УДК 004.822

Б. І. Мороз, доктор технічних наук, декан факультету інформаційних та транспортних систем і технологій Академії митної служби України
Д. Є. Костенко, аспірант кафедри інформаційних систем та технологій Академії митної служби України, провідний фахівець лабораторії інформаційних систем та процесів в митній справі
В. В. Костенко, викладач кафедри інформаційних систем та технологій Академії митної служби України

ОГЛЯД ЗАСОБІВ УПРАВЛІННЯ ЗНАННЯМИ ДЛЯ ПОКРАЩАННЯ ІНФОРМАЦІЙНОЇ СФЕРИ У СУСПІЛЬСТВІ (З УРАХУВАННЯМ ЯКІСНИХ І КІЛЬКІСНИХ ХАРАКТЕРИСТИК ІНФОРМАЦІЇ)

Розглянуто проблеми та можливості управління знаннями на основі масивів неструктурованих текстів у службових документах. Порушується питання про можливості використання онтологій для розв'язання проблем пошуку інформації в службових документах.

Ключові слова: *база даних/знань; дані; документ; інформаційна система; інформація; онтології; пошукові системи; семантичний пошук.*

There are examine the problems and management possibilities by knowledge on the basis of the unstructured texts arrays in official documents. There are affected the question about possibility of the use of ontologies as a decision of problems of information retrieval in official documents.

Key words: *database/knowledge base; data, document; informative system; information; ontologies; searching systems; semantic search.*

Постановка проблеми. Сучасний етап розвитку суспільства супроводжується швидким зростання інформації та інформаційних потоків, які безпосередньо впливають на розвиток і життєдіяльність будь-якого підприємства або організації.

Виникають нові типи відносин між людьми, нові концепції щодо організації соціуму, зокрема інформаційне суспільство, що є соціологічною концепцією, яка визначає головним фактором розвитку суспільства виробництво й використання науково-технічної та іншої інформації. Концепція інформаційного суспільства – це різновид теорії постіндустріального суспільства. Завдяки можливості спілкування (за участю великих масивів інформації) в інформаційному суспільстві класи замінюються "інформаційними спільнотами".

Аналіз останніх досліджень і публікацій. Останнім часом у науково-дослідній літературі все частіше використовується поняття техногенного суспільства або техногенної цивілізації, у дослідженні й розумінні сутності яких виділяють дві головні концепції В. С. Стьопіна та Е. С. Демиденко. Так, розвиток машинної цивілізації В. С. Стьопін пов'язує зі встановленням наукової раціональності, що виникла в Європі в XVII–XVIII ст., новим системним поглядом на природу як на предмет і необмежене джерело для трансформацій, а також зі швидкою зміною техніки й технологій завдяки систематичному використанню у виробництві наукових знань, виникненням капіталістичних відносин і активізацією креативної діяльності людини [1].

Мета статті – дослідження можливостей і перспектив створення спеціалізованої програмної системи з використанням спеціальних методів і алгоритмів морфологічного й синтаксичного аналізу неструктурованих текстів у службових документах, централізованого зберігання отриманої інформації, оптимізації пошуку інформації та її статистичної обробки. Вирішенню підлягають питання, які пов'язані зі старінням інформації. Вбачається важливим використання принципів семантичного пошуку в базах даних/знань для забезпечення точного результату.

Виклад основного матеріалу. З розвитком новітніх технологій помітнішим став їх вплив на людину, суспільство в цілому. Виникають нові типи відносин між людьми, нові концепції щодо організації соціуму, зокрема інформаційне суспільство, що є соціологічною концепцією. Концепція інформаційного суспільства – це різновид теорії постіндустріального суспільства.

За рахунок чого може відбутися перехід до "інформаційного суспільства"? Навіть краще сказати – до повноцінного "інформаційного суспільства". Найочевидніше: за допомогою таких понять, як "спілкування", "передача інформації", "передача даних", "зберігання інформації", "зберігання даних". Звісно, без великих і величезних масивів інформації не обійтися.

Поняття "інформація" у наші часи використовується більше з філософських позицій – як загальна властивість матерії, що характеризує будь-яку взаємодію між об'єктами й навколишнім середовищем. Інформація відображає сутність об'єкта, його властивості, а дані й відомості про об'єкт – це форма прояву цієї сутності (як в економічній теорії: цін на товар багато, а вартість тільки одна).

Інформація може серйозно впливати і впливає на інтенсифікацію та розвиток життя насамперед шляхом забезпечення оптимальних зв'язків між галузями промисловості, окремими індивідами, а також шляхом удосконалення управління, що спирається на інформацію. Від чіткої оперативної інформації значною мірою залежить підпорядкування плановим діям.

Інформація несе в собі відомості про кількісну та якісну оцінку процесів у суспільстві й бере участь у процесі управління, вона є об'єктом збирання, реєстрації, передачі та зберігання.

Як і будь-який об'єкт, інформація має властивості. Одна з незамінних рис інформації – дуалізм: на властивості інформації впливають як вихідні дані, так і методи, що фіксують цю інформацію.

З погляду на інформацію як на невід'ємну частину життя суспільства, найважливішими є і мають бути такі її загальні якісні властивості: об'єктивність, достовірність, повнота, точність, актуальність, корисність, цінність, своєчасність, зрозумілість, доступність, стислість.

Інформація – це відображення зовнішнього об'єктивного світу. Інформація об'єктивна, якщо вона не залежить від методів її фіксації.

Повідомлення "На вулиці тепло" містить суб'єктивну інформацію, а повідомлення "На вулиці 22°С" – об'єктивну, але з точністю, що залежить від похибки засобу виміру. Відображаючись у свідомості конкретної людини, інформація перестає бути об'єктивною, бо перетворюється (більшою або меншою мірою) залежно від думки, судження, досвіду, знань конкретного суб'єкта.

Інформація достовірна, якщо вона показує реальний стан справ. Об'єктивна інформація завжди достовірна, але достовірна інформація може бути як об'єктивною, так і суб'єктивною. Достовірна інформація допомагає прийняти нам правильне рішення. Недостовірною інформація може бути через такі причини:

1. Навмисне перекручування (дезінформація) або ненавмисне перекручування суб'єктивної властивості.

2. Перекручування в результаті впливу перешкод і недостатньо точних засобів її фіксації (тут може йтися про таку проблему, як зберігання інформації в базах даних/знань).

Інформацію можна назвати повною, якщо її вистачає для розуміння і прийняття рішень. Неповна інформація може призвести до помилкового висновку. Точність інформації визначається ступенем її близькості до реального стану об'єкта, процесу, явища тощо.

Нині актуальність інформації дуже важлива. Тільки вчасно отримана інформація може бути корисною.

Корисність (цінність) можна оцінити щодо потреб конкретних її споживачів і за тими завданнями, які можна виконати з її допомогою.

Властивості інформації так чи інакше пов'язані з властивостями її носіїв. Інформація специфічна також з погляду старіння: вона застаріває не одразу, а з появою нової інформації, такої, що заперечує або уточнює.

Майже кожна з якісних властивостей нерозривно пов'язана зі старінням інформації. Узагалі таке поняття, як "старіння інформації", має бути основоположним під час розробки електронних сховищ з даними.

Старіння інформації полягає в зменшенні її цінності з часом. Інформація старіє не через сам час, а з появою нової інформації, яка уточнює, доповнює або відкидає повністю чи частково більш ранню. Фактом є те, що науково-технічна інформація старіє швидше, естетична (твори мистецтва) – повільніше. З часом кількість інформації збільшується, інформація накопичується, відбувається її систематизація, оцінка й узагальнення. На жаль, обсяги збільшуються дуже швидко. Нині це терабайти текстових даних.

Старіння інформації в різні моменту часу формально можна подати у такому вигляді:

$$S_{t'}(t) = M_t - M_{t'},$$

де $S_{t'}(t)$ – старіння інформації (даних), зареєстрованої в момент часу t' на момент t; $M_{t'}$, M_t – значення даних стосовно досягнення мети в момент часу t, t'.

Проблема в тому, що неструктуровані дані становлять більшу частину інформації, з якою мають справу користувачі. Це не менше 90 % усієї інформації, а 10 % – це структуровані дані, що завантажуються в реляційні СУБД (системи управління базами даних).

Знайти в цій плутанині корисні матеріали (саме вчасно знайти) можна лише за допомогою застосування спеціалізованих технологій. Основу розвитку останніх становлять дослідження і досягнення з математичного моделювання. Апроксимуючись у комп'ютерні технології, математичні моделі (досліджуваного об'єкта, предметної області або процесу) набувають якостей і властивостей інформаційної моделі, що допомагає досліджувати особливості об'єкта, а також виконати розробку зміни картини явища в часі й у просторі. З іншого боку, комп'ютерні технології стають дуже важливим інструментом для математичного моделювання.

Відомо, що навіть у величезних масивах інформації мають працювати пошукові системи, які забезпечували б користувачеві швидкий і точний результат. Сучасне інформаційне суспільство використовує ряд аналітичних підходів до подання інформації для забезпечення її подальшого пошуку [2]. Деякі підходи базуються на теорії множин, інші – на елементах векторної алгебри, але всі ефективно реалізовуються в умовах практики.

Найсучасніший математичний апарат, який дозволяє описувати ситуації та приймати рішення в умовах невизначеності, ґрунтується на використанні теорії нечітких множин, яка була запропонована американським математиком Л. А. Заде. В основі підходу – ідея про системи, які залежать від людського сприймання та реакції, нечіткі за своєю природою, тому їх аналіз має ґрунтуватись на понятті нечіткої множини. За допомогою нечітких множин, що оперують лінгвістичними змінними, з'являється можливість оперувати й описувати "людські знання". У даній теорії ступенем невизначеності для оцінки інформації слугує характеристика, що ґрунтується на погляді про "ступінь належності", суть якої в нечіткій формі думки, в теорії нечітких множин називається лінгвістичною змінною. Значенням такої змінної слугують слова або речення мови. Застосування лінгвістичних змінних сприяє можливості приблизної оцінки складних явищ, для яких немає чіткого переходу від одного стану до іншого [3].

Лінгвістичний підхід, який дозволяє перетворювати слова на числа за допомогою модифікованого математичного апарату системного аналізу, зручний для оцінки повідомлення, яке несе в собі розмиту інформацію, тобто ні точну, ні неточну. Наприклад, такі ознаки, як ранг, значущість, час зберігання, використання, не можуть бути описані й визначені з певною точністю, тому що вони залежать від часу, місця, мети використання та накопиченого досвіду користувача. Тому використання апарату нечітких множин є перспективним шляхом у дослідженні характеристик інформації. Для раціонального застосування та визначення методів дослідження інформації все ж необхідно врахувати, яку інформацію та з якою метою потрібно аналізувати й оцінювати. Даний аналіз слід проводити за допомогою розв'язання задачі класифікації інформації, яка функціонує в системах електронно-обчислювальних машин [3].

Щороку збільшується обсяг доступних користувачам масивів текстової інформації на робочому комп'ютері, що сприяє значній актуалізації завдання пошуку необхідних користувачам документів у таких масивах. Для виконання цього завдання застосовуються різні тематичні класифікатори, рубрикатори тощо, які дозволяють шукати (автоматично або вручну) документи в невеликій підмножині бази даних, що відповідає тематиці, яка цікавить користувача [4].

Традиційні засоби контекстного пошуку щодо входження слів у документ найчастіше не забезпечують бажаного результату, незважаючи на те, що сучасні пошукові системи "навчились" автоматично збирати інформацію в базах даних/знань, ураховувати морфологічні особливості й робити своєрідну оцінку значущості знайдених документів. У багатьох пошукових системах нині використовується релевантна модель оцінки відповідності досліджуваного документа пошуковому запиту. Дана модель практично не може розпізнавати омоніми і багатозначні слова. Застосування семантичного пошуку може допомогти уникнути проблеми шляхом "розуміння" змісту тексту в цілому [5].

Найкорисніша інформація – об'єктивна, достовірна, повна й актуальна. При цьому слід ураховувати, що й необ'єктивна, недостовірна інформація (наприклад, художня література) не менш значуща для людини. Соціальна (суспільна) інформація має ще додаткові властивості:

1. Семантичний (значеннєвий) характер, тобто понятійний, тому що саме в поняттях узагальнюються найістотніші ознаки предметів, процесів і явищ навколишнього світу.

2. Мовну природу (крім деяких видів естетичної інформації, наприклад образотворчого мистецтва). Один і той же зміст може бути виражений на різних природних (розмовних) мовах, записаний у вигляді математичних формул тощо.

Для ефективного семантичного пошуку необхідна інформація про предметну область, властиві їй поняття і відносини між ними, а також дані про обмеження, характерні для цих відносин. Таку інформацію прийнято називати онтологією. Онтології містять доступні для комп'ютерної обробки визначення основних понять предметної області, зв'язки між ними [6].

Таким чином, реалізація завдання пошуку інформації на основі онтологій має передбачати наявність:

1. Онтології деякої предметної області, в рамках якої сформульовано пошуковий запит.

2. Колекції документів (природною мовою), онтології яких рівняються з онтологією предметної області.

Якщо онтологія документа – це підмножина онтології предметної області, то документ можна вважати відповідним запиту. Постійне використання онтологій для подання взаємозв'язків між поняттями могло б суттєво поліпшити результат пошуку, зокрема шляхом розширення пошукового запиту еквівалентними за змістом словами.

Нині також залишається актуальним питання дослідження якісно-кількісних характеристик інформації та її оцінки з погляду цінності під час використання такої інформації

для виконання певного завдання в певному часовому інтервалі. Завдання оцінки цінності інформації також потребує попередньої класифікації інформації. Отже, процес класифікації використовується у виконанні багатьох завдань, що виникають у процесі використання та збереження інформації.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямку. Можна зазначити, що у сфері пошуку інформації існує маса проблем, а саме:

1. Схильність до використання звичайного пошуку у складних текстових системах.
2. Помилки під час проектування.
3. Некоректність запитів (у цьому аспекті справа стосується швидше за все користувача, ніж розробника).
4. Неадаптованість словників.
5. Неструктурованість даних.

Онтології можна назвати однією зі спроб порятунку сховищ інформації від тотального переповнення й виникнення помилок. Передбачається, що онтології: можуть і могли б розв'язувати проблему подання знань для виведення інформації, що релевантна запиту користувача, дали б можливість фільтрувати і класифікувати інформацію, створювати загальну термінологію для програмних агентів і користувачів.

Перевіряється також припущення, що онтології могли б розв'язувати таку проблему, як "старіння інформації".

У перспективі планується розробка спеціалізованої програмної системи, що дасть змогу використовувати спеціальні методи й алгоритми морфологічного та синтаксичного аналізу неструктурованих текстів службових документів, централізованого зберігання отриманої інформації, оптимізації пошуку інформації та її статистичної обробки.

Крім того, планується розробка методу та алгоритму семантичного пошуку з урахуванням особливостей обробки масивів неструктурованих текстів у службових документах. Планується розробка системи семантичного пошуку.

Список використаних джерел:

1. Трач Ю. В. Техногенність як одна з ознак розвитку сучасного суспільства [Електронний ресурс] : вибрані матеріали Всеукраїнської науково-практичної конференції "V культурологічні читання пам'яті В. Подкопаєва" (1–2 червня 2007), Український центр культурних досліджень / Трач Ю. В. – Режим доступу до журн. : http://www.culturalstudies.in.ua/sekcia_s_s2_9.php

2. Аникин В. М. Аналитические модели детерминированного хаоса / Аникин В. М. – М. : Физматлит, 2007. – 328 с.

3. Заде Л. А. Понятие лингвистической переменной и его применение к принятию приближенных решений / Заде Л. А. – М. : Мир, 1976. – 168 с.

4. Шатовская Т. Интегрированный подход текстовой кластеризации для неструктурированных документов / Т. Шатовская, И. Каменева // INTERNET–EDUCATION–SCIENCE : материалы 6-й Международной конференции (Винница, Украина, 7–11 октября, 2008 г.). – Винница, 2008. – С. 504–506.

5. Шумейко Ю. Д. Огляд програмних підсистем для розрахунку фізичної взаємодії об'єктів при побудові 3D-сцен / Ю. Д. Шумейко // Системний аналіз та інформаційні технології : матеріали 11-ї міжнародної науково-технічної конф. "САІТ-2009" (Київ, Україна, 26–30 травня, 2009 р.). – К. : НТУУ "КПІ", 2009. – С. 601.

6. Захарова И. В. Об одном подходе к реализации семантического поиска документов в электронных библиотеках / И. В. Захарова // Вестник УГАТУ. – 2009. – Т. 12. – № 1 (30) : Серия "Управление, вычислительная техника, информатика". – С. 133.