На правах рукописи

НЕСТЕРОВА Ольга Андреевна

ТЕХНОЛОГИИ, МОДЕЛИ И АЛГОРИТМЫ ПОИСКА В АРХИВАХ МЕДИЦИНСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ КОНТЕКСТНО-ВРЕМЕННОЙ ОНТОЛОГИИ

05.13.18 – Математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Тюмень – 2011

Работа выполнена на кафедре информационной безопасности Института математики и компьютерных наук ГОУ ВПО Тюменский государственный университет

Научный руководитель: доктор технических наук, профессор
Захаров Александр Анатольевич

Официальные оппоненты: доктор технических наук, профессор
Глазунов Виктор Аркадьевич
доктор технических наук, профессор
Ивашко Александр Григорьевич

Ведущая организация: Томский государственный университет систем управления и радиоэлектроники (ТУСУР)

Защита диссертации состоится «11» марта 2011 г. в 16-00 часов на заседании диссертационного совета Д 212.274.14 при Тюменском государственном университете по адресу, 625003, г. Тюмень, ул. Перекопская, 15а, ауд. 410.

С диссертацией можно ознакомиться в библиотеке Тюменского государственного университета.

Автореферат разослан «10» февраля 2011 г.

Ученый секретарь
диссертационного совета *Н.Н. Бутакова*

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Повышение качества и доступности медицинской помощи – один из приоритетов государственной социальной политики. Подтверждением этого является принятая концепция развития системы здравоохранения в Российской Федерации до 2020 года. Одним из основных направлений решения проблемы информатизации медико-биологических исследований (МБИ) является предоставление информации различным специалистам из тематических электронных архивов историй болезни. Вопросам применения информационных ресурсов в МБИ посвящены работы: Г.И. Назаренко, Г.С. Осипова, А.И. Молодченкова, А.С. Клещева, Ф.М. Москаленко, М.Ю. Черняховской, О.Ю. Ребровой, В.М. Тавровского, В.А. Лищук, С.Е. Бащинского, В.П. Казначеева, Р.М. Баевского, А.П. Берсеневой, В.Н. Евдокименкова, У. Кокрена.

Процессу проведения МБИ присущи задачи сбора, обработки информации и интерпретации результатов. Критический анализ медицинских информационных систем (МИС) – источников информационных ресурсов – выявил ряд проблем информатизации МБИ.

1. Необходимость использования неформализованных данных (неструктурированные текстовые массивы, изображения), для которых применение обычных запросов с использованием предикатной логики является затруднительным, усложняет процесс поиска нужной информации.
2. Широко используемые технологии поиска данных в тексте по точному совпадению слов не подходят для задач кодификации (распознавания) элементов системы (объектов, фактов, событий) в неструктурированных текстовых массивах.
3. Большинство разрабатываемых МИС выполняют только функции учета (хранения) данных, которые имеют заранее определенную структуру. К таким данным невозможно применить произвольный запрос в любой момент времени. Необходимы затраты на сопровождение разработчиками.
4. Решение вопроса интеграции разрозненных данных (территориально, различные разработчики МИС) не только требует финансовых затрат, но и сталкивается с проблемой интеграции семантических данных.
5. Необходимость оперативного доступа к информации, ее интеграция требует особого внимания к обеспечению безопасности с учетом закона о персональных данных.

В рамках одного исследования невозможно решить все сформулированные выше проблемы, поэтому нами определена, на наш взгляд, ключевая проблематика в организации научно-исследовательской деятельности врача по

сбору и анализу данных: оптимизация механизмов поиска и кодификации элементов учетной МИС, содержащихся в неструктурированных текстах медицинских электронных записей.

Теоретическое обоснование методов поиска и анализа текстов рассмотрено в работах Г. Сэлтона, Т. Джойса, Р. Нидхема, К. Маннинга, П. Рагхавана, Г. Шютце. Методы поиска на основе семантической сети находятся еще только в стадии развития. Делаются попытки использования семантических сетей для поиска в сети Internet. Разработке семантических моделей информационного поиска посвящены работы С. Дамайса, Г. Фурнаса, С. Дирвестера, К. Маннинга, Т. Груббера, Е.А. Рабчевского, Н.В. Лукашевича, Б.В. Доброва, Р.В. Шарапова, В.А. Глазунова, Р.Д. Аветисяна.

А. Гладун, Ю. Рогушина, П.С. Шеменков в своих работах отмечают, что в задачах семантического поиска в текстах важным является критерий, представляющий собой оценку информационной потребности пользователя.

Решение задачи связано с проблемой разработки технологии анализа текстовой медицинской информации, которая учитывала бы специфику электронной медицинской информации: разнородность, удаленность, многозначность, неточные формулировки, субъективность, хронологическую последовательность и неформализованное представление в виде неструктурированного текстового массива.

Объект исследования: модели, алгоритмы и технологии информационного поиска в неструктурированных текстах медицинских электронных записей для поддержки медико-биологических исследований.

Предмет исследования: условия и средства организации семантического (смыслового) распознавания различных сведений, данных о соответствующих предметах, явлениях, процессах, отношениях (элементов МИС) в неструктурированных текстовых массивах медицинских электронных записей.

Целью диссертационной работы является совершенствование механизмов информационного поиска медицинских данных для поддержки МБИ посредством обеспечения максимально возможной полноты обзора текстовых информационных ресурсов и точности нахождения информации.

Для достижения поставленной цели в работе решаются следующие **задачи**:

1. Разработка технологии интерпретации смысла текста документов и запросов для представления элементов МИС в неструктурированных текстовых массивах медицинских электронных записей.
2. Разработка метода расчета соответствия образа документа запросу.
3. Разработка алгоритма поиска и сбора данных.

4. Построение модели семантико-энтропийного поиска для организации сбора данных для информационной поддержки медицинских научных исследований.

5. Разработка критерия эффективности поиска.

6. Проектирование архитектуры информационно-поисковой системы (ИПС).

7. Разработка концепции гибридизации МИС.

На рис. 1 приведена структурная схема, отображающая комплексный системный подход к процессу исследования.

Методы исследований. Приведенные в работе методы исследования базируются на использовании методов теории графов, теории принятия решений, теории информации, нечеткой логики, теории вероятности и математической статистики, методов информационного поиска, математического моделирования, графовой кластеризации, модульного и объектно-ориентированного программирования.

Достоверность и обоснованность результатов. Предложенные в диссертационной работе модели и алгоритмы обоснованы теоретическими решениями, не противоречат известным положениям других авторов, определяются методологической базой исследования, сочетанием различных подходов и методов исследования, экспериментальной проверкой теоретических положений и воспроизводимостью результатов.

Положения, выносимые на защиту

- Технология семантико-энтропийного поиска:
  - математическая модель контекстно-временной онтологии;
  - алгоритм поиска и анализа результатов запроса.
- Архитектурная модель информационно-поисковой системы.

Научная новизна работы отражена в следующих результатах.

- Впервые понятия контекстно-временной онтологии (КВО) предметной области применены к информационному поиску в архивах медицинских данных.
- Разработана новая технология семантико-энтропийного поиска с использованием модели КВО.
- Построена новая модель КВО предметной области:
  - введено понятие фактора достоверности, зависящего от времени;
  - предложен метод расчета оценки неопределенности запроса с использованием энтропийной оценки;

Рис. 1. Структурная схема комплексного системного подхода к процессу исследования

– предложен метод расчета оценки релевантности документов с учетом коэффициентов достоверности, как расчет меры близости графов, полученных путем построения семантических сетей документа и запроса на основании построенной экспертом контекстно-временной онтологии.

- Разработан новый алгоритм поиска с обучением с учителем, включающий в себя контекстное индексирование и анализ результатов поиска.

Теоретическая значимость. Стало возможным достижение результатов ряда новых задач.

- Интеграция семантических данных с применением КВО.
- Обработка и анализ семантических данных в системах поддержки принятия решений с использованием КВО.
- Семантико-энтропийный поиск в сети Internet.

Практическая значимость

- Алгоритм поиска с обучением позволяет учитывать соответствие документа информационной потребности пользователя.
- Механизм преобразования общего инвертированного файла (индекса) в контекстный индекс, зависящий от контекста запроса, позволяет получить контекстные образы документа, соответствующие различным запросам.
- Использование разработанной модели поиска позволяет с определенной долей достоверности формализовать семантическую информацию для получения полной выборки данных и дальнейшей обработки данных при проведении МБИ.
- Разработанная архитектурная модель ИПС, состоящая из индексирующего, поискового сервиса и виртуального хранилища данных предоставляет возможность исследователю оперативно получать данные по теме своего исследования из различных источников.
- Предложенный метод перехода от учетных к гибридным ИС позволяет наиболее эффективно использовать имеющиеся данные МИС, предоставляя инструментарий формирования произвольных запросов пользователем, не являющимся IT-специалистом.
- Предлагаемые методические разработки могут быть приняты во внимание разработчиками медицинских информационных систем при проектировании структуры в направлении, рассматриваемом в диссертации.

Реализация и внедрение результатов работы

- Теоретические и практические результаты работы реализованы и внедрены в качестве ИПС для сбора данных и поддержки медицинских исследований в Тюменском кардиологическом центре (ТКЦ). В процессе эксплуатации

представленная система показала свою эффективность. Внедрение системы в ТКЦ подтверждено соответствующими свидетельствами.

- Разработанная ИПС используется при сборе данных в исследованиях по ишемическому инсульту и фибрилляции предсердий, что подтверждается соответствующими публикациями совместно с научными работниками ТКЦ.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на следующих конференциях и семинарах:

III международная научно-практическая конференция «Исследование, разработка и применение высоких технологий в промышленности», Санкт-Петербург, март 2007; III Всероссийская конференция студентов, аспирантов и молодых ученых «Искусственный интеллект: философия, методология, инновации», Москва, ноябрь 2009; II региональная конференция ИМКН ТюмГУ, Тюмень, октябрь 2009; IX международный славянский конгресс «КАРДИОСТИМ-2010», Санкт-Петербург, февраль 2010; 9-я Сибирская научная школа-семинар SIBECRYPT'10, Тюмень, октябрь 2010; IV Всероссийская конференция студентов, аспирантов и молодых ученых «Искусственный интеллект: философия, методология, инновации», Москва, ноябрь 2010; научные семинары НИИ КИТ, кафедры информационной безопасности ТюмГУ, Тюмень, 2006 – 2010.

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.

Этапы исследования. Условно исследование можно разделить на четыре этапа. Первый этап (2006 – 2007 гг.) включал в себя анализ литературы по теме исследования, изучение опыта работы, как в России, так и за рубежом. На втором этапе (2007 – 2008 гг.) разрабатывались организационные модели, отрабатывалось содержание научно-исследовательской деятельности врача. На третьем этапе (2009 г.) велась опытно-экспериментальная работа по изучению возможностей организации гибридной МИС на базе ТКЦ. На четвертом этапе (2010 г.) проводилась обработка и обобщение полученных результатов.

Публикации. Основное содержание отражено в 24 публикациях, из которых 7 свидетельств о государственной регистрации программ для ЭВМ и 4 статьи, опубликованных в изданиях, рекомендованных ВАК.

Структура и объем работы. Приведенные цели и задачи определяют структуру и содержание исследования. Текст диссертации состоит из введения, четырех глав, заключения, списка литературы из 117 наименований работ российских и зарубежных авторов, 4 приложений. Общий объем – 129 страниц, в том числе 5 таблиц, 11 рисунков на 11 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертации, сформулированы цель и задачи работы, научная новизна, теоретическая и практическая значимость, перечислены основные результаты работы.

В **первой главе** рассмотрены принципы, теоретические основы, основные задачи и цели МБИ. Проведен анализ современных типов автоматизированных МИС, сформулирована проблема использования данных учетных МИС для научных исследований. Отмечены перспективные методы развития технологий информационного поиска. Проведена сравнительная оценка вариантов возможных решений исследуемой проблемы, описаны основные принципы распознавания элементов системы в неструктурированных текстовых массивах, статистические, семантические модели информационного поиска и оценки неопределенности.

Автоматизация и поддержка научных исследований в медицине является новым и важным шагом в развитии лечебной, консультативной, профилактической, доказательной, скрининговой и восстановительной медицины. Основное содержание проблемы заключается в необходимости сбора данных электронных архивов МИС для научных исследований. Существующие технологии информационного поиска ориентированы на экономические, маркетинговые информационные системы. Применение подобных технологий для МИС затруднено, потому что медицинские исследования обычно охватывают более широкий временной интервал и большое количество разнообразных категорий данных.

На основе обзора методов анализа текстовой информации, отличающихся в первую очередь используемой моделью, сделан вывод, что для задачи поиска медицинских данных подходят семантические сети, учитывающие информационную потребность пользователя. Это обусловлено тем, что важнее найти не первый документ, релевантный тому или иному запросу, а собрать наибольшее количество документов, удовлетворяющих информационным потребностям пользователя.

Обоснована необходимость выработки унифицированного подхода к построению автоматизированного кодификатора объектов в текстовых массивах на основе математического моделирования и алгоритмических подходов к разработке технологии информационного поиска и способов обработки результатов поиска.

В 2002-м году Р.Д. Аветисян и Д.О. Аветисян показали адекватность энтропийной модели документального поиска. Для семантического поиска энтропийная оценка эффективности еще не применялась. В 2005-м году Г. Зу,

С.Е. Мадником и М.Д. Сайгелом описано использование контекстно-временной онтологии в системе интеграции семантических данных COIN для англоязычных экономических систем. В России таких исследований не проводилось.

Во **второй главе** сформулированы основные принципы семантического поиска, обозначены преимущества такого подхода. Подробно описаны методы, используемые в исследовании. Изложена общая концепция нечеткости и неопределенности. Далее описан процесс построения модели поиска, состоящей из модели представления элементов медицинской информационной системы, основывающейся на использовании онтологии предметной области, алгоритма поиска документов и оценки релевантности (пертинентности). Определение элементов системы терминами и связями между ними выражается с помощью фактора достоверности и темпоральными (временными) характеристиками – принадлежность к некоторому интервалу времени.

Модель КВО определяется следующим образом. Пусть:

$X=\{x_i\}$ – множество понятий ($i=\overline{1,M}$);

$Y=\{y_u\}$ – множество терминов (слово или словосочетание), элементов терминологического словаря ($u=\overline{1,U}$);

$R_t=\{r_k\}$ – множество контекстно-временных отношений между понятиями, определяющих *связи между элементами поиска* ($k=\overline{1,K}$);

$\mathrm{cr}(t)\rightarrow R_{[0;1]}$ – *функция факторов достоверности отношений в момент времени* t, возвращающее в любой момент времени значение в интервале [0;1]: 0 – неизвестно; (0;1) – достоверно в некоторой степени; 1 – отношение достоверно на 100%;

$$\mathrm{cr}(t)=\begin{cases}\mathrm{cr}_h(t), & t\in[t_h^{(1)};t_h^{(2)}];\\ 0, & \text{иначе.}\end{cases}\qquad (t_l^{(1)};t_l^{(2)}]\cap(t_p^{(1)};t_p^{(2)}]=\varnothing,\forall l\neq p;\quad h,l,p=\overline{1,T}, \qquad (1)$$

где T – количество временных интервалов.

Тогда отношение r_k можно определить так: $r_k=<x_i, x_j, y_u, \mathrm{cr}_k(t)>$, где: x_i, x_j – понятия; $\mathrm{cr}_k(t)$ имеет вид (1) – функция фактора достоверности отношения r_k между x_i и x_j, определяемое термином y_u; $i,j=\overline{1,M}$; $k=\overline{1,K}$.

$F_t=<F_n, F_s>$ – множество функций интерпретации.

F_n – функция контекстно-временной нормализации терминов, в любой момент времени для любого термина i-го понятия возвращает номер j-го терма, определяющий элемент поиска с максимальным фактором достоверности:

$$\mathrm{F}_n(N,t)\rightarrow N:\forall i=\overline{1,M},\forall t_0\quad \mathrm{F}_n(i,t_0)=\arg\max_{\forall u=\overline{1,U}}(\mathrm{cx}_{iu}(t_0)), \qquad (2)$$

где $\mathrm{cx}_{iu}(t) \rightarrow R_{[0;1]}$ – функция фактора достоверности u-го термина, определяющего i-е понятие в момент времени t.

F_s – функция контекстно-временной интерпретации термов, в момент времени t ставит в соответствие i-му терму вектор $CX=\{cx_{iu}\}$ факторов достоверности, отражающих степень соответствия u-го термина i-му понятию.

$$\mathrm{F_s}(N,t) \rightarrow R^{U}_{[0;1]}: \quad \forall i = \overline{1,M} \ , \forall t_0 \quad \mathrm{F_s}(i,t_0) = E_u \bullet \mathrm{cx}(t_0) \ , \tag{3}$$

где E_u – матрица $U \times U$, элементы u-го столбца равны 1, остальные равны 0:

$$E_u = \{e_{lp}\}: \quad e_{lp} = \begin{cases} 1, & p = u, \forall l; \\ 0, & p \neq u, \forall l. \end{cases} \tag{4}$$

Pr_t используется для построения правил выводов:

$$Pr_t = \text{ЕСЛИ}(\text{И}(\{r_i,c_i,t_i\}\big|_1^n) \mid \text{ИЛИ}(\{r_j,c_j,t_j\}\big|_1^n) \mid \text{НЕ}(\{r_h,c_h,t_h\}\big|_1^n)) \ \text{ТО}(\{r'_l,c'_l,t'_l\}\big|_1^m), \tag{5}$$

где: r_k – исходные отношения с коэффициентом достоверности c_k в момент времени t_k ($k=\overline{1,K}$); r'_p – выходные отношения с коэффициентом достоверности c'_p в момент времени t'_p ($p=\overline{1,P}$).

В результате получена *модель контекстно-временной онтологии*:

$$O_t = <X, R_t, F_t, Pr_t>, \tag{6}$$

Представление документов в виде набора триплетов образуют в модели подграф, который задает представление документа в данном контексте запроса: $O(D) \in O(Q)$. Узлы соответствуют термам, а ребра – бинарным отношениям между ними.

Веса узлов графа определены как коэффициенты достоверности $\mathrm{cx}(t)$. Для каждого из ребер (x_i, x_j) графа полагается заданным также $(1 \times K)$ – вектор весов $\{\mathrm{cr}_{ijk}(t), k=\overline{1,K}\}$, где $\mathrm{cr}_{ijk}(t)=0$, если термы (x_i, x_j) не связаны между собой отношением r_k, и $\mathrm{cr}_k(t)=\mathrm{cr}_k(t)$ – в противном случае. Здесь $\mathrm{cr}_k(t)$ – заданный вес отношения r_k в онтологии O.

Предложено использование меры соответствия триплетов документов, формализующих близость семантических сетей поисковых образов документа D и запроса Q или, что то же самое, меры близости соответствующих графов $G(D)$ и $G(Q)$, учитывающей веса термов и связей между ними.

Мера близости вершин и ребер графов $G(D)$ и $G(Q)$ определяется как минимальное значение коэффициентов достоверности соответствующих вершин и ребер в любой момент времени t.

Мера близости термов x_i запроса и документа:

$$\overline{cx}(x_{i,D}, x_{i,Q}) = \overline{cx_i}(t) = \min_{\forall t} (cx_{i,D}(t) \ , cx_{i,Q}(t)) . \tag{7}$$

Мера близости ребер r_k запроса и документа:

$$\overline{cr}(r_{k,D}, r_{k,Q}) = \overline{cr}_k(t) = \min_{\forall t} (cr_{k,D}(t), cr_{k,Q}(t)). \quad (8)$$

Тогда пересечение графов можно $G(D) \cap G(Q)$ представить как набор вершин и ребер с коэффициентами достоверности (7) и (8).

Взвешенная мера близости вершин определяется следующим образом:

$$s_x = \frac{2 \cdot \sum_{\alpha, \forall t} \overline{cx}_\alpha(t)}{\sum_{\beta, \forall t} cx_i(t) + \sum_{\gamma, \forall t} cx_j(t)}, \quad (9)$$

где: индекс α пробегает номера узлов, принадлежащих пересечению графов $G(D) \cap G(Q)$, что условно можно записать в виде $\alpha \in [1:n(G(D) \cap G(Q))]$; индексы β, γ пробегают номера узлов $[1:n(G(Q))]$, $[1:n(G(D))]$ соответственно для любого времени t.

Взвешенная мера близости ребер определяется как:

$$s_r = \frac{2 \cdot \sum_{\alpha, \forall t} \overline{cr}_\alpha(t)}{\sum_{\beta, \forall t} cr_i(t) + \sum_{\gamma, \forall t} cr_j(t)}, \quad (10)$$

где, аналогично (9): α пробегает номера ребер пересечения графов $G(D) \cap G(Q)$, индексы β, γ пробегают номера ребер графов $G(Q)$ и $G(D)$ соответственно.

Мера близости графов определяется, как функция полезности мер s_x и s_r (9, 10). Рассмотрена аддитивная свертка мер. Методом половинного деления определен вид скалярной свертки с коэффициентом полезности $\delta \leq 1$:

$$s = \delta s_x + (1 - \delta) s_r. \quad (11)$$

Мера s из (11) принимается за коэффициент достоверности CF – доля уверенности, что определенный документ соответствует смыслу запроса.

С использованием построенной модели КВО и оценки релевантности документов в результате вычислительного эксперимента разработан алгоритм семантико-энтропийного поиска для обучения с учителем системы пониманию смысла запроса. В.В. Иванов в работе «Модели и методы интеграции структурированных текстовых описаний на основе онтологий» предлагает стратегию, в которой понятия тезауруса внедряются в онтологию как экземпляры особого метакласса онтологии. В качестве множества допустимых значений некоторого понятия выступают группы близких понятий тезауруса. На рис. 2 описаны этапы алгоритма семантико-энтропийного поиска.

Этап 1 Подготовка	Шаг 1.1. Первичное индексирование (документ-предложение-слово) Шаг 1.2. Построение онтологии предметной области Шаг 1.3. Построение онтологии времени (выделение понятий времени)	Получение абстрактных образов документов, поступивших в хранилище
Этап 2 Обучение	Шаг 2.1. Построение КВО Шаг 2.2. Построение семантической сети (образа) запроса Шаг 2.3. Оценка неопределенности Шаг 2.4. Оценка обучающей коллекции (полнота, точность, мера информации)	Понижение уровня абстракции, конкретизация контекстно-временных параметров Этап.2 повторяется, пока не будет достигнута желаемая полнота и точность выборки
Этап 3 Тестирование	Шаг 3.1. Вычисление коэффициентов достоверности Шаг 3.4. Определение временных интервалов	
Этап 4 Контекстное индексирование	Шаг 4.1. Поступление новых документов Шаг 4.2. Построение семантического образа документов на основе полученной КВО при обучении системы, т.е. в соответствии с полученным набором термов и отношений	Поиск. Этапы 4 и 5. выполняются для всех поступающих в систему документов на время актуальности запроса
Этап 5 Обнаружение схожих документов	Шаг 5.1. Оценка релевантности	
Этап 6 Анализ результатов	Шаг 6.1. Анализ зависимостей между понятиями КВО, найденными в документах	Предоставление функций советующего справочника

Рис. 2. Этапы алгоритма семантико-энтропийного поиска

Обучение системы нахождению документов, отвечающих заданному в запросе смыслу, заключается в построении обучающей выборки – списка документов, поставленных в соответствие заданному запросу. Процесс построения является итеративным. Эксперт создает некоторый набор терминов, характеризующих смысл, и связей между ними.

Каждое понятие тезауруса, извлеченное из текста запроса, сопоставляется с экземпляром онтологии и используется для построения связного множества триплетов. Эксперт вводит правила, определяет множество терминов и отношений. В результате получаем контекстно-временную онтологию.

Для оценки неопределенности построения запроса при создании обучающей выборки используется понятие меры неопределенности. Количество информации, содержащееся в среднем в одном сообщении о том, каким признан системой очередной документ, вычисляется по формуле:

$$I[sp, up]=H[sp]-H[sp|up]=H[up]-H[up|sp], \qquad (12)$$

где: sp – документ признан системой релевантным запросу; up – документ на самом деле является релевантным запросу; $H[sp]$ – неопределенность того, что наугад взятый документ будет признан релевантным системой; $H[sp|up]$ – неоп-

ределенность того, что документ, признанный системой релевантным на самом деле является релевантным.

С помощью подбора таких параметров, как глубина индексирования (количество документов, которые будет индексировать поисковый сервис) и глубина терминологического наращивания запросов (последовательного/итерационного пополнения словаря терминов, участвующих в данном запросе) обеспечивается понижение меры неопределенности и улучшения коэффициентов полноты и коэффициентов точности, характеризующих соответствие текста заданной теме.

Результатом математического моделирования является построение модели семантического поиска и доказательство ее соответствия поставленной цели исследования. Проведен анализ полученных результатов, исследуется применимость модели в реальном мире. Модель информационно-поисковой системы включает в себя модель интерпретации документов и запросов на основе контекстно-временной онтологии и алгоритм с обучением с учителем для обучения системы контексту заданного запроса. Построенные модели подчиняются всем законам математической логики, способны адекватно описывать исходную ситуацию. Результаты, полученные на основе данных моделей, хорошо отражают действительность в соответствии с выдвинутыми критериями.

Предложенная в работе методика оценки релевантности документов обладает высокой вычислительной сложностью. Подавляющая часть требуемых вычислительных затрат обусловлена выполнением следующих работ.

Во-первых, для каждого из документов D требуется построение соответствующей семантической сети $S(D)$. Если онтология предметной области фиксирована, т.е. «четкая» и не зависит от времени, то эта работа выполняется лишь однажды, при помещении документа в хранилище. Во-вторых, методика требует построения аналогичной семантической сети $S(O)$ онтологии рассматриваемой предметной области. Опять же, если онтология предметной области фиксирована, то эта работа выполняется однократно. В-третьих, в соответствии с методикой для каждого из запросов Q также требуется формирование семантических сетей $S(Q)$. Данная работа должна выполняться системой при обработке каждого из запросов.

Задача определения пертинентности документа является задачей оптимизации. Использованный метод аддитивной скалярной свертки является простейшим и далеко не всегда эффективным методом решения. Поэтому представляет интерес исследование целесообразности использования других, более «тонких» методов решения указанной многокритериальной задачи.

Третья глава содержит описание процесса проектирования и создания технологии распознавания элементов медицинской информационной системы.

Концептуальная модель информационно-поисковой системы приведена на рис. 3.

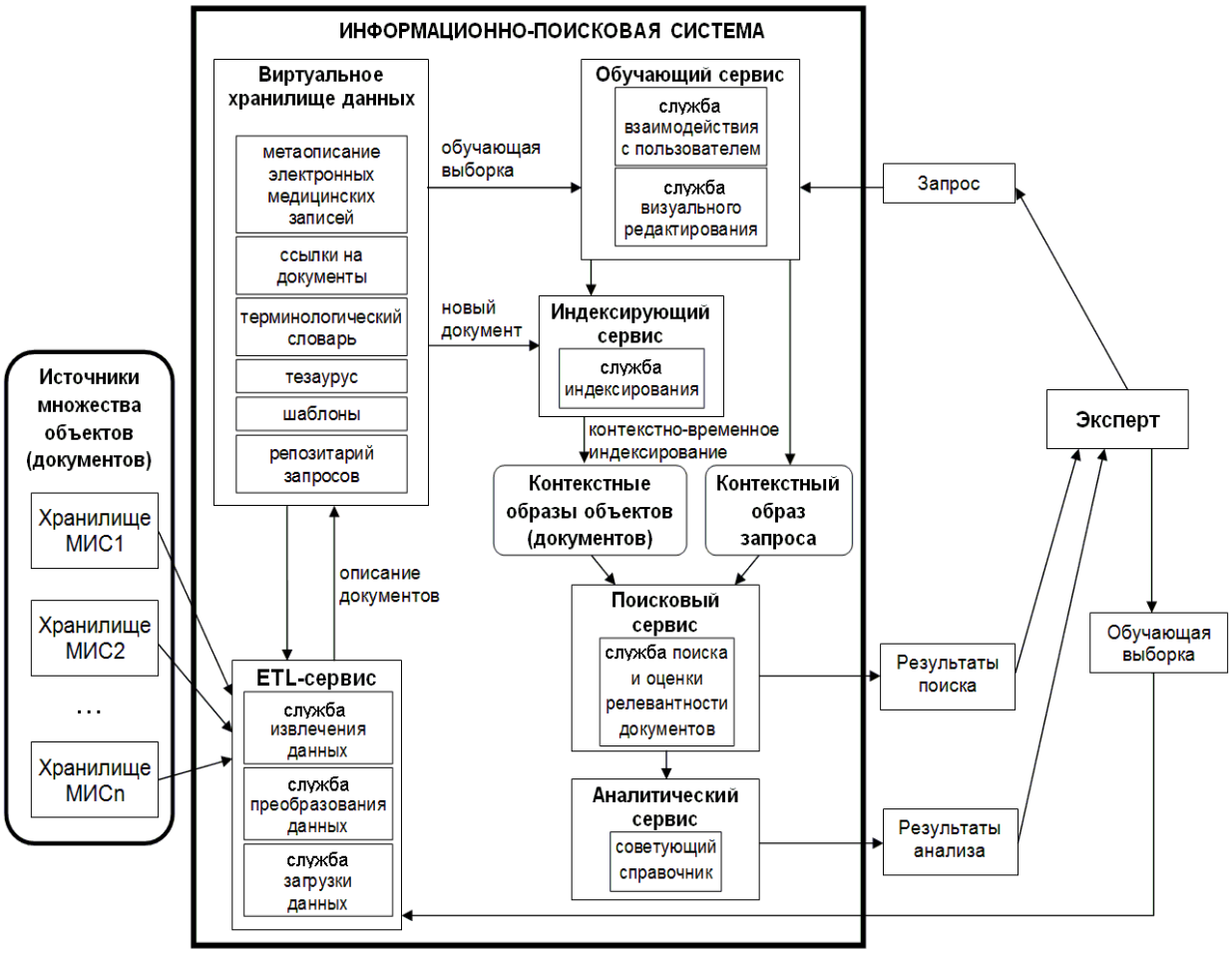

Рис. 3. Концептуальная модель информационно-поисковой системы

Все операции поиска разбиты на пять групп: обучение, хранение информации, поисковые операции, выдача информации, информационный анализ. Система состоит из следующих элементов:

– виртуальное хранилище данных – средство предоставления доступа к распределенным архивам разнородных документов различных МИС; содержит описание электронных медицинских записей, ссылки на документы, терминологический словарь, тезаурус, шаблоны, репозитарий запросов;

– ETL-сервис – содержит инструментарий: извлечения данных из различных источников; преобразования – для первичной индексации и «очистки» данных и инструментарий загрузки данных в хранилище;

– обучающий сервис – средство составления контекстно-временной онтологии, содержит: инструментарий, позволяющий составлять новый запрос с обучением; визуальный редактор, отображающий понятия и связи между ними в удобном для восприятия виде для конструирования запросов к данным;

– индексирующий сервис – средство создания контекстных индексов поступающих в хранилище документов;

– поисковый сервис – средство организации поиска документов;

– аналитический сервис – средство обработки результатов поиска.

Далее предложен подход к реализации ИПС, основанный на создании программных сервисов, отвечающих за выполнение отдельных функций системы и имеющих единый интерфейс взаимодействия. Спроектирована и реализована универсальная программная архитектура ИПС, позволяющая взаимодействовать с разработанными ранее автоматизированными рабочими местами (АРМ) учетной МИС. Схема многоуровневой архитектурной модели информационно-поисковой системы представлена на рис. 4.

Модуль взаимодействия с пользователем использует глубокие знания (представление о пациентах, заболеваниях, клинических тестах) для извлечения дополнительных, более детальных контекстно-временных знаний. На эксперта возлагается задача расширения и уточнения модели онтологии – понижение уровня абстракции. Эта модель затем передается индексирующему сервису. Поведение системы снова анализируется экспертом и обучающим сервисом (энтропийная оценка). Эксперт при необходимости вносит коррективы в онтологию.

Графический интерфейс позволяет эксперту создавать пиктограммы, представляющие элементы запроса, формировать из них графические структуры. Расставляя элементы на экране и вычерчивая связи между ними, эксперт формирует мнемоническую схему взаимосвязей между элементами.

Для более эффективного использования в исследованиях результатов поиска проводится анализ полученных данных. Интеллектуальная обработка результатов поиска, заключается в применении метода графовой кластеризации по алгоритму Буровки.

На следующем этапе определены основные характеристики гибридной информационной системы и предложен метод перехода от учетной системы к гибридной. Разработанная технология поиска в электронных хранилищах МИС позволяет автоматизировать процесс сбора данных для научных исследований и обеспечивает, независимо от структуры и состава МИС, эффективный анализ и обработку данных. Предложенный подход к разработке архитектуры ИПС позволяет использовать ее для гибридизации учетных МИС. Вместо разработки гибридной МИС «с нуля» автором исследования выбран способ повышения интеллектуального уровня ранее разработанной учетной МИС посредством использования технологии информационного поиска.

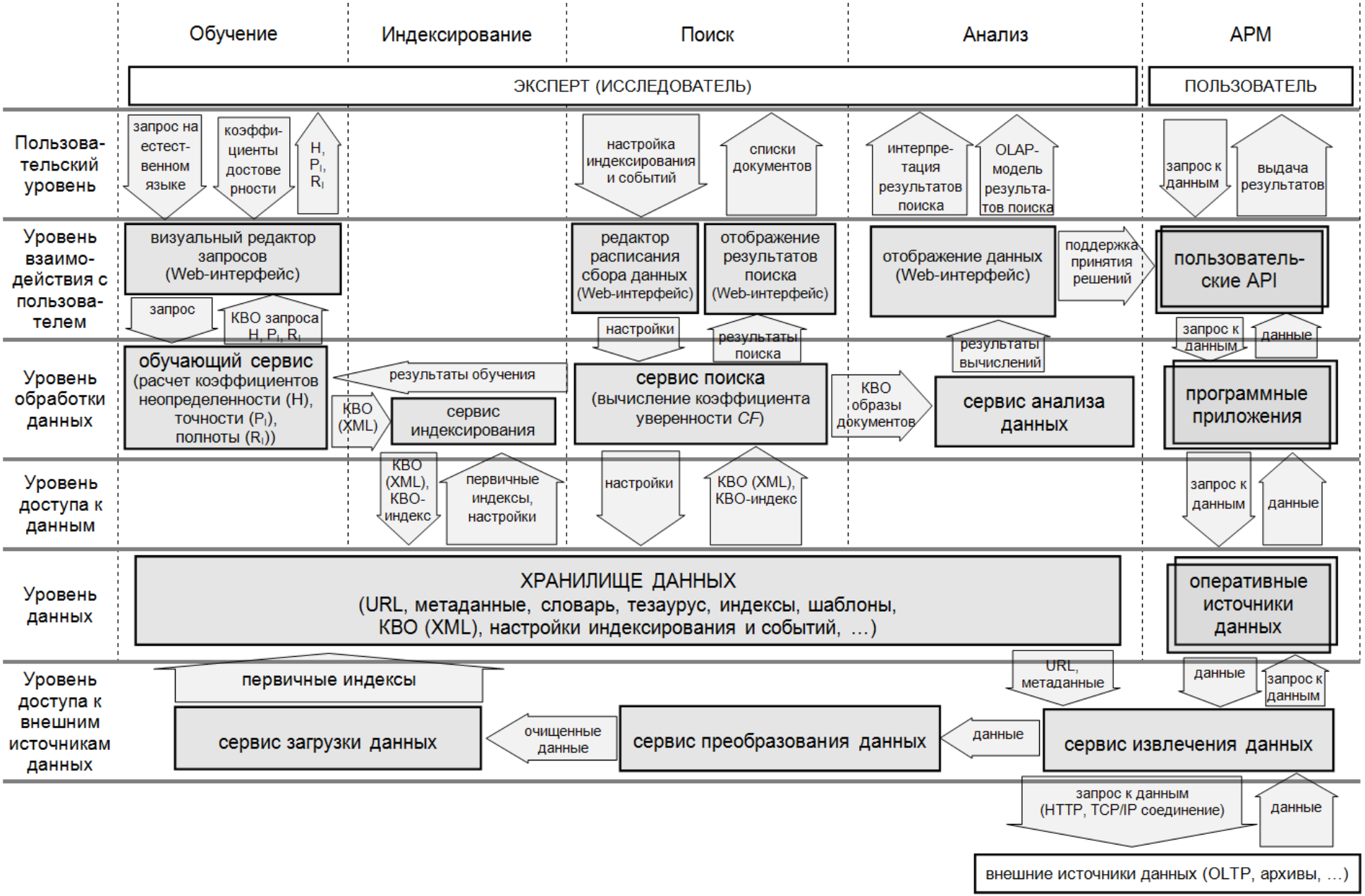

Рис. 4. Схема многоуровневой архитектурной модели информационно-поисковой системы

В **четвертой главе** приведены результаты апробации и статистика работы программного комплекса; описан численный эксперимент по оценке эффективности и полезности ИПС, разработанной на основе модели КВО. Приведены количественные характеристики обучающих и тестовых коллекций и примеры обучения и тестирования. Проведены эксперименты с алгоритмом поиска, позволяющие судить о качестве работы алгоритма по двум основным критериям: полноте и точности.

Исследованы зависимости критериев точности и полноты от следующих параметров: количество документов обучающей выборки; сложность запроса (количество вершин и ребер графа, построенного на основе семантической сети запроса); коэффициент полезности при оценке релевантности по формуле (11).

Экспертами предоставлены выборки из подходящих для исследований документов (историй болезни), выбранных из общего количества за определенный период, на создание которых потрачено несколько месяцев. Выборки разделены на обучающую и контрольную части. Каждой паре «запрос-документ» поставлен в соответствие набор оценок релевантности информационной потребности, представленных в виде бинарных утверждений «релевантный» и «нерелевантный». По каждому запросу вычислены значения коэффициентов точности и полноты выборки для документов с положительной релевантностью запросу. Составы коллекций и средние значения основных характеристик разработанной информационно-поисковой системы приведены в таблице 1.

Таблица 1. **Состав коллекций и средние значения основных характеристик разработанной информационно-поисковой системы**

№	Тематика коллекции	Количество документов		Период, гг.	Время поиска «вручную», мес.	Кол-во тестовых запросов	Критерии оценки (средние значения)			
		всего	выбрано				полнота ($R_{ср}$)	точность ($P_{ср}$)	мера $F_{1ср}$	I[sp,pp]
1	Пациенты с риском ишемического инсульта	14000	200	2007-2009	12	20	0,87	0,81	0,87	0,26
2	Пациенты с симптомом фибрилляции предсердий	7000	250	2007-2008	7	30	0,98	0,95	0,98	0,28

В результате вычислительного эксперимента выявлено, что подходящим значением коэффициента полезности для (11) является $\delta \approx 0{,}63$, следовательно, вершины, определяемые термами запроса, имеют несколько большую значимость, чем связи между термами.

Для наглядного представления и визуального анализа составлены графики. График изменения значений оценок информационно-поисковой системы по запросам коллекции №1 представлен на рис. 5.

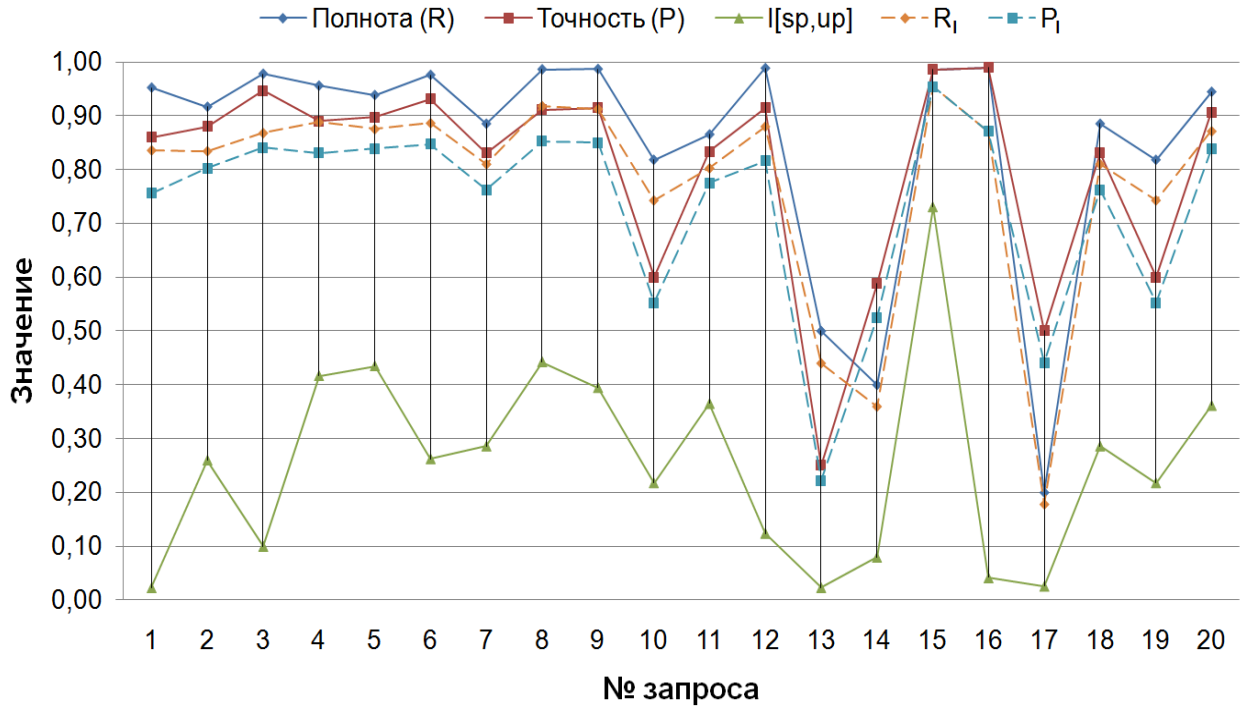

Рис. 5. График изменения значений оценок информационно-поисковой системы по запросам коллекции

Так, например, выполнение запроса №1 *Пациент принимает антикоагулянт* по тестовой коллекции №1 проведено по следующей схеме. Информационная потребность: *Найти истории болезни, в анамнезе упомянуто, что пациенту назначалась антикоагулянтная терапия*. На первом шаге с помощью лингвистической онтологии и логического вывода получаем новые зависимости, которые соответствуют новым триплетам, представляющим соответствующий документ или запрос. На следующем шаге итерации дополняем полученный набор (расширяем лингвистическую онтологию). Можно, например, отождествить *прием* и *назначение препаратов*, так как в определенном контексте одно следует из другого. В результате получаем дополнительный триплет, соответствующий запросу: *Пациенту назначен антикоагулянт*. Далее формируется правило вывода – инструкция, с помощью которой можно получить новую информацию на основе имеющейся. Общий вид: «Если (условие), то (вывод)» или «Условие, следовательно, вывод». В случае запроса №1: *Имеются противопоказания к приему антикоагулятнов, следовательно, пациент не принимает антикоагулянты*.

Для каждого триплета фактор достоверности *CF* определяется экспертом, либо как отношение частоты данного триплета в релевантной выборке к часто-

те во всей совокупности документов. Если в документе или с помощью перечисленных выше операций получен триплет с отрицательным значением, например, *Пациент не принимал варфарин*, то *CF* = 0.

В результате получен ориентированный мультиграф. Представление связей с помощью фактора достоверности представляет собой пропускные способности ребер графа. На рис. 6 отображено графическое представление триплетов запроса и документов.

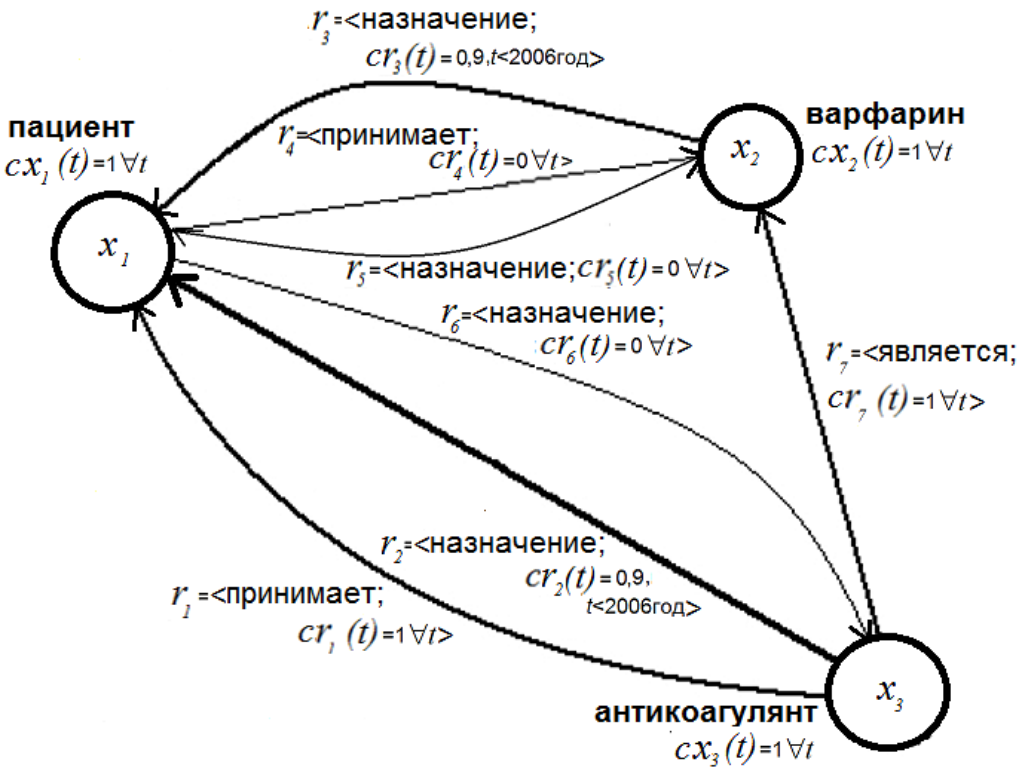

Рис. 6. Графическое представление триплетов запроса и документов
вершины – термы (x_1, x_2, x_3); дуги – триплеты: запроса (r_1, r_2),
документов (r_3, r_4, r_5, r_6); правил вывода (r_7).

Задача сводится к нахождению всех возможных путей от вершины *антикоагулянт* к вершине *пациент*. Соответствие найденных путей (триплетов, представляющих документ) потребности пользователя определяется максимальной близостью значения найденного пути значению пути в графе запроса. Документы, представленные триплетом r_3, удовлетворяют запросу с достоверностью *0,9*; документы, представленные триплетами $r_4 – r_6$, полностью не удовлетворяют запросу (*CF=0*); документы, содержащие триплет запроса r_1, полностью удовлетворяют смыслу запроса (*CF=1*); документы, содержащие триплет запроса r_2, удовлетворяет запросу на 90% (*CF=0,9)*.

В таблице 2 показано изменение значений точности и полноты поиска в зависимости от этапов обучения по результатам пяти итераций.

Таблица 2. **Изменение значений точности и полноты поиска в зависимости от этапов обучения**

№	Итерация	Полнота (R)	Точность (P)	Энтропия (H)
1	Автоматическое построение онтологии по обучающей коллекции документов	0,57	0,90	-
2	Запрос: *Пациент принимает антикоагулянт*	0,51	0,75	0,75
3	Обучение: *До 2005 г. антикоагулянтом назначают варфарин в 90% случаев*	0,69	0,79	1,79
4	Обучение: *Антикоагулянты и противосвертывающие – одно и то же*	0,83	0,81	1,81
5	Обучение: *Если пациенту не противопоказан варфарин и пациент перенес инсульт, то пациент принимает антикоагулянт с уверенностью 90%*	0,95	0,87	0,9

После пятой итерации 63% документов обучающей выборки соответствовали запросу с уверенностью 100%, остальные 37%– с уверенностью 90%. Для уверенности 80% коэффициенты *R* и *P* равны единице. Для технологии полнотекстового поиска MS SQL Server 2008, использующей статистическую модель и ранжированный поиск, получены результаты: найдено 49% документов с релевантностью больше 0,6; *R*=0,62; *P*=0,59.

На рис. 7 показан график динамики значений коэффициентов точности и полноты разработанной ИПС в зависимости от этапов обучения. По тестирующей выборке *R*=0,95 и *P*=0,9 для уверенности 90%. Следовательно, разработанная технология достаточно полно и точно выполняет поиск документов по смыслу.

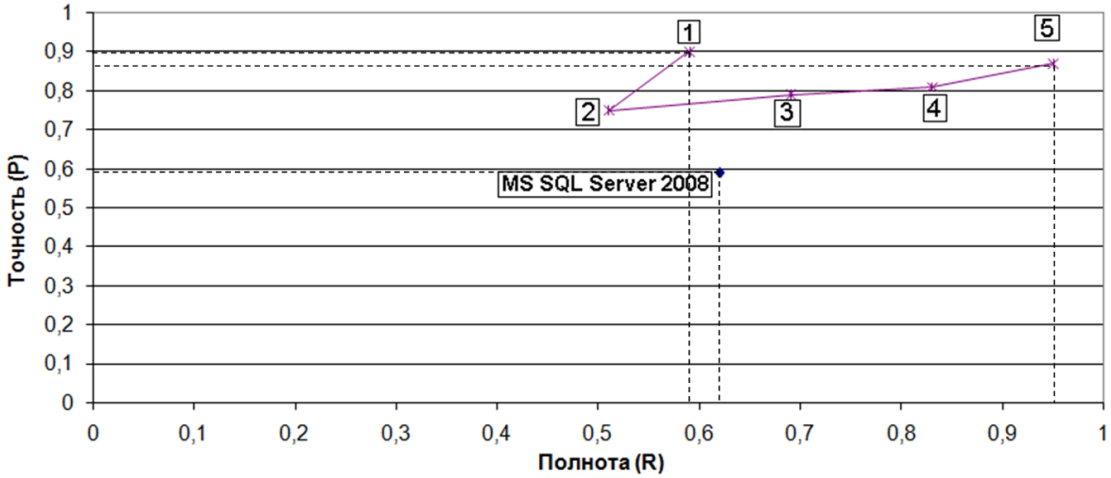

Рис. 7. **График динамики значений коэффициентов точности и полноты в зависимости от этапов обучения**

В исследовании проведена проверка запросов средней сложности: поиск отдельных терминов или параметров онтологии содержат не более 4 – 5 термов. Проведенный анализ полученных результатов подтвердил применимость модели в реальном мире.

В **заключении** приведены основные результаты диссертационной работы.

- Разработанная технология представления элементов МИС в неструктурированных текстовых массивах медицинских электронных записей с использованием дополнительных характеристик онтологических связей и предложенная методика энтропийной оценки неопределенности запроса позволяет осуществлять достаточно точный и полный смысловой поиск в медицинских документах, «слабо» чувствительный к языку, на котором написан документ, что является важным для медицинских документов, содержащих термины на русском языке и на латыни.
- Построенная модель семантического поиска для организации информационной поддержки медицинских научных исследований соответствует рассматриваемой предметной области, является адекватной и непротиворечивой.
- Сформулированная оценка релевантности смысла документов и запроса как мера схожести графов, соответствующих построенным семантическим сетям по созданной в процессе обучения КВО позволяет формировать достаточно полную выборку документов.
- Разработанный алгоритм семантического поиска на основе разработанной модели с обучением с учителем, включающий в себя правила вывода и лингвистическую онтологию для генерации новых онтологических связей позволяет учитывать потребности конкретного пользователя системы.
- Предложенный метод перехода от учетных к гибридным информационным системам позволяет использовать накопленные данные о пациенте для проведения МБИ без существенных затрат на доработку уже внедренных учетных МИС.
- Эффективность информационно-поисковой системы подтверждена в процессе практической эксплуатации программного комплекса для сбора и анализа данных в Тюменском кардиологическом центре.

Приложения содержат исходные данные, результаты численного эксперимента, список терминов, применяемых в данной работе, исходный текст некоторых программных модулей.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ ОПУБЛИКОВАНО В СЛЕДУЮЩИХ РАБОТАХ:

Публикации в рецензируемых журналах, рекомендованных ВАК

1. *Нестерова О.А., Оленников Е.А.* Некоторые подходы к решению проблемы интеграции данных результатов обследований на различном медицинском оборудовании // Вестник Тюменского государственного университета. – Тюмень: ТюмГУ, 2007. – №5. – С. 111 – 115.
2. *Захаров А.А., Нестерова О.А., Оленников Е.А.* Проблемы информационного поиска для научных исследований в медицинских информационных системах // Вестник Тюменского государственного университета. – Тюмень: ТюмГУ, 2009. – №6. – С. 215 – 219.
3. *Рычков А.Ю., Близняков А.А., Хорькова Н.Ю., Нестерова О.А.* Риск тромбоэмболических осложнений и адекватность применения варфарина при фибрилляции предсердий неклапанной этиологии // Вестник аритмологии. – СПб., 2010. – №62. – С. 41 – 44.
4. *Захаров А.А., Нестерова О.А., Оленников Е.А.* Алгоритм информационного поиска в медицинских архивах на основе контекстно-временной онтологии // Вестник Тюменского государственного университета. – Тюмень: ТюмГУ, 2010. – №6. – С. 177 – 182.

Прочие публикации

5. *Нестерова О.А., Петухов А.С.* Программные способы обеспечения безопасности в медицинской информационной системе Тюменского кардиологического центра // Безопасность информационного пространства: Материалы международной научно-практической конференции. – Екатеринбург: ГОУ ВПО УрГУПС, 2006. С. 28.
6. *Захаров А.А., Нестерова О.А., Оленников Е.А.* Медицинская информационная система для Тюменского Кардиологического Центра // Математические методы в технике и технологиях – ММТТ-20: сб. трудов XX Международной научной конференции. – Ярославль: ЯГТУ, 2007. – Т.8. – С. 157 – 161.
7. *Нестерова О.А., Оленников Е.А., Петухов А.С.* Применение INTERNET-технологий в задачах телемедицины // Высокие технологии, фундаментальные и прикладные исследования, образование: сб. трудов III международной научно-практической конференции «Исследование, разработка и применение высоких технологий в промышленности». – СПб.: Политехн. ун-т, 2007. – Т.9. – С. 212 – 213.
8. *Нестерова О.А.* Проблемы безопасности при интеграции данных различных информационных систем в медицинских учреждениях // Безопасность информационного пространства VI: сб. трудов межвузовской научно-практической конференции. – Тюмень: ТюмГУ, 2007. С. 39 – 43.
9. *Нестерова О.А.* Информационное моделирование, разработка и внедрение сервисно- и объектно-ориентированных технологий для использования цифровых и картографических активов в научных исследованиях в медицине // Современные

проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сб. научных трудов I научно-практической региональной конференции. – Тюмень: Вектор Бук, 2008. С. 71 – 75.

10. *Захаров А.А., Нестерова О.А., Оленников Е.А.* Проблемы информационного поиска и анализа данных в медицинских информационных системах // Актуальные проблемы прикладной математики, информатики и механики: сб. трудов международной конференции. – Воронеж: ВГУ, 2009. С. 82 – 85.

11. *Нестерова О.А., Оленников Е.А.* Информационный поиск и интеллектуальный анализ данных в медицинских информационных системах // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сб. научных трудов Второй научно-практической региональной конференции. – Тюмень: Вектор Бук, 2009. С. 80 – 84.

12. *Рычков Ю.А., Близняков А.А., Добрынина Л.А., Нестерова О.А.* Риск ишемического инсульта и профилактическое применение варфарина у пациентов с фибрилляцией предсердий неклапанной этиологии в кардиологической клинике // Инновационные диагностические и лечебные технологии в неврологии: Научно-практический медицинский журнал. – Казахстан, 2009. С. 10.

13. *Нестерова О.А., Оленников Е.А.* Проблема сбора и анализа данных для научных исследований в медицинских информационных системах // Искусственный интеллект: философия, методология, инновации: Материалы III Всероссийской конференции студентов, аспирантов и молодых ученых. – М.: Связь-принт, 2009. С. 371 – 373.

14. *Нестерова О.А., Близняков А.А., Рычков А.Ю., Оленников Е.А.* Разработка технологий онтологического поиска на основе энтропийной модели и их использование в системах поддержки принятия решений // Вестник аритмологии. Материалы IX Международного славянского конгресса по электростимуляции и клинической электрофизиологии сердца «КАРДИОСТИМ-2010». – СПб., 2010. С. 581.

15. *Нестерова О.А.* Использование ориентированных графов для кодификации элементов в неструктурированных текстовых массивах медицинских электронных записей // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сб. научных трудов Третьей научно-практической региональной конференции. – Тюмень: Вектор Бук, 2010. С. 181 – 185.

16. *Захаров А.А., Оленников Е.А, Пуртов В.Г, Нестерова О.А.* Подходы к созданию единого информационного пространства медицинского учреждения // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сб. научных трудов Третьей научно-практической региональной конференции. – Тюмень: Вектор Бук, 2010. С. 94 – 99.

17. *Нестерова О.А.* Контекстно-временная онтология предметной области в информационном поиске медицинских данных // Искусственный интеллект: философия, методология, инновации: Материалы IV Всероссийской

конференции студентов, аспирантов и молодых ученых. – М.: Радио и связь, 2010. – Ч.1. – С. 106 – 109.

Перечень результатов интеллектуальной деятельности

18. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009613527 (30.06.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача отделения ультразвуковой диагностики. Версия 1.0».

19. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009613529 (30.06.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача клинико-диагностической лаборатории. Версия 1.0».

20. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009613528 (30.06.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача клинического отделения. Версия 1.0».

21. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009613530 (30.06.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача-ангиохирурга. Версия 1.0».

22. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009614868 (08.09.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача кардиолога. Версия 1.0».

23. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009614869 (08.09.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача отделения рентгенохирургических методов обследования и лечения. Версия 1.0».

24. *Захаров А.А., Нестерова О.А., Оленников Е.А., Петухов А.С., Пуртов В.Г.* Свидетельство, регистрационный № 2009614867 (08.09.2009). Правообладатель ГОУ ВПО «Тюменский государственный университет» Программа «АРМ врача отделения лечебной физкультуры. Версия 1.0».

Подписано в печать 07.02.2011. Тираж 100 экз.
Объем 1,0 уч. изд. л. Формат 60Ч84/16. Заказ № 125

Издательство Тюменского государственного университета
625003, г. Тюмень, ул. Семакова, 10
Тел./факс (3452) 46-27-32
E-mail: izdatelstvo@utmn.ru