РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2014

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.8 : 510.6

В.В. Подиновский, О.В. Подиновская

Подход теории важности критериев к задачам принятия решений с иерархической критериальной структурой*

Предлагается новая иерархическая модель ситуации принятия решений при многих критериях, образующих многоуровневую систему. Эта модель позволяет применять подходы и методы теории важности критериев для сбора информации о важности критериев и корректного анализа практических многокритериальных задач принятия решений с её использованием.

Ключевые слова: *многокритериальные задачи принятия решений, иерархическая система критериев, шкалы критериев, информация о важности критериев, теория важности критериев*

ВВЕДЕНИЕ

Сложные многокритериальные задачи принятия решений часто формализуются с использованием иерархических структур. Для анализа таких задач в 1980 г. американский ученый Томас Саати (Thomas Saaty) в работе [1] представил разработанный им метод “The analytic hierarchy process” (сокращенно *AHP*, употребляемое нами в статье). Переводчик не точно передал английское название метода, введя наименование «Метод анализа иерархий» [2], однако оно и его сокращение – МАИ – стали у нас общеупотребительными.

За прошедшие десятилетия *AHP* стал одним из самых известных и распространенных методов решения практических многокритериальных задач самого различного характера и сложности. Ему и его приложениям посвящено множество публикаций – обзо-

* Исследование осуществлено в рамках Программы «Научный фонд НИУ ВШЭ» в 2013 г. (проект № 12-01-0059).

ров, монографий, научных статей, а также работ, популяризирующих этот метод. Это объясняется как несомненными привлекательными для его пользователей достоинствами, присущими самому методу, так и наличием отработанных коммерческих компьютерных систем поддержки принятия решений, реализующих *AHP*, а также необыкновенной активностью и удачливостью его автора.

К сожалению, этот, по сути эвристический метод, обладает и рядом принципиальных, причем неустранимых недостатков. Это отмечалось в целом ряде научных работ, в том числе в [3-5]. К основным недостаткам относятся независимость процедур оценивания важности критериев и нормализации критериальных оценок вариантов решений (что нарушает требование математической теории многомерных измерений) и отсутствие точного определения понятия важности критериев.

Однако, в свете современных исследований по теории измерений возможность оценивания предпочтений в шкале отношений (на что претендует *AHP*) исключается [6].

Поэтому в [7] проблема разработки корректного метода анализа многокритериальных задач с иерархической критериальной структурой была указана как одна из актуальных проблем развития теории важности критериев (ТВК), основанной на строгих определениях качественной и количественной их важности.

В настоящей статье предлагается новая математическая модель иерархической многокритериальной структуры.

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПРОБЛЕМНОЙ СИТУАЦИИ

Дальнейшее изложение в статье опирается на следующую математическую модель ситуации принятия индивидуального решения в условиях определенности, принятую в теории важности критериев [8, 9]:

$$\mathbb{M} = < X, f, Z_0, R >,$$

где X – множество вариантов, $f = (f_1, \ldots, f_m)$ – векторный критерий ($m \geq 2$), f_i – частные критерии, Z_0 – общая шкала частных критериев, R – отношение нестрогого предпочтения лица, принимающего решение (ЛПР). Под критерием f_i понимается функция с областью определения X и областью значений Z_0. Будем полагать, что бóльшие значения критерия предпочтительнее меньших. Таким образом, каждый вариант x характеризуется m числами – значениями $f_i(x)$ всех критериев, образующими векторную оценку этого варианта $y = f(x) = (f_1(x), \ldots, f_m(x))$. Сравнение вариантов по предпочтительности сводится к сопоставлению их векторных оценок. Множество всех векторных оценок (как реальных, т.е. соответствующих вариантам из X, так и гипотетических) есть $Z = Z_0^m$. Запись yRz означает, что векторная оценка y не менее предпочтительна, чем z. Отношение R порождает отношение безразличия I и (строгого) предпочтения P:

$$yIz \Leftrightarrow yRz \wedge zRy; yPz \Leftrightarrow yRz \wedge z\not{R}y$$

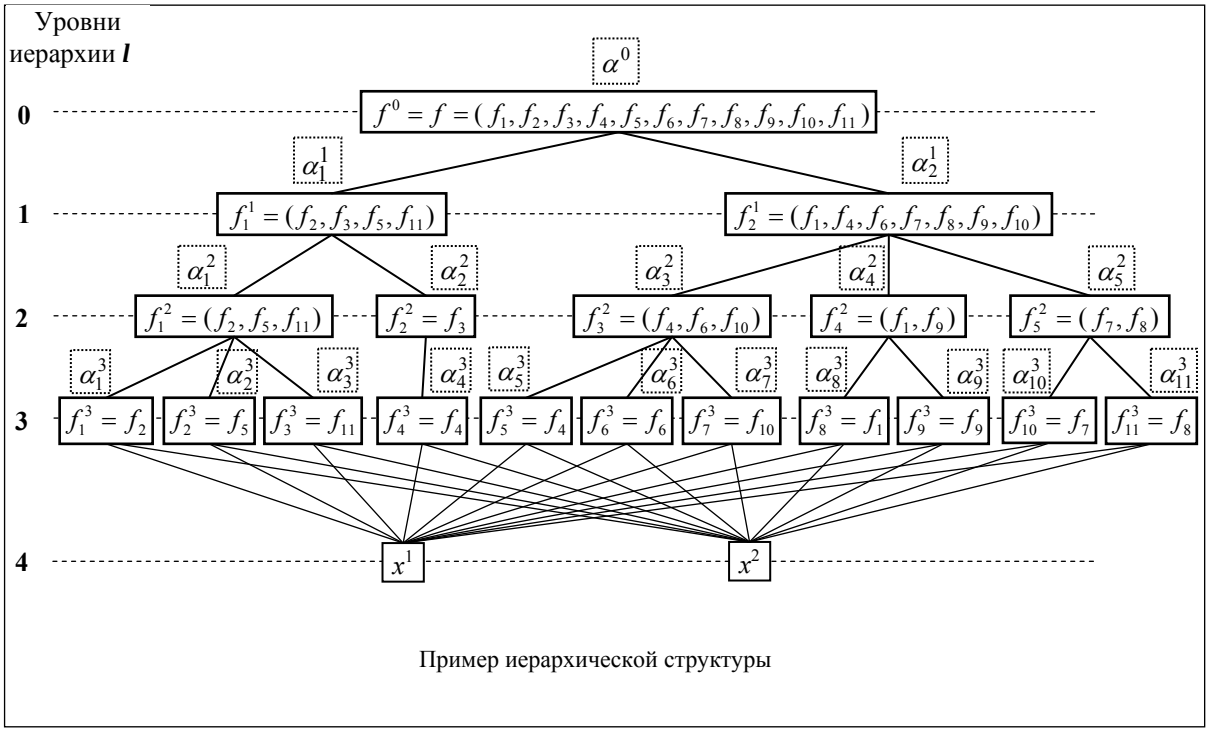
(запись $z\not{R}y$ означает, что zRy – неверно). Предполагается, что отношение R является частичным квазипорядком – оно рефлексивно и транзитивно.

Предположим, что из векторного критерия f, который будем называть критерием верхнего, или нулевого уровня, и обозначать также f^0, путем «разрезания и склеивания» получены векторные критерии первого уровня $f_1^1, \ldots, f_{m_1}^1$, в состав которых входят все m частных критериев, причем каждый их них участвует в формировании одного и только одного критерия первого уровня. Эти критерии называются детьми критерия f^0, а f^0 – их родителем. Критерии второго уровня $f_1^2, \ldots, f_{m_2}^2$ получаются аналогичным образом из критериев первого уровня; в их состав также входят все m частных критериев, причем каждый их них участвует в формировании только одного критерия второго уровня. Критерии второго уровня, полученные из некоторого критерия первого уровня, называются его детьми, а сам этот критерий – их родителем.

И так далее, вплоть до критериев f_i^L уровня L, каждый из которых является одним из частных критериев. На уровне $L + 1$ располагаются варианты, каждый из которых оценивается по всем критериям уровня L. Критериальную систему, состоящую из взаимосвязанных указанным выше образом критериев уровней 0, 1,…, L, будем называть *многоуровневой*, или *иерархической*. В качестве примера на рисунке представлена иерархическая структура для задачи с 11-ю критериями и двумя вариантами (здесь $L = 3$).

При решении практических задач иерархическая критериальная система может формироваться как «сверху вниз», так и «снизу верх», но так, чтобы все критерии всех уровней имели содержательный смысл, понятный ЛПР.

Многоуровневая критериальная система, входящая в построенную иерархическую модель, является базовой для развития ТВК в указанном выше направлении. Подчеркнем, что она принципиально отличается от принятой в *AHP* иерархической системы критериев, так как в *AHP* (как и во всех других ранее предлагавшихся методах анализа иерархических систем) понятие критерия, не находящегося на нижнем уровне, не структурируется, и поэтому, например, непонятно, какие значения он может принимать. Более того, предложенная иерархическая критерильная система позволяет использовать разработанные в ТВК строгие определения важности критериев и развивать основанные на них корректные методы анализа сложных многокритериальных задач.



Пример иерархической структуры

2. ИНФОРМАЦИЯ О ВАЖНОСТИ КРИТЕРИЕВ

Исходная критериальная структура определяет на множестве векторных оценок $Z = Z_0^m$ отношение Парето – отношение строгого предпочтения $P^{\varnothing}$ следующим образом:

$$yP^{\varnothing}z \Leftrightarrow (y_i \geq z_i,\ i = 1,\ldots,\ m,\ y \neq z).$$

Для расширения этого отношения требуется дополнительная информация о предпочтениях ЛПР – сведения о важности критериев и характере изменений предпочтений вдоль шкалы критериев. В иерархической критериальной структуре сравниваются по важности критерии только одного и того же уровня, причем имеющие общего родителя. Для представления результатов такого сравнения введем *коэффициенты важности критериев*: важность критерия f_j^l оценивается положительным числом α_j^l (см. рисунок).

В качестве информации, получаемой от ЛПР, могут выступать сведения о важности критериев и о скорости роста предпочтений вдоль их шкалы [8]. Информация о важности состоит из сообщений, каждое из которых касается результатов сравнения по важности критериев одного уровня (фактически – групп критериев) и имеет точный смысл, определяемый в ТВК. Так, для случая, когда информация о важности критериев – *качественная* (порядковая), исходными являются приведенные ниже два определения [9].

Пусть $A, B \in \{1, \ldots, m\}$ – два непустых непересекающихся набора номеров критериев, которым соответствуют группы критериев $\{f_i\}_{i\in A}$ и $\{f_i\}_{i\in B}$. Обозначим через y^{AB} векторную оценку, полученную из векторной оценки y, в которой все компоненты y_i с номерами из A равны между собой и все компоненты y_i с номерами из B равны между собой (т.е. $y_i = y_j$ при $i, j \in A$ и при $i, j \in B$), заменой каждой компоненты y_j, $j \in B$, на любую компоненту y_i, $i \in A$, и заменой каждой компоненты y_i, $i \in A$, на любую компоненту y_j, $j \in B$. Например, если y = (2, 4, 4, 3, 2, 2), A = {1, 5, 6}, B = {2, 3}, то y^{AB} = (4, 2, 2, 3, 4, 4).

Определение 1. Группы критериев $\{f_i\}_{i\in A}$ и $\{f_i\}_{i\in B}$ *равноважны,* или *одинаково важны*, если векторные оценки y и y^{AB} одинаковы по предпочтительности.

Определение 2. Группа критериев $\{f_i\}_{i\in A}$ *важнее* группы критериев $\{f_i\}_{i\in B}$, если из пары векторных оценок y и y^{AB} первая предпочтительнее второй, когда компонента y_i с номером i из A больше компоненты y_i с номером i из B.

Отметим, что, если множества A и B одноэлементны, то эти определения говорят об упорядочении по важности отдельных критериев [10].

В ТВК разработаны способы получения и анализа качественной информации о важности критериев, основанные на этих определениях, и методы использования такой информации для решения многокритериальных задач. Для представления и анализа информации о важности групп критериев введем (*множественные*) *коэффициенты важности*, т.е. важность группы критериев $A = \{f_i\}_{i\in A}$ будем оценивать положительным числом α_A. Эти коэффициенты обладают следующим ординальным свойством: если $A \supset B$, то $\alpha_A > \alpha_B$. Но не предполагается наличие свойства аддитивности, т.е. не требуется, чтобы для непересекающихся групп критериев $\{f_i\}_{i\in A}$ и $\{f_i\}_{i\in B}$ (т.е. при $A \cap B = \varnothing$) выполнялось равенство $\alpha_{A\cup B} = \alpha_A + \alpha_B$.

В задачах с иерархической критериальной структурой под критериями f_j^l можно понимать также группы частных критериев, входящих в их состав, и

тогда α_j^l – (множественные) коэффициенты важности этих групп.

При анализе практических задач с иерархической структурой вопросы ЛПР можно задавать о сравнении по важности критериев одного и того же уровня с общим родителем. При этом таким критериям можно приписывать значения шкалы частных критериев следующим образом: считать, что значение критерия f_j^l соответствует градации $s \in Z$, когда значения всех входящих в его состав частных критериев равно s. Например, если шкала Z является лингвистической и оценка s – это «хорошо», то при выполнении указанного условия критерию-родителю можно приписать «общую» («итоговую») оценку «хорошо».

В результате проведенных сравнений критериев по важности для каждых двух критериев f_j^l и f_t^l с общим родителем будет установлена справедливость одного из трех соотношений: критерий f_j^l важнее критерия f_t^l (так что $\alpha_j^l > \alpha_t^l$), критерий f_t^l важнее критерия f_j^l (и поэтому $\alpha_j^l < \alpha_t^l$) или что критерии f_t^l и f_j^l равноважны (и тогда $\alpha_j^l = \alpha_t^l$). Так, для критериальной структуры (уровня 0-1), показанной на рис. 1, может быть установлено, например, что:

для уровня $l = 1$: критерий f_2^1 важнее критерия f_1^1;

для уровня $l = 2$: критерий f_1^2 важнее критерия f_2^2; критерии f_3^2 и f_4^2 –равноважны и каждый из них менее важен, чем f_5^2;

для уровня $l = 3$: критерии f_1^3 и f_2^3 – равноважны и каждый из них менее важен, чем f_3^3; критерий f_6^3 важнее каждого из равноважных критериев f_5^3 и f_7^3; критерии f_8^3 и f_9^3 – равноважны; критерий f_{10}^3 важнее критерия f_{11}^3.

Эта информация представляется такими соотношениями между коэффициентами важности критериев:

$l = 1$: $\alpha_1^1 < \alpha_2^1$;

$l = 2$: $\alpha_1^2 > \alpha_2^2$, $\alpha_3^2 = \alpha_4^2 < \alpha_5^2$;

$l = 3$: $\alpha_1^3 = \alpha_2^3 < \alpha_3^3$, $\alpha_6^3 > \alpha_5^3 = \alpha_7^3$, $\alpha_8^3 = \alpha_9^3$, $\alpha_{10}^3 > \alpha_{11}^3$.

Для получения и использования количественной информации о важности критериев следует, аналогично [11], ввести в рассмотрение N-модель, но теперь уже тоже иерархическую.

3. ОТНОШЕНИЯ ПРЕДПОЧТЕНИЯ И БЕЗРАЗЛИЧИЯ

Каждое из сообщений о важности критериев или характере роста предпочтений вдоль их шкалы задает на множестве векторных оценок Z отношение предпочтения или безразличия. На основе этой информации Г на Z определяется отношение нестрогого предпочтения R^Γ как квазитранзитивное замыкание объединения всех этих отношений и отношения Парето. Отношение R^Γ индуцирует аналогичное по смыслу отношение R_Γ на множестве вариантов:

$$x' R_\Gamma x'' \Leftrightarrow f(x') R^\Gamma f(x'').$$

Оно непосредственно используется для решения поставленной задачи. Например, один наилучший вариант следует выбрать из множества вариантов, недоминируемых по отношению строгого предпочтения P_Γ, которое порождается на X отношением R_Γ.

Для построения отношения R^Γ можно, в принципе, воспользоваться общим методом из [12]. К сожалению, уже при небольшой размерности задачи указанный метод оказывается весьма обременительным даже для использования современной вычислительной техники. Поэтому актуальной является проблема разработки эффективных методов сравнения вариантов по предпочтительности, не ограничивающих допустимую размерность решаемых практических задач.

ЗАКЛЮЧЕНИЕ

В статье предложена новая модель многоуровневой критериальной структуры. Эта модель позволяет использовать подходы и методы теории важности критериев (ТВК) для анализа практических многокритериальных задач принятия решений.

Основными направлениями дальнейшего развития ТВК применительно к многокритериальным задачам принятия решений с иерархической структурой, по мнению авторов статьи, являются:

- разработка аналитических и эффективных алгоритмических решающих правил, обобщающих созданные ранее в ТВК для задач с одноуровневой критериальной системой, том числе из [9–11; 13–16], для различных сочетаний видов информации о важности критериев и изменении предпочтений вдоль их шкалы;
- разработка методов объяснения результатов, получаемых при анализе задач, причем понятных для ЛПР (эти методы должны предусматривать, в частности, построение объясняющих цепочек [8]).
- разработка методов оценки чувствительности (устойчивости) получаемых решений (к изменению не только одного, но и нескольких параметров модели предпочтений), опирающихся на подход из [17, 18].
- Развитие основанных на результатах из [19–21] теории и методов ТВК для задач принятия решений при риске и неопределенности.
- Развитие теории и методов ТВК для иерархической модели с размытыми (нечеткими – fussy) критериальными оценками, оценками важности критериев и отношениями предпочтения.

Решение указанных проблем позволит создать новые компьютерные информационно-аналитические системы поддержки принятия решений для задач с иерархической структурой.

СПИСОК ЛИТЕРАТУРЫ

1. Saaty T.L. The analytic hierarchy process. – New York: McGraw Hill, 1980. – 287 с.
2. Саати Т. Принятие решений. Метод анализа иерархий: пер. с англ. – М.: Радио и связь, 1993. – 278 с.
3. Belton V., Stewart T.J. Multiple criteria decision analysis. An integrated approach. – Boston: Cluwer, 2003. – 374 с.
4. Подиновская О.В. Метод анализа иерархий как метод поддержки принятия многокритериальных решений // Информационные технологии моделирования и управления. –2010. – № 1 (60). – С. 71-80.
5. Подиновский В.В., Подиновская О.В. Еще раз о некорректности метода анализа иерархий // Проблемы управления. – 2012. – № 4. – С. 75-78.
6. Barzilai J. Preference function modelling: The mathematical foundations of decision theory // Trends in multiple criteria decision analysis / eds. M. Ehrgott, J.R. Figueira, S. Greco. – New York: Springer, 2010. – P. 57-86.
7. Подиновский В.В., Потапов М.А. Важность критериев в многокритериальных задачах принятия решений: теория, методы, софт и приложения // Открытое образование. – 2012. – № 2. – С. 55-61.
8. Подиновский В.В. Введение в теорию важности критериев в многокритериальных задачах принятия решений: Учебное пособие. – М.: Физматлит, 2007. – 64 с.
9. Подиновский В.В. Аксиоматическое решение проблемы оценки важности критериев в многокритериальных задачах принятия решений // Современное состояние теории исследования операций / под ред. Н.Н. Моисеева. – М.: Наука, 1979. – С. 117-145.
10. Подиновский В.В. Коэффициенты важности критериев в задачах принятия решений. Порядковые, или ординальные коэффициенты важности // Автоматика и телемеханика. – 1978. – № 10. – С. 130-141; Podinovskii V.V. Importance coefficients of criteria in decision making problems. Serial, or ordinal importance coefficients // Automation and remote control. – 1978. – Vol. 39. – P. 1514-1524.
11. Podinovski V.V. The quantitative importance of criteria for MCDA // Journal of multi-criteria decision analysis. – 2002. – Vol. 11. – P. 1-15.
12. Осипова В.А., Подиновский В.В., Яшина Н.П. О непротиворечивом расширении отношений предпочтения в задачах принятия решений // Журнал вычислительной математики и математической физики. – 1984. – № 6. – С. 831-840; Osipova V.A., Podinovskii V.V., Yashina N.P. On non-contradictory extension of preference relations in decision making problems // USSR Computational Mathematics and Mathematical Physics. – 1984. – Vol. 24, № 3. – P. 128-134.
13. Podinovski V.V. On the use of importance information in MCDA problems with criteria measured on the first ordered metric scale // Journal of multi-criteria decision analysis. – 2009. – Vol. 15. – P. 163-174.
14. Нелюбин А.П., Подиновский В.В. Алгоритмическое решающее правило, использующее ординальные коэффициенты важности критериев со шкалой первой порядковой метрики // Журнал вычислительной математики и математической физики. – 2012. – Т. 52. – № 1. – С. 43-59; Nelyubin A.P., Podinovski V.V. Algorithmic decision rule using ordinal criteria importance coefficients with a first ordinal metric scale // Computational Mathematics and Mathematical Physics. – 2012. – Vol. 52, № 1. – P. 48-65.
15. Нелюбин А.П., Подиновский В.В. Аналитические решающие правила, использующие упорядоченность по важности критериев со шкалой первой порядковой метрики // Автоматика и телемеханика. – 2012. – № 5. – С. 84-96; Nelyubin A.P., Podinovski V.V. Analytical decision rules using importance-ordered criteria with a scale of the first ordinal metric // Automation and Remote Control. – 2012. – Vol. 73, № 5. – P. 831-840.
16. Подиновский В.В., Подиновская О.В. Новые многокритериальные решающие правила в теории важности критериев // Доклады академии наук. – 2013. – Т. 451. – № 1. – С. 604 – 606; Podinovski V.V., Podinovskaya O.V. New multicriterial decision rules in criteria importance theory // Doklady Mathematics. – 2013. – Vol. 88, № 1. – P. 486-488.
17. Podinovski V.V. Sensitivity analysis for choice problems with partial preference relations // European journal of operational research. – 2012. – Vol. 221. – P. 198-204.
18. Подиновский В.В. Анализ устойчивости результатов выбора при частичном отношении предпочтения // Искусственный интеллект и принятие решений. – 2009. – № 4. – С. 45 – 52; Podinovski V.V. Analyzing the stability of choice for the partial preference relation // Scientific and technical information processing. – 2011. – Vol. 38, № 5. – P. 1-7.
19. Подиновский В.В. Теория важности критериев в многокритериальных задачах принятия решений при неопределенности. I. Исходные положения // Информационные технологии моделирования и управления. – 2010. – № 5 (64). – С. 599-607.
20. Подиновский В.В. Теория важности критериев в многокритериальных задачах принятия решений при неопределенности. II. Задачи с количественной информацией о важности критериев и вероятностях значений неопределенного фактора //

Информационные технологии моделирования и управления. – 2012. – № 2 (74). – С. 131-137.

21. Подиновский В.В. Теория важности критериев в многокритериальных задачах принятия решений при неопределенности. III. Задачи с качественной информацией о важности критериев и количественной информацией о вероятностях значений неопределенного фактора // Информационные технологии моделирования и управления. – 2012. – № 3 (75). – С. 186-193.

Материал поступил в редакцию 30.08.13.

Сведения об авторах

ПОДИНОВСКИЙ Владислав Владимирович - доктор технических наук, профессор, Национальный исследовательский университет «Высшая школа экономики», Москва
e-mail: podinovski@mail.ru

ПОДИНОВСКАЯ Ольга Владиславна – младший специалист, информационно-аналитический отдел, Мак-Кинзи и Компания СиАйЭс, Москва
e-mail: podinovskaya@mail.ru

УДК 004.89 : 510.6

И.Б. Яшков

Отбор признаков с помощью деревьев решений в задаче ДСМ-классификации

Рассматривается метод отбора признаков для обучения алгоритма классификации на основе ДСМ-метода. Метод основан на применении деревьев решений, включает в себя построение максимального дерева с помощью алгоритма C4.5 и получение ряда усеченных деревьев на основе критерия минимальной цены-сложности. Сравниваются результаты и параметры работы ДСМ-классификатора при различном количестве отобранных признаков.

Ключевые слова: *машинное обучение, сокращение размерности, отбор параметров, ДСМ-метод, классификация, деревья решений, энтропия*

Задача машинного обучения состоит в поиске и генерации новых знаний о событиях и феноменах на основе существующих данных, чтобы можно было классифицировать или предсказать будущие события. Массивы данных состоят из множества векторов, каждый из которых соответствует некоторому событию: каждый вектор состоит из некоторого числа признаков (или переменных). В общем случае неизвестно, какие из этих признаков имеют значение для предсказания.

Отбор значимых признаков для машинного обучения актуален в областях и задачах с массивами данных, содержащими десятки, сотни или даже тысячи осей в признаковом пространстве. Это такие области как обработка текстов, биоинформатика и комбинаторная химия. Отбор признаков преследует три цели:

1. Улучшить качество классификации.
2. Ускорить процесс обучения.
3. Повысить понятность сгенерированных правил для экспертов.

Еще в конце XX в. проблема сокращения размерности в практических задачах стояла не так остро, так как редко можно было встретить задачу, в которой объекты содержали бы более 40 признаков [1]. Однако к середине 2000-х гг. ситуация кардинально изменилась, стали появляться области, в которых алгоритмам машинного обучения приходилось работать с сотнями, тысячами и десятками тысяч признаков. При этом большинство из них зачастую являлись иррелевантными или избыточными.

Предсказательная сила алгоритма напрямую зависит от того, насколько эффективно он умеет распознавать закономерности, содержащиеся в данных. Иррелевантные и избыточные признаки увеличивают размерность пространства поиска, в результате поиск закономерностей затрудняется (это утверждение справедливо не только для машинного обучения, но и для экспертов-людей). Кроме того, чем больше признаков у объектов, тем выше риск переобучения алгоритма. Вероятность того, что некоторые признаки случайно проявят закономерность увеличивается, если объем выборки не растет экспоненциально вместе с ростом вектора признаков. Более того, в большинстве практических задач стоит цель узнать те ключевые признаки, которые оказывают наибольшее влияние на классификацию.

Таким образом, применение методов сокращения размерности, и в частности, отбора признаков имеет множество потенциальных преимуществ для достижения следующих целей:

– упрощение представления и повышение читабельности данных,

– сокращение объемов данных,

– сокращение времени обработки данных и обучения,

– устранение проблем, связанных с так называемым «проклятием размерности» [2].

В настоящей работе приводятся результаты экспериментов, цель которых – оценить применимость методов отбора признаков при порождении гипотез с помощью ДСМ-системы. Известная проблема, связанная с применением ДСМ-метода на практике – высокая вычислительная сложность алгоритма построения решетки формальных понятий [3]. Поэтому проблема сокращения размерности стоит перед исследователем, решившим воспользоваться ДСМ-методом, особенно остро.

МЕТОДЫ ОТБОРА ПРИЗНАКОВ

Основная идея всех методов отбора состоит в выборе признаков, максимизирующих точность предсказания или классификации. Они основаны на принципе бритвы Оккама [4]. Он гласит, что следует стремиться к моделям с минимально возможным числом параметров, адекватно представляющим существующие данные. В работе [5], например, цитируются слова Эйнштейна о том, что «все следует делать как можно проще, но не более того». Этот принцип, однако, сложно применить к отбору признаков. Доказано, что выбор лучшего подмножества признаков является NP-полной задачей [6].

Среди методов отбора признаков выделяют два больших подкласса – *фильтры* и *обертки*. Рассмотрим их подробнее.

Фильтры чаще всего основаны на *ранжировании признаков* в качестве основного механизма отбора. Алгоритмы ранжирования используют некоторую функцию оценки *S(i)*, где *i* – индекс оцениваемого признака. Общепринятая интерпретация этой функции следующая: чем выше ее значение, тем полезнее мы считаем данный признак. Таким образом, посчитав *S(i)* для каждого *i*, имеем все признаки, ранжированные по данному критерию.

В качестве функций ранжирования часто используются: коэффициент корреляции Пирсона, критерий взаимной информации, критерий хи-квадрат, t-критерий Стьюдента, U-критерий Манна-Уитни. Все они считаются для интересующего нас признака относительно целевого.

Методы ранжирования, которые основаны на функции оценки признаков относительно ключевого признака, не позволяют уловить зависимости, установившиеся между признаками. В работе [1] показано, что признаки могут быть бесполезны сами по себе, но могут давать существенный прирост в качестве предсказания, если использовать их совместно с другими признаками. И наоборот, признаки, которые оценены как полезные, могут на самом деле быть избыточными. Кроме того, в общем виде непонятно, где ставить отсечку для выбора признаков из ранжированной последовательности.

Итеративные методы отбора, называемые *обертками*, позволяют учитывать синергию атрибутов. Обертки рассматривают обучаемую систему как черный ящик, которому на вход подаются данные с отфильтрованными признаками, а на выходе имеется алгоритм предсказания, для которого можно посчитать точность. Для обертки необходимо определить следующие параметры: 1) как обходить множество всех подмножеств атрибутов; 2) как оценивать качество предсказания. Если признаковое пространство не слишком большое, то в качестве алгоритма обхода можно использовать полный перебор. Использование алгоритма обучения как черного ящика позволяет оберткам быть универсальным и простым механизмом.

Обертки часто критикуются за то, что они используют метод «грубой силы», требующий больших объемов вычислений, но это не обязательно так. Для них можно разработать эффективные стратегии поиска, использование таких стратегий не всегда означает снижение качества предсказания. Напротив, в некоторых случаях грубые стратегии приводят к переобучению (см., например, [7]).

В последнее время также предлагаются *гибридные алгоритмы* отбора признаков, комбинирующие преимущества фильтров и оберток. Основная их идея состоит в том, что сначала к полному множеству признаков применяется фильтр, после чего из отобранного подмножества выбираются признаки с помощью алгоритма-обертки.

Кроме того, в отличие от обертки, воспринимающей алгоритм обучения как черный ящик, существуют также *встраиваемые методы отбора признаков*, которые: а) работают быстрее, так как не требуют обучения каждый раз после отбора признаков; б) бережливее относятся к исходным данным, так как не требуют разделения исходного множества объектов на обучающее и тестовое подмножества. Недостаток встраиваемых методов – в сильной зависимости от выбранного метода обучения.

Деревья решений. Алгоритм C4.5

Применяемый нами метод отбора признаков основан на *деревьях решений*. Дерево решений является структурой, широко применяемой в машинном обучении. Каждая нетерминальная вершина такого дерева представляет собой проверку значения некоторого признака, каждая терминальная вершина соответствует некоторому классу объекта или значению целевого признака, каждая ветка – значению или интервалу значений признака, проверяемому в вершине, из которой исходит эта ветка. Таким образом, для классификации нового объекта с помощью дерева решений нужно пройти путь от корня дерева до терминала по соответствующим веткам.

Максимальным деревом решений называется такое дерево, которое разбивает обучающую выборку на чистые классы, т. е. для каждого объекта обучающей выборки оно должно проводить корректную классификацию.

Построение максимального дерева на основе обучающей выборки служит первым шагом нашего алгоритма отбора. Для этой задачи мы выбрали *алгоритм C4.5* [8]. Приведем его описание.

Пусть нам задано множество примеров T, где каждый элемент этого множества описывается m атрибутами. Количество примеров в множестве T будем называть мощностью этого множества и обозначать $|T|$.

Пусть метка класса принимает следующие значения: $C_1, C_2, \ldots, C_k$.

Задача – построить иерархическую классификационную модель в виде дерева из множества примеров S. Процесс построения дерева идет сверху вниз. Сначала создается корень дерева, затем потомки корня и т.д.

На первом шаге имеем пустое дерево (точнее, дерево из единственного корня) и исходное множество S, ассоциированное с корнем. Требуется разбить исходное множество на подмножества. Для этого выбирается один атрибут, после чего создаются n разбиений, где n – число различных значений этого атрибута. Затем эта процедура рекурсивно применяется ко всем подмножествам (потомкам корня).

Рассмотрим подробнее критерий выбора атрибута, по которому на текущем шаге пойдет ветвление. Имеется m возможных вариантов, из которых мы должны выбрать наиболее подходящий. Некоторые алгоритмы исключают повторное использование атрибута при построении дерева, но C4.5 не накладывает подобных ограничений, т.е. любой из атрибутов можно использовать неограниченное число раз.

Пусть мы имеем проверку X (в качестве проверки может быть выбран любой атрибут), которая принимает n значений $A_1, A_2, \ldots, A_n$. Критерий, который мы берем за основание, использует информацию о том, как распределены классы в множестве S и подмноже-

ствах, получающихся при разбиении. Этот критерий называется *приростом информации* и равен разности энтропии множества S и разбиения S по атрибуту X.

Для начала приведем определение *энтропии* (степени неопределенности) множества примеров:

$$Ent(S) = -\sum_i p(C_i)\log_2 p(C_i)$$

где $p(C_i)$ – отношение мощности класса C_i к мощности S. Когда $Ent(S) = 0$, множество S является идеально классифицированным, т.е. все его объекты принадлежат одному классу.

Энтропией разбиения $\{S_i\}$ называют взвешенную сумму степеней неопределенности всех множеств, принадлежащих разбиению:

$$E(\{S\_i\};S) = \sum_i \frac{|S_i|}{|S|} Ent(S_i)$$

Таким образом, интересующий нас критерий считается следующим образом:

$$Gain(X)=Ent(S)=E(X;S),$$

где $E(X; S)$ – энтропия разбиения S по значениям атрибута X. Выбирается такой X, который максимизирует величину $Gain(X)$ для данного S. Атрибут X будет являться проверкой в текущем узле дерева, затем по этому атрибуту будет производиться дальнейшее построение дерева.

Те же рассуждения рекурсивно применяем к полученным подмножествам $S_1, S_2, ..., S_n$ и продолжаем процесс построения дерева до тех пор, пока в узле не окажутся примеры из одного класса.

Если в процессе работы алгоритма получен узел, ассоциированный с пустым множеством (т.е. ни один пример не попал в данный узел), то он помечается как лист, и в качестве решения листа выбирается наиболее часто встречающийся класс у непосредственного предка данного листа.

Основным принципом построения дерева является максимизация прироста информации, так как из свойств энтропии известно, что максимально возможное значение энтропии достигается в том случае, когда все сообщения равновероятны. В нашем случае энтропия достигает своего максимума, когда частота появления классов в примерах множества S равновероятна. Нам же необходимо выбрать такой атрибут, чтобы при разбиении по нему один из классов имел наибольшую вероятность появления. Это возможно в том случае, когда энтропия будет иметь минимальное значение и, соответственно, прирост информации достигнет максимума.

Алгоритм C4.5 умеет обрабатывать и непрерывные атрибуты – для них выбирается порог, с которым в данном узле будет сравниваться значение атрибута.

Пусть непрерывный атрибут имеет конечное число известных нам значений. Обозначим их $\{v_1, v_2, ..., v_n\}$. Предварительно отсортируем все значения. Тогда любое значение, лежащее между v_i и v_{i+1}, делит все примеры обучающей выборки на два множества: те, которые лежат слева от этого значения $\{v_1, v_2, ..., v_i\}$, и те, что справа $\{v_{i+1}, v_{i+2}, ..., v_n\}$. В качестве порога можно выбрать среднее между значениями v_i и v_{i+1}:

$$Ent(S) = -\sum_i p(C_i)\log_2 p(C_i)$$

$$TH_i = \frac{v_i + v_{i+1}}{2}\ .$$

Таким образом, нужно рассмотреть только $n - 1$ потенциальное значение атрибута: $TH_1, TH_2, ..., TH_{n-1}$. Для каждого из этих значений последовательно вычисляется прирост информации и среди них выбирается то, которое его максимизирует. Далее значение прироста сравнивается с максимальными значениями, полученными для других атрибутов. Выбирается такой атрибут и такое пороговое значение, которые максимизируют этот критерий. Отметим, что все проверки числовых атрибутов будут бинарными, т.е. разделят узел на две ветви.

Усечение максимального дерева на основе минимальной цены-сложности

Вторым шагом алгоритма отбора признаков является получение из построенного на первом шаге максимального дерева последовательности усеченных деревьев. Вообще идея усечения состоит в том, чтобы «отвязать» дерево решений от обучающей выборки, так как эта структура известна своей тенденцией переобучаться. Алгоритмы усечения используют то свойство деревьев решений, что наиболее значимые атрибуты в них располагаются ближе к корню. Соответственно, если убрать «нижние» ветки, наличие которых зачастую обусловлено только данной конкретной выборкой, остаются только ветки, имеющие высокую предсказательную силу.

Естественно, чем более сильно усечено дерево, тем важнее оставшиеся в нем атрибуты, однако, слишком сильно усеченное дерево (скажем, до единственной вершины – корня) может оказаться недостаточным для хорошей классификации. Соответственно, нашей задачей было сравнить между собой фильтры, полученные на основе деревьев различной степени усечённости.

Алгоритм, который позволяет получить последовательность таких деревьев, называется *усечением на основе минимальной цены-сложности* (*minimal cost-complexity pruning*) [9]. Приведем его описание.

Пусть T – некоторое дерево решений. *Ценой подставки* $R(T)$ называется вероятность того, что пример из обучающей выборки данным деревом будет классифицирован неправильно. У максимального дерева $R(T) = 0$, так как процесс построения всегда доводится до точки, в которой всем листьям соответствуют чистые множества примеров. Если заменить некоторую ветку дерева T на лист, классифицирующий пример максимально представленным в этой ветке классом, $R(T)$ возрастет.

Теперь определим метрику *цены-сложности*: для каждого поддерева $T < T_{max}$, определим сложность $|T|$ как количество листьев в дереве T. Пусть $\alpha \geq 0$ – действительное число, назовем его *параметром слож-*

ности и определим метрику *цены-сложности* $R_\alpha(T)$ следующим образом:

$$R_\alpha(T) = R(T) + \alpha|T|$$

Чем больше листьев у дерева, тем выше его сложность, так как существует больше возможностей для разбиения пространства на подпространства, а, значит, больше возможностей подстроиться под обучающую выборку. Параметр α позволяет настраивать значимость размера дерева.

$R_\alpha(T)$ по сути представляет собой цену подстановки, дополненную штрафом за сложность дерева. Это функция, которую мы будем минимизировать в процессе усечения дерева.

Таким образом, при заданном α задача усечения состоит в поиске поддерева $T(\alpha)$, минимизирующего $R_\alpha(T)$:

$$R_\alpha(T(\alpha)) = \min_{T \le T_{\max}} R_\alpha(T)$$

Такое дерево всегда существует для любого α, так как множество поддеревьев дерева конечно.

Процесс нахождения ряда вложенных поддеревьев и соответствующего им параметра сложности с гарантией монотонного возрастания этого параметра называется *усечением по слабому звену*. Он основан на подсчете метрики цены-сложности для каждой вершины t и ветки T_t дерева T:

$$R_\alpha(\{t\}) = R(t) + \alpha$$
$$R_\alpha(T_t) = R(T_t) + \alpha|T_t|$$

При $\alpha = 0$ выполняется неравенство $R_0(T_t) < R_0(\{t\})$. Так как $R_\alpha(T_t)$ увеличивается быстрее (коэффициент при α большем, чем $R_\alpha(\{t\})$ при постепенном увеличении α, то при некотором α мы получим $R_\alpha(T_t) = R_\alpha(\{t\})$). Вершина, которая достигает этого равенства при наименьшем α, называется *слабым звеном*. Возможна ситуация, при которой несколько вершин достигают равенства одновременно, в этом случае имеем несколько слабых звеньев.

Решив неравенство, получим

$$\alpha < \frac{R(T) = R(T_t)}{|T_t| - 1}$$

В правой части находится отношение разницы в цене замещения к разнице в сложности. Оно гарантированно положительное, так как числитель и знаменатель положительные.

Определим функцию g1(t), t ∈ T1:

$$g_1(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{|T_t| - 1}, t \notin \widetilde{T_1} \\ +\infty, t \in \widetilde{T_1} \end{cases}$$

Слабое звено в T_1 минимизирует функцию $g_1(t)$, параметр сложности α_2 для него равен этому минимальному значению. Соответственно, чтобы получить оптимальное поддерево, соответствующее этому параметру сложности, нужно заменить слабое звено на терминальную вершину. Если слабых звеньев на этом шаге несколько, заменяем все эти ветки. α_2 является первым значением параметра после $\alpha_1 = 0$, которое соответствует оптимальному поддереву, строго содержащемуся в T_1. Для всех $\alpha_1 \le \alpha \le \alpha_2$ минимальное оптимальное поддерево совпадает с T_1.

Пусть $T_2 = T_1 - T_t$, где t – слабое звено в T_1. Чтобы получить ряд усеченных поддеревьев в порядке, соответствующем монотонно возрастающему α, используем T_2 вместо T_1, находим его слабое звено и заменяем его на терминальную вершину, и так далее рекурсивно, пока не достигнем дерева, состоящего из единственной вершины.

Алгоритм отбора признаков

Последний шаг алгоритма отбора я – использование каждого из деревьев в качестве фильтра атрибутов.

Таким образом, общий алгоритм отбора следующий:

1. Построить максимальное дерево решений на основе обучающей выборки.

2. Получить последовательность усеченных деревьев от максимального (исходного) до минимального (единственная вершина, соответствующая корню исходного).

3. Каждое из деревьев этой последовательности считать фильтром, отбирающим только те атрибуты, которые в нем проверяются.

Это метод, не падающий однозначно ни в категорию фильтров, ни в категорию оберток. В отличие от фильтра, он не дает возможности упорядочить атрибуты по значимости, однако, в отличие от обертки, не требует обязательного переобучения основного алгоритма, чтобы сделать следующий шаг поиска. Подобные методы успешно применялись в задачах обучения (см., например, [10, 11]).

Таким образом, описанный алгоритм позволяет получить последовательность вложенных подмножеств атрибутов. Заметим, что он не позволяет ответить на вопрос, какое из этих подмножеств выбрать – схожая проблема возникает с использованием классических ранжирующих фильтров, для которых непонятно, сколько лучших атрибутов отбирать. Здесь на помощь могут прийти классические критерии остановки отбора для методов оберток, либо собственный критерий исследователя, который он может определить в зависимости от решаемой им задачи.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Мы попробовали провести обучение ДСМ-системы на каждом из получившихся подмножеств признаков и сравнить метрики качества, скорости и «понятности» (в данном случае – количества гипотез). В качестве тестовой задачи был взят открытый набор данных из репозитория UCI [12]. Для дискретизации непрерывных параметров использовалась энтропийная дискретизация [13].

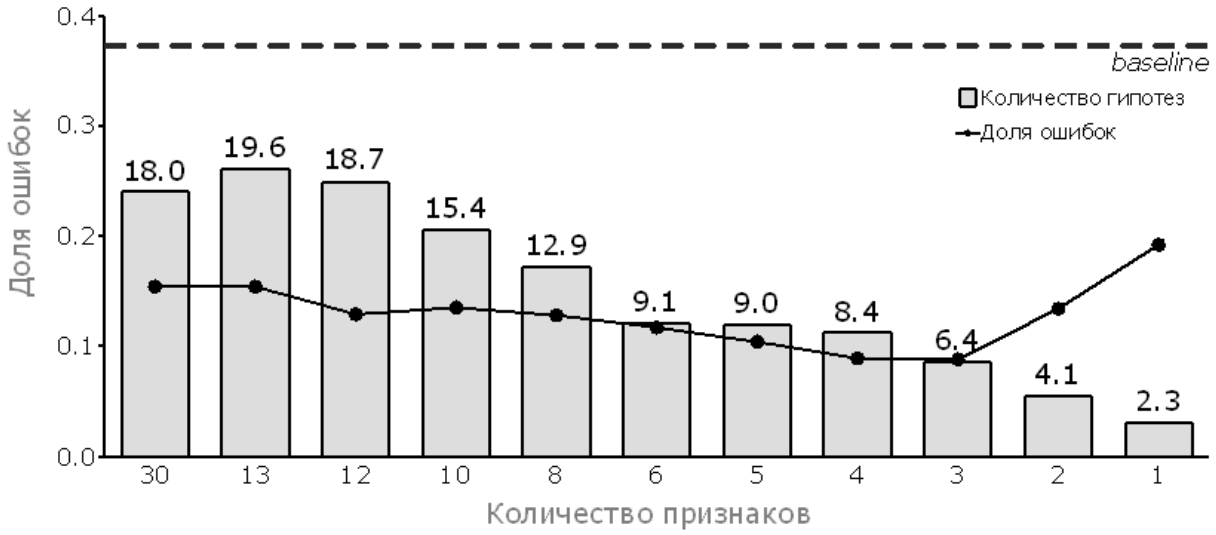
Цель данной задачи – определить тип опухоли: злокачественная (M = malignant) или доброкачественная (B = benign). Объектом является снимок пробы клеточной ткани: признаки описывают характеристики трех клеточных ядер, по 10 признаков на каждое, всего 30 атрибутов. Количество объектов равно 569. Характеристики ядер следующие:

1. Радиус.
2. Текстура (стандартное отклонение оттенков серого на снимке).
3. Периметр.
4. Площадь.
5. Гладкость (дисперсия радиуса).
6. Компактность (переметр^2 / площадь – 1).
7. Вогнутость.
8. Частота точек перегиба.
9. Симметричность.
10. Фрактальная размерность.

Результаты экспериментов приведены в таблице и на рисунке. Каждая строчка таблицы (или точка горизонтальной оси на рисунке) описывает 10-кратную перекрестную валидацию ДСМ-классификации с указанным количеством атрибутов со случайным разбиением обучающей выборки в отношении 1:1 (284 обучающих примера, 285 тестовых). Указаны среднее количество гипотез, среднее значение и дисперсия доли ошибок на тестовой выборке, а также время относительно времени обучения системы при наличии всех атрибутов (шаг №0).

Таблица

№ шага	Количество атрибутов	Количество гипотез	Доля ошибок	Относительное время
0	30	18	0,154 ± 0,034	100%
1	13	19,6	0,154 ± 0,034	8%
2	12	18,7	0,129 ± 0,036	6%
3	10	15,4	0,135 ± 0,033	5%
4	8	12,9	0,128 ± 0,028	3%
5	6	9,1	0,117 ± 0,030	2%
6	5	9	0,104 ± 0,023	2%
7	4	8,4	0,089 ± 0,018	1%
8	3	6,4	0,088 ± 0,022	1%
9	2	4,1	0,134 ± 0,024	1%
10	1	2,3	0,192 ± 0,063	1%



Лучшие результаты показали классификаторы с 3-4 признаками. Им удалось достигнуть точности 91-92% при 6-9 ДСМ гипотезах и времени обучения в 100 раз меньшем, чем при использовании всех 30 атрибутов (в абсолютном выражении в тестовой среде – около 30 мс против 3 с).

В завершение приводим пример списка из 9 гипотез, которые получены с помощью ДСМ-метода при использовании всего 3 признаков из 30 доступных:

1) периметр второго ядра (**2-3**);
2) частота точек перегиба у второго ядра (**2-8**);
3) радиус первого ядра (**1-1**).

i. **2-3** < *101,95* & **2-8** < *0,1096* → **B (136)**
ii. **2-3** > *120,35* → **M (78)**
iii. **2-8** < 0,1096 **& 1-1** ∈ [13,155; 15,025] → **B (32)**
iv. **2-3** < 101,95 **& 1-1** ∈ [13,155; 15,025] → **B (27)**
v. **2-3** < 101,95 **& 2-8** ∈ [0,1096; 0,1417] → **B (16)**
vi. **2-8** > 0,1417 **& 1-1** ∈ [13,155; 15,025] → **M (15)**
vii. **2-8** > 0,1417 **& 1-1** ∈ [15,025-16,925] → **M (14)**
viii. **2-8** > 0,1417 **& 1-1** < 13,155 → **M (5)**
ix. **2-8** < 0,1096 **& 1-1** ∈ [15,025-16,925] → **M (5)**

В скобках после гипотезы указан ее вес, равный количеству объектов из обучающей выборки, на которых она подтверждается. В случае, если не удается применить ни одну гипотезу, выполняется классификация по максимальному классу.

Приведенные гипотезы позволяют получить точность классификации 93,4%.

* * *

Автор работы выражает благодарность за ценные обсуждения Д.В. Виноградову, В.К. Финну, Е.С. Панкратовой, О.М. Аншакову и В.А. Ковтуну.

СПИСОК ЛИТЕРАТУРЫ

1. Guyon I., Elisseeff A. An introduction to Variable and Feature Selection // Journal of Machine Learning Research. – 2003. – Vol. 3, №3. – P. 1157-1182.
2. Bellman R.E. Adaptive Control Processes: A Guided Tour. – Princeton, NJ: Princeton University Press. – 1961.
3. Kuznetsov S.O., Obiedkov S.A. Comparing Performance of Algorithms for Generating Concept Lattices // Journal of Experimental and Theoretical Artificial Intelligence. – 2002. – Vol. 14. – P. 189-216.
4. Bell D.A., Wang H. A formalism for relevance and its application in feature subset selection // Machine Learning. – 2004. – Vol. 41, №2. – P. 175-195.
5. Parzen E. ARARMA models for time series analysis and forecasting // Journal of Forecasting. – 1982. – Vol. 1, № 1. – P. 67-82.
6. Peng H., Long F., Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min redundancy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2005. – Vol. 27, №8. – P. 1226-1238.
7. Reunanen J. Overfitting in making comparisons between variable selection methods // Journal of Machine Learning Research. – 2003. – Vol. 3, №3. – P. 1371-1382.
8. Quinlan J. R. C4.5: Programs for Machine Learning. – San Mateo: Morgan Kaufmann Publishers Inc, 1993.
9. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. – Monterey, CA: Wadsworth. – 1984.
10. Grabczewski K., Jankowski N. Feature Selection with Decision Tree Criterion // Fifth International Conference on Hybrid Intelligent Systems. – 2005. – P. 212-217.
11. Yacob Y.M., Sakim H.A.M., Isa N.A.M. Decision Tree-based Feature Ranking using Manhattan Hierarchical Cluster Criterion // World Academy of Science, Engineering and Technology. – 2012. – № 62. – P. 765-771.
12. UC Irvine Machine Learning Repository. – URL: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/.
13. Яшков И.Б. Сравнение методов дискретизации и отбора непрерывных параметров для логико-комбинаторной классификации // Научно-техническая информация. Сер. 2. – 2013. – № 8. – С. 26-31.

Материал поступил в редакцию 06.09.13.

Сведения об авторе:

ЯШКОВ Игорь Борисович – аспирант ВИНИТИ РАН, Москва
e-mail: IgorYashkov@gmail.com

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.891.3 : 616 – 006.81

Д.А. Добрынин, Л.В. Демидов, А.Ю. Барышников, И.Н. Михайлова

Интеллектуальная компьютерная система для анализа клинических данных*

Описано применение Интеллектуальной системы ДСМ для анализа клинических данных больных с опухолью меланома. Работа велась совместно с ФГБУ «РОНЦ им. Н.Н. Блохина» РАМН. Система является помощником врача-онколога и предназначена для прогнозирования времени до возможного прогрессирования опухоли (в данном случае меланомы) на основе извлеченных из базы фактов закономерностей причинно-следственного типа.

Ключевые слова: *интеллектуальная система, ДСМ-метод, медицинская диагностика, меланома, доказательная медицина*

ВВЕДЕНИЕ

Современные достижения в области молекулярной биологии создали предпосылки для разработки новых, более селективных и эффективных методов противоопухолевой терапии. Одним из таких методов, вызывающих в настоящее время наибольший интерес, является вакцинотерапия, включающая различные модификации клеточной технологии. Применение данного вида лечения наиболее целесообразно для тех злокачественных заболеваний, которые являются иммуногенными (например, меланома, рак почки) и в силу этого способны регрессировать под влиянием специфического воздействия иммунной системы. Однако прогноз при проведении противоопухолевой вакцинотерапии остается неясным. Для решения этой проблемы **можно использовать Интеллектуальную систему** (ИнтДСМ[1]) [1], которая является инструментом поддержки медицинских исследований со сложно структурированными данными и множеством фактов, необозримых без использования компьютерных технологий и является средством усиления интеллектуальных возможностей экспертов-медиков и включает:

ИнтДСМ = Решатель задач + Информационная среда (база фактов (БФ) и база знаний (БЗ)) + Интеллектуальный интерфейс (диалог + представление результатов + научение работе с системой).

Решатель ИнтДСМ основан на ДСМ-методе автоматического порождения гипотез [1], реализующем автоматизированные правдоподобные рассуждения. Разрабатываемые правдоподобные рассуждения образуют формализованную эвристику извлечения зависимостей причинно-следственного типа из баз структурированных фактов.

ДСМ-рассуждение на основании анализа объектов с заранее известными свойствами осуществляет правдоподобный вывод, результатом которого является информация о наличии или отсутствии свойств у объектов. Для работы ДСМ-системы требуется набор объектов, про которые известно, что они обладают каким-либо свойством («+»-объекты), и набор объектов, про которые известно, что они не обладают соответствующим свойством («-»-объекты). Задается также набор объектов, наличие свойств которых требуется определить («τ»-объекты).

ДСМ-метод автоматического порождения гипотез состоит из трех познавательных процедур – индукции (порождения причин эффектов на основе обнаруженных сходств фактов), аналогии (правдоподобных выводов, использующих наличие положительных или отрицательных причин в фактах с неопределенной оценкой, требующей уточнения – наличия или отсутствия изучаемого эффекта) и, наконец, абдукции [1] (принятия гипотез посредством объяснения начального состояния базы фактов с помощью (±)–причин, т. е. гипотез, ответственных за наличие эффекта ((+)-причины) и за отсутствие эффекта ((-)-причины).

ДСМ-система работает в три этапа.

На первом этапе посредством специально созданной процедуры индукции, основанной на сходстве объектов, порождаются гипотезы о (±)-причинах, извлекаемые из БФ.

На втором этапе осуществляется правдоподобный логический вывод, реализующий доопределение «τ»-объектов на основании «+»- и «-»-гипотез, порожденных на первом этапе. После этого процесс может повторяться, в случае если остались не-

* Работа выполнена при поддержке программы фундаментальных исследований Президиума РАН «Математическое моделирование и интеллектуальные системы» на 2013 г. (проект №209) и РФФИ (проект № 11-07-00618а).

[1] Название в честь английского мыслителя Джона Стюарта Милля.

доопределенные объекты, до стабилизации исходной выборки.

На третьем этапе ДСМ-рассуждение завершается абдуктивной процедурой – объяснением начального состояния БФ, которое является или достаточным основанием для принятия гипотез, или средством расширения БФ для итерации ДСМ-рассуждения, если существуют необъясненные факты из БФ.

Гипотезы о (±)-причинах, извлекаемых из БФ, порождаются посредством специально созданной процедуры индукции, основанной на сходстве объектов - источнике или причине наличия (отсутствия) изучаемого эффекта. Предсказание этого эффекта осуществляется посредством аналогии, использующей гипотезы о (±)-причинах, содержащихся в базе знаний и порожденных индукцией. И, наконец, ДСМ-рассуждение завершается абдуктивной процедурой – объяснением начального состояния БФ, которое является или достаточным основанием для принятия гипотез, или средством расширения БФ для итерации ДСМ-рассуждения, если существуют необъясненные факты из БФ [2-4]. Извлечение знаний из БФ типа «причина-следствие» основано на принципе: «сходство фактов порождает наличие (отсутствие) эффектов и их повторяемость» (этот принцип отличен от вероятностного подхода к анализу данных: «повторяемость эффектов определяет сходство фактов»). Следует отметить преимущества ДСМ-метода: возможность работать на малых массивах данных, учитывать индивидуальные особенности изучаемых объектов исследования и работать с открытыми массивами данных (а не с замкнутыми таблицами!).

Область анализа клинических данных больных меланомой отвечает условиям применимости ДСМ-метода:

1) возможность структурирования данных и формального определения сходства фактов из БФ;

2) наличие положительных и отрицательных примеров (±)-примеров в БФ;

3) наличие в БФ неявно заданных зависимостей причинно-следственного типа ((±)-причины изучаемых эффектов).

ПОСТАНОВКА ЗАДАЧИ

Совокупность результатов специальных обследований больного иногда бывает трудно проанализировать. Наш опыт общения с экспертами-врачами показывает, что им хотелось бы иметь интеллектуальную систему, помогающую в принятии решения. Работа с врачами-экспертами начинается с постановки задачи. Система является помощником врача-онколога и предназначена для прогнозирования времени до возможного прогрессирования опухоли (в данном случае меланомы) на основе извлеченных из БФ закономерностей причинно-следственного типа. Работа по созданию интеллектуальной системы типа ДСМ для анализа клинических данных больных меланомой велась совместно с Российским онкологическим научным центром им. Н.Н. Блохина РАМН [5]. В нашей предыдущей работе нам удалось создать систему прогнозирования продолжительности жизни больных меланомой в зависимости от экспрессии протеина S100 [6]. В настоящем исследовании задачей интеллектуальной системы являлся прогноз заболевания больных меланомой кожи в зависимости от уровня цитокинов и биомаркеров на фоне вакцинотерапии. Нам предлагалось определить время до прогрессирования меланомы (больше 5 месяцев или меньше 5 месяцев) на основании анализа некоторого количества признаков больного. Для оценки эффективности вакцинотерапии мы использовали цитокиновый профиль и биомаркеры

Для выяснения возможности применения ДСМ-метода к решению перечисленных задач проводится совместная работа с экспертами-врачами. Врач перечисляет все возможные признаки описания больного или описания конкретного анализа больного, по которым надо сделать прогноз. Задача врача - представить достаточное количество примеров положительного и отрицательного типа в терминологии ДСМ-метода, а также не пропустить значимых для диагноза признаков. Задача разработчиков структурировать эти признаки таким образом, чтобы к ним было возможно применять операцию сходства и определять отношение вложения. Эффективность анализа результатов различных исследований зависит от максимально полного описания обследований больного.

Первым этапом совместной работы была разработка подсистемы представления знаний и создание базы фактов – одной из составных частей ИнтДСМ. Атрибутами этой БФ являются сведения о больных, имеющиеся в медицинских картах и описанные в соответствии с языком представления данных. База данных – это описания историй болезней на специальном формальном языке [7] представления знаний и, кроме клинических и лабораторных данных, включает цитокины и биомаркеры, прогностическое значение уровня которых для прогрессирования меланомы кожи на фоне иммунотерапии предстоит оценить. По согласованию с врачами используются 36 признаков, характерных для исследуемой болезни:

1) возраст
2) пол
3) локализация первичной опухоли
4) проведенное лечение
5) есть ли признаки болезни на момент проведения первой вакцинации
6) метастазы
7) стадия болезни на момент вакцинации
8) количество линий химиотерапии
9) количество циклов химиотерапии
10) IL-2
11) IL-12
12) IFNy
13) IL-4
14) IL-10
15) S100
16) CD-44
17) VEGF-A
18) TGF-b-2
19-27) замер тех же показателей через некоторый промежуток времени
28-36) изменение значений этих показателей

37) время до прогрессирования
38) продолжительность жизни.

Каждый из перечисленных признаков имеет некоторое количество значений. Для каждого конкретного больного выбираются значения признаков из заранее заданного списка. В настоящее время создана экспериментальная БФ, содержащая 108 историй болезней.

НАСТРОЙКА НА ПРЕДМЕТНУЮ ОБЛАСТЬ

Для применения ИнтДСМ к конкретной задаче необходима настройка на предметную область [5], включающая:

(1) разработку языка представления данных;
(2) определение понятия «объект» и «свойство» в терминологии ДСМ-метода;
(3) задание операции сходства;
(4) задание отношения вложения;
(5) определение аксиом предметной области.

(1) Первый пункт настройки – разработка языка представления данных. В конкретной ИнтДСМ для представления признаков, определенных врачами, используются три типа данных из семи возможных, описанных в [7, 8]:

Первый тип данных. Указывается **один из возможных качественных признаков,** например, признак №11 «Тип роста первичной опухоли»:

1. Поверхностно распространяющаяся меланома.
2. Узловая меланома.
3. Меланома по типу злокачественного лентиго.
4. Акролентигинозная меланома.

Второй тип данных. В кортеже длины «*n*» (где *n* - количество элементов списка признаков) указываются **присутствующие качественные признаки,** например, признак №4 «Лечение»:

1. Химия
2. Иммунотерапия

Третий тип данных. Указываются конкретные **признаки иерархической структуры,** например признак №14 «На момент первой вакцинации»:

14.1 Нет признаков болезни
14.2 Есть первичная опухоль
14.3 Есть поражение региональных лимфоузлов
14.4 Есть отдаленные метастазы
14.4.1 кожа, мягкие ткани
14.4.2 отдаленные лимфоузлы
14.4.3 легкие
14.4.4 печень
14.4.5 кости
14.4.6 ЦНС
14.4.7 надпочечники
14.4.8 селезенка.

(2) Второй пункт настройки – определение понятий «объект» и «свойство». Объект представляет собой кортеж из 36 описанных выше признаков. Свойство в данном случае - время до прогрессирования опухоли меланома.

(3) Третий пункт настройки – в зависимости от типа данных, которому принадлежат признаки, операция сходства определяется поэлементно:

1) для первого типа данных результатом операции сходства является совпадение признаков (в случае несовпадения – пустой элемент);

2) для второго типа данных результатом операции сходства является кортеж из присутствующих в обоих элементах признаков;

3) для третьего типа данных результатом операции сходства является кортеж из присутствующих в обоих элементах признаков иерархической структуры.

(4) Четвертый пункт настройки – отношение вложения ($El' \subseteq El''$) определяется поэлементно и зависит от типа данных:

1) для 1-го типа данных вложение равносильно совпадению значений атрибутов;

2) для 2-го типа данных требуется присутствие в El'' каждого из элементов El';

3) для 3-го типа данных требуется присутствие в иерархической структуре El'' каждого из элементов иерархической структуры El'.

(5) Последний, пятый пункт настройки на предметную область – определение аксиом предметной области.

Для каждой задачи все перечисленные признаки, входящие в объект, возможно разделить с учетом знаний о предметной области на три группы:

1. Группа необходимых признаков – признаки, без наличия которых гипотеза не имеет смысла. Например, нельзя прогнозировать результат лечения без наличия признака «терапия».

2. Группа существенных признаков – признаки, без наличия хотя бы одного из которых, гипотеза не имеет смысла.

3. Группа сопутствующих признаков – признаки не входящие в группы 1 и 2, например, пол, возраст.

Таким образом, можно сформулировать следующую аксиому:

В потенциальную гипотезу должны входить все признаки группы 1 и хотя бы один из признаков группы 2.

Исходя из этой аксиомы, вводятся понятия фильтров. Конъюнктивный фильтр – требование вхождения в потенциальную гипотезу всех перечисленных в нем признаков. Дизъюнктивный фильтр – требование включения в потенциальную гипотезу хотя бы одного из перечисленных в дизъюнктивном фильтре признаков.

Кроме того, логично сформулировать следующую аксиому:

В потенциальную гипотезу должен входить хотя бы один признак со знаком «+».

СТРУКТУРА КОМПЬЮТЕРНОЙ СИСТЕМЫ

Для эффективного проведения экспериментальных исследований при различных типах и количестве входных признаков, система имеет специальный промежуточный язык для описания этих признаков. Все признаки описываются как множества с одним или многими элементами, а также как деревья. Для каждого типа признаков система автоматически генерирует свои процедуры поиска сходства при пере-

сечениях, а также значения для выбора конъюнктивных и дизъюнктивных фильтров.

В целом компьютерная система состоит из редактора признаков, программной части для генерации гипотез и системы просмотра полученных результатов. В редакторе признаков оператор вводит для каждого пациента значения выбранных заранее признаков. После ввода всех параметров, проводится получение гипотез о причинах, при этом состав признаков, которые будут учитываться, можно менять. Использование конъюнктивных и дизъюнктивных фильтров позволяет резко сократить количество получаемых гипотез и улучшить качество предсказания. Система просмотра гипотез позволяет оценивать полученные гипотезы и проводить их анализ.

НАСТРОЙКА СИСТЕМЫ НА ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

В конкретном экспериментальном исследовании признаки 1-36 входят в понятие «объект», «свойством» же является время до прогрессирования опухоли: больше пяти месяцев (положительные примеры в терминологии ДСМ-метода) и меньше пяти месяцев (отрицательные примеры). Этот интервал был выбран исходя из того, что в представленной БФ было наибольшее количество положительных и отрицательных примеров.

Настройка системы на экспериментальное исследование включает:

1. Выбор стратегии: простой метод сходства, метод сходства с запретом на контрпримеры (невложение полученных методом сходства гипотез в исходные примеры противоположного знака) отдельно для (+)- и (-)-примеров.

2. Подбор нужного количества «родителей» (минимальное количество примеров, формирующих гипотезу).

3. В фильтре конъюнктивных признаков перечисление признаков, которые должны присутствовать в каждой гипотезе.

4. В фильтре дизъюнктивных признаков перечисление признаков, по крайней мере, один из которых должен присутствовать в гипотезе.

Возможный критерий оценки настройки системы - применение процедуры «доопределение по одному»: последовательно каждому объекту выборки присваивается значение «неизвестно», производится доопределение этого объекта средствами ДСМ-системы с выбранными параметрами и сравнивается доопределенное значение с существующим в БФ. Подсчитывается общее количество правильных и неправильных доопределений. Выбираются параметры пп.1-4, при которых при применении процедуры «доопределение по одному» будет наилучший результат, т.е. наибольшее количество правильных доопределений и наименьшее количество неправильных.

Для нашего экспериментального исследования были выбраны параметры пп.1-4, при которых общее количество правильных доопределений – 20 и неправильных – 4 доопределения:

1) для (+)-примеров используется метод сходства с запретом на контрпримеры; для (-)-примеров используется простой метод сходства;

2) количество «родителей» для (+)-примеров – 6; количество родителей для (-)-примеров – 4;

3) конъюнктивных фильтров нет;

4) дизъюнктивный фильтр для (+)-примеров – CD44 на момент вакцинации ИЛИ показатель CD44 на момент 5(.) вакцинации; дизъюнктивный фильтр для (-)-примеров –S100 на момент вакцинации ИЛИ показатель S100 на момент 5(.) вакцинации.

Приведем пример работы системы.

Предположим, что системе неизвестно время до прогрессирования опухоли некоторого больного. Система породила 78 положительных и 26 отрицательных гипотез. Система доопределила время до прогрессирования опухоли этому больному меньше пяти месяцев шестью гипотезами. Перечислим их:

1. Показатель S100 на момент 5 (.)=201-500.
2. Старше 50 лет.
   Показатель S100 на момент 5 (.)=201-500.
3. Показатель S100 на момент 5 (.)=201-500.
   IL-12 на момент вакцинации =31-40.
4. Показатель S100 на момент 5 (.)=201-500.
   Вторичная стадия на момент вакцинации III.
5. Показатель S100 на момент 5 (.)=201-500.
   IL-2 на момент вакцинации =16-20.
6. Показатель S100 на момент 5 (.)=201-500.
   Показатель IFNy на момент вакцинации =0-5.
   Показатель IFNy на момент5(.) вакцинации =0-5.

Реально этот больной прожил менее пяти месяцев, что, к сожалению, соответствовало прогнозу ДСМ-системы.

Следующая больная жива в настоящее время, что согласуется с прогнозом системы. Перечислим пять гипотез-причин этого прогноза:

I. На момент вакцинации нет признаков болезни.
   Количество отдаленных метастазов 0.
   Показатель IL-4 на момент (5) 6-10.
   Показатель CD44 на момент (5) 601-800.

II. Старше 50 лет.
   На момент вакцинации нет признаков болезни.
   TGF-b-2 на момент вакцинации 500-1000.
   Показатель CD44 на момент (5) 601-800.

III. Показатель IL-4 на момент (5) 6-10.
   Показатель CD44 на момент (5) 601-800.

IV. Общее количество циклов химиотерапии 6-8.
   Показатель CD44 на момент (5) 601-800.
   На момент вакцинации нет признаков болезни.
   Количество отдаленных метастазов 0.

V. Старше 50 лет.
   Показатель CD44 на момент (5) 601-800.
   На момент вакцинации нет признаков болезни.
   Количество отдаленных метастазов 0.

ЗАКЛЮЧЕНИЕ

Описанная компьютерная интеллектуальная система обладает высоким быстродействием, что позволяет проводить большой объем экспериментов по прогнозированию свойств в реальном времени.

С помощью созданной нами компьютерной интеллектуальной системы на основе клинических характеристик и большой панели цитокинов и биомаркеров, возможно прогнозирование течения заболевания. Данная оценка является особенно актуальной при назначении различных видов адъювантной терапии: интерферонов, вакцин и химиотерапии. Контроль над течением заболевания позволяет своевременно на этапе доклинических проявлений менять тактику, выбирая более комплементарный тип терапии, что соответственно определит эффективность лечения.

Результаты применения ДСМ-метода как средства анализа онкологических данных демонстрируют полезность этого метода, являющегося новым инструментом доказательной медицины, т.е. выбора решений на основе анализа фактов.

Эта система является инструментом поддержки медицинских исследований со сложно структурированными данными и множеством фактов, необозримых без использования компьютерных технологий, и является средством усиления интеллектуальных возможностей экспертов-медиков.

Система может быть применена при соответствующей базе данных и для анализа фактов других онкологических заболеваний.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. Искусственный интеллект: методология, применения, философия. – М.: КРАСАНД, 2011.
2. Финн В.К., Блинова В.Г, Панкратова Е.С., Фабрикантова Е.Ф. Интеллектуальные системы для анализа медицинских данных. Ч. 1 // Врач и информационные технологии. – 2006. – № 5. – С.62-70.
3. Финн В.К., Блинова В.Г., Панкратова Е.С., Фабрикантова Е.Ф. Интеллектуальные системы для анализа медицинских данных. Ч. 2 // Врач и информационные технологии. – 2006. – № 6. – С. 50-60.
4. Финн В.К., Блинова В.Г., Панкратова Е.С., Фабрикантова Е.Ф. Интеллектуальные системы для анализа медицинских данных. Ч. 3. // Врач и информационные технологии. – 2007. – №1. – С. 51-57.
5. Михайлова И.Н., Парсункова К.А., Решетникова В.В., Финн В.К., Е.С. Панкратова, Михеенкова М.А., Шестерникова О.П. Компьютерная интеллектуальная система прогнозирования клинического течения меланомы. // Свидетельство о гос. регистрации программ для ЭВМ. – № 2012615098. – 10 апреля 2012г.
6. Михайлова И.Н., Панкратова Е.С., Добрынин Д.А., Самойленко И.В., Решетникова В.В., Шелепова В.М., Демидов Л.В., Барышников А.Ю., Финн В.К. О применении интеллектуальной компьютерной системы для анализа клинических данных больных меланомой.// Российский Биотерапевтический Журнал. – 2010. – № 2 – Т.9. – С. 54.
7. Панкратова Е.С. Интеллектуальная система типа ДСМ для анализа клинических данных // Научно-техническая информация. Сер.2. – 2011. – № 4. – С. 8-16.
8. Панкратова Е.С., Виноградов Д.В. Формальное описание настройки интеллектуальных ДСМ-систем на область клинической и лабораторной диагностики // Научно-техническая информация. Сер.2. – 2011. – № 9. – С. 1-5.

Материал поступил в редакцию 05.05.13.

ДОБРЫНИН Дмитрий Анатольевич – кандидат технических наук, научный сотрудник ВИНИТИ РАН, Москва.
e-mail: dobr@viniti.ru

ДЕМИДОВ Лев Вадимович – доктор медицинских наук, профессор, заведующий отделением Биотерапии опухолей Российского онкологического научного центра им. Н.Н. Блохина РАМН, Москва.

БАРЫШНИКОВ Анатолий Юрьевич – доктор медицинских наук, профессор, зав. лабораторией экспериментальной диагностики и биотерапии опухолей Российского онкологического научного центра им. Н.Н. Блохина РАМН.

МИХАЙЛОВА Ирина Николаевна – доктор медицинских наук, ведущий научный сотрудник Российского онкологического научного центра им. Н.Н. Блохина РАМН.
e-mail:irmikhaylova@gmail.com

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.65

М.С. Шматова

Проектирование типологической базы данных количественных конструкций (NNC-database)

Описывается проект типологической базы данных количественных конструкций (NNC-database[1]). Целью базы данных является обобщение и документирование данных о поведении конструкций «количественное числительное + существительное» в различных языках мира. Настоящая база данных призвана стать онлайн-инструментом, при помощи которого исследователи смогут не только получать информацию об упомянутых конструкциях, но и делиться собственными данными с коллегами.

Ключевые слова: *база данных, типология, количественные конструкции, числительное, существительное*

1. ВВЕДЕНИЕ

Типологическая база данных (ТБД) количественных конструкций создается по аналогии с уже существующими базами данных, описывающими конкретные явления в языках мира (подробнее о таких ТБД см., например, [1, 2]). Основными целями описываемого проекта стали систематизация информации о количественных конструкциях в языках мира и разработка теоретически-нейтральной пополняемой ТБД количественных конструкций в языках мира, доступной широкому кругу заинтересованных пользователей в сети Интернет.

Как отмечено в [3], ТБД является хорошим инструментом для создания полного описания того или иного явления в языках мира. Пополняемость типологической базы данных, т. е. возможность вносить в нее новую и корректировать уже имеющуюся информацию, становится для нас ключевым средством в борьбе за наиболее полное и достоверное описание этого явления. Таким образом мы надеемся получить данные о количественных конструкциях не только из уже имеющихся грамматик и статей, но также непосредственно от ученых, занимающихся конкретным языком, а наличие структурированной схемы описания, как мы надеемся, будет способствовать унификации этих данных.

Сами по себе количественные конструкции представляют интерес с типологической точки зрения. Как правило, в рамках грамматик и грамматических очерков языков такие конструкции рассматривают как частный случай именных групп, не останавливаясь подробно на их описании. Существуют также отдельные статьи, в которых подробно рассматриваются особенности количественных конструкций в одном или нескольких языках, как правило, в свете отдельно взятой теории (см., например: [4] – о роли числительного в именной группе, [5] – о согласовании числительного и существительного в русском языке, [6] – об именных группах с числительными в валлийском языке, [7] – о классификаторах в количественных конструкциях корейского языка и др.). Однако полного и вместе с тем теоретически нейтрального типологического описания данного явления, насколько нам известно, пока не существует. Именно этот факт и послужил поводом к созданию ТБД количественных конструкций.

Далее в статье речь пойдет о структуре описываемой ТБД, ее системе параметров и средствах реализации. В Разделе 2 мы остановимся на проблеме описания конструкций «числительное + существительное» в языках мира и ограничим предметную область ТБД. В Разделе 3 мы перейдем к описанию структуры и дизайна ТБД количественных конструкций.

2. КОЛИЧЕСТВЕННЫЕ КОНСТРУКЦИИ В ЯЗЫКАХ МИРА

Как и во многих других случаях, феномен, для рассмотрения которого разрабатывается типологическая БД, может называться в разных работах поразному. Это может быть связано как с теорией, в рамках которой он рассматривается, так и с авторским определением понятия. В таких случаях часто прибегают к уточнениям вроде «в широком/узком смысле». Например, в ТБД по редупликации [8] объектом описания является феномен редупликации, но не в общем случае, а только на уровне морфологии, т.е. редупликация понимается создателями БД в одном из узких смыслов.

[1] Поскольку база данных создается для международного пользования, ее основным языком является английский, хотя русский интерфейс также разрабатывается.

В широком смысле количественная конструкция в языке – это любая конструкция, обозначающая количество, как точное, так и приблизительное. В ее состав могут входить числительные, так называемые количественные слова, или квантификаторы (наречия, существительные), и существительные. В рамках нашего исследования будут рассмотрены только конструкции с (количественными) числительными, т. е., вообще говоря, подкласс всех возможных количественных конструкций. «Количественная конструкция – синтаксическая конструкция типа ”два солдата”, ”пять книг”, ”сто двадцать четыре этажа” и т. п., формируемая исчисляемым существительным и числительным» [9].

Рассматриваемый феномен в различных русскоязычных работах именуется также «количественной группой» [10], «конструкцией числительное плюс существительное» [11], «конструкцией числительное – существительное» [12]. У всех перечисленных названий существуют свои плюсы и минусы[2], и для русскоязычного названия данного проекта была выбрана именно формулировка «количественные конструкции» с обозначенными выше оговорками. Далее остановимся на особенностях количественных конструкций.

2.1. Состав и структура количественных конструкций

Итак, как уже было сказано, достаточно полного типологического описания феномена количественных конструкций не существует, однако в работе [13] приводится типологически нейтральное описание структуры числительных и их поведения в отношении существительных в европейских языках, которое было взято за основу для системы параметров описываемой ТБД.

Рассмотрим сначала некоторые особенности количественных числительных. Дж. Р. Харфорд подразделяет все числительные на простые лексические (однокорневые, чаще всего выражающие значения чисел от одного до десяти) и сложные (состоящие из двух и более корней простых числительных), для построения которых используются в основном операции сложения и умножения. В [14] последняя группа разделена еще на две по структурному признаку: сложные числительные (слова со сложной основой) и составные числительные (числительные, состоящие из нескольких слов). В некоторых случаях взаимодействие существительных со сложными (как собственно сложными, так и составными) числительными отличается от взаимодействия с простыми. В конструкциях с составными числительными различия касаются прежде всего порядка слов, что будет продемонстрировано ниже.

Говоря о взаимодействии простых числительных с существительными внутри одной именной группы, Харфорд замечает, что количественные конструкции не всегда состоят только из числительного и существительного – в их состав также могут входить предлоги и классификаторы.

Так, например, в валлийском языке (индоевропейская семья) существуют две возможные структуры количественных конструкций – без предлога и с предлогом.

ВАЛЛИЙСКИЙ ЯЗЫК ([15])

(1) а. naw o ddynion
девять из человек.PL
девять человек

б. naw dyn
девять человек
девять человек

Согласно [15] и [6], никакой разницы в значении между (1а) и (1б) не существует, но наблюдается тенденция к употреблению числительных, поведение которых более похоже на поведение существительных, в предложных конструкциях.

К таким же трехчленным конструкциям можно отнести и французские конструкции с числительными, референтами которых являются большие числа.

ФРАНЦУЗСКИЙ ЯЗЫК ([13])

(2) quatre million-s de personnes
четыре миллион-PL PREP человек-PL
четыре миллиона человек

Другим видом «трехкомпонентных» количественных конструкций являются конструкции с классификаторами. В ряде языков мира (например, в австронезийских языках, подробнее см. [16]) при существительном в составе количественной конструкции употребляются «специальные синтаксические лексемы (реже морфемы)» [17] – (количественные) классификаторы[3].

2.2. Классификаторы в количественных конструкциях

Особенностям функционирования классификаторов в отдельных языках мира, а также в рамках отдельных языковых семей и ареалов посвящен целый ряд работ (см., например, [18] (японский язык), [19] (венгерский язык), [20] (языки восточной и юго-восточной Азии) и [16] (языки западной Индонезии)). Первой работой, описывающей поведение классификаторов с типологической точки зрения, стала [21].

В монографии [22] представлен подробный типологический обзор классификаторов как средств категоризации существительных в языках мира (рассматриваются классификаторы в более чем 500 языках мира). А. Айхенвальд использует слово «классификатор» как «зонтичный термин для ряда инструментов категоризации имен существительных» [22]. Она

[2] Так, два последних названия, в отличие от первого, однозначно исключают из рассмотрения конструкции с количественными словами вроде *много столов*, но в то же время «ограничивают» количество членов конструкции до двух – числительного и существительного, тогда как в некоторых языках в таких конструкциях присутствует еще и классификатор.

[3] О том, почему, говоря о классификаторах, употребляют уточняющее определение «количественные», будет сказано ниже.

выделяет несколько типов классификаторов, среди которых, в частности, количественные классификаторы. К этому типу относятся классификаторы, употребляющиеся только в количественных конструкциях (они могут также указывать на одушевленность, размер и форму существительного, входящего в состав конструкции). Именно они и представляют интерес в рамках настоящего исследования.

Наличие количественных классификаторов является чаще всего свойством целого языкового ареала (например, Восточной и Юго-Восточной Азии). Они характерны в основном для изолирующих языков, реже встречаются в агглютинативных (японский, некоторые языки семьи нигер-конго), полисинтетических (языки нижнего течения Амазонки) и в некоторых флективных языках (языки индийской и южнодравидийской языковых подсемей/ветвей). Также наличие этого типа классификаторов в языке зависит от количества числительных в языке: чем больше система числительных, тем вероятнее наличие классификаторов.

Морфологически количественные классификаторы могут быть выражены отдельными лексемами, аффиксами числительных, клитиками, относящимися к числительным, аффиксами существительных или клитиками, относящимися к существительным.

Так, в узбекском языке (тюркская семья) представлено 14 количественных классификаторов, являющихся отдельными лексемами. В примерах (3а,б) классификаторы для людей и круглых объектов занимают место между числительным и существительным.

УЗБЕКСКИЙ ЯЗЫК ([22])

(3) а.
bir	nafar	âdam
один	CL:Human	человек

один человек

б.
bir	bâs	kâram
один	CL:Head.Shaped	капуста

один (кочан) капусты

Во многих языках классификаторы представляют собой суффиксы, реже префиксы, а в отдельных случаях – и инфиксы числительных. Так, в конструкции (4) из языка юкуна (аравакская семья, Южная Америка) классификатор, обозначающий одушевленность, присоединяется к числительному «один».

ЯЗЫК ЮКУНА ([22])

(4)
pajluhua-na	yahui
один-CL:Anim	собака

одна собака

В других языках, как, например, в корейском (изолят), классификатор следует сразу за числительным в виде клитики и принимает на себя показатель падежа.

КОРЕЙСКИЙ ЯЗЫК ([22])

(5)
sey	calwu-uy	yenphil-lul	sa-ss-ta
три	NUM.CL: Long. Slender-ATT	карандаш- ACC	купить-PST-DEC

(Я) купил три карандаша

Согласно Айхенвальд, единственными известными языками, где классификаторы синтаксически относятся не к числительному, а к существительному, являются языки подсемьи агони (семья нигер-конго, Нигерия), в частности, язык кана. Так, в конструкции (6б) классификатор *kà* употребляется при существительном (предшествует ему), а не при числительном.

ЯЗЫК КАНА ([22])

(6) а.
í	núú
DIM	крыса

маленькая крыса

б.
zìì	í	kà	núú
один	DIM	CL:Generic	крыса

одна маленькая крыса

Необходимо также отметить, что в одном языке возможны разные виды морфологического представления классификаторов.

Согласно описанию количественных классификаторов в рамках WALS ([23]) в работе [24], в языках, где присутствуют такие классификаторы, их употребление в количественных конструкциях может быть как обязательным, так и факультативным.

Так, в малайском языке (австронезийская семья, Суматра) грамматичной является конструкция как с классификатором, так и без него.

МАЛАЙСКИЙ ЯЗЫК ([25])

(7) а.
dua	buah	buku
два	CL	книга

две книги

б.
dua	buku
два	книга

две книги

В [25] отмечается также, что в языках, где в общем случае классификаторы являются обязательными членами количественных конструкций, в некоторых контекстах их употребление может становиться факультативным. Так, к примеру, в корейском языке несколько одушевленных существительных могут употребляться с числительными без классификатора (среди них *salam* «человек», *haksayng* «студент», *kay* «собака») [25].

2.3. Порядок слов в количественных конструкциях

Порядок слов в конструкциях «числительное + существительное» в типологической перспективе рассмотрен в [26] в рамках обширной типологической базы данных WALS ([23]). М. Драйер выделяет три возможных стратегии образования количественных конструкций в языках мира (в скобках приведем количество языков, использующих ту или иную стратегию)[4]:

[4] Также существует два языка, в которых числительное может модифицировать только глагол (каро (арара), семья тупи, вари, чапакурская семья). В данной работе мы их рассматривать не будем, так как непосредственно количественные конструкции вида «числительное + существительное» в них не представлены.

- в большинстве или во всех случаях числительное предшествует существительному (далее NumN) (479 языков);
- в большинстве или во всех случаях числительное следует за существительным (далее NNum) (608 языков);
- возможны оба вышеуказанных порядка слов и ни один из них не является преобладающим (65 языков)[5].

Как видим, наиболее частотным в языках мира является порядок NNum. Он преобладает в абсолютном большинстве языков Африки и Центральной и Юго-Восточной Азии, а также в некоторых индейских языках. Ниже приведем пример из языка Пуми тибето-бирманской подсемьи сино-тибетской семьи языков (распространен в Китае).

ЯЗЫК ПУМИ ([26])

(8) qüa xué
свинья восемь
восемь свиней

Вторым по распространению внутри количественной конструкции является порядок слов NumN, который наблюдается, в частности, в индоевропейских языках.

НЕМЕЦКИЙ ЯЗЫК

(9) acht Schweine
восемь свинья
восемь свиней

Такой же порядок слов является преобладающим и в египетском и марокканском вариантах арабского языка, однако в этих случаях имеются исключения из общего правила. Так, в марокканском варианте числительное «один» следует за существительным, которое оно модифицирует, а в египетском варианте это правило распространяется на числительные «один» и «два».

АРАБСКИЙ ЯЗЫК (ЕГИПЕТСКИЙ ВАРИАНТ) ([26])

(10) а. binteen ʔitneen
девочка.DU два
две девочки

б. talat banaat
три девочка.PL
три девочки

Существуют также языки, в которых оба порядка следования членов конструкции равновозможны и выбор между ними не всегда очевиден. В [26] приведено несколько примеров из таких языков.

ЯЗЫК АКОМА ([26])

(11) а t^{y}ának'a wáaštíṭa
четыре детеныш
четыре детеныша

б. kák^{h}ana t^{h}úẃé
волк два
два волка

Согласно Драйеру, в языке акома (кересанская подсемья, Нью-Мехико) неясно, на каком основании для конструкции (11а) выбран порядок NumN, а в случае (11б) – порядок NNum.

ЯЗЫК НИАС ([26])

(12) а. öfa geu m-baβi s=afusi
четыре CL ABS-свинья REL=белый
четыре белые свиньи

б. baβi-ra s=afusi si=öfa geu
свинья- REL=белый REL=четыре CL
3PL.POSS
их четыре белые свиньи

Как видно из (12), в языке ниас (австронезийская семья, Суматра) определенная именная группа следует за числительным (12б), которое ее модифицирует, тогда как неопределенная – предшествует ему (12а).

В рамках настоящего исследования нас также интересует порядок слов в трехчленных количественных конструкциях. Если предлог в описанных в [13] языках всегда занимает фиксированную позицию (после числительного и перед существительным - NPrepNum), то в случае с классификаторами возможны варианты.

Если рассматривать все теоретически возможные перестановки членов такой конструкции, получим шесть вариантов перестановок: [NumCl]N, N[NumCl], [ClNum]N, N[ClNum], ClNNum и NumClN. Гринберг отмечает, что в языках мира встречаются только первые четыре комбинации, так как в двух других «числительное и классификатор разделены существительным» [27]. Айхенвальд, однако, показывает, что пятый порядок (ClNNum) возможен по крайней мере в одном языке – языке эйагхам (бенуэ-конго, Нигерия), где количественный классификатор образует составляющую с существительным, а не с числительным (по крайней мере, просодически) [22].

Первые два из перечисленных выше возможных порядков слов в трехчленной количественной конструкции встречаются намного чаще, чем вторые два. Как было отмечено выше, существуют языки, допускающие варьирование порядка следования числительного и существительного в рамках количественной конструкции. Согласно [27], это только языки, в которых классификатор следует за числительным или является его аффиксом.

Здесь следует также отметить особенности порядка слов в конструкциях с составными числительными, о которых уже было упомянуто выше. Несмотря на то, что чаще всего составные числительные ведут себя как простые и находятся по одну сторону существительного, возможны исключения из этого правила. В шотландском языке, согласно [28], составные числительные могут разрываться существительным. Так, в (13а) слово «мужчина» находится в середине конструкции.

[5] Если в языке присутствует лишь один или несколько случаев исключений из доминирующего правила построения количественных конструкций, такой язык отнесен в данной классификации к первому или второму типу.

ШОТЛАНДСКИЙ ЯЗЫК ([13])

(13) а. ceithir fichead fear 's a naoi-deug
четыре двадцать мужчина плюс девять-десять
девяносто девять мужчин

б. naoi-deug 's ceithir fichead fear
девять-десять плюс четыре двадцать мужчина
девяносто девять мужчин

В [26] также отмечается влияние системы счисления, принятой в языке, на подобные исключения. Так, в языке оболо (бенуэ-конго, Нигерия) составные числительные «десять», «двадцать» и «четыреста» всегда предшествуют модифицируемому существительному, тогда как все остальные следуют за ним. Согласно [26], это связано с тем, что в языке оболо используется двадцатеричная система счисления[6]. В рамках этой системы составные числительные имеют своей базой вышеперечисленные слова. Кроме того, в количественной конструкции с составным числительным в языке оболо его первое слово предшествует существительному, тогда как все остальные его части следуют за существительным.

ЯЗЫК ОБОЛО ([26])

(14) а. óbóp úwù
четыреста дом
четыреста домов

б. úwù ílé ítá
дом большой три
три больших дома

в. étîp úwù mè gò
двадцать дом и пять
двадцать пять домов

Стоит отметить, что в классификации Драйера этот язык отнесен ко второму классу – классу языков с порядком NNum, так как большинство несоставных числительных следуют именно за существительным. Это кажется уместным в большом атласе, целью которого является продемонстрировать статистику по данному вопросу. Разрабатываемая нами база данных, являясь «частной» типологической БД, должна предоставлять возможность пользователю получать данные о том или ином языке вплоть до мелочей, поэтому представляется необходимым рассматривать «разрывность» составных числительных в рамках отдельного параметра и, таким образом, увеличить число стратегий образования количественных конструкций.

2.4. Отношения между членами количественных конструкций

Итак, как было показано выше, состав количественной конструкции (без учета порядка следования ее членов) можно описать по следующей схеме:

(15) N(P) + Num (+ Cl)

Помимо особенностей синтаксиса, описанных выше, в некоторых языках мира наблюдаются нетривиальные синтаксические отношения между членами конструкции.

Ранее для описания отношения числительного (или числительного с классификатором) к существительному мы во всех случаях использовали слово «модифицирует», что может ввести в заблуждение, давая понять, что существительное всегда является главным словом, а числительное (или числительное с классификатором) – модификаторами существительного. На самом же деле отношения между членами количественной конструкции гораздо сложнее.

Г. Корбетт в [32] формализовал свои наблюдения о количественных конструкциях (в основном в славянских языках) в виде так называемых универсалий. Первая универсалия заключается в том, что простые количественные числительные оказываются (по своему поведению) между числительными и существительными. А согласно второй универсалии, если поведение числительных различно, то чем больше число-референт числительного, тем более «субстантивно» ведет себя числительное в количественной конструкции. В монографии [31] Харфорд утверждает, что подобные обобщения все же не являются универсальными и нуждаются в дополнительной проверке и «реорганизации».

Автор приводит два типа базовых синтаксических структур, которые возможны для количественных конструкций (в своей работе он, как и некоторые другие исследователи (см. [33-35] и др.), в общем случае называет их именными группами (Noun Phrase, (NP)). Заметим, что на месте существительного на рис. 1 может находиться не только существительное само по себе, но и любая неколичественная именная группа.

На рис. 1 видно, что, согласно Харфорду, в конструкциях «числительное + существительное» встречается два типа синтаксических связей: согласование и управление, которые Харфорд описывает в традиционных терминах «вершины-модификатора» (в случае согласования) и «контролера» (в случае управления). В ходе согласования вершина определяет падеж, а также иногда число и род модификатора. Контролер же приписывает падеж, но не число или род управляемого слова.

Количественные конструкции, утверждает Харфорд, способны «сочетать» в себе оба типа синтаксической связи. Тогда обе приведенные на рис. 1 структуры не вступают в конфликт, а дополняют друг друга.

На рис. 2 продемонстрировано, как именно оба члена конструкции «числительное + существительное» в русском языке получают свои грамматические признаки. Слово *сосна* употребляется в родительном падеже единственного числа, как того требует числительное *два*, тогда как само это числительное получает признак женского рода, так как слово *сосна* женского рода.

[6] Подробнее о системах счисления см. [29, 30]. С точки зрения формального устройства системы счисления также были рассмотрены в [31].

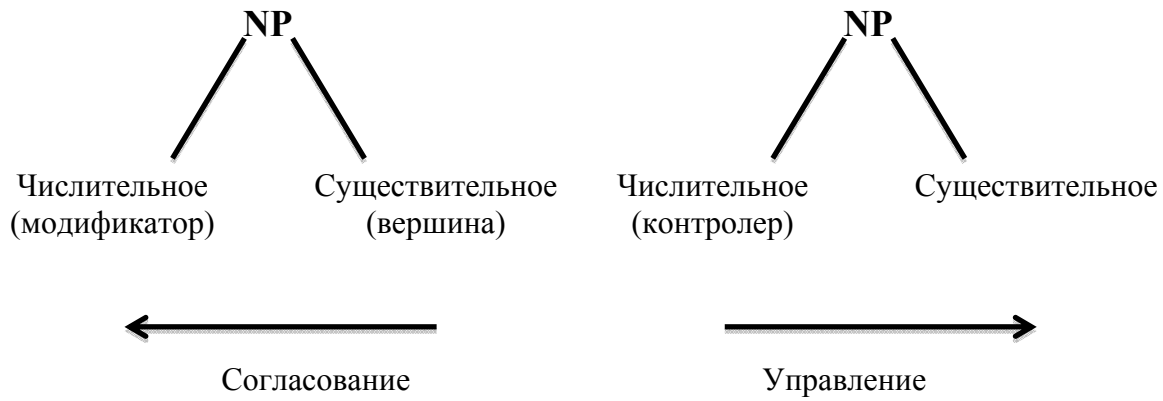


Рис. 1. Две интерпретации синтаксической структуры по Харфорду

В работе [13] предлагается рассматривать отношения между членами конструкций, перечисляя возможные грамматические категории, с помощью которых они могут быть выражены, а именно определенность, падеж, число и род.

Категория определенности приписывается исключительно из семантических или прагматических соображений. Согласно [13], показатель определенности может маркировать как всю именную группу целиком (что выражается в маркировании самого левого или самого правого члена конструкции), так и, реже, числительное внутри группы.

БАСКСКИЙ ЯЗЫК ([13])

(16) а. gizon bat
мужчина один
один мужчина

б. gizon bat-a
мужчина один-DEF
один (этот) мужчина

в. bi gizon
два мужчина
двое мужчин

г. bi gizon-a-k
два мужчина-DEF-PL
эти двое мужчин

БОЛГАРСКИЙ ЯЗЫК ([13])

(17) а. knigi-te
книга.PL-DEF
эти книги

б. dve-te knigi
два.FEM-DEF книга.PL
две эти книги

Так, в баскском языке (изолят) определенность, как и прочие именные категории, всегда маркируется на самом правом члене количественной группы (в (16б) это числительное, тогда как в (16г) – существительное). В болгарском языке один и тот же маркер определенности используется и для числительных, и для существительных, а в случае количественной конструкции он маркирует числительное.

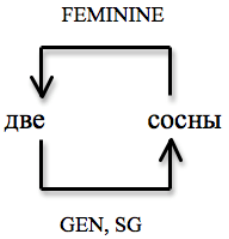


Рис. 2. Два типа синтаксических связей в рамках количественной конструкции в русском языке [36]

Маркирование падежа в тех случаях, когда он приписывается всей группе на основании ее места в предложении, может осуществляться как на обоих членах конструкции (например, в греческом языке (18)), так и на одном из них (в немецком языке (19)).

ГРЕЧЕСКИЙ ЯЗЫК ([13])

(18) а. pendakosi-on jinek-on
пятьсот-GEN женщины-GEN
пятисот женщин

б. tesar-on jinek-on
четыре-GEN женщины-GEN
четырех женщин

НЕМЕЦКИЙ ЯЗЫК

(19) die Freundschaft vier-er Frau-en
DEF дружба четыре-GEN женщина-PL
дружба четырех женщин

Примером, иллюстрирующим приписывание падежа внутри количественной группы, а также согласование числительного с существительным по роду, может служить русская количественная конструкция на рис. 2 (числительное приписывает падеж существительному, в то же время согласуясь с ним в роде).

Числовое маркирование существительных в количественных конструкциях может осуществляться согласно двум противоположным стратегиям (см. [37]).

При использовании первой стратегии существительное в количественной конструкции маркируется числовым показателем в тех случаях, когда речь идет о более чем одном объекте. Вторая стратегия предполагает в этом случае отсутствие числового показателя на существительном. Так, в хантыйском языке (уральская семья) в общем случае существительное внутри количественной конструкции употребляется в единственном числе (20)[7], тогда как числительное в английском языке (21) приписывает существительному множественное число.

ХАНТЫЙСКИЙ ЯЗЫК ([12])
(20) χoɬum jaj
три брат
три брата

АНГЛИЙСКИЙ ЯЗЫК
(21) three brother-s
три брат-PL
три брата

В некоторых случаях числительное в количественной конструкции также может быть маркировано показателем числа. Такое маркирование наблюдается, к примеру, в русском языке в конструкциях типа *одни сани*, где числительное *один* согласуется в числе с существительным, имеющим только форму множественного числа [13].

Стоит отметить, что между маркированием различных грамматических категорий существуют взаимодействия. Так, возвращаясь к примерам (16в,г), заметим, что в неопределенных количественных группах баскского языка именная группа употребляется в единственном числе, в то время как в определенных (кроме числительного «один») – во множественном.

Таким образом, для полного описания количественной конструкции в языке необходимо рассматривать все вышеперечисленные параметры, которые и будут включены в создаваемую БД.

3. СТРУКТУРА И ДИЗАЙН ТИПОЛОГИЧЕСКОЙ БАЗЫ ДАННЫХ

В работе [39] авторы выделяют несколько этапов процесса разработки лингвистической БД, основываясь на подходе разработки БД «сверху вниз» (англ. top-down), предложенном в [40]. Среди этих этапов, помимо непосредственно изучения явления и выделения основных моментов в их описании (чему был посвящен Раздел 2), называются также этапы собственно разработки – разработка абстрактного и логического дизайна ТБД, выбор программных средств для ее проектирования и пользовательских интерфейсов. Эти вопросы и будут освещены в данном разделе.

3.1. Программные средства

Для разработки настоящей ТБД была выбрана модель RDF (Resource Description Framework, см. [41]) и соответствующий ей язык запросов SPARQL. Хотя такая модель БД не является общепринятой и самой популярной среди разработчиков лингвистических баз данных (большая часть существующих на данный момент баз данных в лингвистике имеет реляционную структуру и управляется при помощи одной из реализаций языка SQL), она имеет ряд неоспоримых преимуществ перед реляционными БД и SQL для целей проекта.

ТБД призвана хранить большой объем структурированных данных для каждого языка. Как уже отмечалось в Разделе 2, количественная конструкция состоит из двух или трех членов, а для ее полного описания используется вплоть до 20 параметров (их количество зависит от языка). К тому же, как было показано выше, в одном языке может быть представлено несколько вариантов конструкций, различающихся порядком слов, типом синтаксических отношений между их членами, значением и т. д. Также необходимо место для отглоссированных примеров и данных о языке (подробнее об этом будет сказано ниже). Умножая все это на количество языков, мы получаем очень сложную структуру данных. Кроме того, всегда есть возможность увеличения и преобразования структуры данных с добавлением в ТБД информации о новых языках.

Поэтому решение использовать для хранения таких данных формат RDF, в рамках которого все отношения представлены в виде триплетов «субъект – предикат – объект», описывающих связь, направленную от субъекта к объекту, а не в виде связанных между собой таблиц (как это происходит в реляционных БД), представляется уместным.

Так, на рис. 3 приведен пример одного из триплетов, имеющегося в описываемой в настоящей работе ТБД. Здесь в роли объекта оказывается параметр «Система счисления», в роли субъекта – одно из его значений, «Десятичная система счисления», а в роли связывающего их предиката – отношение «является подклассом».

В такой БД, во-первых, не нужно создавать новую таблицу или использовать внешние ключи для создания нового отношения. Достаточно просто добавить новые триплеты, определив субъекты, объекты и предикаты, как наглядно продемонстрировано в работе С. Морана, посвященной разработке типологической базы данных по фонологии [42].

Во-вторых, запросы к базе данных RDF-формата задаются непосредственно при помощи протокола HTTP (протокола передачи гипертекста, повсеместно используемого в настоящее время для обмена данными в сети Интернет), что позволяет напрямую встраивать их в сервисные архитектуры [43]. Тогда как для интеграции реляционной БД в какой-либо онлайн-сервис потребуется разработка дополнительного программного компонента, что может осложнить работу с БД, а также привести к потере надежности и производительности сервиса.

[7] Согласно [38], в этом случае наблюдается следование принципу экономии и «семантически пустой» показатель числа не употребляется, так как количество референтов существительного уже определено числительным.

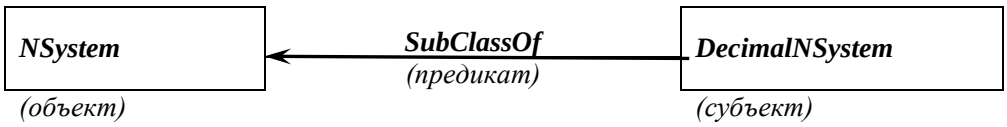


Рис. 3. Представление триплета для отношения «Значение ”Десятичная система счисления” (*DecimalNSystem*) является подклассом (*SubClassOf*) параметра ”Система счисления” (*NSystem*)».

В-третьих, безусловным преимуществом выбранного подхода является возможность более простой интеграции создаваемой ТБД с другими, уже имеющимися ТБД в рамках межуниверситетского проекта Typological Database System ([44]) по объединению всех ТБД в систему с объединенным запросным интерфейсом (о проекте см. [45–47] и др.), а также в прочими лингвистическими ресурсами в рамках проекта Linguistic Linked Open Data Cloud (LLODC, [48]) по интеграции всех лингвистических данных в сети Интернет в единую онтологию (о проекте см. [49]). К тому же уровень стандартизации RDF гораздо выше, чем в случае с реляционными БД, где из-за различий в реализациях языка SQL могут возникнуть сложности с интеграцией данных или переносом их из одного хранилища в другое.

Что касается интерфейса, было принято решение создать два типа максимально наглядных пользовательских интерфейсов: интерфейс для внесения информации в ТБД и запросный интерфейс. Если к первому получают доступ только зарегистрированные пользователи, то второй доступен для всех пользователей сети Интернет.

Для реализации доступа к ТБД использованы средства Django, свободного фреймворка для вебприложений на языке Python ([50]). Выбор в пользу данного фреймворка был сделан в первую очередь из-за принятого его создателями принципа разработки программного обеспечения DRY (англ. *Don't Repeat Yourself*, что означает «не повторяйся»), нацеленного на снижение количества повторений данных, что особенно важно для реализации ТБД, где количество информации разного рода весьма велико.

3.2. Структура ТБД

Настоящая ТБД представляет собой набор описаний количественных конструкций в языках мира. Для каждого языка в БД проектируется онтологическая структура, которую можно условно поделить на три смысловых блока: общая информация о языке, инвентарь параметров (грамматических и семантических признаков и их значений) для этого языка, описание количественной конструкции (или нескольких количественных конструкций) в языке.

На рис. 4 схематично показан фрагмент описания одного языка в ТБД. Прежде чем переходить к подробному описанию каждого из трех блоков, стоит обратить внимание на связи между ними. Линии на рис. 4 не являются отражением конкретных предикатов в RDF-структуре ТБД, а представляют блоки таких предикатов по отношению к блокам объектов[8]. Отметим также, что, в отличие от блоков информации о языке и описания количественных конструкций, блок инвентаря параметров является вспомогательным – он служит для ограничения круга значений грамматических и некоторых семантических признаков для конкретного языка, из которых пользователь сможет выбирать, внося в ТБД информацию о количественной конструкции в этом языке.

В блок общей информации о языке входят такие характеристики языка, как его альтернативные названия, генетическая и ареальная принадлежность, а также его код в справочнике «Этнолог: Языки Мира» (*Ethnologue: Languages of the World*, [51]). Для описания ареальной принадлежности языка в ТБД предусмотрено четыре поля – *area* («ареал»; значения Африка, Америка, Азия, Европа или Океания), *subarea* («субареал»; к примеру, для макрорегиона «Африка»: «Восточная Африка», «Северная Африка», «Южная Африка», «Западная Африка» и «Центральная Африка»), *country* («страна») и *region* («регион»; используется для уточнения региона распространения языка внутри страны). Данные для заполнения перечисленных полей взяты из онлайн-версии упомянутого выше справочника. Четвертый параметр – «регион» – может быть также уточнен пользователем.

Информация о генетической принадлежности языка взята из электронного каталога языков мира «Глоттолог» (*Glottolog*, [52]), где генетическая информация представлена в виде иерархической структуры (от семьи к языку). Предпочтение было отдано именно этому электронному ресурсу, так как он предоставляет данные о генетической классификации языков в формате RDF, что способствует их интеграции в настоящую ТБД.

Как уже было упомянуто выше, второй блок, блок инвентаря параметров, носит вспомогательный характер и содержит перечень некоторых особенностей конкретного языка и значений грамматических и семантических признаков числительных и существительных, необходимых для описания количественных конструкций. Блок включает в себя такие параметры, как *classifier* («классификатор», т. е. наличие или отсутствие классификатора в языке), *numeral system* («система счисления»), а также параметры возможных значений категорий числа, падежа, рода, определенности и посессивности (при условии наличия указанных категорий в языке).

[8] Так, не существует триплета вида, к примеру, *language – has – general_language_information* («язык – иметь – общая информация о языке»), но есть триплет вида *language – belong_to – language_family* («язык – принадлежать – языковая семья»), где объект «языковая семья» входит в блок общей информации о языке.

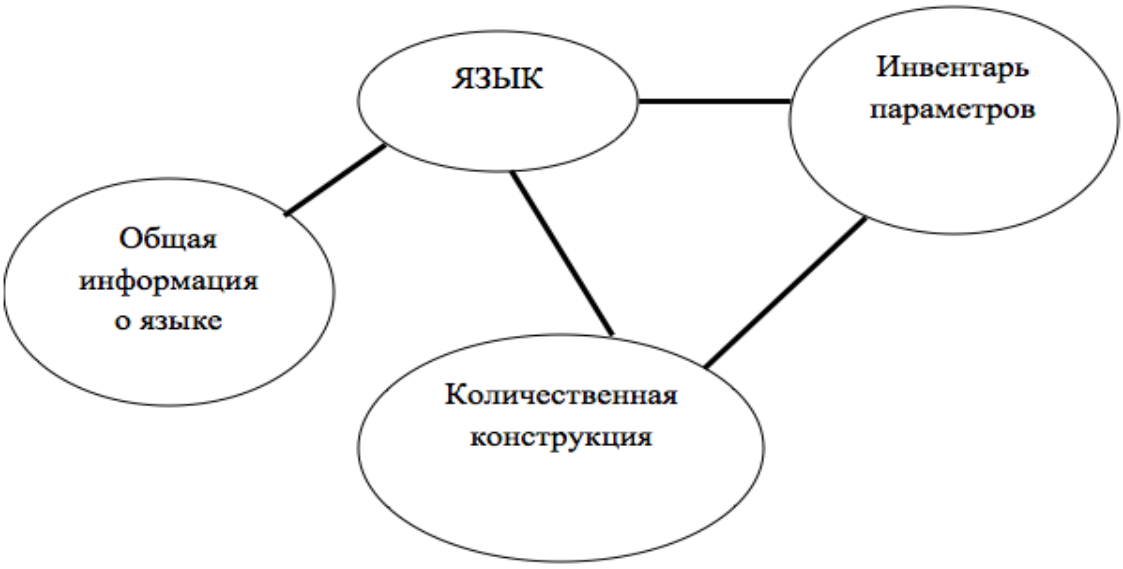


Рис. 4. Схема построения части ТБД количественных конструкций (для одного языка)

Решение в пользу наличия такого блока в структуре ТБД было принято в первую очередь для удобства работы пользователя, вносящего в базу данных новую информацию. Ограничения, накладываемые выбором значений для большей части параметров на этапе характеристики языка в целом, задают круг возможных значений параметров для описания непосредственно количественной конструкции. Как именно происходит процесс ввода языковых данных в ТБД и каким образом ограничение на значения признаков облегчает работу пользователей, будет описано ниже, когда речь пойдет о пользовательских интерфейсах.

Однако стоит также заметить, что наряду с этим наличие блока инвентаря параметров позволяет нам собрать еще один структурированный набор типологических данных, который можно назвать своего рода базой данных возможных значений грамматических и семантических признаков числительных и существительных в языках мира. И, хотя создание такого набора не является целью описываемой ТБД, он также может послужить хорошим источником информации для исследователей.

Наконец, блок описания количественной конструкции включает в себя данные о количественной конструкции в заданном языке. Таких блоков может быть несколько в зависимости от количества реализаций таких конструкций. Так, например, в Разделе 2 данной работы упомянуты языки акома (11) и ниас (12), которые имеют два варианта реализации количественных конструкций, различающиеся порядком слов. Соответственно, для описания количественных конструкций в каждом из этих языков в ТБД будет создано как минимум по два блока.

В рамках этого блока дается описание грамматических и семантических признаков каждого члена конструкции (включая предлог и классификатор, если они есть), а также направление связи между ними. В терминах триплетов отношения между членом конструкции и значением того или иного признака на нем описывается при помощи предиката *has_value* («имеет значение»). Если же член конструкции приписывает некое значение другому ее члену (как числительное *две* приписывает единственное число и генитив существительному *сосны* на рис. 2), для описания используется предикат *assigns_value* («приписывает значение»), а в случае согласования – предикат *shares_value* («распространяет значение»).

Помимо синтаксических ограничений на употребление того или иного члена конструкции в той или иной форме существуют также семантические ограничения. Список семантических признаков, которые могут понадобиться для описания тех или иных ограничений, не всегда легко определить заранее, поэтому мы не ставим задачу описать их все в рамках блока инвентаря параметров, а определяем их при необходимости уже в блоке описания конструкций. Например, такая необходимость может возникнуть при описании количественной конструкции в мужевском говоре коми-зырянского языка[9], когда множественное число в количественной конструкции приписывается только существительному со значением «человеческое существо», тогда как всем остальным существительным приписывается единственное число.

Нередко для описания количественной конструкции также необходимо выбрать значения числительного, так как от этого может зависеть синтаксис количественной конструкции. Так, в славянских языках значение числительного влияет на синтаксические отношения между членами конструкций (см. [32]). Например, в русском языке синтаксически различаются конструкции с числительным *один*, *два, три* и *четыре* (и всеми составными числительными, окан-

[9] Говор коми-зырян села Мужи Шурышкарского р-на Ямало-Ненецкого автономного округа.

чивающимися на перечисленные слова) и всеми остальными числительными.

Также в описании каждой количественной конструкции имеются поля для комментариев пользователей, ссылок на источник информации[10] и отглоссированных языковых примеров. Мы призываем исследователей использовать Международный фонетический алфавит (IPA, см. [53]) для фонетической транскрипции языковых примеров и Лейпцигскую систему правил глоссирования ([54]), чтобы примеры были единообразными и понятными как можно большему кругу исследователей. В рамках настоящей ТБД поддерживается одна из общепринятых и стандартизированных кодировок текста, UTF-8, в рамках которой есть возможность ввода большого количества языковых символов, которые встречаются в различных языках мира.

Перейдем теперь к описанию интерфейсов ТБД и возможностей, которые получат ее пользователи.

3.3. Интерфейсы ТБД

Как уже указывалось выше, для удобного доступа пользователя к ТБД было принято решение создать два пользовательских интерфейса: интерфейс ввода данных и интерфейс запросов к ТБД. Первый из упомянутых интерфейсов доступен только зарегистрированным пользователям (это необходимо, чтобы контролировать ввод данных и следить за их достоверностью), тогда как второй открыт для всех пользователей интернета.

3.3.1. Интерфейс ввода данных

Итак, после того как пользователь прошел регистрацию (с указанием как минимум своего логина, инициалов, адреса электронной почты и места работы/учебы), он получает доступ к редактированию БД, и каждое внесенное им изменение будет записано в БД с пометкой в виде логина пользователя и времени редактирования. Это поможет управлять данными как редакторам ТБД, так и самому пользователю.

Каждая страница интерфейса представляет собой набор полей, которые пользователю необходимо заполнить. В большинстве случаев это заполнение сводится к выбору подходящих значений каждого параметра из выпадающего списка. Только в тех случаях, когда по тем или иным причинам нужного значения нет в списке, пользователь имеет возможность добавить его, указав в комментарии причину такого добавления. Для параметров вида «наличие/отсутствие» (например, *Classifier* («Классификатор»)) пользователю необходимо выбрать один из двух вариантов, отметив его флажком. Существенно отличается только интерфейс выбора порядка слов в количественной конструкции, описание которого приведено ниже.

В качестве первого шага пользователю предлагается ввести общую информацию о языке. Ему необходимо выбрать язык, а также ареал, субареал(ы) и страну или страны его распространения из имеющихся списков, полученных из каталога «Этнолог: Языки Мира» [51], и, если это необходимо, уточнить регион распространения языка. Код языка по упомянутому каталогу, а также генетическая информация о языке заполняются автоматически (с использованием RDF-представления каталога «Глоттолог» ([52])).

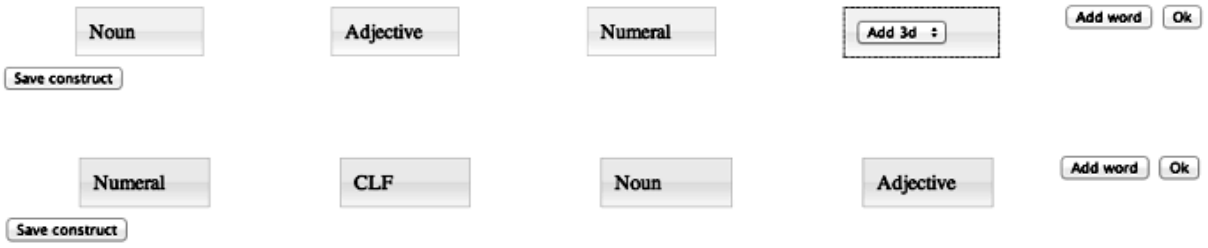
После ввода общей информации о языке пользователю предлагается ограничить инвентарь параметров, которыми он сможет воспользоваться при описании количественных конструкций. На этом этапе необходимо отметить флажками наличие или отсутствие в описываемом языке классификаторов, категорий посессивности, одушевленности, определенности, рода, числа и падежа, выбрать возможные значения для ряда категорий (например, выбор значений для категории числа происходит из следующего списка: *singular* («единственное»), *dual* («двойственное»), *trial* («тройственное»), *paucal* («паукальное»), *plural* («множественное»), *general* («общее»)), а также значение типа системы счисления (*decimal* («десятеричная»), *vigesimal* («двадцатеричная») или *other* («другая»), причем последний параметр не может быть выбран без пояснения). Когда все предложенные поля заполнены, пользователь может переходить к описанию непосредственно количественной конструкции.

Интерфейс описания количественных конструкций отличается от описанного выше тем, что сначала пользователю предлагается определить порядок слов в конструкции при помощи перемещения прямоугольников с названиями членов конструкции в ряду друг относительно друга. Для описания некоторых условий употребления (например, разрыва составного числительного в языке оболо (14)) существует возможность добавления в строку дополнительных прямоугольников (рис. 5).

После выбора порядка слов пользователю необходимо указать связи между ними при помощи стрелок, а также выбрать или ввести значения грамматических и семантических признаков для каждого из членов конструкции. Выбор значений, как уже было упомянуто выше, будет осуществляться из инвентаря, указанного пользователем на предыдущем этапе ввода данных в ТБД, путем выбора значений из выпадающих списков.

Отличается в данном случае выбор возможных значений для числительного. Пользователь имеет возможность выбрать как диапазон числительных, так и набор отдельных числительных. Так, для описания четырех типов количественных конструкций в русском языке пользователь выберет числительные и диапазоны «1 и все составные числительные, оканчивающиеся на 1», «2 и все составные числительные, оканчивающиеся на 2», «3, 4 и все составные числительные, оканчивающиеся на 3 или 4», «все числительные, кроме вышеперечисленных». Сделать свой выбор пользователь сможет, выбрав числительные или их диапазоны на числовой матрице (рис. 6).

[10] В тех случаях, когда те или иные данные получены не самим пользователем в ходе изучения языка, а из уже опубликованного источника, мы ожидаем, что пользователь укажет этот источник, чтобы другие также могли обратиться к нему в том случае, если их так или иначе заинтересуют эти данные.



а.

б.

Рис. 5. Выбор порядка слов в интерфейсе ТБД для количественных конструкций в языках а. оболо (см. (14б)) и б. ниас (см. (12а))

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Рис. 6. Фрагмент числовой матрицы, описывающей третий тип количественных конструкций в русском языке (*три сестры, четыре мальчика*).

После того как выбраны параметры и, при необходимости, прокомментирована структура количественной конструкции, следует дополнить описание хотя бы одним языковым примером, для которого в интерфейсе предусмотрено специальное поле. После этого описание считается законченным, и, если пользователь уверен в своих данных и не хочет больше вносить исправления, он может сделать информацию доступной другим пользователям при помощи кнопки *Submit* («Подтвердить»).

Если же в языке существуют другие способы выражения количественной конструкции, пользователю предлагается добавить еще одно или несколько описаний, воспользовавшись кнопкой *Add NNC* («Добавить количественную конструкцию») в интерфейсе описания количественных конструкций, и таким же образом заполнить поля для нее.

3.3.2. Интерфейс запросов к ТБД

Интерфейс запросов к ТБД представляет собой форму для выбора признаков всех параметров, имеющихся в ТБД, из выпадающих списков. При запросе к БД можно выбирать один признак, а можно комбинировать их.

Например, для того чтобы получить ответ на вопрос «В каких языках есть классификаторы?», достаточно выбрать значение *Yes* («Да») для параметра *Classifier* («Классификатор») и подтвердить запрос при помощи кнопки *Submit* («Подтвердить»). А для того чтобы узнать, к примеру, в каких языках уральской семьи существительное в количественной конструкции может стоять в форме множественного числа, пользователю необходимо выбрать значение *Uralic* («Уральская») для параметра *Language Family* («Языковая семья»), а затем – значение *Plural* («Множественное число») для параметра *Number (Noun)* («Число существительного»).

Как уже отмечалось выше, в ТБД также будет возможным получить информацию о наличии/отсутствии тех или иных значений грамматических категорий имен существительных и числительных в языках, представленных в базе данных. Так, пользователь может задавать запросы, не относящиеся напрямую к структуре количественных конструк-

ций, например, «В каких языках есть паукальное число?» или «В каких языках категория рода имеет два значения?».

ЗАКЛЮЧЕНИЕ

В данной работе была описана Типологическая база данных количественных конструкций, а также были освещены некоторые особенности феномена количественной конструкции, требующие внимания при его всестороннем описании. Настоящая ТБД призвана служить не только и не столько готовым источником информации о конкретном явлении, сколько именно инструментом, позволяющим исследователям делиться информацией с коллегами, используя унифицированную модель описания. На данный момент база данных находится на стадии тестирования и доступ к ней открыт ограниченному числу пользователей. Однако мы надеемся в скором времени выложить базу данных в открытый доступ и тем самым продолжить развитие проекта по сбору информации о количественных конструкциях в языках мира.

СПИСОК ЛИТЕРАТУРЫ

1. Haspelmath M. The typological database of the World Atlas of Language Structures // The Use of Databases in Cross-Linguistic Studies. – 2009. – Т. 41. – С. 283.
2. Brown D. P., Tiberius C., Corbett G. G. A typological database of agreement. – 2002. // Proceedings of the Third International Conference on Language Resourses and Evaluation. – Т. 4. – С. 1843-1846.
3. The use of databases in cross-linguistic studies / ed. M. Everaert, S. Musgrave, A. Dimitriadis. – Walter de Gruyter, 2009. – Т. 41.
4. Zweig E. Nouns and adjectives in numeral NPs // PROCEEDINGS-NELS. – 2005. – Т. 35, №. 2. – С. 663.
5. Bošković Ž. Case and agreement with genitive of quantification in Russian // Agreement systems. – 2006. – С. 99–121.
6. Mittendorf I., Sadler L. Numerals, nouns and number in Welsh NPs // Proceedings of the LFG05 Conference. CSLI Publications. Stanford, California. – 2005.
7. Lee C. Numeral classifiers, (in-)definites, and incremental themes in Korean // Ms., Seoul National University, Seoul, Korea. – 1999.
8. Graz Reduplication Database. – URL: http:// reduplication.uni-graz.at/ db.html (дата обращения – 04.07.2013).
9. НОСС: Новый объяснительный словарь синонимов русского языка. 2-е изд. / под общим рук. Ю. Д. Апресяна. – М.: Языки славянской культуры, 2004.
10. Гращенков П. В. Родительный падеж при русских числительных: Типологическое решение одной "сугубо внутренней" проблемы // Вопросы языкознания. – 2002. – №. 3. – С. 74–119.
11. Янко Т. Е. Русские числительные как классификаторы существительных // Русский язык в научном освещении. – 2002. – Т. 3, №. 1. – С. 168–181.
12. Шматова М. С. Конструкции «числительное – существительное» в хантыйском языке: случаи маркирования показателями числа //Acta Linguistica Petropolitana. – СПб: Наука. – 2010. – Т. 4. – Ч. 3.
13. Hurford J. The interaction between numerals and nouns // Noun phrase structure in the languages of Europe. – Walter de Gruyter. – 1998. – С. 561-620.
14. Шведова Н. Ю. Русская грамматика. Т. 2. – М.: Наука, 1980.
15. Thorne D. A. A comprehensive Welsh grammar. – Cambridge, MA: Blackwell, 1993.
16. Lander Y. Western Indonesian prenominal modifiers and compositional obligatoriness // Proceedings of the International Conference on Far East, South-East Asia and West Africa Languages. – 2009. – С. 242-257.
17. Плунгян В. А. Общая морфология: Введение в проблематику.– М.: Эдиториал УРСС, 2000.
18. Downing P. Numeral classifier systems: The case of Japanese. – John Benjamins Publishing, 1996. – Т. 4.
19. Csirmaz A., Dékány É. Hungarian classifiers // Unpublished manuscript, University of Utah and University of Tromsø, 2010.
20. Bisang W. Classifiers in East and Southeast Asian languages: Counting and beyond // Trends in Linguistic Studies and Monographs. – 1999. – Т. 118. – С. 113–186.
21. Greenberg J. H. Numeral Classifiers and Substantival Number: Problems in the Genesis of a Linguistic Type // Working Papers on Language Universals. – 1972. – № 9.
22. Aikhenvald A. Y. Classifiers: A Typology of Noun Categorization Devices. – Oxford University Press, 2000.
23. The World Atlas of Language Structures Online. – URL: http://wals.info/ (дата обращения: 05.07.2013).
24. Gil D. Numeral classifiers // The World Atlas of Language Structures Online. – URL: http://wals.info/chapter/55/ (дата обращения: 05.07.2013).
25. Nomoto H. Number in Classifier Languages. – University of Minnesota, 2013.
26. Dryer M. S. Order of numeral and noun // The World Atlas of Language Structures Online. – URL: http://wals.info/chapter/89/ (дата обращения: 05.07.2013).
27. Greenberg J. H. On language: selected writings of Joseph H. Greenberg. – Stanford University Press, 1990.

28. Paterson J. M. Gaelic made easy: A gaelic guide for beginners. – Glasgow, 1968.
29. Comrie B. Some problems in the theory and typology of numeral systems // Proceedings of LP'96. Praha. – 1997. – T. 56. – C. 41-56.
30. Comrie B. Endangered numeral systems // Bedrohte Vielfalt: Aspekte des Sprach(en)tods [Endangered diversity: Aspects of language death]. – 2005. – C. 13–15.
31. Hurford J. R. Language and number: The emergence of a cognitive system. – Oxford, England: Basil Blackwell, 1987.
32. Corbett. G. G. Universals in the syntax of cardinal numerals // Lingua. – 1978. – T. 46. – № 4 – C. 355–368.
33. Мельчук И. А. Поверхностный синтаксис русских числовых выражений: The surface syntax of Russian numeral expressions. – Institut für Slavistik der Universität Wien. – 1985.
34. Babby L. H. Case, prequantifiers, and discontinuous agreement in Russian // Natural Language & Linguistic Theory. – 1987. – T. 5, №. 1. – C. 91–138.
35. Pesetsky D. Path and categories. – MIT, Cambridge, 1982.
36. Corbett G. G. Agreement: a partial specification, based on Slavonic data // Linguistics. – 1986. – T. 24, №. 6. – C. 995–1024.
37. Corbett. G. G. Number. Cambridge textbooks in linguistics. – Cambridge: CUP – 2000.
38. Ionin T., Matushansky O. The composition of complex cardinals // Journal of Semantics. – 2006. – T. 23, №. 4. – C. 315–361.
39. Dimitriadis A., Musgrave S. Designing linguistic databases: A primer for linguists // The Use of Databases in Cross-Linguistic Studies. – 2009. – T. 41. – C. 13.
40. Connoly N., Hegarty M. Review of ICZM Methodologies. – The Coastal Resourses Centre, University College Cork, Ireland. – 1999.
41. Lassila O. et al. Resource Description Framework (RDF) model and syntax specification. – 1998. [Электронный ресурс] – URL: http://www.w3.org/TR/WD-rdf-syntax (дат обращения: 05.07.2013).
42. Moran S. Using linked data to create a typological knowledge base // Linked Data in Linguistics. – Springer Berlin Heidelberg, 2012. – C. 129–138.
43. Головков В., Портнов А., Чернов В. RDF – инструмент для неструктурированных данных // Открытые системы. – 2003. – T. 9. – C. 46–69.
44. The Typological Database System project. – 2009. – URL: http://languagelink.let.uu.nl/tds/index.html/ (дата обращения: 04.07.2013).
45. Monachesi P. et al. The typological database system // Proceedings of the IRCS workshop on linguistic databases. – 2001. – C. 181–186.
46. Dimitriadis A., Monachesi P. Integrating different data types in a Typological Database System // Proceedings of the International Workshop on Resources and Tools in Field Linguistics, Las Palmas, Spain. – 2002.
47. Dimitriadis A. et al. How to integrate databases without starting a typology war: The Typological Database System // The Use of Databases in Cross-Linguistic Studies. – Walter de Gruyter. – 2009. – C. 155–207.
48. The Linking Open Data cloud diagram. – URL: http://lod-cloud.net/ (дата обращения: 29.07.2013).
49. Chiarcos C., Hellmann S., Nordhoff S. Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group // TAL. – 2011. – T. 52. – №. 3. – C. 245–275.
50. Django Software Foundation. – URL: https://djangoproject.com (дата обращения 01.2013).
51. Ethnologue: Languages of the World, Seventeenth edition. – URL: http://www.ethnologue.com (дата обращения: 05.07.2013).
52. Glottolog 2.0. – URL: http://glottolog.org (дата обращения: 28.07.2013).
53. International Phonetic Association (ed.). Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. – Cambridge University Press, 1999.
54. Conventions for interlinear morpheme-by-morpheme glossing. – URL: http://www.eva.mpg.de/lingua/resources/glossing-rules.php (дата обращения: 05.07.2013).

Материал поступил в редакцию 09.07.13.

Сведения об авторе

ШМАТОВА Мария Сергеевна – аспирант Московского государственного университета им. М.В. Ломоносова. Филологический факультет, кафедра теоретической и прикладной лингвистики

e-mail: mashashma@gmail.com

БАЗА ДАННЫХ ВИНИТИ РАН

ВИНИТИ предлагает к использованию через WWW-сервер (**http://www.viniti.ru**) крупнейшую Федеральную базу отечественных и зарубежных публикаций по естественным, точным и техническим наукам. БД ВИНИТИ РАН генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. БД ВИНИТИ представлена ретроспективными тематическими фрагментами и единой политематической БД (ретроспектива с 2001 г.), объединяющей все тематические фрагменты БД ВИНИТИ.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ – http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов могут быть предоставлены **на CD-ROM в поисковой системе (ИПС) "Сокол"**, обеспечивающей все поисковые функции, доступные в режиме on-line:

- Поиск можно вести в годовом или ретроспективном массиве (за несколько лет сразу) в одном или нескольких тематических фрагментах .
- Поиск по словам и любым словосочетаниям из заглавия, реферата, ключевых слов.
- Использование года, языка, рубрик, шифров тематических разделов БД для уточнения поиска.
- Поиск по словарю, выполняющему функции многоаспектного указателя, в том числе авторского, предметного, источников, индексов МПК, номеров патентных документов и депонированных рукописей и т.д.
- Возможность запоминания запросов для последующего их использования и/или редактирования.
- Чтение документов не только как в РЖ (последовательный просмотр документов одного номера за другим), но и чтение документов нужных тематических фрагментов (разделов) по оглавлению за весь период заказанной ретроспективы.

ИПС "Сокол" является прикладной программой Microsoft Windows.

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов могут быть подготовлены **в коммуникативных форматах** ISO-2709, МЕКОФ, txt на любых видах электронных носителей.

Продукты предоставляются на договорной основе.

Информационная служба БД ВИНИТИ: 125190, Москва, ул. Усиевича 20, ВИНИТИ
Телефон: (499) 155-45-01, 155-45-02, **Факс:** (499) 152-62-31 **e-mail:** csbd@viniti.ru

УВАЖАЕМЫЕ ЧИТАТЕЛИ!

ЦЕНТР НАУЧНО-ИНФОРМАЦИОННОГО ОБСЛУЖИВАНИЯ ВИНИТИ РАН

ПРЕДОСТАВЛЯЕТ КОПИИ ПЕРВОИСТОЧНИКОВ

ВИНИТИ РАН осуществляет обслуживание копиями первоисточников, хранящихся в фонде научно-технической литературы ВИНИТИ, в фондах других библиотек, а также в доступных ВИНИТИ электронных ресурсах.

Фонд научно-технической литературы ВИНИТИ включает более 2 млн изданий по точным, естественным и техническим наукам, в т.ч.:

– отечественные и иностранные периодические и продолжающиеся издания – с 1987 г.;

– отечественные книги – с 1987 г.;

– иностранные книги – с 1991 г.;

– рукописи, депонированные в ВИНИТИ, – с 1962 г.

Заказы на бумажные или электронные копии первоисточников принимает Центр научно-информационного обслуживания (ЦНИО) ВИНИТИ. ЦНИО ВИНИТИ обслуживает коллективных (организации и учреждения) и индивидуальных пользователей.

Формы обслуживания:

– абонементная (на основе договоров и предоплаты);

– разовые заказы (с предоплатой заказа по счету);

– индивидуальная форма обслуживания в читальном зале ЦНИО ВИНИТИ.

На сайте ВИНИТИ (http://www.viniti.ru) представлен полный Электронный каталог научно-технической литературы (http://catalog.viniti.ru), зарегистрированной в ВИНИТИ с 1994 г. Доступ для просмотра и поиска по Каталогу свободный. Постоянные абоненты ЦНИО ВИНИТИ, имеющие логин и пароль для работы с Каталогом, могут делать заказ копий непосредственно через Каталог.

Услуги по изготовлению копий первоисточников из фондов других библиотек предоставляются только постоянным абонентам. Место хранения первоисточников указывается в Электронном каталоге.

За подробной информацией обращаться по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН, ЦНИО

Телефоны: 8 (499)155-42-43, 155-42-09, 152-54-59

Факс: 8 (499) 943-00-60

E-mail: cnio@viniti.ru; **URL:** http://www.viniti.ru