Применение знаний по механизмам обработки информации нейронами, почерпнутых из биологии, могут привести в дальнейшем к разработке моделей интеллектуальных процессов и к расширению набора функциональных элементов при создании «мыслящих» ЭВМ и роботов.

Секрет эффективности современных ЭВМ состоит в том, что все сложные вычислительные и логические операции можно разбить на элементарные действия, с которыми могут справиться очень простые устройства. Можно ожидать, что характер этих элементарных действий будет со временем изменяться. Но не изменится то, что составляет действительно фундаментальную особенность всего класса ЭВМ, – то, что они достигают своих поразительных результатов путем выполнения огромного числа очень простых операций. Этими же словами, по-видимому, можно было бы охарактеризовать и сущность работы мозга. В вычислительной технике преобладает тео-

рия, каким образом можно разбивать искусственные интеллектуальные процессы на простые операции. В этой связи разрабатываются машинные программы, проводятся эксперименты на роботах и интерпретируются полученные результаты. Однако ни по принципам своего действия, ни по своим функциональным свойствам современные компьютеры не похожи на «природные компьютеры» – нервные системы.

Как и какими способами решается задача сближения свойств искусственных машин – компьютеров и нервных систем живых организмов? Основные области исследований касаются биологии, вычислительной техники, программирования. Препятствием является необходимость освоения большой суммы специальных знаний в каждой из этих областей, требующихся для решения указанной проблемы.

СПИСОК ЛИТЕРАТУРЫ

- Edelman G.M. Group selection and phasic re-entrant signaling; A theory of higher function // *Mindful Brain* / ed. G.M. Edelman, V.B. Mountcastle. – Cambridge, MA: MIT, 1978; перевод: Эдельман Дж. Селекция групп и фазовая повторная сигнализация, теория высших функций мозга // Эдельман Дж., Маунткэстл В. Разумный Мозг. – М.: Мир, 1981.
- Mountcastle V.B. An organizing principle for cerebral function // *Mindful Brain* / ed. G.M. Edelman and V.B. Mountcastle. Cambridge, MA: MIT, 1978; перевод: Маунткэстл В. Организующий принцип функции мозга – элементарный модуль и распределенная система // Эдельман Дж., Маунткэстл В. Разумный Мозг. – М.: Мир, 1981.
- Поспелов Д.А. Ситуационное управление: теория и практика. – М: Наука, 1986. – 288с.
- Поспелов Д.А. Становление информатики в России. – М., 1997. – URL: <http://raai.org/about/persons/pospelov/pages/stanovl.htm>.
- Рапопорт Г.Н., Герц А.Г. Искусственный и биологический интеллект. Общность структуры, эволюция и процессы познания. – М.: КомКнига, 2005. – 312 с.
- Zhang B. et al. A review of the commercial Brain – Computer Interface technology from perspective of industrial robotics // *IEEE Int. Conf. Autom. Logistics*. – Aug., 2010. – P. 379 – 384.
- Вулдридж Д. Механизмы мозга. – М.: Мир, 1965. – 314 с.
- Жданов А.А. Автономный искусственный интеллект. 2-е изд. – М.: БИНОМ. Лаб. знаний, 2009. – 360 с.
- Хьюбел Д. Глаз, мозг, зрение / пер. с англ. – М.: Мир, 1990. – 239 с.
- Becedas J. Brain – Machine Interfaces: Basis and Advances // *IEEE Trans. Syst., Man, and Cybern. C*. – 2012. – Vol. 42, № 6. – P. 825 – 836.
- Lebedev M., Nicolelis M. Brain – Machine Interfaces: Past, present and future // *Trends Neurosci.* – 2006. – Vol. 29, № 9. – P.536 – 546.
- Редько В.Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики. Изд. 8-е. – М.: Либроком, 2012. – 220 с. – (Синергетика: от прошлого к будущему, № 23).
- Мак-Каллок У.С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // *Нейрокомпьютер*. – 1992. – № 3,4. – С.40 – 53.
- Фон Нейман Дж. Теория самовоспроизводящихся автоматов – М.: Мир, 1971. – 382 с.
- Розенблатт Ф. Принципы нейродинамики (персептрон и теория механизмов мозга). – М.: Мир, 1965.
- Смолин Д.В. Введение в искусственный интеллект: Конспект лекций. 2-е перераб. изд. – М.: Физматлит, 2007. – 260с.
- Hopfield J.J., Tank D.W. Computing with neural circuits: A model // *Science*, 1986. – Vol.233, № 464. – P. 625– 633.
- Kohonen T. Self-organized formation of topologically correct feature maps // *Biol. Cybern.* – 1982. – Vol. 43, № 1. – P. 56–69.
- Fukushima K. Neocognitron: A hierarchical neural network capable for visual pattern recognition // *Neural networks*. – 1988. – Vol.1, № 2. – P. 119–130.
- Фролов А.А., Муравьев И.П. Информационные характеристики нейронных сетей. – М.: Наука, 1988. – 160 с.
- Krichmar J.L. Edelman G.M. Brain-based devices; Intelligent systems based on principles of the nervous system // *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. – 2003. – P. 940–945.
- Izhikevich E.M. Which model to use for cortical spiking neurons? // *IEEE Transactions on Neural Networks*. – 2004. – Vol. 15, № 5. – P. 1063–1070.
- Гаазе-Рапопорт М.Г., Поспелов Д.А. От амёбы до робота: модели поведения. – М.: URSS, 2011. – 296 с.
- Guillot A., Meyer J.-A. From SAB94 to SAB2000: What's new, Animat? // *From Animals to Animats 6. Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* / eds. Meyer J.-A. et al. The MIT Press, 2000/ See also: animatlab.lip6.fr/index.en.html.
- Непомнящих В.А. Поиск общих принципов адаптивного поведения живых организмов и автоматов // *Новости искусственного интеллекта*. – 2002.– № 2.– С. 48–53.

Материал поступил в редакцию 24.02.15.

Сведения об авторе

ПЕТРИН Александр Анатольевич – главный специалист ООО «АКА-Контроль», Москва
e-mail: alptr@rambler.ru

Автоматический поиск соответствий между системами классификации на примере сопоставления рубрик МПК и ГРНТИ*

Разработан метод автоматического поиска соответствий между классами двух систем классификации. Метод позволяет выявлять отношения тематического включения и пересечения между классами с помощью индикаторов, численно характеризующих отношения между множествами классифицированных документов. Выполнен поиск отношений включения и пересечения между рубриками Международной патентной классификации (МПК) и Государственного рубрикатора научно-технической информации (ГРНТИ).

Ключевые слова: системы классификации документов, ГРНТИ, МПК, сопоставление систем классификаций, таблицы соответствия

ВВЕДЕНИЕ

Всего в России в области научной информации и в смежных сферах применяется около 20 тематических классификаций и рубрикаторов, такие как ГРНТИ, УДК, ББК, МПК, WoS, ОЭСР, SCOPUS, ОКСВНК и др. [1].

Создание таблиц соответствия между различными системами классификации позволяет повысить эффективность решения следующих задач:

- информационный обмен между организациями НТИ;
- поиск в разнородных базах данных, использующих различные системы классификации;
- совершенствование систем классификации НТИ.

Различные подходы к сопоставлению классификаторов можно условно разделить на три категории:

- 1) экспертные;
- 2) статистические;
- 3) комбинированные (экспертно-статистические).

В первом случае эксперты сопоставляют тематические рубрики вручную, опираясь на своё знание предметной области и описания рубрик. В основе методов второй категории лежат различные частоты, вычисляемые на множестве документов, индексированных кодами рубрик. Статистическая обработка выборочных массивов документов позволяет выявить различные отношения соответствия между классификаторами (пересечение, включение, эквивалентность тематики). В комбинированных подходах на первом этапе применяются статистические методы,

позволяющие найти гипотетические соответствия, на втором этапе специалисты выполняют экспертизу найденных соответствий. Цель экспертизы не только принять или отвергнуть, но также уточнить вид соответствия между рубриками.

Выбор статистического метода полностью зависит от наличия или возможности получения выборочного массива документов. Если в распоряжении исследователей находится значительный по объёму массив документов, индексированных кодами обеих систем классификации, то самый прямой путь к сопоставлению классификаторов – использовать частоту совместной встречаемости классификационных кодов в описаниях документов. Если такого массива документов нет, то к цели придётся идти окольным путём: вначале присвоить рубрикам ключевые слова, затем сопоставить рубрики посредством сопоставления ключевых слов, присвоенных рубрикам. Для этого необходимо иметь два массива документов, оснащённых ключевыми словами и кодами рубрик – по одному массиву на каждую систему классификации.

Настоящая публикация посвящена автоматизации поиска соответствий между ГРНТИ и МПК «впрямую», т.е. на основании результатов статистической обработки электронного массива патентных документов, классифицированных кодами обеих систем классификации.

ХАРАКТЕРИСТИКА СОПОСТАВЛЯЕМЫХ СИСТЕМ КЛАССИФИКАЦИИ

Государственный рубрикатор научно-технической информации (ГРНТИ) представляет собой универсальную иерархическую классификацию областей знаний, принятую для систематизации всего потока научно-технической информации. На основе Рубри-

* Работа выполнена в рамках НИР, поддержанной субсидией Минобрнауки России (шифр лота 2014-14-573-0024, № госрегистрации 114120470044 от 04.12.2014).

котора построена система локальных (отраслевых, тематических, проблемных) рубрикаторов, используемых в органах научно-технической информации для организации информационных потоков и поиска в базах данных. ВИНТИ РАН является разработчиком большей части разделов ГРНТИ, выполняет задачи поддержания и развития ГРНТИ [2].

Международная патентная классификация, созданная на основании Страсбургского соглашения 1971 г., предусматривает иерархическую систему независимых от языка символов для классификации патентов и полезных моделей в зависимости от различных областей технологии, к которым они относятся [3].

Развитием и публикацией новых версий Международной патентной классификации (МПК) занимается Всемирная организация интеллектуальной собственности (ВОИС) – World Intellectual Property Organization (WIPO). Эталонные классификации версии публикуются на английском и русском языке. Переводом и публикацией версий МПК на русском языке занимается Федеральный институт промышленной собственности (ФИПС), официальный сайт <http://www1.fips.ru>.

По состоянию на 2014 г. Государственный рубрикатор научно-технической информации (ГРНТИ) содержит 8 273 рубрики.

Действующая версия Международной патентной классификации – МПК-2015.01 – вступила в силу 1-го января 2015 г. и содержит 71 738 рубрик [3, 4].

Такая значительная разница в количестве рубрик объясняется тем, что ГРНТИ имеет всего лишь три уровня иерархии, в то время как МПК – пять уровней. В результате, несмотря на более широкий охват отраслей науки и техники, ГРНТИ имеет в 8,7 раз меньше рубрик, чем МПК, которая более развита в глубину.

Классификационная схема МПК в 2006 г. была разделена на два уровня – базовый и расширенный. К базовому уровню относились рубрики первых трёх уровней иерархии, остальные рубрики относились к расширенному уровню. Впоследствии, в 2011 г., разделение на базовый и расширенный уровень было отменено, и МПК вновь стал публиковаться в виде полной схемы. Однако первые три уровня МПК остаются более стабильными во времени. В предшествующий выходу версии 2015.01 годовой период пересмотра МПК в её текст было введено 345 новых рубрик и аннулировано 39. При этом на первых трёх уровнях иерархии были введены всего лишь две рубрики – это класс В33, который содержит подкласс В33У [4].

Относительная стабильность во времени первых трёх уровней МПК и обозримое экспертами количество рубрик (639 рубрик на третьем уровне) привели к следующему решению: на данном этапе работы ограничиться сопоставлением рубрик МПК третьего уровня (МПК-3) с рубриками ГРНТИ третьего уровня (ГРНТИ-3).

Все рубрики МПК-2015.01, расположенные на первых трёх уровнях иерархии, были загружены в рабочую базу данных с сайта ФИПС <http://www1.fips.ru>. В результате в БД оказались 777 рубрик с названиями и примечаниями на русском языке, из которых 639 рубрик относятся к третьему уровню МПК.

ПОДГОТОВКА ВЫБОРОЧНОГО МАССИВА ДОКУМЕНТОВ

Для любого статистического метода необходимо сформировать выборку. В нашем случае выборка – это электронный массив описаний патентных документов, оснащённых рубриками МПК и ГРНТИ. Источником выборки послужила технологическая база данных ВИНТИ, используемая для подготовки реферативных журналов и баз данных. В ней оказался значительный объем патентных документов, имеющих коды МПК и ГРНТИ.

Формирование выборочного массива документов было выполнено в три этапа:

- 1) отбор патентных документов из технологической БД (около 100 тыс. записей);
- 2) исключение из выборки тех документов, описания которых не содержат хотя бы один код ГРНТИ;
- 3) приведение кодов рубрик МПК к нормализованному виду.

Последний этап нуждается в пояснении. Рубрики МПК в массиве документов имели различный вид. Приведём некоторые примеры:

Исходный вид	Нормализованный вид
B23Q 11/10 (2006.01)	B23Q 11/10
G01S17/00	G01S 17/00
A 61 K 31/716	A61K 31/716
C 07 H 1/08 (2006.01)	C07H 1/08
B60K 6/40 (2007.10)	B60K 6/40
B 09 C 1/06 (2007.01)	B09C 1/06
{8} A23G 3/52	A23G 3/52
{7} B 23 K 10/00	B23K 10/00
{7}G01F 23/296	G01F 23/296

Из этих примеров видно следующее:

- практика употребления пробелов в кодах МПК со временем менялась;
- номера версий МПК от года к году различаются, причём обозначаются по-разному: в фигурных скобках перед кодом МПК (в версиях до 2006 г.) или в круглых скобках после кода МПК.

Исключение номера версии и унификация применения пробелов (базовый уровень отделён пробелом от расширенного уровня) позволили привести все коды МПК к нормализованному виду. Таким образом, каждый код МПК в выборочном массиве документов представлен в единственном виде.

В результате был сформирован выборочный массив из 96321 патентных документов, пришедших в ВИНТИ в течение 2010–2014 гг. и отразившихся в РЖ ВИНТИ. Все документы массива снабжены хотя бы одним кодом МПК и хотя бы одним кодом ГРНТИ.

СТАТИСТИЧЕСКИЙ МЕТОД ПОИСКА ОТНОШЕНИЙ ВКЛЮЧЕНИЯ

В [1] для рассматриваемой рубрики ГРНТИ в каждой сопоставляемой классификации предлагается отыскивать рубрики, в определённой мере соответствующие по смыслу. При этом устанавливаются три меры смыслового соответствия:

- 1) эквивалентность по содержанию (совпадение объёмов понятия);

2) иерархическая подчинённость (включение объёмов понятия);

3) ассоциативная связь (значительное пересечение объёмов понятий).

Мы используем те же смысловые соответствия между рубриками, только для удобства для них введены более краткие названия и система обозначений. Итак, требуется статистическим методом выявить следующие вероятные отношения между тематиками рубрик А и В двух различных систем классификации:

- равенство тематики А и В, обозначим как $A = B$.

- включение тематики А в тематику В, обозначим как $A \subseteq B$;

- включение тематики В в тематику А, обозначим как $B \subseteq A$;

- пересечение тематик А и В, обозначим как $A \cap B$.

Фразу «тематика рубрики А включена в тематику рубрики В» можно сформулировать по-другому: «тематика рубрики В полностью охватывает тематику рубрики А». Как известно, между отношениями включения и равенства существует логическая связь, выраженная следующим соотношением:

$$A \subseteq B \text{ \& } B \subseteq A \leftrightarrow A = B.$$

Была поставлена задача – разработать и применить на практике статистический метод выявления гипотетических отношений включения между рубриками МПК третьего уровня (МПК-3) и рубриками ГРНТИ третьего уровня (ГРНТИ-3). Другими словами, для кодов МПК-3 требовалось найти коды ГРНТИ-3, которые, вероятно, полностью охватывают их тематику. Аналогично для кодов ГРНТИ-3 требовалось найти коды МПК-3, которые, вероятно, полностью, охватывают их тематику.

В основе разработанного метода три индикатора, численно характеризующие отношения между множествами документов, классифицированных рубриками. Индикаторы вычисляются для каждой пары кодов (МПК, ГРНТИ), совместно встретившихся в описании хотя бы одного документа выборки.

1. «Мощностью пересечения» назовём следующую абсолютную частоту:

$$\omega = |M \cap G|,$$

где М – множество документов с данным кодом МПК, G – множество документов с данным кодом ГРНТИ.

Поскольку мощность любого конечного множества равна количеству элементов этого множества, величина ω равна количеству документов. Таким образом, ω является абсолютной частотой совместной встречаемости кодов МПК и ГРНТИ в описаниях документов.

2. «Мерой включения МПК в ГРНТИ» назовём следующую относительную частоту:

$$\alpha = |M \cap G| : |M|.$$

3. «Мерой включения ГРНТИ в МПК» назовём следующую относительную частоту:

$$\beta = |M \cap G| : |G|.$$

Меры включения имеют вероятностный смысл. Мера включения МПК в ГРНТИ – это статистическая оценка вероятности события «случайно выбранный документ с кодом МПК относится к тематике кода ГРНТИ». Аналогично мера включения ГРНТИ в МПК – это статистическая оценка вероятности события «случайно выбранный документ с кодом ГРНТИ относится к тематике кода МПК»

На основании введённых мер включения выдвигаются следующие гипотезы, подтвердить или опровергнуть которые для каждой пары (МПК, ГРНТИ) может только экспертная оценка:

$$\alpha \approx 1 \rightarrow \text{МПК} \subseteq \text{ГРНТИ},$$

$$\beta \approx 1 \rightarrow \text{ГРНТИ} \subseteq \text{МПК},$$

$$\alpha \approx 1 \text{ \& } \beta \approx 1 \rightarrow \text{МПК} = \text{ГРНТИ},$$

$$\alpha > 0 \vee \beta > 0 \rightarrow \text{МПК} \cap \text{ГРНТИ}.$$

Необходимо пояснить причину использования знака приближённого равенства $\approx$ вместо знака точного равенства. Дело в том, что результат интеллектуальной операции классифицирования документа не является строго детерминированным. Он зависит от квалификации и усердия специалиста, выполняющего классификацию, от знания и опыта применения классификационной схемы, знакомства с предметной областью, умения выделить несколько тематических аспектов в политематическом документе. При этом любой специалист не застрахован от ошибок. Следовательно, для выявления максимального количества отношений включения необходимо выдавать на экспертизу не только те пары (МПК, ГРНТИ), у которых мера включения равна единице, но и те, у которых мера включения близка к единице.

В любой поисковой системе необходимо соблюдать разумный компромисс между полнотой и точностью поиска. В нашем случае выполняется поиск гипотетических отношений включения. Необходимо задать поисковой системе пороговые значения индикаторов, при достижении которых пара (МПК, ГРНТИ) выдаётся на экспертизу.

В общем случае необходимо выбрать три пороговых значения: для индикаторов ω , α , β . В нашем случае было решено использовать одинаковое пороговое значение для индикаторов α и β и поэтому достаточно было выбрать два пороговых значения.

Увеличение пороговых значений ведёт к возрастанию точности поиска и падению полноты поиска. Соответственно, уменьшение пороговых значений ведёт к росту полноты поиска, т.е. будет найдено больше гипотетических отношений включения. Однако точность поиска при этом падает, так как уменьшается доля истинных отношений включения во множестве найденных гипотетических отношений.

Рис. 1 иллюстрирует описанную выше зависимость количества найденных пар (МПК-3, ГРНТИ-3) от пороговых значений индикаторов ω , α , β .

Экспериментально были подобраны пороговые значения $\alpha \geq 0,7$, $\beta \geq 0,7$, $\omega \geq 5$, при которых найдены 324 гипотетических отношения включения между тематикой рубрик МПК-3 и ГРНТИ-3. Из них 291 – включения ГРНТИ-3 в МПК-3 и 41 – включения МПК-3 в

ГРНТИ-3. Такой результат поиска убедительно показал, что рубрики МПК-3 имеют в среднем более широкий тематический охват, чем рубрики ГРНТИ-3. Иными словами, коды МПК-3 примерно в семь раз чаще «охватывают» коды ГРНТИ-3, чем коды ГРНТИ-3 «охватывают» коды МПК-3. Впрочем, этого следовало ожидать, так как количество кодов ГРНТИ-3 более чем в 10 раз превышает количество кодов МПК-3. У восьми найден-

ных пар, представленных в таблице, обе меры включения превысили 70%.

Несмотря на то, что в найденных гипотетических отношениях включения участвуют только 153 уникальных кода МПК-3 и 307 уникальных кодов ГРНТИ-3, указанные коды МПК-3 охватывают 66% патентных документов, которые ВИНТИ РАН и отражает в своих реферативных базах данных.

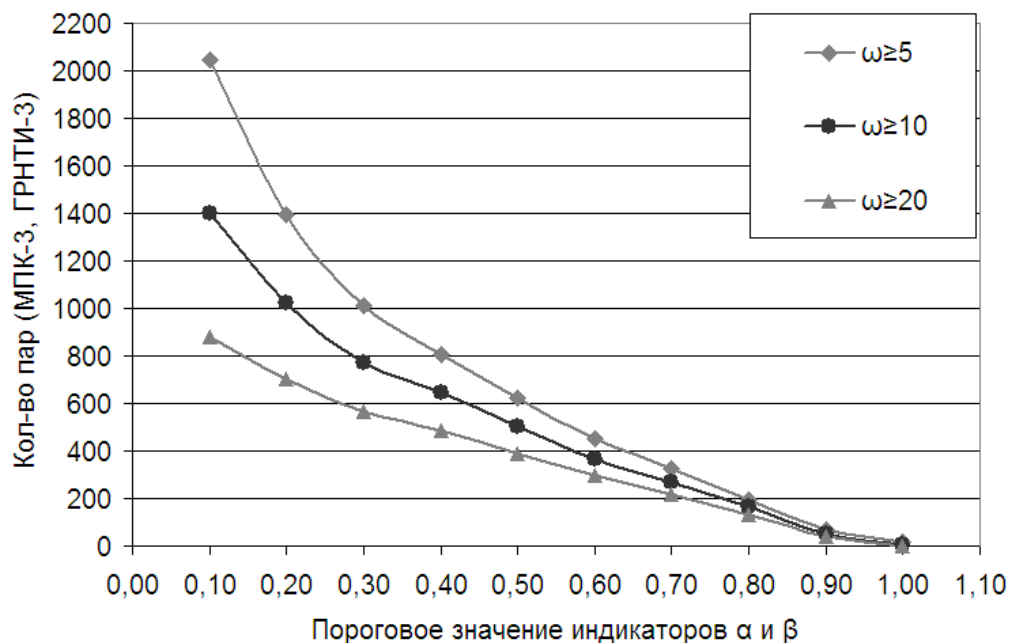


Рис. 1. Зависимость количества включений от пороговых значений

Отношения взаимного включения

МПК	Название в МПК	ГРНТИ	Название в ГРНТИ	α	ω	β
B41F	Печатные машины	60.29.17	Печатные процессы и оборудование	0,943	182	0,795
B66B	Подъемники; эскалаторы или движущиеся дорожки	55.51.29	Лифты и подъемники периодического действия	0,762	240	0,825
C07D	Гетероциклические соединения	61.45.31	Органические синтетические лекарственные вещества	0,707	1534	0,722
C12C	Пивоварение	65.43.01	65.43 = Пивоваренная промышленность 65.43.01 = Общие вопросы	0,875	210	0,995
C14C	Обработка голья, шкур или кож химическими средствами, энзимами или микроорганизмами, например дубление, пропитка или отделка; устройства для этого; составы для дубления	64.35.13	Общая технология кожевенной промышленности	0,727	24	0,706
D05B	Шитье	55.59.35	Машины и оборудование для швейной промышленности	0,816	62	0,729
D06F	Стирка; сушка; глажение; прессование или складывание текстильных изделий	55.67.33	Машины и приборы для обработки белья, ухода за одеждой и обувью	0,772	436	0,906
G06E	Оптические вычислительные устройства	50.33.41	Оптические ЭВМ	0,735	25	0,781

Если поставить задачу найти не только включения, но и сильные пересечения между тематиками рубрик, то необходимо снизить пороговые значения мер включения. Например, при пороговых значениях $\alpha=0,5$, $\beta=0,5$, $\omega=5$ было найдено 625 отношений между рубриками. В этих отношениях участвовали 271 уникальный код МПК-3 и 554 уникальных кода ГРНТИ-3, при этом тематический охват потока патентов кодами МПК-3 составил 87,6%.

На рис. 2 представлен фрагмент таблицы включений, выдаваемой на экспертизу. Эксперты могут просматривать описания включений в табличной форме, как это видно на рисунке, или в экранной форме, которая показывает описание одного включения, но в более полном виде.

Задача эксперта – ввести в таблицу признак, обозначающий наличие или отсутствие отношения, а также вид отношения между рубриками. Эксперт может вводить один из пяти символов, которые обозначают следующие отношения:

- < – включение тематики рубрики МПК в рубрику ГРНТИ;
- > – включение тематики рубрики ГРНТИ в рубрику МПК;
- = – равенство тематики рубрик МПК и ГРНТИ;
- X – пересечение тематик рубрик МПК и ГРНТИ;

0 – отсутствие каких-либо отношений между тематикой рубрик МПК и ГРНТИ.

ПЕРСПЕКТИВЫ

Разработанный метод предназначен для автоматизации поиска отношений между классификаторами на основании выборочного массива документов, оснащённых кодами обоих классификаторов. К сожалению, такие массивы документов удаётся получить не для всех систем классификации. Значительно чаще встречается другая ситуация: в распоряжении исследователей оказывается два массива документов, индексированных кодами двух различных систем классификации, а также ключевыми словами и словосочетаниями (КС). В этом случае для каждого класса (по возможности) создаётся нечёткое множество КС, характеризующих тематику класса. Затем выполняется сравнение классов посредством сравнения множеств КС, присвоенных классам.

В ВИНТИ в течение ряда лет ведутся работы по созданию массива ключевых слов и словосочетаний по тематике ВИНТИ. В основу положены ключевые слова, составляющие поисковые образы документов. На основе созданной статистической базы терминов составляются списки ключевых слов для выпусков Реферативных журналов (РЖ), разделов выпусков РЖ, разделов рубрикатора ВИНТИ [5].

КодМПК	НазваниеМПК	КодГРНТИ	НазваниеГРНТИ	МПКвГРНТИ	МощнПерес	ГРНТИвМПК
G01S	Радиопеленгация; радионавигац	47.49.29	Радиолокационные системы, станции	0,217	110	0,821
G01S	Радиопеленгация; радионавигац	47.49.41	Акустические системы локации	0,039	20	0,741
G01S	Радиопеленгация; радионавигац	47.49.43	Оптические системы локации	0,028	14	0,824
G01T	Измерение ядерных излучений и	90.27.38	Измерения ионизирующих излучений и	0,513	40	0,741
G01V	Геофизика; гравитационные изме	38.57.23	Разведочная геофизика	0,399	145	0,833
G01V	Геофизика; гравитационные изме	52.13.03	Рудничная геология	0,058	21	0,84
G01V	Геофизика; гравитационные изме	38.57.15	Геологические и геохимические методь	0,039	14	1
G05B	Регулирующие и управляющие с	50.43.17	Системы автоматического управления,	0,042	11	0,786
G06E	Оптические вычислительные ус	50.33.41	Оптические ЭВМ	0,735	25	0,781
G06F	Обработка цифровых данных с п	50.41.15	Операционные системы	0,089	127	0,784
G06F	Обработка цифровых данных с п	50.09.31	Арифметические и логические устройс	0,051	73	0,88
G06F	Обработка цифровых данных с п	50.33.39	Процессоры	0,027	39	0,886
G06F	Обработка цифровых данных с п	50.41.23	Программное обеспечение вычислитель	0,024	34	0,829
G06F	Обработка цифровых данных с п	50.09.53	Интерфейсное оборудование	0,021	30	0,732
G06F	Обработка цифровых данных с п	50.07.05	Теория вычислительных систем высок	0,017	24	0,857
G06F	Обработка цифровых данных с п	50.07.03	Теория и моделирование вычислитель	0,016	23	0,852
G06F	Обработка цифровых данных с п	49.38.49	Связь по глобальным компьютерным с	0,013	19	0,731
G06F	Обработка цифровых данных с п	50.09.33	Микропроцессоры	0,012	17	0,85
G06F	Обработка цифровых данных с п	50.05.13	Технология программирования. Автома	0,011	16	1
G06F	Обработка цифровых данных с п	50.05.17	Теоретические основы системного прог	0,011	16	0,941
G08B	Сигнальные устройства или уст	81.92.86	Пожарная сигнализация	0,136	22	0,846
G11B	Накопление информации, основ	47.53.31	Оптические системы записи и воспроиз	0,586	99	0,952
G11B	Накопление информации, основ	50.11.33	Оптические запоминающие устройства	0,308	52	0,981
G11B	Накопление информации, основ	47.53.29	Магнитные системы записи и воспроиз	0,077	13	0,765
G21C	Ядерные реакторы (термоядерн	44.33.29	Атомные электростанции	0,893	158	0,532
H01B	Кабели; проводники; изоляторы;	45.47.33	Силовые кабели	0,061	10	0,909
H01B	Кабели; проводники; изоляторы;	45.47.35	Кабели связи	0,061	10	0,833
H01F	Магниты; индуктивности; трансф	45.33.29	Трансформаторы	0,25	56	0,727
H01F	Магниты; индуктивности; трансф	45.33.31	Электрические реакторы	0,054	12	0,857
H01J	Электрические газоразрядные и	47.03.03	Теоретические основы электровакуумн	0,08	15	0,789
H01J	Электрические газоразрядные и	47.29.37	Электровакуумные приборы сверхвысо	0,053	10	0,833
H01L	Полупроводниковые приборы; эл	47.33.33	Оптоэлектронные приборы	0,103	57	0,905

Рис. 2. Фрагмент списка найденных отношений включения

До сих пор тематические списки КС создавались для достаточно крупных предметных областей, при этом предварительные списки, полученные из статистической базы данных, модифицировались и утверждались экспертами. Главной целью создания подобных списков было повышение качества индексирования документов, а главными пользователями списков были специалисты ВИНТИ, выполняющие индексирование.

Индексирование ключевыми терминами документа для его последующего поиска и индексирование класса для его последующего сопоставления с другими классами – разные задачи, требующие различного подхода к отбору КС. Операция индексирования документа с контролем по тематическому тезаурусу требует, чтобы тезаурус содержал все термины, которые допустимо использовать при создании поискового образа документа в данной предметной области, в том числе значительное количество терминов, употребление которых рассеяно по другим предметным областям. Операция автоматического сопоставления классов требует, чтобы классу были присвоены избирательные термины, характеризующие предметную область именно этого класса. Иными словами, предпочтение следует отдавать таким специальным терминам, которые часто встречаются в документах данного класса, но редко встречаются в документах, относящихся к другим классам. Такой подход позволяет отсеять неизбирательные термины и повысить точность автоматического сопоставления классов.

Поскольку количество классов в развитых системах классификации исчисляется сотнями и даже тысячами, то присвоение им ключевых терминов силами экспертов вряд ли возможно по причине огромной трудоёмкости этой работы. Здесь на помощь может прийти компьютерный отбор КС на основании частоты их использования в документальных базах данных. Компьютерной программе, выполняющей отбор КС, необходима характеристическая функция, позволяющая численно определить включение тематики КС в тематику класса, т.е. дать числовую оценку отношению $КС \subseteq КЛАСС$. Один из подходов – использовать «меру включения термина в тематику класса», которая вычисляется аналогично описанным выше мерам включения, измеряющим отношения $МПК \subseteq ГРНТИ$ и $ГРНТИ \subseteq МПК$. Как было отмечено ранее, меры включения являются относительными частотами. В данном случае «мера включения КС в КЛАСС» – это статистическая оценка вероятности события «случайно выбранный документ, имеющий КС в поисковом образе, относится к тематике данного класса». Дальнейшее сравнение множеств КС, присвоенных классам, также предполагается выполнять автоматически.

ЗАКЛЮЧЕНИЕ

Полученные результаты позволили в значительной степени автоматизировать создание таблицы соответствий между двумя системами классификации (МПК и ГРНТИ) благодаря наличию в ВИНТИ

РАН массива документов, индексированных параллельно кодами обеих систем классификации.

Применяемые в работе индикаторы – мощность пересечения, меры включения – могут служить основой для отбора и присвоения классам ключевых терминов, характеризующих тематику класса. Это позволит реализовать на практике другой подход к сопоставлению классификаторов: сравнение классов посредством сравнения множеств терминов, присвоенных классам. Такой подход более универсальный, так как не требует документов, индексированных параллельно кодами различных систем классификации. Достаточно для каждого классификатора иметь единственный массив описаний документов, включающий в себя классификационные коды и ключевые слова.

СПИСОК ЛИТЕРАТУРЫ

1. Никольская И. Ю., Антошкова О. А., Белозеров В. Н., Гоннова С. М. Концепция разработки системы соответствий между классификациями научно-технической информации // Восемнадцатая международная конференция и выставка "LIBCOM-2014" «Информационные технологии, компьютерные системы и издательская продукция для библиотек», г. Суздаль 10–14 ноября 2014 г. – URL: <http://www.gpntb.ru/libcom14/tezis/004.pdf>.
2. Страница «Государственный рубрикатор НТИ» на сайте ВИНТИ РАН. – URL: http://www2.viniti.ru/index.php?option=com_content&view=article&id=39:rubrikator-nti.
3. International Patent Classification (IPC). – URL: <http://www.wipo.int/classifications/ipc/en/>.
4. Международная патентная классификация 2015, полная схема в формате XML. – 2015. – URL: http://www1.fips.ru/wps/wcm/connect/content_ru/ru/inform_resources/international_classification/inventions.
5. Малинина К. О., Шапкин А. В. Массив ключевых терминов как часть информационного обеспечения аналитико-синтетической обработки научно-технической литературы: механизмы формирования, источники данных, установление связей между терминами // НТИ-2012. Актуальные проблемы информационного обеспечения науки, аналитической и инновационной деятельности. Материалы VIII Международной конференции. – М.: ВИНТИ, 2012. – С. 127–128.

Материал поступил в редакцию 14.04.15.

Сведения об авторе

ФЕДОРЕЦ Олег Владимирович – кандидат технических наук, старший научный сотрудник Отдела программных систем ВИНТИ, Москва.
e-mail: ovf@viniti.ru