7. Madnick, S. Wang, R. and Xian, Xiang. The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality//Journal of Management Information Systems. 2004. Vol. 20. № 3. Pp. 41-69.

8. Свидетельство о государственной регистрации программ для ЭВМ № 2010615437 от 24.08.2010 АИС «Межевик» (Роспатент).

9. Грищенко В.С. Метрики репутации: построение открытых информационных сред: Дис. … канд. физ.-мат. наук. Екатеринбург, 2007. 124 с.

Александр Анатольевич ЗАХАРОВ —
зав. кафедрой информационной безопасности,
доктор технических наук, профессор
azaharov@utmn.ru

Ольга Андреевна НЕСТЕРОВА —
аспирантка кафедры информационной безопасности
o-nesteroma@mail.ru

Евгений Александрович ОЛЕННИКОВ —
доцент кафедры информационной безопасности,
кандидат технических наук
olennikov@utmn.ru

Институт математики и компьютерных наук
Тюменского государственного университета

УДК 004.852

*АЛГОРИТМ ИНФОРМАЦИОННОГО ПОИСКА В МЕДИЦИНСКИХ АРХИВАХ НА ОСНОВЕ КОНТЕКСТНО-ВРЕМЕННОЙ ОНТОЛОГИИ**

ALGORITHM OF INFORMATION RETRIEVAL IN MEDICAL RECORDS BASED ON CONTEXT-TEMPORAL ONTOLOGY

АННОТАЦИЯ. *Рассматривается использование онтологии для решения проблемы информационного поиска медицинских данных, предлагается алгоритм поиска на основе разработанной семантической модели поиска для поддержки медико-биологических исследований.*

SUMMARY. *The use of ontology for the solution of medical data retrieval problem is considered, the algorithm of semantic retrieval is proposed for medical-biological research support.*

КЛЮЧЕВЫЕ СЛОВА. *Информационный поиск, медицинская информационная система, временная семантика, онтология, контекст*

KEY WORDS. *Information retrieval, medical information system, temporal semantics, ontology, context.*

Повышение качества и доступности медицинской помощи — один из приоритетов государственной политики, благодаря которой развиваются такие направления, как скрининговая, восстановительная и доказательная медицины [1], [2]. Эти направления основываются на результатах медико-биологических

* Работа выполнена при поддержке гранта Министерства образования и науки РФ «Проведение научных исследований в области экологии языка и смежных наук» (ГК № 02.740.11.0594).

исследований (МБИ), которые во многом ориентированы на поиск, сбор и обработку данных о здоровье и лечении пациентов за несколько лет, что обуславливается актуальностью исследований, связанных с оптимизацией механизмов поиска информации в медицинских архивах [3] (рис. 1).

Рис. 1. Область применения контекстно-временной онтологии

Извлечение нужных данных из медицинских архивов осложнено и тем, что информация только частично формализована, а частично представлена в виде произвольного текста. Это не позволяет использовать простой «точный поиск» не учитывающий смысловое значение терминов и информационную потребность пользователя [4].

Отметим, что разнородность, многозначность, неточность, субъективность и неформализованное представление медицинской информации также осложняет возможности использования технологий семантического поиска на основе онтологий, которые позволяют распознавать смысл [5]. Принципиально важно отметить, что при использовании онтологий для поиска данных, например, в массиве электронных историй болезни, приходится учитывать ряд особенностей — онтология должна быть большой, а понятия онтологии должны быть связаны с языковыми единицами — медицинскими терминами. При этом, поскольку онтология для любой реальной предметной области не может быть полной, то необходимо дополнительно использовать пословные методы в едином поисковом механизме.

При разработке алгоритма поиска в архиве электронных историй болезни нами были учтены не только информационные потребности пользователя и его субъективные знания, которые в настоящий момент времени не могут быть выражены в запросе к информационно-поисковой системе, но и такие факторы как «естественный» язык в описании истории болезни, зависимость

смысла терминов от контекста и времени, полнота коллекции подходящих для настройки алгоритма документов.

В отличие от обычного поиска, требующего мгновенного точного ответа, при поиске данных в медицинских архивах мы придерживались следующих приоритетов: во-первых, полнота поиска, во-вторых, точность поиска, в-третьих, отслеживание новых документов в течение длительного времени (возможно, в течение нескольких лет).

Алгоритм семантического поиска (обучение с учителем):

ШАГ 1. Подготовка коллекции документов. Простое общее индексирование по словам. Структура индекса: документ-предложение-слово. Автоматическое построение онтологии по обучающей коллекции документов.

ШАГ 2. Обучение системы на основе обучающей выборки. Оценка неопределенности поиска. Эксперт вводит необходимые правила. Определение множества терминов, отношений, функций интерпретации. Создание онтологий.

ШАГ 3. Контекстное индексирование — в соответствии с полученной онтологией (полученным набором терминов и отношений).

ШАГ 4. Выбор схожих документов. Оценка релевантности документов.

Онтология определяет общий словарь для пользователя и поисковой системы, а также машинно-интерпретируемые формулировки основных понятий предметной области и отношения между ними. Онтологии могут быть представлены в виде набора ориентированных мультиграфов. Для выбора схожих документов используется графовый алгоритм кластеризации [4], [6].

Определим понятие модели *контекстно-временной онтологии (КВО)* следующим образом. Пусть:

$X = \{x_i\}$ — конечное множество терминов (слово или словосочетание), определяющих *элементы поиска* ($i=\overline{1,M}$);

$R=\{r_k\}$ — конечное множество терминов (слово или словосочетание), определяющих *связи между элементами поиска* ($k=\overline{1,K}$);

$cr_{kl}(t) \to R_{[-1;1]}$ — *конечное множество функций факторов достоверности отношений в момент времени t*, возвращающее в любой момент времени значение в интервале [-1;1]: для *i*-го и *j*-го термов: -1 — абсолютно невероятно; (-1;0) — отношение маловероятно; 0 — неизвестно; (0;1) — достоверно в некоторой степени; 1 — отношение достоверно на 100%;

$$cr_{\mathrm{k}}(\mathrm{t}) = \begin{cases} cr_{\mathrm{kh}}(t), t \in [t_{kh}^{(1)}; t_{kh}^{(2)}]; \\ 0, иначе. \end{cases} \quad (t_{kl}^{(1)}; t_{kl}^{(2)}] \cap (t_{kp}^{(1)}; t_{kp}^{(2)}] = \varnothing, \forall l \neq p; \quad h,l,p = \overline{1,H},$$

H — количество временных интервалов.

Тогда можем определить:

$R_c(t) = \langle x_i, x_j, r_k, cr_k(t) \rangle$ — *конечное множество контекстно-временных отношений* между терминами ($i,j=\overline{1,M}$; $k=\overline{1,K}$).

$F_c(t)=\langle F_n(N,t), F_s(t), F_l(t) \rangle$ — множество функций интерпретации, где: $F_n(N,t)$ — контекстная нормализация терминов; $F_s(N,t)$ — контекстная интерпретация термов; $F_l(t)$ — правила выводов.

$F_n(N,t)$ в любой момент времени для *i*-го термина возвращает номер *j*-го терма, определяющий элемент поиска с максимальным фактором достоверности:

$$F_n(N,t) \to N \ : \forall i = \overline{1,M}, i \in N \quad , \forall t_0 \quad F_n(i,t_0) \to \arg\max_{\forall j=\overline{1,M}}(cl_{ij}(t_0)),$$

где $cl_{ij}(t) \to R_{[-1;1]}$ — *факторы достоверности* того, что термины x_i и x_j определяют один и тот же элемент поиска (принадлежат одному семантическому полю) в момент времени t.

$F_s(N,t)$ в момент времени j-му терму ставит в соответствие вектор $V=\{v_{ij}\}$ коэффициентов уверенности, отражающих степень соответствия i-го термина j-му терму $(i=\overline{1,M})$.

$$F_s(N,t) \to R^M_{[-1;1]}: \quad \forall i = \overline{1,M}, i \in N \quad , \forall t_0, F_s(i,t_0) \to < v_{ij} >, j = \overline{1,M}.$$

$$F_s(j,t_0) = E_j \bullet cl(t_0), \text{ где } E_j = \{e_{lp}\}, \text{ матрица } M \times M,$$

$$(\text{j-й столбец} = 1); \ e_{lp} = \begin{cases} 1, \ p = j, \forall l; \\ 0, \ p \neq j, \forall l. \end{cases}$$

F_l(t) используется для построения правил выводов на определенный момент времени:

F_l = *ЕСЛИ<И({x1,c1,t1},{x2, c2 ,t2},...,{xn, cn ,tn})|ИЛИ({x1,c1 , t1},{x2, c2, t2},..., {xn, cn, tn}) | НЕ({x1,c1 , t1},{x2, c2 , t2},...,{xn, cn , tn}) > ТО <({y1, cy1 ,ty1},{y2, cy2 ,ty2},..., {ym, cym, tym})>*

В результате получаем *модель контекстно-временной онтологии*:

$$O = <X, R_c(t), F_c(t)>.$$

Описанный алгоритм реализован нами в программном комплексе для поиска документов в информационном хранилище Тюменского кардиологического центра. Ниже описывается один из численных экспериментов.

Врач-исследователь представил выборку из 200 подходящих для исследования историй болезни из более 7000 историй за предыдущий год. На создание этой выборки им было потрачено 7 месяцев. Необходимо было отобрать истории болезни, где в анамнезе (произвольное описание истории развития заболевания) было упомянуто, что пациенту была назначена антикоагулянтная терапия.

Выборка была разделена на обучающую и контрольную.

После пятой итерации 63% документов обучающей выборки соответствовали запросу с уверенностью 100%, остальные 37% документов — с уверенностью 90%. Для уровня 90% коэффициенты полноты и точности поиска по обучающей выборке равны 1.

Для тестирующей выборки коэффициент полноты оказался равен 0,95 (95% документов были признаны соответствующими запросу с уровнем уверенности 90%), коэффициент точности равен 0,9 (90% документов, признанные системой соответствующими, на самом деле оказались соответствующими запросу). Отсюда получаем, что рассматриваемая технология достаточно полно (95%) и точно (87%) осуществляет поиск в соответствии с потребностью пользователя.

Для сравнения: с помощью полнотекстового поиска средствами MS SQL Server 2008, использующего векторную модель, были получены следующие значения: было найдено 49% документов с релевантностью больше 0,6. При этом полнота r=0,62, точность p=0,59.

Таблица 1

Изменение значений точности и полноты поиска

№	Итерация	Релевантность	Полнота (r)	Точность (p)
1	Автоматическое построение онтологии по обучающей коллекции документов	>0,6 (для 63% док-тов)	0,7	0,9
2	Запрос: Пациент принимает антикоагулянт	>0,6 (для 37% док-тов)	0,51	0,75
3	Обучение: до 2005 года антикоагулянтом назначают варфарин в 90% случаев	>0,6 (для 52% док-тов)	0,69	0,79
4	Обучение: Антикоагулянты и противосвертывающие — одно и то же	>0,6 (для 86% док-тов)	0,83	0,81
5	Обучение: Если пациенту не противопоказан варфарин и он перенес инсульт, то пациент принимает антикоагулянт с уверенностью 90%	=1 и =0,9 (для 63% и 37% док-тов соотв.)	0,95	0,87

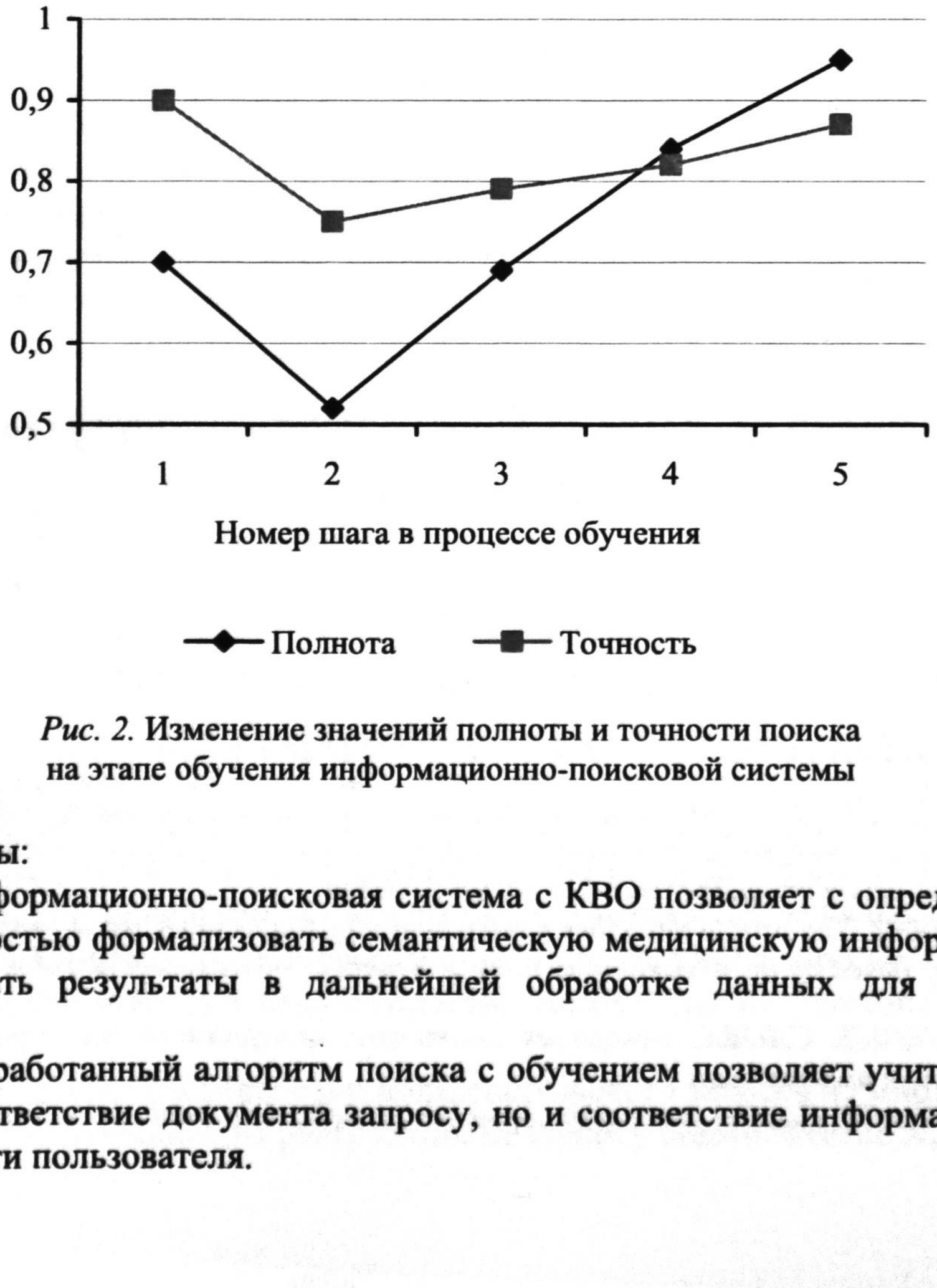

Рис. 2. **Изменение значений полноты и точности поиска на этапе обучения информационно-поисковой системы**

Выводы:

— Информационно-поисковая система с КВО позволяет с определенной достоверностью формализовать семантическую медицинскую информацию и использовать результаты в дальнейшей обработке данных для проведения МБИ.

— Разработанный алгоритм поиска с обучением позволяет учитывать не только соответствие документа запросу, но и соответствие информационной потребности пользователя.

СПИСОК ЛИТЕРАТУРЫ

1. Казначеев В.П., Баевский Р.М., Берсенева А.П. Донозологическая диагностика в практике массовых обследований населения. М.: Медицина, 1992. 208 с.

2. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 1. М.: ФИЗМАТЛИТ, 2005. 144 с.

3. Евдокименков В.Н. Компьютерные технологии сбора, обработки и анализа данных медико-биологических исследований: Учебное пособие. М.: Изд-во МАИ, 2005. 436 с.

4. Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schutze Schutze. Introduction to Information Retrieval. Cambridge UP. Online edition (c), 2009. 544 c.

5. Нестерова О.А., Оленников Е.А. Проблема сбора и анализа данных для научных исследований в медицинских информационных системах / Искуственный интеллект: философия, методология, инновации: мат-лы III Всерос. конф. студентов, аспирантов и молодых ученых. М.: Связь-принт, 2009. С. 371-373.

6. Аветисян Р.Д., Аветисян Д.О. Теоретические основы информатики. М.: Вильямс, 2002. 168 с.

Игорь Николаевич ГЛУХИХ —
зав. кафедрой мат. методов, статистики
и информационных технологий в экономике
доктор технических наук, профессор
igluhih@utnm.ru

Михаил Владимирович ГУБИН —
ассистент кафедры мат. методов, статистики
и информационных технологий в экономике
Mikhail.gubin@gmail.com

Институт математики и компьютерных наук
Тюменского государственного университета

УДК 004.023: 004.031.42

ОПТИМИЗАЦИЯ ВЫДАЧИ КОНТЕНТА В ДИНАМИЧЕСКОМ ИНТЕРНЕТ-ПРОЕКТЕ НА ОСНОВЕ АНАЛИЗА ВЫБОРОВ НЕИДЕНТИФИЦИРОВАННЫХ ПОЛЬЗОВАТЕЛЕЙ

OPTIMIZATION OF CONTENT DELIVERY IN THE DYNAMIC INTERNET PROJECT ON THE BASIS OF SELECTION ANALYSIS OF UNIDENTIFIED USERS

АННОТАЦИЯ. В статье рассматриваются модели и алгоритмы, предназначенные для управления выдачей пользователям контента на основе учета действий множества неидентифицированных пользователей.

SUMMARY. The article describes models and algorithms designed to manage the content delivery based on the account of actions of a number of unidentified users.

КЛЮЧЕВЫЕ СЛОВА. Генерация контента, гипертекстовая статья, анализ данных, поиск информации.

KEY WORDS. Content generation, hypertext article, data analysis, search of information.