УДК 02(477):004

Тетяна ШЕРЕПА,
мол. наук. співробітник НБУВ

Інформаційна технологія виділення та обробки знань у CDS/ISIS-сумісних базах даних

Визначено методику інтелектуальної обробки бібліографічної інформації у CDS/ISIS-сумісних базах даних для виділення нових знань з використанням пакета прикладних програм IDAMS, який вільно поширюється UNESCO й орієнтований на проведення статистичних досліджень і аналізу даних у великих інформаційних масивах.

Ключові слова: сховища даних, бази даних, бази знань, аналіз даних, класифікація, кластеризація, екстракція знань, CDS/ISIS, UNESCO.

У системі документних комунікацій «*джерело інформації – видавництво – документ – бібліотека – споживач інформації*» основною вважається суперечність між обсягом документних ресурсів і можливістю їх використання суспільством. Значно сприяють вирішенню цієї проблеми створення програмних засобів, які призначені для збору, зберігання, організації даних в електронній формі і забезпечення доступу до них з використанням комп'ютерних мереж для пошуку, отримання і обробки необхідної інформації.

Сьогодні актуальним стає вже не розв'язання суперечності в рамках підсистеми «*документ – споживач інформації*», а більш загальна проблема включення всього інтелектуального потенціалу суспільства в систему документних комунікацій і забезпечення його подальшого ефективного використання [4].

Комп'ютерні системи повинні не лише зберігати і використовувати великі обсяги інформації, а й ефективно допомагати користувачам знаходити нові шляхи вирішення проблем. Щоб перебороти обмеження існуючих систем, необхідно розробити методи для з'єднання людського інтелекту і комп'ютерних систем. Для вирішення цього питання призначені аналітичні технології – методики, які на основі певних моделей, алгоритмів, математичних теорем дозволяють по відомих даних оцінити значення невідомих характеристик і параметрів. Технології інтелектуального аналізу використовують складний статистичний аналіз і моделювання для знаходження моделей і відношень, прихованих у базі даних. Метою інтелектуальних технологій є знаходження нового знання, яке користувач може надалі застосувати для поліпшення результатів своєї діяльності [1].

Автоматична обробка текстових і графічних файлів з наступною майже миттєвою передачею на відстань є основою сучасних інформаційних технологій. Але, незважаючи на бурхливий розвиток інтернету, цифрових телекомунікацій, все більш «розумних» операційних систем та прикладних програм, автоматизована обробка і швидка передача даних («сирої інформації») є лише середовищем-носієм *знання*, а не самим знанням в змістовній формі. Сутність інформаційної революції, яка відбувається зараз, є новий, якісний стрибок – перехід від автоматизованої обробки інформації до комп'ютерного представлення і обміну *чистим знанням* [6].

У даній роботі ставиться ціль визначити методику інтелектуальної обробки бібліографічної інформації у CDS/ISIS-сумісних базах даних для виділення та обробки нових знань з використанням пакета прикладних програм IDAMS, який вільно поширюється UNESCO, орієнтованого на проведення статистичних досліджень і аналізу даних у великих інформаційних масивах.

Елементи автоматичної обробки і аналізу даних, що називають *Data Mining* (знаходження знань) стають невід'ємною частиною концепції інформаційних сховищ даних (data warehouse) та організації інтелектуальних обчислень. Сховище даних – це предметно-орієнтований, інтегрований, прив'язаний до часу, незмінний набір даних для підтримки процесу прийняття рішень [1]. Простий доступ користувача до сховища даних забезпечує тільки отримання відповідей на питання, котрі були задані, в той час, як технологія data mining дозволяє побачити («знайти») приховані правила і закономірності у наборах даних, які користувач не може передбачити, і застосування яких може сприяти виявленню більш ефективного результату.

Інформація в сховищі об'єднується в цілісну структуру по різних рівнях деталізування, що забезпечує необхідні користувачам міри узагальнення даних. У цій концепції центральне місце займають метадані – дані про дані. Управління метада-

ними забезпечує автоматизацію процесу збору і обробки інформації. При цьому в сховищі також вміщуються результати перетворення даних, їх сумаризації і верифікації.

Чим більше аналітик може «грати» з даними, будувати моделі, оцінювати результати, тим кращий може бути результат. Робота з даними стає більш ефективною, коли можлива інтеграція таких компонентів: візуалізація, графічний інструментарій, засоби формування запитів, оперативна аналітична обробка, що дозволяють зрозуміти дані й інтерпретувати результати, і, нарешті, самі алгоритми, які будують моделі [1].

З основних видів моделей, які використовуються для виявлення й аналізу знань на основі даних інформаційного сховища, варто виділити принаймні шість методів [1]:

1) *класифікація* (виявлення ознак, котрі характеризують групу, до якої належить той чи інший об'єкт, за допомогою аналізу вже класифікованих об'єктів і формулювання деякого набору правил);

2) *кластеризація* (виділення різних однорідних груп даних, відрізняється від класифікації тим, що самі групи заздалегідь не задані);

3) *регресія* (кількісне вираження відношення між змінними у виді деякої комбінації цих змінних, яке використовується для передбачення значення, що може приймати цільова змінна, яка обчислюється на заданому наборі значень вхідних змінних);

4) *прогнозування часових послідовностей* (побудова математичної моделі за «історичною» інформацією, що зберігається в інформаційних сховищах у вигляді часових рядів);

5) *асоціація* (має місце в тому випадку, якщо кілька подій пов'язані між собою);

6) *послідовність* (має місце, коли існує ланцюжок пов'язаних у часі подій).

Перші три використовуються, головним чином, для передбачення, у той час, як останні зручні для опису існуючих закономірностей в даних.

Зараз відбувається стрімкий ріст числа програмних продуктів, котрі використовують нові технології з організацією інтелектуальних обчислень, а також типів задач, застосування яких надає значного ефекту.

Одним з них є пакет прикладних програм IDAMS, призначений для валідації, маніпулювання і статистичного аналізу даних. IDAMS виробляється та вільно поширюється UNESCO. Він включає в себе інструменти маніпулювання й аналізу даних, які є доступними через інтерфейс користувача та командну мову. Однією з особливостей IDAMS є проведення вичерпної валідації даних (перевірки їх коректності та логічності) перед проведенням аналізу [8].

IDAMS дозволяє підраховувати базові статистичні параметри вибірки – середні, частотні характеристики, кореляції та ін. Основний набір статистичних процедур включає також декілька важливих видів аналізу, таких як кластерний (підтримується шість алгоритмів), дискримінантний, факторний (метод головних компонент і аналіз відповідностей), регресійний та дисперсійний.

Декілька процедур IDAMS дозволяють побудувати різноманітні узагальнення регресійної моделі, призначених для виявлення внутрішніх взаємозалежностей і зв'язків у структурі даних. Це множинний класифікаційний аналіз та деякі інші тести із множини прогнозування та класифікації.

Крім тестів, які виконуються за допомогою командного синтаксису, частину важливих процедур можна підраховувати інтерактивно з використанням зручних діалогових вікон WinIDAMS. Таких типів аналізу три: багатовимірні таблиці, інтерактивне графічне дослідження та блок аналізу часових рядів [3].

Для того, щоб знайти нове знання на основі даних великого сховища недостатньо просто взяти алгоритми Data Mining, запустити їх і чекати появи цікавих результатів. Знаходження нового знання – це процес, котрий містить у собі кілька кроків, кожний з яких необхідний для ефективного застосування засобів інтелектуальних обчислень:

1) визначення проблеми (постановка задачі, визначення мети майбутнього аналізу);

2) збір та підготовка даних (оцінка даних, об'єднання й очищення, відбір й перетворення даних);

3) побудова моделі (оцінка й інтерпретація, зовнішня перевірка);

4) використання моделі;

5) спостереження за моделлю.

Однією з найперспективніших сфер застосування вищезгаданих алгоритмів є електронні бібліотеки, які містять великі обсяги даних і відповідають концепціям інформаційних сховищ даних:

- предметна орієнтація (дані, об'єднані в категорії);
- інтегрованість (наявність єдиної централізованої сукупності даних);
- прив'язка до часу (сховище можна розглядати як сукупність «історичних» даних);
- незмінність (дані у сховище лише додаються).

Наукова електронна бібліотека НБУВ – це велике сховище баз даних, які об'єднують в собі такі інформаційно-ресурсні компоненти: електронний каталог НБУВ, загальнодержавну реферативну базу

даних, фонд електронних документів з повними текстами.

Особливістю наукової електронної бібліотеки НБУВ є наявність процесу реферування наукових видань України. У результаті аналітико-синтетичної переробки вхідного документного потоку отримується якісно новий науково-інформаційний продукт, який є по своїй суті метаданими, що можуть значно підвищити повноту й оперативність задоволення інформаційних потреб користувачів.

Пошукова система електронних колекцій бібліотек НБУВ розроблена на базі пакету прикладних програм CDS/ISIS. CDS/ISIS (Computer Documentation System / Integrated System Information Services) є універсальним інструментарієм для створення автоматизованих систем бібліотек, архівів і музеїв – для обробки структурованих нечислових баз даних. Головною особливістю CDS/ISIS є автоматичне створення і підтримка файлів швидкого доступу до кожної бази даних, яка забезпечує максимальну швидкість пошуку навіть за великих об'ємів даних [10].

Бібліотечні бази накопичують масиви документів, але єдине, що користувачі хочуть від них одержати – це корисна інформація. Традиційно в бібліотеках для пошуку літератури в великих масивах інформації використовувались предметні, тематичні і алфавітні карточні каталоги. До числа основних задач, які розв'язуються на основі електронних бібліотек входять: інформаційний пошук, класифікація і кластеризація документів.

Центральна проблема інформаційного пошуку формулюється просто – допомогти користувачеві знайти саме ту інформацію, в якій він зацікавлений. Однак, описати інформаційні потреби користувача не так просто. Як правило, цей опис формулюється як деякий запит, що являє собою набір ключових слів, які характеризують потреби користувача. Класичною задачею інформаційного пошуку є пошук документів, що відповідають запиту, в рамках деякої статичної колекції документів [2]. Критеріями, які характеризують якість інформаційного пошуку, виступають точність і повнота видачі результатів пошуку.

Класичні моделі інформаційного пошуку розглядають документи як множини ключових слів (*термів*), які представляють ці документи. Як правило, терм – звичайне слово (термін), семантика якого дозволяє описати основний зміст документа.

З точки зору інформаційного пошуку існують два типи класифікацій: *класифікації термів*, метою яких є групування термінів у синонімічні класи для підвищення співпадання термінів запиту і документа, і *класифікації* документів, які здатні покращити результати і оперативність пошуку за рахунок звернення тільки до відповідних частин масиву документів.

За допомогою класифікації першого типу можна згрупувати різноманітні низькочастотні споріднені терміни в спільні класи тезауруса. При цьому терміни, які входять до одного класу, можуть заміняти один одного в процесі пошуку, і при використанні такої класифікації можна очікувати підвищення повноти видачі результатів пошуку. У свою чергу класифікація документів дозволяє звузити область пошуку до більш вагомих класів документів і забезпечити тим самим високу точність видачі [3].

Електронна колекція документів може бути представлена матрицею *терм-документ*, яка містить у собі частоти використання деякого терміна в кожному з документів колекції. Зі сукупності документів формується список всіх термінів електронної колекції документів, з якого вилучається другорядні частини мови (сполучники, прийменники та ін.), загальні дієслова, прикметники та прислівники (бути, знати, робити, великий, малий та ін.), займенники, терміни, які використовуються в усіх документах та терміни, які використовуються лише в одному документі. За допомогою одержаного списку може бути побудована матриця терм-документ.

Використання основ слів в якості термів сприяє підвищенню ефективності числових методів. Мовознавці дослідили, що загальновживані слова становлять у наукових текстах до 80% загальної кількості слів. Звичайно, в різних науках по-різному. Математика, наприклад, їх потребує найменше, інші науки – більше. У будь-якому випадку загальновживані слова дають найбільшу кількість помилок [7].

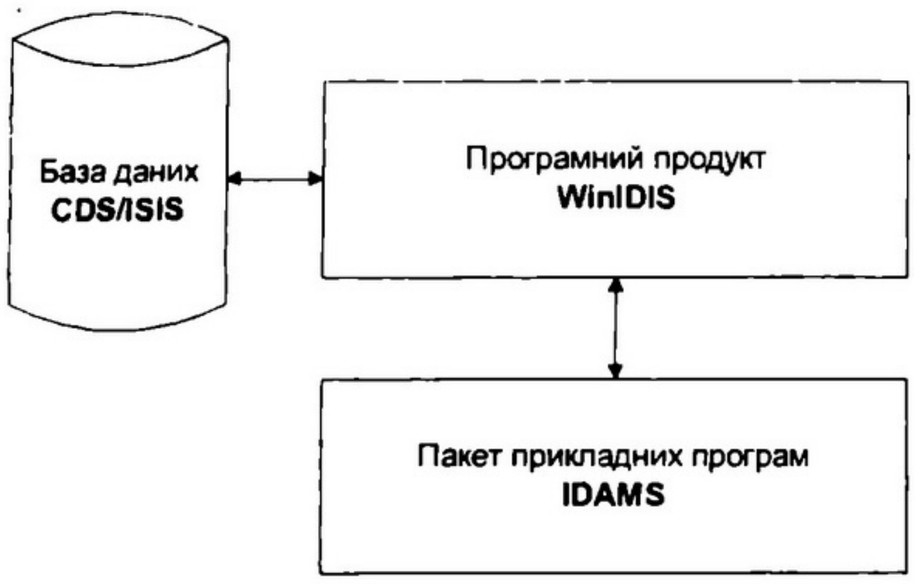
Отже, для розбиття колекції електронних документів на класи (кластери) за допомогою IDAMS необхідно сформувати текстовий файл з матрицею терм-документ.

Метою кластеризації документів є автоматичне виділення семантично схожих документів серед заданої фіксованої множини документів. Групи формуються тільки на основі попарної схожості описів документів, і ніякі характеристики цих груп попередньо не задаються. Для підвищення ефективності та швидкості інформаційного пошуку, запит користувача може порівнюватись з центрами побудованих кластерів чи груп.

Для зберігання даних IDAMS не має спеціального внутрішнього формату – таблиці зберігаються в звичайних текстових файлах, в яких змінні

(стовпці) займають фіксовані позиції. Саме текстовий формат з відокремлювачами чи стовпцями фіксованої ширини є основним для обміну даними IDAMS з зовнішніми прикладними програмами (схема 1). В IDAMS є припустимими дані двох типів – числові і текстові. Для обміну даними між CDS/ISIS та IDAMS існує окрема програма WinIDIS, яка готує опис даних і виконує передачу даних.

Схема 1



Структурна схема виділення та обробки знань

Відомо, що пошуковий алгоритм CDS/ISIS базується на індексації, тобто потребує створення допоміжного файла («індексу»), який має спростити й прискорити пошук. Цей файл називається словником пошукових термінів, і вміщує всі терміни, які можуть бути використані під час пошуку в базі даних. Для кожного терміна словник має список вказівників на записи бази даних, з яких цей термін було виділено, та частоту використання терміна у відповідному документі [10].

Якщо проіндексувати частину електронної колекції документів, що підлягає аналізу, методом індексування по окремих словах, попередньо включивши до словника стоп-слів загальновживані терміни, побудований словник пошукових термінів буде містити терміни для подальшого аналізу. Таким чином, зі словника пошукових термінів можуть бути відібрані терми з вказівниками на документи, які їх містять.

Так, матриця терм-документ може бути побудована у текстовому файлі з попередньо сформованого словника пошукових термінів відповідної бази даних CDS/ISIS за допомогою ISIS_DLL, прикладного програмного інтерфейсу ISIS для операційних систем Windows та Linux, що розроблений та вільно поширюється UNESCO [9]. Описана взаємодія зображена на схемі 2.

Після імпортування до пакета IDAMS матриці терм-документ у вигляді текстового файла з відокремлювачами, на основі отриманих даних необхідно створити словник даних IDAMS, що визначає типи даних та правила їх валідації. На базі словника даних будується файл даних IDAMS, який і буде підлягати обробці і аналізу.

Для проведення кластеризації колекції електронних документів на класи (кластери) за допомогою IDAMS необхідно сформувати командний файл, в який записується необхідна послідовність команд, що може включати в себе перевірку даних, перекодування, визначення методу та основних параметрів кластеризації, формат виводу результатів та ін.

Результатом виконання аналізу IDAMS має бути вихідний файл з виводом у відповідному форматі отриманих кластерів.

Порівнюючи отримані кластери документів з існуючим розбиттям (класифікацією) документів варто зробити висновок про семантичну близькість деяких тематик, на стику яких у подальшому може з'явитись нова галузь знання.

Метою аналізу часових рядів записів бібліотечних баз може бути прогноз загальної кількості документів або кількість документів відповідних тематик на наступні періоди.

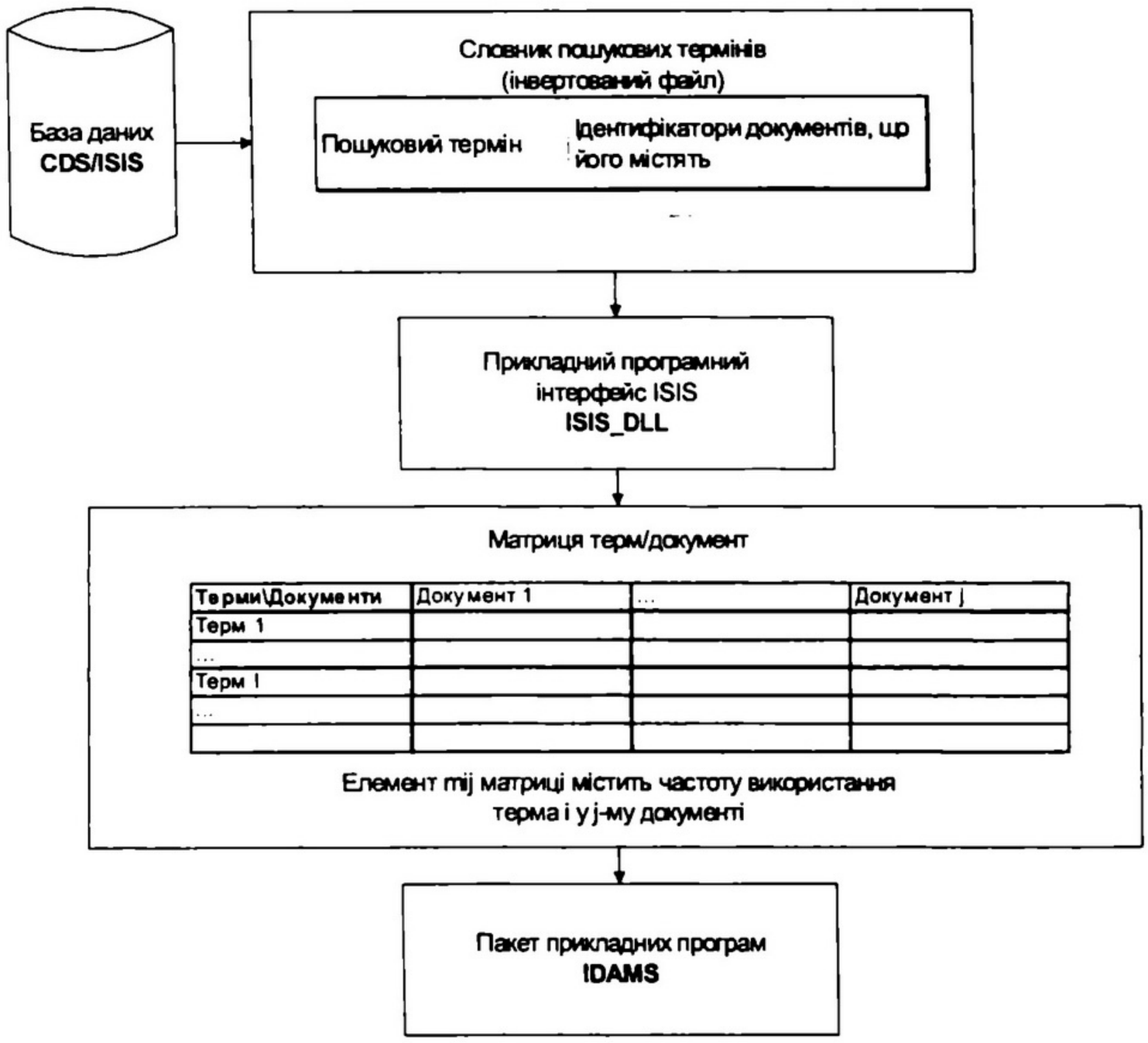
Наведений нижче графік ілюструє аналіз даних бази даних НБУВ, яка містить автореферати дисертацій, захищених в Україні в 2000–2004 рр. На основі «історичних» даних за допомогою методу найменших квадратів побудовано лінійний тренд та отримано прогноз кількості захищених дисертацій на наступний рік.

Практичне застосування методів інформометричного аналізу електронних бібліотек може включати авторубрикацію повних текстів, класифікацію і кластеризацію документів, відслідковування змін у часі термінів предметних галузей, уточнення пошукових запитів та інтелектуалізацію пошуку. Ці дослідження мають вивести інформаційні системи бібліотек на якісно новий рівень і сприяти їх трансформації в інтелектуальні системи, що проводитимуть бібліометричні, інформометричні та наукометричні дослідження у великих масивах інформації й дозволять творити нові знання.

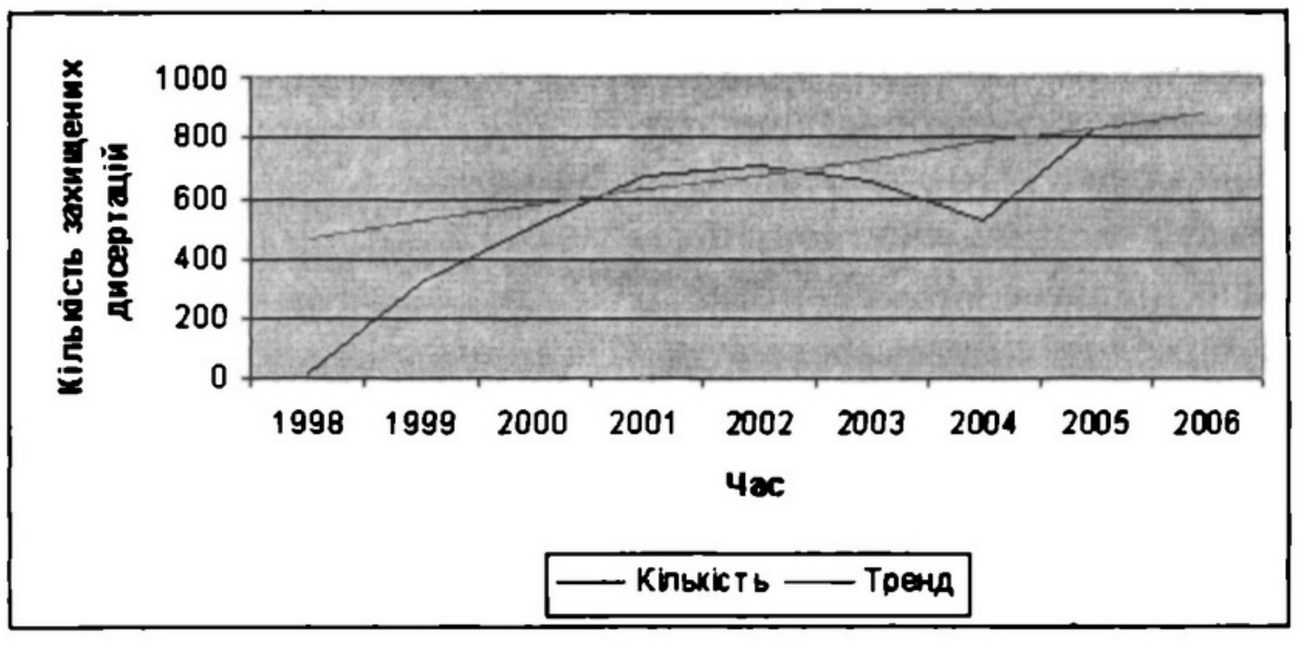
Висновки

1. Сьогодні актуальним стає вже не розв'язання суперечності в рамках підсистеми «документ– споживач інформації», а більш загальна проблема включення всього інтелектуального потенціалу суспільства в систему документних комунікацій і забезпечення його подальшого ефективного використання. Магістральний напрям вирішення цієї проблеми – це застосування технологій інтелектуального аналізу, які базуються на статистичному аналізі та моделюванні для знаходження моделей і

Схема 2



Функціональна схема виділення та обробки знань



Динаміка захисту дисертацій спеціальності «Технічні науки»

відношень, котрі містяться в інформаційних масивах у неявному вигляді. Метою інтелектуальних технологій є знаходження нового знання, яке користувач може надалі застосувати для поліпшення результатів своєї діяльності.

2. Елементи автоматичної обробки і аналізу даних, які називають Data Mining (знаходження знань), стають невід'ємною частиною концепції інформаційних сховищ даних (data warehouse) та організації інтелектуальних обчислень. Електронні бібліотеки, у яких накопичені великі обсяги даних, відповідають концепціям інформаційних сховищ даних.

3. До числа основних задач, які розв'язуються на основі електронних бібліотек, входять: інформаційний пошук, класифікація і кластеризація документів. Класичні моделі інформаційного пошуку розглядають документи як множини ключових слів (термів), які представляють ці документи. Використання граматичних основ слів в якості

термів веде за собою підвищення ефективності числових методів. Електронна колекція документів може бути представлена матрицею терм-документ, яка містить в собі частоти використання термів у кожному з документів колекції. Ця матриця може слугувати базою для проведення класифікації та кластеризації документів електронної бібліотеки.

4. Пакет прикладних програм IDAMS призначений для валідації, маніпулювання і статистичного аналізу даних виробляється та вільно поширюється UNESCO. IDAMS включає в себе інструменти маніпулювання та аналізу даних, які є доступними через інтерфейс користувача та командну мову. Аналізуючи матрицю терм-документ, слід наголосити, що цей пакет дозволяє, зокрема, й проводити класифікацію та кластеризацію документів у структурованих електронних колекціях документів.

5. Практичне застосування методів інформометричного аналізу електронних бібліотек може включати авторубрикацію повних текстів, класифікацію і кластеризацію документів, відслідковування змін у часі термінів предметних галузей, уточнення пошукових запитів та інтелектуалізацію пошуку.

Література

1. Курс лекцій «Організація інтелектуальних обчислень» [Electronic Resource]. – Way of access: URL: http://www.victoria.lviv.ua/html/oio/. – Title from the screen.

2. *Некрестьянов И. С.* Тематико-ориентированные методы информационного поиска [Electronic Resource]. – Way of access: URL: http://meta.math.spbu.ru/~igor/thesis.

3. *Носов К.* Компьютерная статистика – доступная и полноценная [Electronic Resource]. – Way of access: URL: http://itc.ua/article.phtml?ID=19049&IDw=-29&pid=18.

4. *Павлуша І.* А. Електронні бібліотеки: зарубіжний досвід, питання розробки української концепції // Бібл. вісн. – 1999. – № 4. – С. 13–24.

5. *Солтон Дж.* Динамические библиотечно-информационные системы. – М.: Мир, 1979. – 558 с.

6. *Сороколетов П. В.* Мир на пороге четвертой информационной революции [Electronic Resource]. – Way of access: URL: http://www.ifap.ru/library/book020.doc.

7. Як дібрати С Л О В О ? [Electronic Resource]. – Way of access: URL: http://dict.linux.org.ua/dict/other/SSR/RE1.html.

8. IDAMS statistical software [Electronic Resource]. – Way of access: URL: http://www.unesco.org/webworld/idams. – Title from the screen.

9. ISIS Application Program Interface ISIS_DLL User's Manual Preliminary Version BIREME, Sro Paulo, July 2001 [Electronic Resource]. – Way of access: URL: http://www.bireme.br/.

10. UNESCO CDS-ISIS databases [Electronic Resource]. – Way of access: URL: http://www.unesco.org/. – Title from the screen.

УДК 004(477)+02(477):004

Галина ЯДРОВА,
директор Центру інформаційних технологій Міжвузівського Центру «Крим»

Роль інформаційних центрів у підвищенні економічної ефективності застосування комп'ютерних технологій

У статті описаний науковий експеримент з нарощування інформаційного потенціалу навчальних і наукових бібліотек кримського регіону, показана економічна ефективність застосування корпоративних технологій в організації роботи різних установ.

Ключові слова: корпоративні об'єднання, електронна колекція, інформаційний центр, електронний ресурс, фінансові витрати.

Законом України «Про пріоритетні напрями розвитку науки і техніки» [5] визначено, що одним з найважливіших напрямів розвитку науки є нові комп'ютерні засоби і технології інформаційного суспільства. Тим часом в умовах реформування суспільства державні структури, які відповідають за просування інформаційних технологій, знаходяться в складному періоді реорганізації і пошуку нових форм функціонування і фінансування. Тому проблеми просування і застосування комп'ютерних технологій у різні сфери життєдіяльності суспільства зараз найефективніше вирішують корпоративні об'єднання, які функціонують за принципом розподілених структур [13].

Мета цього дослідження – описати вивчений досвід і показати, як форма представлення елект-