УДК 681.3

Д.В. ЛАНДЕ, *доктор технічних наук, старший науковий співробітник*
В.М. ФУРАШЕВ, *кандидат технічних наук, доцент,*
старший науковий співробітник

СИСТЕМИ МОНІТОРИНГУ, ВИТЯГУ ФАКТІВ, ПОБУДОВИ ЗВ’ЯЗКІВ НА ОСНОВІ АНАЛІЗУ НЕСТРУКТУРОВАНИХ ТЕКСТІВ

Анотація. *Огляд поширеніших систем моніторингу інформаційних ресурсів з Інтернет-простору, глибинного аналізу текстів та побудови зв`язків понять, які екстрагуються з неструктурованих текстів.*

Аннотация. *Обзор наиболее распространенных систем мониторинга информационных ресурсов Интернет-пространства, глубинного анализа текстов и построения связей понятий, которые экстрагируются из неструктурированных текстов.*

Summary. *Outlook of most popular systems of monitoring of cyberspace informational resources, deep analysis of texts and construction of communications of concepts which are extracted from the unstructured texts.*

Ключові слова. Інформаційні ресурси, засоби пошуку, синтезу та аналізу текстів.

Розбудова та вдосконалення демократичних засад самоорганізації суспільства, процеси глобалізації соціально-економічного середовища світу супроводжуються збільшенням кількості законодавчих ініціатив, судових розглядів, змінами в законодавстві, зростанням вимог до якості та оперативності процедур прийняття рішень. Наслідком цих процесів є поширення впровадження систем управління юридично значущою інформацією та механізмів e-discovery – засобів пошуку документів юридичної спрямованості. На думку аналітиків Forrester Research, витрати на механізми e-discovery зростуть з 1,4 млрд. доларів США у 2006 році до 4,8 млрд. доларів США у 2011-му. Зокрема, програмне забезпечення компанії Interwoven (у 2009 р. її купила інша компанія, про яку мова нижче, – Autonomy) використовують 1200 провідних юридичних фірм. З його допомогою здійснюється підтримка близько 100 тис. сайтів Extranet і Intranet.

Сьогодні існує ряд систем, які виконують окремі функції, необхідні для побудови комплексної системи моніторингу інформаційних ресурсів юридичної спрямованості, витягу фактів, побудови зв’язків на основі аналізу неструктурованих текстів. У роботі наведено короткий огляд таких систем.

1. RCO

Російська система, основна функціональність якої – виділення на основі аналізу текстів російською мовою змістовної суті понять з цих текстів і зв’язків між ними [1].

Система є найрозвиненішою в Росії в цьому напрямі та надається у вигляді готового програмного забезпечення, а також бібліотек програм для розробників у Windows-середовищі. Інструментарій розробника: RCO Morphology Professional SDK – підтримує всі можливості граматичного аналізу будь-якого слова російської мови: визначення граматичних характеристик слова, приведення до нормальної форми, отримання необхідних словоформ. Ціна (1 процесор) – 278000 руб., річна підтримка – 61160 руб. RCO Fact Extractor SDK – для розробки інформаційно-пошукових і аналітичних систем, що вимагають лінгвістичного аналізу тексту російською мовою. Ціни на останній продукт публічно не розголошуються.

Готовий інструментарій аналітика: RCO КАОТ – інформаційно-аналітична система для роботи в локальній мережі на базі MS Windows і MS Internet Information Server, яка реалізує функції “інтелектуального” аналізу і пошуку текстової інформації з підтримкою інтерфейсу у вигляді веб-сервера. RCO Fact Extractor – персональне застосування для Windows, яке призначене для аналітичної обробки тексту російською мовою і виявлення фактів різного типу, пов’язаних із заданими об'єктами – персонами і організаціями.

Клієнти – Банк Росії, ВАТ “Газпром”, “Тюменська Нафтова Компанія”, “МДМ-Банк”, “Альфа-банк”, “ІТАР-ТАРС”, Федеральна служба безпеки РФ.

Ціна на 1 професійне робоче місце – 104000 руб., річна підтримка – 22880 руб.

2. Галактика-Zoom

Система російської компанії “Галактика” [2].

“Галактика-Zoom” – це система, основна функціональність якої – пошук інформації у великих інформаційних масивах, а також виявлення значущих слів і словосполучень документа, що відображають його сенс, порівняння документів, виявлення схожості, відмінності, аномалій об’єктів, що вивчаються. За допомогою комплексу можна будувати так званий “інформаційний портрет”, сукупність ключових слів, пов’язаних із заданим запитом.

Система не забезпечує моніторингу веб-ресурсів, але може охоплювати потік, що сканується з Інтернету зовнішніми модулями.

Система “Галактика-Zoom” функціонує в архітектурі “клієнт/сервер” на сучасних платформах Windows. Існує російська і англомовна версії системи. Клієнти системи “Галактbка-Zoom”: телекомпанії НТВ, РТР, компанії “ТВЕЛ”, “ЮКОС”, а також Федеральна служба податкової поліції, Федеральна служба безпеки РФ, Центральна виборча комісія РФ і її регіональні відділення.

Базова версія програмного забезпечення системи “Галактика-Zoom” залежно від функціональності коштує від 6000 до 20000 доларів США.

3. Семантичний архів

Розробник – компанія “Аналітичні бізнес-рішення” [3].

Інформаційно-аналітична система “Семантичний архів” є інструментом для створення інтегрованого сховища інформації з можливістю зберігання досьє на об’єкти моніторингу, події, що відбуваються, а також текстові документи. Крім цього, система дозволяє зберігати інформацію, імпортовану з різних реляційних баз даних. Користувачі-аналітики мають можливість шукати інформацію, виявляти взаємозв’язки між об’єктами і подіями, генерувати аналітичні звіти. Система має такі можливості, як завантаження документів з пошукових серверів, пошти за допомогою Інтернет-роботів (проте без форматування документів); виділення згадок об’єктів, відносин і подій в текстових документах; візуалізація інформації у вигляді семантичної мережі; створення текстових звітів і анотування статей. Включає як окремий блок систему автоматизованої перевірки фізичних і юридичних осіб (ІАС “Перевірка позичальників”).

Система “Семантичний архів” є клієнт-серверним застосуванням, що працює з СУБД MS SQL Server 2000/2005 (ОС Windows). Існує інтерфейс з CRM-системою Oracle Siebel. Серед замовників – Рахункова палата РФ, Інститут суспільного проектування, Ситуаційний центр Санкт-Петербурга, Інститут безпеки бізнесу, “Альфа-банк”, АКБ “Ак Барс”, УГМК, ВАТ “Атомредметзолото” та ін. (70 корпоративних замовників).

Ціна системи залежить в основному від двох чинників: повноти комплектації і кількості ліцензій. Ще один чинник, що впливає на ціну, – наявність компонента, який дозволяє в призначеному для користувача режимі змінювати структуру бази досьє.

Базова версія для великої кількості користувачів коштує 10 тис. – 40 тис. доларів США, версія для одного користувача удвічі менше. Річний технічний супровід становить 15 % від вартості купленого комплекту програмного забезпечення.

4. Медіалогія

Медіалогія – це система моніторингу та аналізу преси, ТБ і радіо, електронних ЗМІ та блогів у режимі реального часу [4].

Можливості фільтрів дозволяють відбирати потрібні користувачам повідомлення із ЗМІ, сортувати повідомлення моніторингу за важливістю. База ЗМІ – блізько 4000 джерел, зважених по регіонах і галузях. Реалізовано: пошук інформації, виділення об’єктів (компаній, персон, брендів); оцінка помітності публікації; оцінка “негатив/позитив”; групування за темами, виявлення зв’язків між об’єктами; порівняння об’єктів.

Клієнти системи – РОСНАНО, Газпром, Уряд РФ, Міністерство охорони здоров’я та соцрозвитку РФ, ВТБ, Білайн, Міненерго РФ, Мінпромторг РФ, Російські залізниці, Raiffaisen BANK та ін.

Система функціонує у середовищі ОС Microsoft Windows (2000, XP, Vista).

Ціна постачання ПО залежить від конкретного впровадження. Приклад: бюджет сумісного проекту навчальних ситуаційних центрів у Московському державному інституті міжнародних відносин склав близько 53 млн. руб. (серпень 2009 р.)

5. PolyAnalyst

Розробник – компанія Megaputer Intelligence Inc. (США) [5].

Система PolyAnalyst дозволяє вирішувати проблеми прогнозування, класифікації, кластеризації, групування об’єктів, аналізом зв’язків, багатовимірного аналізу й інтерактивного створення звітів.

Система PolyAnalyst (та її компонента – система TextAnalyst) забезпечує лінгвістичний і семантичний аналізом тексту, виявлення суті, візуалізацію зв’язків, систематизацію документів, резюмування та обробку запитів природною мовою.

На базі PolyAnalyst компанія Megaputer Intelligence Inc. пропонує програму аналізу діяльності конкурентів, яка завантажує необхідні текстові матеріали, відбирає з них потрібну інформацію у вигляді звітів.

Система PolyAnalyst реалізована у вигляді клієнт-серверного рішення на платформі ОС Windows.

Megaputer Intelligence Inc. має більш як 500 компаній-клієнтів по всьому світу, серед яких: Національний комітет безпеки перевезень США (NTSB), Electronic Data Services (EDS), International Air Transport Association (IATA), A State Medicaid Agency, A Police Department, Southwest Airlines і так далі.

Інформація по цінах: PolyAnalyst 5.0 повний пакет, ліцензія на 1 робоче місце: 147623 руб.; робоче місце (PolyAnalyst Workplace): 8613 руб.; PA Text Analysis TA (англійська мова): 11584 руб.; PA Text Categorizer TC – каталогізатор текстів (англійська мова): 6980 руб. Для повнофункціонального рішення необхідно ще приблизно 30 модулів вартістю в середньому 10000 руб. за кожний.

6. RetriewalWare

На даний час програмний продукт RetrievalWare належить компанії Fast (Microsoft Subsidiary) Search&Transfer [6].

Родина продуктів RetrievalWare (RW) забезпечує пошук і аналіз інформації за запитами природною мовою. Інформація може бути представлена як в неструктурованому вигляді, так і в формалізованих базах даних, в локальній мережі організації або в Інтернеті.

У RW реалізована технологія “нечіткого” пошуку, асоціативного пошуку на основі семантичної мережі, можливість витягу суті з текстів, що дозволяє автоматично знаходити документи за термінами, зв’язаними за змістом із заданим запитом.

RW має можливість включати в базах даних інформацію, представлену як у файловій системі, так і у СУБД, поштових системах і системах документообігу, індексувати видалені базах даних. Це властивість RW дозволяє створювати єдиний корпоративний інформаційний простір. RW забезпечує користувачеві перегляд документів більш ніж у 250 форматах, серед яких як широко відомі: doc, rtf, txt, pdf, html, так і специфічні формати. У RW реалізована функція крос-мовного пошуку. В даний час проводяться роботи із створення українського семантичного сервера.

Серед користувачів RW (всього їх понад 5000) – уряди Росії, США, Великобританії, Ізраїлю, Польщі, Чехії, Угорщини і Швеції; патентні відомства Швейцарії, Велікобританії, США, Узбекистану і Росії; банки і компанії – Worldbank, ЦБ Росії, Зовнішторгбанк Росії, Swiss Bank, Boeing Company, General Electric, Intel, Ford Motor Company, Visa International.

Ціни: RetrievalWare Server – перший процесор 2227680 руб., додатковий процесор 2227680 руб., лінгвістичні процесори, семантичні й таксономічні картріджи – перший процесор 222768 руб.

Повна функціональність забезпечується набором з ще близько 100 модулів. Ціни на технічний супровід не вказані.

7. InfoStream

InfoStream – це система контент-моніторингу інтернет-ресурсів, розроблена українською компанією “Інформаційний центр “Електронні вісті” [7].

На основі системи InfoStream побудована однойменна технологія, що охоплює за станом на грудень 2009 року понад 4000 інформаційних веб-сайтів, представлених українською, російською, білоруською та англійською мовами. Щодня за допомогою технології InfoStream сканується понад 60 тис. нових документів.

Система InfoStream забезпечує доступ до оперативної інформації з єдиного інтерфейсу (по мірі появи у веб-просторі) в пошуковому режимі з урахуванням можливого дублювання і семантичної близькості документів, мовних версій, розмірів документів, їх цифрової насиченості тощо; доступ до ретроспективного фонду, що охоплює близько 80 млн. документів; підтримку аналітичної роботи в режимі реального часу: побудову сюжетних ланцюжків, дайджестів, діаграм тематик, що зустрічаються, і таблиць взаємозв'язків понять, медіа-рейтингів.

Система InfoStream функціонує на серверній платформі під управлінням ОС типу UNIX (FreeBSD). Доступ користувачів до системи здійснюється через веб-інтерфейс або електронною поштою в режимі підписки.

Користувачами системи і сервісу на основі технології InfoStream є понад 500 державних організацій і комерційних структур, таких як РНБО України, Міністерство еко-

номіки України, Антимонопольний комітет України, Національний центр з питань євроатлантичної інтеграції України, Українське національне інформаційне агентство “Укрінформ”, Фонд “Відродження”, Ощадбанк, Брокбізнесбанк, Кредобанк, Райффайзен Банк Аваль, Морський транспортний банк та ін.

Вартість мінімальної базової версії програмного забезпечення системи InfoStream становить 150 тис. грн. Вартість доступу до системи InfoStream в режимах онлайн і підписки електронною поштою становить від 240 до 660 грн. на місяць.

8. Інтегрум

Російське інформаційно-аналітичне агентство “Інтегрум” надає сервіс із забезпечення користувачів необхідною інформацією [8, 9].

“Інтегрум” може розглядатися не як постачальник спеціалізованого програмного забезпечення, а виключно як найбільша в Росії служба доступу до електронної інформації. Інформаційні сховища “Інтегрум” містять понад 500 000 000 оцифрованих матеріалів з 10000 джерел, доступ до яких можна отримати в режимі пошуку (за допомогою пошукової системи Artefact) і підписки.

Передплатники служби “Інтегрум” – компанії і держустанови – можуть отримувати такі послуги, як пошук в ЗМІ – в пресі, на ТБ і радіо; пошук компаній, зокрема фінансовій і організаційній інформації; моніторинг ЗМІ (дослідження медіасередовища); бізнес-аналітика (з рекомендаціями експертів); консалтинг (маркетингові дослідження, бізнес-планування, ТЕО).

Серед тисяч зарубіжних клієнтів служби “Інтегрум” можна виділити Британську бібліотеку, Публічну бібліотеку Квінса, Оксфордський університет, Шанхайський університет іноземних мов, Інститут міжнародних досліджень (Монтерей), ООН, The Wall Street Journal, Радіо “Свобода”, ARD.

Базова вартість підписки – 30 000 руб. на місяць. Передплатники “за умовчання” дістають можливість пошуку в режимі он-лайн.

9. Інші системи

Autonomy IDOL Server – одна з найбільших у світі систем обробки неструктурованої, текстової та звукової інформації з різних джерел з подальшою її обробкою, аналізом і управлінням [10]. Autonomy охоплює рішення для управління інформаційними потоками та пов'язаними з ними ризиками. Технологія Autonomy дозволяє автоматично розбирати смислове значення неструктурованої інформації шляхом використання математичних алгоритмів для визначення основних концепцій, що містяться у фрагментах інформації. Компанія обслуговує понад 20 тис. клієнтів, у тому числі British Telecom, France Telecom, General Motors, Reuters, BBC, British Airways, але поки не досить активно представлена на ринках Росії і України. Ціни на програмні продукти Autonomy відповідають цінам її основного конкурента – системи RetrievalWare.

Webscan – російська служба моніторингу інтернет-ресурсів Webscan Technologies [11]. На даний час сканується близько 1000 інформаційних веб-сайтів. Інформація надається користувачам за підпискою на запити. Базова вартість підписки складає приблизно 200 доларів США на місяць (при підписці на рік). Вартість підписки за 1 запит коливається від 10 до 120 доларів США.

Fast Search & Transfer (FAST) – виробник систем масштабу підприємства для пошуку, вибірки й аналізу інформації в реальному масштабі часу [12]. Продукти і техноло-

гії FAST забезпечують інтелектуальну обробку великих масивів структурованої і неструктурованої інформації.

Attensity suite – технологія витягу інформації з неструктурованих текстів. Вона дозволяє виявляти інформацію, приховану в неструктурованому тексті, та переводити її в структуровані дані, що мають зв’язки, які можуть бути проаналізовані тими ж методами, що й інші види структурованих даних [13].

STATISTICA Text Miner – розширення програми STATISTICA Data Miner, призначене для перетворення неструктурованих текстів в інформацію, придатну для ухвалення рішень.

AeroText – програма, яка дозволяє витягувати такі елементи інформації, як суть (entities), взаємини (relationships) і події (events) в неструктурованих текстах, а також виявляти приховані взаємозв’язки та події в текстах [14].

Businessobjects Text Analysis – програма, яка дозволяє витягувати інформацію щодо 35 типів об’єктів і подій, включаючи людей, географічні назви (топоніми), компанії, дати, грошові суми, email-адреси та виявляти зв’язки між ними [15].

WordStat – програма, яка базується переважно на статистичному аналізі слів у неструктурованих текстових документах і дозволяє витягувати інформацію з інцидент-звітів, книг скарг, обробляти результати опитів, розробляти таксономії тощо [16].

Висновки.

Роль інформації у повсякденному житті людини, підприємства, галузі господарства, системи управління, суспільства в цілому невпинно зростає. І ця тенденція буде “нарощувати оберти”.

Обсяги доступної по різни джерела, інформації також невпинно зростають і з кожним днем, роком будуть зростати ще більше.

Із зростанням обсягів інформації також стрімко зростають обсяги “зайвої” інформації.

Людські можливості по пошуку, синтезу та аналізу інформації досить обмежені. Потрібні допоміжні засоби.

На цей час існує велика кількість систем, які можуть застосовуватися для задач e-discovery, допомагати вирішувати задачі пошуку документів визначеної спрямованості, управління масивами таких документів. Найпотужніші з цих систем мають занадто велику вартість з точки зору вітчизняних користувачів. Більшість з названих систем зовсім не адаптовані або погано адаптовані до роботи з документами, написаними українською мовою.

Разом з тим, навіть застосовуючи окремі доступні інформаційні системи, можна вирішувати завдання моніторингу, витягу фактів, побудови зв’язків на основі аналізу слабко структурованих текстів за визначеною, в тому числі й юридичною, тематикою. На практиці не завжди доречне придбання великих систем, часто досить скористатися сервісними можливостями.

Розвиток інформаційної інфраструктури України, пошуки шляхів виходу з кризової ситуації в економіці мають сприяти доступу фахівців, які працюють в наукових, правоохоронних та інших установах, до сучасних систем пошуку та управління документами необхідної спрямованості.

Електронні ресурси

1. Режим доступу: //www.rco.ru
2. Режим доступу: //www.galaktika-zoom.ru
3. Режим доступу: //www.anbr.ru

4. Режим доступу: //www.mlg.ru
5. Режим доступу: //www.megaputer.ru
6. Режим доступу: //www.convera.com
7. Режим доступу: //www.infostream.ua
8. Режим доступу: //www.integrum.ru,
9. Режим доступу: //www.integrumworld.com
10. Режим доступу: //www.autonomy.com
11. Режим доступу: //www.webscan.ru
12. Режим доступу: //www.microsoft.com/pathways/fast
13. Режим доступу: //www.attensity.com
14. Режим доступу: //www.lockheedmartin.com/products/AeroText/products.html
15. Режим доступу: //www.businessobjects.com/product/catalog/text_ analysis/features.asp
16. Режим доступу: //www.provalisresearch.com/wordstat/wordstat.html

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~