УДК 613.62:004.4

ПРОБЛЕМА ПОШУКУ МЕДИЧНОЇ ІНФОРМАЦІЇ В ЕЛЕКТРОННИХ ФАЙЛОВИХ АРХІВАХ ТА СТВОРЕННЯ ПОВНОТЕКСТОВИХ БАЗ ДАНИХ В КЛІНІЦІ ПРОФЕСІЙНИХ ЗАХВОРЮВАНЬ

А. Б. Ященко, Д. А. Ященко

Український НДІ промислової медицини

У статті окреслені труднощі автоматизованого інформаційного пошуку, представлений досвід реєстрації, збереження медичної інформації в електронних файлових архівах, структуризації та аналізу інформації в них. Проведене тестування неспеціалізованих інформаційно-пошукових систем для створення повнотекстових баз даних історій хвороб з урахуванням релевантності та експертної пертинентності.

Ключові слова: повнотекстові бази даних, інформаційно-пошукові системи, релевантність, медичні інформаційні системи.

ПРОБЛЕМА ПОИСКА МЕДИЦИНСКОЙ ИНФОРМАЦИИ В ЭЛЕКТРОННЫХ ФАЙЛОВЫХ АРХИВАХ И СОЗДАНИЕ ПОЛНОТЕКСТОВЫХ БАЗ ДАННЫХ В КЛИНИКЕ ПРОФЕССИОНАЛЬНЫХ ЗАБОЛЕВАНИЙ

А. Б. Ященко, Д. А. Ященко

Украинский НИИ промышленной медицины

В статье очерчены основные проблемы автоматизированного информационного поиска, внедрения баз данных и медицинских информационных систем в практическую деятельность лечебных учреждений, представлен опыт регистрации, сохранения медицинской информации в электронных базах данных, структуризации и анализа информации в них. Проведено тестирование неспециализированных информационно-поисковых систем для поиска медицинской информации в электронных файловых архивах и создания полнотекстовых базах данных историй болезни с учетом релевантности найденной информации. Описаны достоинства и недостатки различных информационно-поисковых систем, в частности при поиске информации в русско- и украиноязычных полнотекстовых информационных массивах. В качестве объекта исследования были использованы электронные файловые архивы и базы данных Украинского НИИ промышленной медицины по нозологиям «радикулопатия» и «радикуломиелопатия».

Ключевые слова: полнотекстовые базы данных, информационно-поисковые системы, релевантность, медицинские информационные системы.

THE PROBLEM OF MEDICAL INFORMATION SEARCH IN ELECTRONIC FILE ARCHIVES AND FULL-TEXT DATABASES CREATION IN THE CLINIC OF OCCUPATIONAL DISEASES

A. B. Yashchenko D. A.Yashchenko

Ukrainian Research Institute of Industrial Medicine

The article outlines main problems of automated data search, application of databases and medical information systems in practice of medical institutions is presented. The practice of registration, storage of health information in electronic databases, structuring and analysis of data available is presented. Unspecialized retrieval systems for searching medical information in electronic file archives and creating full-text databases of medical records considering the relevance of found information have been tested. The advantages and disadvantages of various retrieval systems are reviewed, particularly at searching for Russian and Ukrainian full-text data arrays. The electronic file archives and databases of Ukrainian Research Institute of Industrial Medicine after nosology "radiculopathy" and "radiculomyelopathy" were used as the object of research.

Key words: full-text databases, retrieval systems, relevance, medical information systems.

Вступ. Безпрецедентні темпи розвитку обчислювальної техніки і мережевих технологій привели до того, що з 90-х років минулого століття без комп'ютерів практично всі сфери діяльності людини неможливі. Розвиток Інтернету останніми роками привів до інформаційного вибуху - явища, добре знайомого не лише фахівцям, але і простим користувачам комп'ютерної техніки. Система електронних інформаційних комунікацій, що формується на очах нинішнього покоління, кардинально змінює ситуацію у сфері збору, зберігання і обробки даних. В узагальненому вигляді такі підходи до стандартизації інформації сьогодні трактуються як створення «цифрових» або «електронних» бібліотек і повнотекстових баз даних (БД).

Під електронною бібліотекою або повнотекстовою БД розуміють розподілену інформаційну систему, що дозволяє нагромаджувати, зберігати і використовувати всілякі колекції електронних документів. Сьогодні на зміну інформаційному обслуговуванню на друкованих носіях приходить забезпечення користувачів на основі доступу до електронних документів через глобальну мережу передачі даних - Інтернет.

Медицина не стала винятком, і одним з найактуальніших завдань в сучасній охороні здоров'я є реєстрація, аналіз і зберігання медичної інформації в комп'ютерних системах. Ця проблема може вирішуватись шляхом застосування медичних інформаційних систем (МІС).

«У нинішньому розумінні, медичною інформаційною системою називається комплекс методологічних прийомів, технічних засобів і алгоритмів керування, призначених для збору, зберігання, обробки й передачі інформації в лікувально-профілактичних установах» [1].

Сьогодні більшість лікувальних установ в тій чи іншій мірі обладнані комп'ютерною технікою, але переважно вона використовується як заміна друкарської машинки. Лише в деяких лікувально-профілактичних установах нашої держави впроваджуються МІС, в основному це клініки медичних університетів та науково-дослідних установ. Бурхливий розвиток комп'ютерних та інформаційних технологій призводить до того, що пересічний користувач без спеціальної освіти не може осягнути та повністю використовувати їх можливості.

Крім освітніх проблем, таке становище пов'язане з проблемами, які ускладнюють впровадження МІС в практику охорони здоров'я: моральна та фізична застарілість МІС, що перестали підтримуватись виробниками, відірваність лікаря від працюючої МІС (доступ через оператора), необхідність автоматизувати кожне робоче місце та дорожнеча цього процесу тощо [2, 3].

Як початковий етап впровадження інформаційних технологій в медичну практику для збереження інформації в електронному вигляді пропонуємо створення електронних архівів медичних документів (описів хворих, щоденників, даних додаткових методів обстеження та виписок із історій хвороби). Це не потребує значних додаткових ресурсів, лише певної організації в структуризації та зберіганні інформації.

В неврологічній клініці Українського НДІ промислової медицини така робота триває з 2002 року по теперішній час. У зв'язку з виконанням науково-дослідної роботи (НДР): «Професійна радикулопатія у гірників залізорудної промисловості (розповсюдженість, особливості перебігу, прогноз)», номер державної реєстрації 0105U006083, виникла необхідність створення повнотекстових БД за нозологіями «радикулопатія» та «радикуломієлопатія» професійного і непрофесійного ґенезу. При пошуку інформації в поширеному електронному файловому архіві виникли проблеми вибору пошукових інформаційних систем та оцінювання релевантності і пертинентності знайденої інформації. Різні алгоритми роботи пошукових систем давали різні результати навіть серед однорідних за форматом файлів.

Метою дослідження було тестування й оцінювання можливості використання неспеціалізованих інформаційно-пошукових систем для пошуку текстової медичної інформації та створення повнотекстових баз даних за різними нозологіями з електронних файлових архівів, визначення найбільш придатної для цих потреб інформаційно-пошукової системи шляхом оцінювання релевантності та експертної пертинентності знайденої медичної інформації. В якості експертів виступали виконавці зазначеної НДР.

Матеріали та методи дослідження. Для організації електронного архіву історій хвороб пацієнтів неврологічного стаціонару клініки Українського НДІ промислової медицини медична документація, що роздруковувалася, зберігалася також в електронному вигляді. До неї входили: опис хворого, щоденники, дані додаткових методів обстеження, консультацій фахівців, виписка з історії хвороби тощо. Файли в електронному архіві розподілені за роками: з 2002 по 2009 рік. Електронний архів структурований за принципом: одна історія хвороби - один файл. Електронний варіант історії хвороби зберігався в файлах форматів doc або rtf. Назва файлу вміщала в себе прізвище, ініціали пацієнта, дату госпіталізації. Всього електронний архів розміром 876 Мб складався з 15 036 файлів. Усі документи невеликі за розміром, від 35 до 138 кБ. З 2002 по 2004 рік історії хвороби велися російською мовою, з 2005 по теперішній час - українською.

Відбір пошукових систем проводився за такими їх характеристиками, як: відповідні апаратні вимоги та програмна платформа; наявність пробної версії й її обмеження; максимальний об'єм бази даних (Гб); максимальна кількість баз даних; максимальна кількість файлів в БД; типи індексованих файлів (обов'язково doc, txt та rtf); швидкість індексування, Мб/хв; максимальний об'єм одного індексованого документа; тип пошуку: по документах, по сторінках, по всьому текстовому полю БД; організація БД і видачі: подокументна, посторінкова, пофрагментна; середній час пошуку за запитом в БД об'єму близько 1 ГБ; можливий об'єм пошукового запиту (кількість слів); можливість установки відстані між словами при пошуку, використання логічних операторів в запиті, пошук по атрибутах документів (дата, автор, розмір, назва і ін.), сортування результатів пошуку, наявність морфологічного словника, наявність додаткового лінгвістичного забезпечення; пошук по сенсу.

Проводилося тестування неспеціалізованих пошукових систем.

Для наукових досліджень необхідно було відібрати пацієнтів з радикулопатіями і радикуломієлопатіями професійного та непрофесійного ґенезу Даними статистичного відділу клініки НДІ користуватися було неможливо, тому що облік проводився тільки за основним діагнозом, тобто поза увагою залишалися хворі з супутніми захворюваннями. За допомогою пошуку в електронному архіві можливо було визначити пацієнтів, у яких радикулопатія була супутнім діагнозом.

Терміни «релевантність» та «пертинентність» тлумачаться по різному [4]. Тому адекватна оцінка результатів роботи пошукових систем у вказаному вище медичному електронному архіві потребує термінологічного уточнення. Так релевантність (лат. relevo - піднімати, полегшувати) в інформаційному пошуку - це семантична відповідність пошукового запиту і пошукового образу документа. У більш загальному сенсі, одне з найближчих понять до «релевантності» - «адекватність», тобто не лише оцінка ступеня відповідності, але і ступеня практичної застосовності результату [5]. Тобто релевантність - стосовно результатів роботи пошукової системи - ступінь відповідності запиту і знайденого, доцільність результату, згідно з міждержавним стандартом цей термін визначається, як «відповідність отриманої інформації інформаційному запиту» [6]. Для реалізації поставленої мети автори під релевантністю розуміли формальну відповідність знайденої інформації заданим ключовим словам «радикулопатія», «радикуломієлопатія», «мієлопатія». Пертинентність (лат. pertineo - торкаюся, відношуся) - відповідність знайдених інформаційно-пошуковою системою документів інформаційним потребам користувача [6], незалежно від того, як повно і як точно ця інформаційна потреба виражена в тексті інформаційного запиту. Інакше кажучи, це співвідношення об'єму корисної інформації до загального об'єму отриманої інформації. Для реалізації поставленої мети автори під пертинентністю (експертною пертинентністю) розуміли адекватність знайденої інформації потребам створення повнотекстових БД пацієнтів.

Результати та їх обговорення. Критеріями для включення до БД була наявність у графі "діагноз" назв нозологій: «радикулопатія» чи «радикулопатия», «радикуломієлопатія» чи «радикуломиелопатия», «мієлопатія» чи «миелопатия». За результатами пошуку програмними засобами, після ручної перевірки файлів створено дві повнотекстові БД історій хвороб пацієнтів з радикулопатією - кількість файлів 9926 (далі БД «Радикулопатія»), та з радикуломієлопатією чи мієлопатією - кількість файлів 107 (далі БД «Мієлопатія»). До 2007 року пошук текстової інформації в електронному архіві проводився за допомогою стандартних засобів операційної системи Microsoft Windows XP Professional Service Pack 1 та 2, а також програми AV Search (версія 3.13), шляхом перебору файлів. Далі, в зв'язку зі збільшенням об'єму файлового архіву, пошук став тривалішим, займав від 8 до 20-25 хвилин за одним запитом, стали з'являтися помилки, а програми, що використовувались, видавали за однаковими пошуковими термінами різні результати. Дані пошуку інформації цими засобами для створення БД за нозологіями «радикулопатія», «мієлопатія» представлені в таблиці 1.

За результатами пошуку цими програмними засобами видно, що кількість правильно знайдених файлів для БД «Радикулопатія» коливається від 79,2 до 81,4, для БД «Мієлопатія» від 40,2 до 74,8 %.

Результати пошуку засобами Windows XP підтверджують дані інших фахівців, що словоформи української мови не враховуються, а російської мови враховуються лише за рахунок відкидання закінчень при контекстному пошуку [7]. Проте, навіть при такому пошуку не виявлена велика кількість файлів з діагнозами радикулопатія та мієлопатія. Деякі видані документи взагалі не відповідали запиту, ні за змістом слів, ні за сенсом.

При використанні логічних операторів програма AV Search не визначила наявних файлів, найімовірніше за рахунок невідповідного аналізу української морфології. Інтерфейс програми кращий ніж у стандартному пошукувачу Windows XP, на екран виводиться текст знайдених файлів та відзначаються входжен-

Таблиця 1. Результати пошуку програмами, що проводять перебір файлів

Пошукові терміни запиту	Кількість знайдених файлів	
	Пошук засобами Windows	AV Search
миелопатия	3G	41
мієлопатія	5G	2
миелопатия OR мієлопатія	*	-
миелопат	61	41
мієлопат	-	2
радикулопатия	1954	21G9
радикулопатія	4234	21
радикулопатія OR радикулопатия	*	-
радикулопат	7S61	SG7S

* використання логічних операторів не підтримується

ня. Але під час пошуку цією програмою при кількості знайдених файлів близько 8000 комп'ютер видавав повідомлення «Out of system resourses».

Тривалий і незадовільний пошук інформації цими засобами спричинив необхідність перегляду підходу до нього: був проведений огляд доступних в Інтернеті інформаційно-пошукових систем, що спочатку проводять індексування масивів інформації, а потім пошук за індексом, таким чином значно скорочують час пошуку. Огляд Інтернет-ресурсів показав, що існує декілька десятків колективів, які проводять роботу з розроблення систем пошуку інформації не лише на рівні програмування, а й на серйозному науковому рівні з дослідженнями в області лінгвістики, семантики, аналізу сенсу тексту тощо, необхідних для адекватної роботи інформаційно-пошукових систем. Результати тестування пробними версіями цих пошукових систем представлені в таблиці 2.

Таблиця 2. Результати пошуку інформаційно-пошуковими системами, які індексують масив даних

Пошукові терміни запиту	Кількість знайдених файлів					
	Архіваріус 3GGG	Copernic Desktop Search	Windows Desktop Search 4.G	MBD SE 2.2	Google Desktop	Advansed Document Setup
миелопатия	3G	7	7	7	7	3G
мієлопатія	54	11	12	71	11	55
миелопатія OR мієлопатія	46	1S	19	77	*	S3
миелопат	1G6	1S	27	71	-	42
мієлопат	1G6	26	18	-	-	66
радикулопатия	3197	2S99	2S96	3G37	2931	3G36
радикулопатія	669S	6325	6526	6756	6327	6463
радикулопатія OR радикулопатия	9954	91G6	93G2	9669	*	9549
радикулопат	9954	9525	9725	6756	-	1GG42

* використання логічних операторів програмою не підтримується

Тривалість створення індексу програмами, вказаними в таблиці 2, для електронного архіву розміром до 1 Гб відрізнялась не суттєво, складала близько 7-12 хвилин. Пошук інформації за одним пошуковим терміном тривав близько 1 с. Із таблиці 2 видно, що у більшості програм знайдена кількість файлів далека від належної: для БД з мієлопатією - 107 файлів, для БД з радикулопатією - 9926.

Так, словоформи української та російської мови враховувалися недостатньо та лише за рахунок відкидання закінчень при контекстному пошуку. Дані таблиці 2 свідчать, що навіть використання ручного обрізання слів було неефективне. Слід зазначити, що при пошуку з запитами без закінчень, при можливості, використовувався пошук за маскою і в таблицях представлені кращі результати. В програмах Windows Desktop Search 4.0, Copernic Desktop Search-Home версія 3.4.0, Google Desktop Search відсутні морфологічні словники української та російської мов, тому релевантність їх видачі низька. Тобто зарубіжні програми не завжди адекватно використовують лінгвістичне і програмне забезпечення для аналізу текстів українською та російською мовами, також у програмах Copernic Desktop Search-Home, Google Desktop Search не передбачені функції копіювання знайдених файлів.

Результати пошуку MBD Search Engine 2.2 кращі, але для БД «Мієлопатія» релевантність 72 % занизька з точки зору якісного пошуку. Хоча в програмі є досить зручні опції з пошуку тексту на певній відстані від 1 до 100 слів, та можливість посторінкового пошуку, що вигідно відрізняють її серед інших програм. Добрі результати пошуку показала програма Advansed Document Setup 1.2, але остання версія програми вийшла ще в 2002 році, вона не підтримується розробником і не має можливості переміщення та копіювання знайденої інформації.

Найбільш ефективний пошук проводила пошукова система Архіваріус 3000, в якій передбачене морфологічне оброблення і російської, і української мов. Найбільша точність видачі досягається при булевому пошуку з використанням логічних операторів та за рахунок відкидання закінчень. Так для БД «Радикулопатія» знайдено 100 % файлів, неточність з надлишком у 28 файлів (коефіцієнт пошукового шуму - 0,3 %) спричинена наявністю пошукових термінів у розділі анамнезу, та атрибутах файлів, тобто пертинентність для неї складала 99,7 %. Для БД «Мієлопатія» релевантність склала 99,1 % за рахунок орфографічних помилок. Пертинентність в даному випадку практично збіглася з релевантністю. Програма «Архіваріус 3000» також надала можливість пошуку та позбавлення дублікатів файлів, швидше за інші програми проводила індексування масиву файлів. За її допомогою можливо проводити досить точний аналіз медичної текстової інформації в БД, наприклад пошук інформації про пацієнтів з певними клінічними відмінностями захворювань, діагностичними та лікувальними процедурами, віковими та стажовими даними, професійними шкідливостями тощо.

Отже, створення комп'ютерних електронних архівів і повнотекстових БД медичної документації без додаткових коштів і при мінімальних затратах робочого часу може використовуватись як засіб реєстрації, аналізу і зберігання медичної інформації лікувально-профілактичних установ, зокрема в клініці професійних захворювань. На початковому етапі інформатизації медичної допомоги населенню вони можуть бути заміною МІС, впровадження яких в подальшому полегшиться, враховуючи можливість використання вже збереженої в електронному вигляді інформації. Це вкрай важливо з огляду на реформування медичної галузі та впровадження засад сімейної медицини, де лікарю прийдеться зустрічатися з різноманітними клінічними випадками. Використання пошукових систем та повнотекстових БД в діагностично-лікувальному процесі дозволить проводити пошук у великих масивах інформації, швидко знаходити подібні складні клінічні випадки захворювань, полегшити вибір лікування пацієнтів.

Висновки. 1. Застосування неспеціалізованих інформаційно-пошукових систем дозволяє індексувати чисельні електронні файлові архіви, створювати повнотекстові БД та проводити швидкий пошук медичної інформації.

2. Тестування пошукових програм у роботі з україномовними медичними текстовими інформаційними масивами даних розміром близько 1 Гб виявило найбільшу ефективність у «Архіваріусу 3000» з релевантністю та пертинентністю більше 99 %.

3. Створення повнотекстових БД «Радикулопатія» та «Мієлопатія» та застосування автоматизованої пошукової системи дозволило проводити аналіз інформаційно-пошукового масиву файлів історій хвороби за професіями пацієнтів, ускладненнями захворювань, супутньою патологією.

4. Функціонування повнотекстових БД робить можливим оцінювання й аналіз великих масивів медичної інформації за обмежений проміжок часу та суттєво полегшує встановлення діагнозу, вибір адекватного лікування, а також розв'язання інших практичних та наукових завдань з діагностики та профілактики захворювань, у тому числі і професійних.

Література

1. Мінцер О. П. Інформатика та охорона здоров'я / О. П. Мінцер // Медична інформатика та інженерія. - 2010. - N° 2. - С. 8-12.
2. Болгов М. Ю. Проблеми впровадження медичних інформаційних систем в Україні / М. Ю. Болгов, М. Д. Тронько // Медична інформатика та інженерія. - 2009. - N° 2. - С. 37-42.
3. Качмар В. О. Медичні інформаційні системи - стан розвитку в Україні / В. О. Качмар // Український журнал телемедицини та медичної телематики. - 2010. - Т. 8., №9 1.
4. Основы информационной культуры. Версия 1.0 [Электронный ресурс]: электрон. учеб. пособие / В. П. Казанцева, Т. А. Вольская, Е. М. Згурская и др. ; ред. Е. Г. Кривоносова. - Электрон. дан. (3Мб). - Красноярск: ИПК СФУ 2008. - Основы информационной культуры : УМКД № 208. - 2007. - 142 с.
5. Словарь по кибернетике; под ред. В. С. Михалевича. - Изд. 2-е. - К. : аРед. УСЭ, 1989. - 751 с.
6. Поиск и распространение информации. Термины и определения : ГОСТ 7.73 -96. - [Введен01.01.1998]. - М, 1998. - 15с. - (Межгосударственный стандарт. Система стандартов по информации, библиотечному и издательскому делу).
7. Захарченко В. М. Программы поиска информации в полнотекстовых базах данных. Аналитический обзор; Рос. гос.технолог. ун-т - М. : МАТИ, 2005. - 24 с. - http://rstu.ru/metods/books/zakhar.pdf.