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

Ольга Андреевна НЕСТЕРОВА —
аспирант 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
Тюменский государственный университет

УДК 004.852

ПРОБЛЕМЫ ИНФОРМАЦИОННОГО ПОИСКА ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ В МЕДИЦИНСКИХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

THE PROBLEMS OF RETRIEVAL FOR RESEARCH STUDIES IN MEDICAL INFORMATION SYSTEMS

АННОТАЦИЯ. В статье рассматриваются подходы к организации хранилища данных для решения проблемы информационного поиска, сбора и анализа данных в медицинских информационных системах для обработки в научных исследованиях.

SUMMARY. The article describes the approaches to organization of data warehouse for resolving of problem of retrieval, data gathering and data analysis in medical information systems for research studies.

КЛЮЧЕВЫЕ СЛОВА. Информационный поиск, медицинская информационная система, хранилище данных, авторубрикация

KEY WORDS. Retrieval, medical information system, data warehouse, autorubricating.

Научные исследования в медицине нередко связаны со сбором и обработкой большого объема медико-биологической информации, что определяет высокий спрос на использование уже накопленных данных.

Активное внедрение в лечебные учреждения медицинских информационных систем (МИС) и перевод данных в цифровой формат, казалось бы, должны существенно сократить время, затрачиваемое на поиск и обработку необходимой информации, обеспечить удобный и оперативный доступ к данным и их автоматизированную обработку [1].

Однако на практике этого не происходит. Большинство существующих МИС изначально разрабатывались и разрабатываются как учетные системы, призванные автоматизировать деятельность лечебного учреждения (ведение истории болезни пациентов, автоматизация статистической отчетности, оптимизация планирования лечебных процессов, управление финансовыми потоками организации) и ориентированы в первую очередь на ввод данных и формирование стандартной отчетной документации [2].

Инструменты, обеспечивающие поддержку процесса научных медицинских исследований, в них отсутствуют. И если отсутствие инструментов обработки данных (например, статистика) можно заменить существующими пакетами сторонних разработчиков, то отсутствие инструментов для нефор-

мализованного поиска и извлечения необходимой исследователю информации из электронного архива заменить нечем.

Таким образом, возможность извлечения данных из электронных хранилищ в научных целях ограничена функционалом, заложенным в самих информационных системах, и врач не может самостоятельно производить нерегламентированный поиск актуальной для него информации. Порой на это тратится столько же времени, сколько необходимо для просмотра бумажных архивов, или даже больше.

Для каждой новой задачи по сбору электронных данных приходится разрабатывать дополнительные программные модули. Все это невозможно без постоянного участия IT-специалиста, т.к. нельзя заранее сформулировать все задачи поиска для того, чтобы создать регламентированные запросы, удовлетворяющие всем потребностям научного исследователя [3].

Таким образом, можно констатировать, что проблема автоматизированного информационного поиска для научных исследований в больших архивах цифровых данных является актуальной.

Основную сложность по сбору данных представляет собой поиск того или иного существенного факта или события в неформализованных данных, которые содержат описание этих фактов в произвольном виде, например, перенесенное заболевание в анамнезе пациента и т.п. К таким данным невозможно применить простые запросы, основанные на логике предикатов. А полнотекстовый поиск на большом объеме данных малоэффективен, когда необходимо найти документы, которые не просто содержат ключевые слова, а имеют определенный смысл.

Поэтому возникает необходимость использования технологий, которые позволяют без вмешательства оператора с определенной долей вероятности относить текстовые данные к заранее сформулированным темам или областям знаний, а в нашем случае — к неявному упоминанию того или иного факта или события, например, если необходимо найти в описании анамнеза пациента упоминание о том или ином заболевании или осложнениях.

В данной работе предлагается использовать метод авторубрикации (автоматического отнесения текста к определенной теме), который применяется в библиотечно-справочных системах, в патентных системах, системах фильтрации спама и пр. [4].

Обоснованно предполагаем, что основной характеристикой текста, содержащего описание необходимых фактов или событий, в задаче поиска медицинских данных является набор терминов и их синонимов, встречающихся в тексте с определенной долей значимости — весом.

Обозначим через:

T — пространство признаков (терминов);

h — вес признака, для определения наиболее существенных признаков наблюдаемого класса, учитывающий его относительную ценность (значимость для искомого факта или события).

В основе технологии обучения рубрикатора лежит построение обучающей выборки — набора текстовых данных, поставленных в соответствие искомым фактам. Процесс построения является итеративным и заключается в построении экспертом некоторого набора терминов, характеризующих эти факты или события.

В результате автоматического морфологического и синтаксического анализа получаем набор терминов, их вес (для определения, какие признаки наиболее существенны для данной выборки) и частоту появления в тексте. Этот набор должен с достаточной полнотой описывать тот или иной факт, который необходимо найти.

Критерием того, что в тексте содержится упоминание искомых фактов или событий, будет являться статистическая близость частотного распределения терминов, выделенных в процессе обучения вектору частот обучающей выборки. Для термина T_i вектора $V_i=\{v_{ik}\}$ это частоты появления термина в k-м документе обучающей выборки (k-м испытании из n документов).

Будем использовать следующую модель задачи классификации:

Ω — множество текстов.

$\omega : \overline{\omega} \in \Omega$ — объект авторубрикации (документ).

$g(\omega) : \Omega \to M, M = \overline{1,m}$ — индикаторная функция, разбивающая пространство образов Ω на m непересекающихся классов $\Omega^1,\Omega^2,\ldots, \Omega^m$ (искомых тем или фактов). Индикаторная функция неизвестна наблюдателю.

$v(\omega):\Omega->V$ — функция, ставящая в соответствие каждому объекту ω точку $v(\omega)$ в пространстве признаков V. Вектор $v(\omega)$ — это образ объекта, воспринимаемый наблюдателем. В пространстве признаков определены непересекающиеся множества точек K_i, где $i = \overline{1,m}$, соответствующие шаблонам одного класса.

$V = (v(\omega_i), g(\omega_i)), i=1..N$ множество прецедентов.

$\hat{g}(t) : V \to M$ — решающее правило — оценка для $g(\omega)$ на основании $v(\omega)$, то есть $\hat{g}(v) = \hat{g}(v(\omega))$.

Пусть $v_j = v(\omega_j), j = \overline{1,N}$ — доступная наблюдателю информация о функциях $g(\omega)$ и $v(\omega)$, но сами эти функции наблюдателю неизвестны.

Тогда $(g_j, v_j), j = \overline{1,N}$ – есть множество прецедентов, то есть шаблонов текста, правильная классификация которых известна.

Задача заключается в построении такого решающего правила $\hat{g}(v)$, чтобы распознавание проводилось с минимальным числом ошибок.

$\hat{g}(v)$ — решающее правило, тогда $\hat{g}(v) : T \to M$.

Качество решающего правила измеряют частотой появления правильных решений. Поэтому его оценивают, ставя в соответствие множеству объектов Ω некоторую вероятностную меру.

Выбор решающего правила исходит из минимизации $d(g,\hat{g}) \to min$, где d — метрика, мера близости функций $g(\omega)$ и $\hat{g}(t(\omega))$. Построение $\hat{g}$ — это и есть задача обучения. $\hat{g}$ — это ученик, процедура формирования — это учитель, прецеденты — это обучающая последовательность.

Добавим каждому признаку, термину, свой вес. Тогда полностью степень близости прецедента по всем признакам можно вычислить, используя обобщенную формулу вида:

$$\frac{\sum_i h_i \cdot sim(v_{ij}, v_{ik})}{\sum_i h_i}$$

где h_i — вес i-го признака, sim — функция подобия (метрика), v_{ij} *и* v_{ik} — значения признака (частота термина) v_i для текущего случая (j-го события) и k-го прецедента, соответственно [5].

В нашем случае с учетом главных целей исследования в качестве метрики выбираем расстояние — χ^2:

$$sim(k,j)=\sqrt{\sum_{i=1}^{m}\frac{(v_{ik}-E(v_{ik}))^2}{E(v_{ik})}+\sum_{i=1}^{m}\frac{(v_{ij}-E(v_{ij}))^2}{E(v_{ij})}}.$$

Данное выражение отражает степень схожести исследуемого документа с документами обучающей выборки и принимает значение [0,1].

Таким образом, после вычисления степеней близости все прецеденты выстраиваются в единый ранжированный список, это и будет являться для них оценкой релевантности (соответствия) найденного текста запросу.

Для оптимизации поиска ключевых терминов в большом объеме текстовых данных необходимо создание специализированного хранилища.

Хранилище данных содержит:

• модуль извлечения из оперативных источников данных, позволяющий интегрировать данные различных информационных систем;

• метатаблицы, содержащие описание формализованных данных и текстовых полей с неструктурированным текстом, а также ссылки на текстовые массивы — документы, экспортированные из внешних МИС;

• модуль преобразования данных, включающий в себя механизм индексации текстовых документов, что существенно сокращает время выполнения поиска, в том числе и авторубрикации;

• модуль загрузки данных в хранилище;

• визуальный редактор, отображающий медицинские факты и связи между ними в удобном для восприятия виде для конструирования запросов к данным.

Часть данных преобразовывается по принципу выписного эпикриза, т.к. именно этот документ является основным для врача, что позволяет отфильтровать наиболее значимые факты для поиска и отбросить второстепенные.

Визуальный редактор включает в себя инструментарий для авторубрикации неструктурированных данных, проиндексированных в процессе преобразования.

Предложенный подход сокращает время и повышает качество поиска, доступность и достоверность информации, позволяет получать более точную выборку схожих документов на одних и тех же наборах данных. Благодаря этому повышается качество самих научных исследований, т.к. научный работник сможет больше времени уделять поиску новых связей, зависимостей и т.п.

СПИСОК ЛИТЕРАТУРЫ

1. Нестерова О.А., Оленников Е.А. Информационный поиск и интеллектуальный анализ данных в медицинских информационных системах (статья) // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сборник научных трудов. Тюмень: Вектор-Бук, 2009.

2. Эльянов М.М. Медицинские информационные технологии. Каталог. Вып. 7. М.: Третья медицина, 2007. 300 с.

3. Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч. 2. Исследование медицинских технологических процессов на основе интеллектуального анализа данных. М.: Физматлит, 2006. 144 с.

4. Титов Ю.В., Фарсобина В.В. Сравнительное тестирование авторубрикаторов // «Интеллектуальные информационные технологии. Концепции и инструментарий» Сб. тр. Института системного анализа РАН, 2005. Mode access: http://www.cognitive.ru/innovation/sbornic7/index.htm.

5. Косинов Д.И. Использование статистической информации при выявлении схожих документов // Интернет-математика 2007: сб. работ участников конкурса науч. проектов по информ. поиску. Екатеринбург, 2007. С. 205-207.

Анатолий Юрьевич ОЩЕПКОВ —
аспирант кафедры информационных систем
aoschepkov@gmail.com

Александр Григорьевич ИВАШКО —
зав. кафедрой информационных систем,
доктор технических наук, профессор
ivashco@mail.ru

Институт математики и компьютерных наук
Тюменский государственный университет

УДК 519.152

ПОСТРОЕНИЕ АЛГОРИТМА НАХОЖДЕНИЯ «ОПТИМАЛЬНОГО УЗЛА» ДЛЯ РАСПРЕДЕЛЕНИЯ ТРАФИКА «WEB-КОНФЕРЕНЦИИ» В РАСПРЕДЕЛЕННОЙ СИСТЕМЕ

ALGORITHM DEVELOPMENT FOR FINDING THE "OPTIMAL NODE'S" FOR TRAFFIC DISTRIBUTION OF "WEB CONFERENCE" IN CLUSTERING SYSTEM

АННОТАЦИЯ. В статье рассматривается построение алгоритма балансера в кластерной системе для проведения «WEB-конфереций», основанного на теории массового обслуживания.

SUMMARY. The article describes the development of algorithm balancer in the cluster system for «Web Conference», based on the theory of waiting lines.

КЛЮЧЕВЫЕ СЛОВА. Веб-конференция; балансер, СМО.
KEY WORDS. Web Conference; Load Balancing; theory of waiting lines.

В последнее время, в связи с внедрением в учебный процесс ВЕБ-2.0 технологий, широкое распространение получили так называемые «вебинары», или «веб-конференции». Среди основных возможностей конференцсвязи можно выделить: проведение слайдовых презентаций; VoIP (аудиосвязь через компьютер в режиме реального времени; видео в режиме реального времени); Whiteboard (электронная доска для комментариев); Screen sharing (удаленный рабочий стол) — совместное использование приложений)).

Для университета использование веб-конференций — это возможность для преподавателя работать удаленно так, чтобы студенту казалось, что преподаватель находится рядом. Однако применение этих технологий затруднено в связи с необходимостью использования сервера с высокой производительностью и широкополосного интернет-канала.

Одним из эффективных способов повышения производительности вебинаров является их кластеризация.

В кластерной системе можно выделить следующие группы участников (рис. 1):