УДК 004.91

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

УПРАВЛІННЯ ЗНАННЯМИ НА ОСНОВІ МАСИВІВ НЕСТРУКТУРОВАНИХ ТЕКСТІВ У СЛУЖБОВИХ ДОКУМЕНТАХ

Розглядаються проблеми і можливості управління знаннями на основі масивів неструктурованих текстів у службових документах. Пропонуються конкретні процедури і технології побудови проблемно-орієнтованих баз знань (експертних систем), які доцільно використовувати для інформаційних та аналітичних служб.

Рассматриваются проблемы и возможности управления знаниями на основании массивов неструктурированных текстов в служебных документах. Предлагаются конкретные процедуры и технологии построения проблемно-ориентированных баз знаний (экспертных систем), которые целесообразно использовать для информационных и аналитических служб.

There are examine problems and management possibilities of knowledge on the basis of arrays of the unstructured texts in official documents. There are offer the concrete procedures and technologies of construction of the problem-oriented bases of knowledge (expert system) that it is expedient to use for informative and analytical services.

Ключові слова. База знань, дані, дескриптор, документ, знання, інформаційна система, класифікатор знань, мікротезаурус, пошукові системи, технології управління знаннями.

Вступ. Важливим аспектом інформаційного суспільства є комунікації та інформаційні технології, зокрема – технології управління знаннями. Саме ці технології зумовлюють незворотні цивілізаційні зміни, які полягають в усвідомленні важливості інформації, у значному поширенні інформаційних технологій, що активно підтримують державні інституції.

Про процеси управління знаннями доречно говорити лише як про процедуру вилучення, документування, візуалізації й доступу до знань.

Постановка завдання. Мета цієї статті – дослідження можливостей та перспектив створення спеціалізованої програмної системи з використанням спеціальних методів і алгоритмів морфологічного й синтаксичного аналізів неструктурованих текстів у службових документах, централізованого зберігання отриманої інформації, оптимізації її пошуку та статистичної обробки. При цьому основною проблемою залишається розробка методів автоматичної сегментації документів і способів програмування, що дозволяють мінімальними зусиллями виконання поставлене завдання щодо текстів службових документів довільної (або майже довільної) синтаксичної складності, а також побудова прикладної моделі лінгвістичного процесора, що відповідає описовому, пояснювальному та стимулівному принципам.

Результати дослідження. Неструктуровані дані становлять не менше 90 % інформації, з якою мають справу користувачі. Знайти в таких даних щось цінне можна лише за допомогою спеціалізованих технологій.

Сучасний рівень інформаційно-аналітичної роботи, як і прогрес цивілізації загалом, показує тенденцію зменшення ролі природного інтелекту в результатах інтелектуальної діяльності, перекладання її на автомати, а також підвищення інтелекту в системах, які мають допомагати, а потім і спрямовувати дослідження даних залежно від їхнього контенту.

Сучасний світ розвивається нерівномірно. За підвищеного темпу виникнення радикальних змін послаблюється зв'язок із минулим. Традиційне мислення руйнується. Зростає спеціалізація наук і технологій. Власні знання виявляються недостатніми для адекватного вирішення елементарних питань. І навіть на основі доступного чужого досвіду стає важче приймати рішення, адекватні безпрецедентним проблемам, що виникають.

Спроби використовувати комп'ютер для керування знаннями стикаються з нерозв'язуваним традиційними методами завданням доступу до досвіду і знань, викладеним у неосяжних потоках (масивах) текстів.

Знання не лише самостійна цінність, вони породжують мультиплікативний ефект відносно інших чинників виробництва, впливаючи на рівень ефективності їх застосування. Таким чином, у сучасній економіці джерелом конкурентних переваг стає не вигідна ринкова позиція, а складні для реплікації знання. Причому в центрі уваги тут не створення знань, а їх рух і використання в організації [1].

Інформація і знання становлять основу інтелектуального капіталу, вони мають низку специфічних характеристик, на відміну від грошових, природних, трудових і технічних ресурсів організації:

1) цінність знань полягає в їх достатності, тоді як інші ресурси обмежені;

2) у структурі собівартості "матеріалізованого знання" (наукоємних товарів і послуг) переважає тенденція до накопичення витрат на початковій стадії виробництва;

3) між витратами знань на вході та їх обсягом на виході немає суттєвої економічної відповідності.

Нині організаційні знання розглядаються як інформаційний запас і як потік (рух цієї інформації) водночас.

Девенпорт і Прусак зазначають, що "знання – це рідка суміш оформленого досвіду, цінностей, контекстної інформації і поглядів експерта, яка дає схему для оцінки і об'єднання нового досвіду й інформації. В організаціях вони частенько потрапляють не лише в документи або сховища, але й в організаційні процедури, процеси, практику і норми" [2].

Ікуджіро Нонакою розроблено спіраль знань – модель, що пояснює, як під час створення нових знань явні і неявні знання взаємодіють в організації завдяки чотирьом процесам їх перетворення [3]:

1) соціалізації (перетворенню неявних знань на неявні);

2) комбінації (перетворення явних знань на явні);

3) екстерналізації (перетворенню неявних знань на явні);

4) інтерналізації (перетворенню явних знань на неявні).

Система управління знаннями – це набір регулярно повторюваних управлінських процедур для підвищення ефективності збору, зберігання, поширення й використання цінної для компанії інформації. У своїй статті "Концепція управління знаннями в сучасних організаціях" Мільнер виділяє три основні компоненти системи управління знань, а саме [4]: людські, технологічні, організаційні.

Культура є найважливішою проблемою у сфері знань, оскільки саме людський чинник (цінності, рівень зв'язків або ізольованості в організації) створює чи руйнує систему управління знаннями.

Технологія не може самостійно розв'язати проблему знання або створити атмосферу для обміну знаннями, хоча і є дуже важливим елементом системи управління ними. Використання сучасних інформаційних технологій у жодному випадку не повинно виключати необхідні елементи звичайного міжособистісного спілкування, адже саме вони роблять процеси обміну знаннями в організації інтенсивнішими. У зв'язку з цим необхідно приділяти увагу не лише матеріально-технічній частині, але й головним чином організаційним моментам.

Структура організаційних знань складається з практичних, теоретичних, стратегічних,

комерційних і виробничих знань. Організація вилучає інформацію, робить висновки і генерує нові знання з метою підвищення якості виробів, що випускаються, і послуг, що надаються, а отже, й конкурентної позиції фірми. Управління кожним із зазначених елементів у складі системи управління знаннями базується на використанні вже розглянутих процесів – створенні, зберіганні, використанні та поширенні знань.

Категорії природної мови ("слово" і "текст") за своєю природою відрізняються від відповідних ментальних категорій "поняття" і "зміст" (це так звана проблема "зміст – текст").

Комп'ютер знаннями керувати не може – це під силу лише людині. Однак пізнавальні можливості людини також обмежені: вона не здатна проаналізувати ту частку потоку інформації, що зумовлює інтерес. Робота стає монотонною і дуже нагадує пошук голки в сіні. Для своєчасного означення проблеми й пошуку варіантів розв'язання необхідно забезпечити автоматизоване спостереження за змінами в області зацікавленості особистості або органу керування, а також пошук і вивчення матеріалів усього набутого досвіду.

Передача знань відбувається тільки під час узаємодії між конкретними людьми, формування співтовариства як середовища людей, об'єднаних загальними професійною зацікавленістю чи метою, що дає можливість установити контакт між тими, хто шукає знання та джерелом знань в умовах довіри й з використанням сформованих особистих взаємовідносин. Перешкодою впровадження методик управління знаннями може стати внутрішня конкуренція. Для формування атмосфери спілкування в співтоваристві, корпоративної культури необхідно враховувати цю особливість людей і сприяти тому, щоб вони обмінювалися знаннями з радістю. Якщо основним мотивом співробітника є не індивідуальне лідерство, а досягнення мети, то колектив здатний за сприятливих умов досягти вищих результатів, ніж загалом під час відсутності кооперації. Рішення у сфері інформаційних технологій (ІТ-рішення) підтримують правила, що супроводжують процес управління знаннями, допомагають зняти бар'єри у формуванні єдиного робочого середовища, реалізації механізму відчуження, нагромадження, використання й модифікації знань, підтримки інновацій і доведення цієї інформації до всіх зацікавлених співробітників.

Однак ІТ-рішення не відіграють домінуючу роль у методиках управління знаннями: якщо на фірмі не вживатимуться заходи щодо формування культури спільної роботи й загального доступу до даних, то ніякі ІТ-рішення не дозволять отримати відчутні результати. Так, і використання лише гуманітарних технологій без залучення інформаційних не дозволить ефективно управляти знаннями. Але існують задачі, які неможливо розв'язати без використання рішень у сфері інформаційних технологій для управління знаннями. Система управління знаннями зберігає знання в контексті вирішення завдань, виконання проектів і відносин між людьми. Контекст показує діловий процес, що привів до бажаного результату, розкриває фонову інформацію, випробувані альтернативи, а також причини, через які вони не принесли бажаних результатів. Знання, які можна використати для удосконалення ділового процесу, переносяться в нові продукти та послуги. Система управління знаннями направляє дії користувачів на розміщення інформації за визначеними правилами, що дозволяють у майбутньому успішно її знаходити і застосовувати. Стає можливим використання зв'язків у системі "люди – зміст". Навіть якщо ви не знайшли в системі знання у повному обсязі, що ідеально підходять для виконання вашого нового завдання, ви можете скористатися зв'язком "людина – зміст" і знайти, таким чином, людину, що є носієм необхідних вам знань. Це зумовлює зменшення залежності знань від людей, що володіють ними. Ви можете відчути це, вводячи у справу нових співробітників. Крім того, зводяться до мінімуму втрати, пов'язані зі звільненням співробітників в інші компанії (утрати знань, важливих для ведення бізнесу; зв'язків із ключовими клієнтами/постачальниками). Заочні комунікації не тільки зменшать необхідність витрачати час на особисті зустрічі, знання, отримані в процесі

персональних заочних консультацій зберігатимуться в системі разом із контекстом і можуть використовуватися й надалі усім співтовариством чи групою. Доступ у будь-який час, у будь-якому місці не створює обмежень тривалості заочних комунікацій і гарантує, що ви зможете отримати накопичені компанією знання тоді, коли зручно вам, а не тільки в момент персонального спілкування чи заходів, що забезпечують групові комунікації [5].

Використання багатомірних класифікаторів знань (подібних до УДК – міжнародної Універсальної Десяткової Класифікації) і механізму змістовної фільтрації потоків інформації дозволяє створювати, розбудовувати й підтримувати в режимі реального часу систематизовані бази даних (СБД).

Предметами класифікування можуть бути звичайні тексти й будь-які гіпермедіаоб'єкти (звук, графіка, фото, їхні комбінації), забезпечені текстовими анотаціями.

Змістовна фільтрація текстів досягається автоматичним виконанням запитів повнотекстових пошукових систем. Як зовнішні пошукові системи можуть використовуватись локальні пошукові системи (персональний комп'ютер), корпоративні розвідувачі (локальна або розподілена комп'ютерна мережа компанії) і пошукові сервери Інтернет.

Систематизовані бази текстів можна розглядати як функціональний аналог того, що фахівці в галузі штучного інтелекту називають "базами знань", точніше, експертними системами типу "висновок, заснований на прецедентах". Тож на основі відкритих колекцій текстів бази знань – це рекламний фантом.

Коректне й практичне розв'язання проблеми адекватного вираження думки й розуміння тексту забезпечується застосуванням підготовленим персоналом комплексу прикладних методів вилучення знань із текстів.

Спеціалізовані пошукові шаблони дозволяють шукати безпосередньо в текстах терміни для уточнення їх змісту та виявлення пов'язаних понять.

Техніка виявлення вичерпного переліку лексичних образів (словоформ) конкретних понять або їх частин (правових актів, дат, географічних об'єктів, персоналій) дозволяє шукати, класифікувати і стежити за об'єктами, що цікавлять, за внутрішніми й відкритими джерелами. Часто використовують пошукові шаблони. Наведемо різноманіття тільки цифрового написання дати "10 березня 2012 року" в базах даних: 10.03.12, 10.03.2012, 10/03/12, 10-03-12, 03-10-12, 03-10-2013.

Крім того, існує безліч аналогічних ситуацій для юридичних та економічних понять, коли застосовуються слова-синоніми.

Зауважимо, що й лексика, і склад елементарних понять у вхідній і аналогічній ситуаціях кардинально відрізняються.

Систематизація знань заданої проблемної області ведеться від вичерпних реєстрів термінів, що позначають поняття з класифікатора. Терміни поєднуються в мікротезауруси окремих понять. Поняття уточнюються, тобто будується словник, упорядковуються й оформлюються у вигляді класифікатора. Проектування класифікатора СБД і розробка відповідного пакета запитів доступних пошукових систем для пошуку всіх об'єктів (текстів) або їх фрагментів, що належать поняттю або ситуації (рубриці) класифікатора, забезпечує можливість повної автоматизації збору й рубрикації інформації із зовнішніх і внутрішніх джерел.

Документування й резервування знань захищає інвестиції організації у виробництво корпоративної бази знань і знижує кадрові ризики.

Окремо зазначимо, що тексти в корпоративній інформаційній системі ніяк не переробляються, а лише доповнюються метаданими. Пошукова система тільки позначає в них усі написання шуканого поняття або судження (як комбінації понять). Така змістовна розмітка текстів дозволяє надалі забезпечити швидкий доступ до відповідних фрагментів потрібних об'єктів, а в результаті їх сприйняття й розуміння споживачем – і до необхідних

знань. Якщо на процес змістовної розмітки тексту подивитися з точки зору комп'ютерної лінгвістики, то це всього лише формування для кожного поняття або судження вичерпної тезаурусної статті для конкретної природної мови з її оформленням за правилами мови запитів конкретної пошукової системи.

Методом прочитання можна одержати тільки 5 % потрібних документів, простим пошуком за "ключовими" словами або через існуючі класифікатори – до 25 % потрібної інформації, тобто необхідних текстів, що доставляються автоматично з використанням пропонованої технології. Таким чином, з'являється можливість якісно (не більше 10 % інформаційного шуму) класифікувати необмежені масиви текстів.

У процесі вилучення й документування знань розробляються й активно використовуються одиниці технологічної документації. Також доцільно увести окрему базу даних документації до класифікаторів.

Об'єкт документування (дескриптор) – пошуковий мікротезаурус, відповідальний за окреме поняття з корпоративного класифікатора або реєстру, оформлений у вигляді системи запитів пошукових систем. По суті, це самостійні пошукові запити, що включають сам дескриптор поняття ("головне" найменування поняття), його синоніми, антоніми, підпорядковані поняття, поширені неправильні написання, інші написання, асоційовані терміни, однокореневі слова тощо. Щоб уникнути виходу за встановлений визначенням обсяг поняття, будь-які терміни наводяться або у відповідних контекстах, або без невідповідних контекстів [6].

У разі потреби для кожного слова в запит включаються всі словоформи. Об'єкти документування відповідають за виявлення в текстах фрагментів необхідного змісту й коректно розв'язують проблему "зміст – текст". Особливість оформлення пошукових мікротезаурусів – вони завжди (якщо дозволяє пошукова система) включені у логічні дужки. Таким чином досягають можливості, маніпулюючи об'єктами документування вручну й автоматично, "збирати" пошукові запити для проблемно-орієнтованого пошуку, тобто застосовувати об'єктно-орієнтований підхід, що став класичним у традиційному програмуванні.

Для кожного поняття (або ситуації) з корпоративного класифікатора маніпулювання (пошук, класифікація, візуалізація) текстами за допомогою конкретної природної мови створюється пошуковий мікротезаурус, а саме:

1) пишеться система пошукових запитів для збору необхідної інформації з будь-яких відкритих джерел (зовнішній контур корпоративної бази знань, застосовується технологія виробництва малозатратних систем комп'ютерної конкурентної розвідки – "універсальна пошукова специфікація");

2) розробляється єдиний пошуковий запит внутрішньої (корпоративної) пошукової системи для автоматичної класифікації інформації (внутрішній контур корпоративної бази знань) [7].

Словник (тезаурус бази знань) містить словесні визначення всіх термінів (об'єктів документування), що використовуються у класифікаторах і реєстрах. Кожне визначення доповнюється актуальним і повним списком визначень терміна, знайденого у відкритих джерелах. У разі багатозначності терміна обов'язково наводиться обґрунтування прийнятого визначення. Словник забезпечує можливість уточнити обсяг поняття і зміст терміна під час документування знань, формування пошукових запитів, інтерпретації й аналізу знайденого, пошуку розв'язків і проектуванні класифікаторів. Окрема стаття словника може включати посилання на інші поняття (підпорядковані, абстрактніші, протилежні за змістом, асоційовані), блок аналітичних (оглядових) статей і новини про поняття.

Клас понять – елементарний тип суттєвих для споживача (корпорації) понять, що становлять проблему або проблемну ситуацію. Приклади класів: об'єкт, процес (відношення), абстрактний суб'єкт (активний/пасивний), ознака, обставина часу, місця тощо, аспект (точка

зору), контекст (галузь знань), спосіб, визначення, правило, теорія, функція.

У певних випадках класи також можуть упорядковуватися за окремим класифікатором верхнього рівня. По суті, такий класифікатор описує структуру фасетної класифікації, заснованої на одночасному логічному розподілі матеріалу (множини), що систематизується з кількома класифікаційними ознаками одночасно. Так створюється багатомірне уявлення простору проблемних ситуацій (кожний клас понять або фасет – окрема вісь або вимір).

Класифікатор – проблемно-орієнтований ієрархічний класифікатор понять обраного класу. Через багатомірність класів понять одного класифікатора ніколи не вистачає. Як наслідок система класифікаторів і реєстрів утворює не плоску (деревоподібну/ієрархічну), а об'ємну (фасетну) семантику області:

1) інтересів корпорації (ризики й можливості);

2) компетенцій персоналу (посадових обов'язків співробітників).

Це забезпечує можливість пошуку й аналізу за всіма й будь-якими вагомими для навігації класами понять у будь-якій їхній комбінації. Множинність і очевидність правил (стратегій) вибірки необхідної інформації із систематизованої бази даних помітно прискорює й підвищує якість процесу доступу й аналізу проблемно-орієнтованої інформації, синтезу нових знань і підготовки нестандартних рішень.

Реєстр – алфавітний або відсортований на іншій основі перелік дескрипторів понять обраного класу, систематизація (встановлення родо-видових відносин) яких неможлива або недоцільна. Реєстри звичайно створюються щодо конкретних понять (об'єктів, суб'єктів та інших класів): підприємств, персон, нормативних актів, адрес тощо.

Правила вибірки задають алгоритм перетворення "проблеми, як вона подана споживачем" у запит для пошуку необхідної інформації.

Галузі застосування технологій керування знаннями:

- підтримка переходу корпорацій, громадських організацій і органів влади до керування на основі знань (knowledge based management), керування корпоративними знаннями та інтелектуальним капіталом;
- реалізація проектів типу "Електронний уряд" (e-government), "Електронна митниця", створення міжнародних, федеральних, муніципальних, регіональних і галузевих довідково-інформаційних служб, поширення й підтримка публічних Інтернет-порталів, розв'язання проблеми "цифрової нерівності";
- виробництво корпоративних експертних систем типу "висновок на основі прецедентів" та інших типів; публічний доступ до товарів, послуг і знань, консалтінг, електронна торгівля, корпоративні портали, центри автоматичної обробки телефонних викликів, патентна розвідка;
- виборчі технології, прикладна політологія й соціологія, бойовий PR (зв'язки із громадськістю), протидія й ведення інформаційної війни, ситуаційні центри, ділова й конкурентна розвідки та контррозвідка, керування підприємницькими й іншими ризиками, пошук нових можливостей для бізнесу, стратегічний менеджмент і маркетинг;
- розв'язання правових проблем і прикладні методи законотворчості, протидія криміналу, охорона правопорядку й профілактика злочинності, підвищення культури та соціальна реабілітація;
- очне й дистанційне навчання, постійна сертифікація персоналу;
- поширення навчальних енциклопедій у некомп'ютерних формах;
- підтримка пошуку і прийняття нестандартних рішень, творчість, фундаментальні й прикладні дослідження та ін.

Висновки. Запропоновано оригінальні підходи для зниження вартості побудови

семантичних мереж на основі повнотекстової неструктурованої інформації, а також дослідження можливостей використання теорії фреймів. На рівні корпоративних інформаційних систем потрібно виконувати завдання проблемно-орієнтованого пошуку аналогій за умови, що інформацію про конкретний прецедент розподілено по групі змістовно зв'язаних документів (традиційні пошукові системи шукають окремі документи, що задовольняють усі умови пошукового запиту одночасно). Для аналізу вірогідності інформації багатьох текстових джерел результуючих рекомендацій планується вивчити можливості автоматизації побудови систем лінгвістичних змінних і використання математичного апарату нечіткої логіки.

У перспективі планується розробка спеціалізованої програмної системи, що дасть можливість використовувати спеціальні методи і алгоритми морфологічного й синтаксичного аналізів неструктурованих текстів службових документів, централізованого зберігання отриманої інформації, оптимізації пошуку інформації та її статистичної обробки. Однією з основних вимог до системи є необхідність забезпечення доступу до вже отриманої інформації широкому колу користувачів.

Програмний продукт, що реалізовуватиме зазначені вище можливості, зараз перебуває на стадії проектування. Обробка кожного документа здійснюватиметься в три етапи. Перші два з них передбачають збір даних (морфологічний і синтаксичний аналізи). На третьому – виділяються групи статистично однорідних документів.

Морфологічний розбір планується здійснювати в автоматизованому режимі на основі словника з найчастіше вживаними в службових документах словоформами.

Для оптимізації пошуку й статистичної обробки зібраної інформації передбачається застосування низки методів компонентного, кластерного аналізів, штучні нейронні мережі. Блок статистичної обробки реалізується як легкорозширюваний набір бібліотек, що динамічно компонуються.

Крім того, у системі планується автоматизація синтаксичного розбору на основі апарату найуживаніших виразів, реалізованих з використанням кінцевих автоматів.

Література

1. Teece D. J. Firm organization, industrial structure and technological innovation / D. J. Teece // Journal of Economic Behavior and Organization. – 2001. – № 2. – P. 193–224.

2. Davenport T. Working knowledge: how organizations manage what they know / T. Davenport, L. Prusak. – Harvard Business School Press, 1998. – P. 200.

3. Нонака И. Компания – создатель знания. Зарождение и развитие инноваций в японских фирмах / И. Нонака, Х. Такеучи ; [пер. с англ. А. Трактинского]. – М. : Олимп-Бизнес, 2011. – 384 с.

4. Мильнер Б. З. Управление знаниями: эволюция и революция в организации / Мильнер Б. З. – М. : Инфра-М, 2007. – 177 с.

5. Гапоненко А. Л. Управление знаниями. Как превратить знания в капитал / А. Л. Гапоненко, Т. М. Орлова. – М. : Эксмо, 2008. – 400 с.

6. Кузнецов С. В. Классификация: системно-морфологический подход / С. В. Кузнецов, В. В. Титов. – М. : РНЦ "Курчатовский институт", 1998. – 476 с.

7. Кузнецов С. В. Технологии управления, основанного на знаниях / С. В. Кузнецов // Проблемы теории и практики управления. – 2004. – № 6. – С. 85–89.

8. Кормен Т. Алгоритмы: построение и анализ / Кормен Т., Лейзерсон Ч., Ривест Р. ; [пер. с англ. А. Шеня]. – М. : МЦНМО, 2007. – 893 с.