УДК 004.827

В. О. Яковенко, доктор технічних наук,
доцент кафедри інформаційних систем
та технологій Академії митної служби України
І. В. Петренко, аспірант кафедри
інформаційних систем та технологій
Академії митної служби України

ЗАДАЧІ ПОШУКУ ТА КЛАСИФІКАЦІЇ СЛУЖБОВИХ ДОКУМЕНТІВ В ІНФОРМАЦІЙНИХ СИСТЕМАХ МИТНОЇ СЛУЖБИ

Розглянуто важливе практичне завдання автоматизації та управління митними документами. Зокрема, висвітлено питання пошуку та класифікації службових документів для створення автоматизованої системи документообігу, а також основні етапи такої системи, зважаючи на її використання в митній системі. Приділено увагу основним вимогам до системи та етапам процесу класифікації документів.

Рассмотрена важная практическая задача автоматизации и управления таможенными документами, в частности вопросы поиска и классификации служебных документов с целью создания автоматизированной системы документооборота, а также основные этапы такой системы с учетом её использования в таможенной системе. Уделено внимание основным требованиям к системе и этапам процесса классификации документов.

It is considered the important practical task automation and management of customs documents. In particular, the retrieval and classification of official documents for the creation of the automated workflow system, and the main stages of this system with regard to its use in the customs system. Paid attention to the basic system requirements and stages of the documents classification process.

Ключові слова. Методи класифікації, документообіг, митна служба.

Вступ. Для органів державної служби актуальне дослідження процесів пошуку і класифікації даних у системах обробки та передачі інформації. Інтереси держави захищаються за допомогою використання передових інформаційних технологій для забезпечення оперативного і кваліфікованого реагування на події. Ефективність прийняття управлінського рішення безпосередньо залежить від швидкості та своєчасності отримання інформації, тобто від якості інформаційного пошуку і класифікації. Для вирішення цього питання слід розробити та впровадити автоматизовану систему оперативного інформаційного пошуку та обміну, яка дозволить контролювати не тільки внутрішню документацію, але й забезпечити в реальному режимі часу доступ до нормативно-правових документів, пов'язаних із поточним документом. Останнім часом зростають обсяги та номенклатура митної інформації [1], проте сучасні методи роботи з нею не досить ефективні у зберіганні, оперативному пошуку, класифікації й обміні документів.

Актуальними залишаються такі завдання:

1) наявні системи пошуку документації не завжди мають позитивні наслідки;

2) класифікація документів здійснюється не автоматично – кожен норматив стосується певного аспекту діяльності митних органів, тому їх класифікуємо лише з цього погляду;

3) у сучасних автоматизованих системах пошуку та класифікації інформації не використовуються критерії цінності чи старіння інформації.

Розглянуто методи розв'язання перших двох задач для створення автоматизованої системи документообігу в митній службі.

Нині існує велика кількість спеціальних програмно-апаратних засобів, які беруть на себе основні аспекти роботи зі зберігання, класифікації, обробки документів [2; 3]. У задачах управління інформацією важливе отримання та пошук інформації.

Однією з частин цієї проблеми є завдання класифікації документів, що полягає в їх поділі за тематиками для кожного документа.

Усі методи класифікації використовують один і той же узагальнений алгоритм, який має такі етапи:

1) побудова описів для всіх тематик;

2) побудова опису даного документа;

3) визначення оцінок близькості між описами тематик та описом документа і вибір найбільш схожих тематик.

Постановка завдання. Проаналізуємо задачі пошуку та класифікації службових документів в інформаційних системах митної служби, з'ясуємо основні завдання та можливість моделювання процесів пошуку і класифікації службових документів для застосування цієї моделі під час розробки інформаційної технології пошуку текстів і надалі, їх автоматичної класифікації та виявлення супровідних документів.

Результати дослідження. В митних органах виокремлюють такі види документації: вхідна, вихідна, внутрішня, конфіденційна, звернення громадян, обіг ВМД. Вхідна й вихідна документація реалізується за допомогою електронної пошти. Комп'ютерні (автоматизовані) технології обробки документаційної інформації мають відповідати вимогам державних стандартів та Примірної інструкції з діловодства у міністерствах, інших центральних органах виконавчої влади, а також Раді міністрів Автономної Республіки Крим, місцевих органах виконавчої влади.

Вхідні документи в митній службі обробляються за такою схемою: загальний відділ → керівник → загальний відділ → структурний підрозділ → керівники структурного підрозділу; вихідні – у зворотному порядку. Тож логічно сформулювати основні вимоги до автоматизованої системи:

– класифікація документів;

– ведення журналу реєстрації та обліку документів;

– організація взаємодії між відділами відповідно до схеми;

– контроль за виконанням документів;

– можливість оперативного пошуку документів за певними реквізитами та можливість працювати із супровідними документами;

– забезпечення зв'язку із супровідними документами;

– архівація та захист даних.

Основні методи виконання цих вимог:

– обробка вхідної, вихідної, внутрішньої документації;

– зберігання файлів на сервері та передача їх за необхідністю на робочу станцію;

– зберігання документів у файлі типу rtf;

– СУБД – Oracle 8.0 і вище;

– ОС – родина Win32;

– використання мови розмітки xml або створення спеціального файлового типу для обміну файлами між робочою станцією та сервером;

– зберігання документів як окремих файлів.

Розглянемо ці проблеми детальніше. Нині існує декілька способів показу інформації в базах даних для забезпечення її подальшого пошуку. Найпоширеніші підходи для пошуку і подання текстової інформації, що динамічно надходить у бази даних інформаційно-пошукових систем, ґрунтуються на:

1) основі теорії множин;

2) векторній алгебрі;

3) теорії вірогідності.

Ці підходи досить ефективні на практиці, проте в канонічному вигляді вони мають недоліки, зумовлені припущенням, що основний зміст документа визначається множиною складових ключових слів-термінів і понять. Звичайно ж, такий підхід частково спричиняє втрату змістових відтінків документів, проте дозволяє виконувати швидкий пошук і групування документів за формальними ознаками.

Усі документи, що функціонують в митній системі, сформульовано природною мовою з використанням специфічних термінів. Як відомо, природна мова – це універсальна знакова система для обміну інформацією між людьми. Оскільки документи на вході документально-пошукових інформаційних систем записані природною мовою, виникає питання щодо використання природної мови як основного засобу подачі інформації під час усього циклу функціонування документально-інформаційних пошукових систем. Відповідь буде позитивною, якщо йдеться про ті інформаційно-пошукові системи, в яких відповідність між запитом і документом встановлює людина. Проте в сучасних документально-пошукових інформаційних системах цю операцію виконує комп'ютер, що практично виключає застосування природної мови як основного засобу подачі інформації. Це пояснюється значними недоліками природної мови з погляду машинної технології обробки інформації.

1. Семантична неоднозначність.

2. Парадигматичні відношення між словами.

3. Текстуальні відношення між словами.

4. Синонімія.

Такі чинники призводять іноді до неоднозначного розуміння окремих слів природної мови і тексту в цілому. Багатозначність слова виникає у процесі розвитку мови, тому вона має два різновиди – полісемію та омонімію. Полісемія – це здатність одного слова передавати різну інформацію про предмети та явища позамовної дійсності. Наприклад, у слова "горло" 4 значення: передня частина шиї; порожнина позаду рота; верхня звужена частина судини; вузький вихід із затоки, гирло. У багатьох мовах багатозначні слова переважають над однозначними. Полісемію слів прийнято відмежовувати від омонімії, оскільки значення багатозначного слова взаємопов'язані загальними семантичними елементами (семантичними ознаками) і утворюють певну семантичну єдність (семантичну структуру слова). Розрізняються первинні і вторинні (похідні) значення, які іноді розуміють як прямі й переносні значення. Первинні значення, як правило, меншою мірою залежать від контексту. Співвідношення між первинними і вторинними значеннями з часом може змінюватися. Особливості об'єднання значень в межах одного слова більшою мірою визначають своєрідність словникового складу кожної мови. Багатозначність властива і граматичним формам слова та синтаксичним конструкціям. Документально-пошукові інформаційні системи оперують письмовими повідомленнями природною мовою, внаслідок чого мовна фонетика не змінює сенс таких повідомлень, через це омографи можуть прирівнюватися до омонімічних слів.

Розглянемо завдання класифікації митних документів для автоматизованої системи. Завдання полягає у визначенні для кожного документа, що надходить у систему, однієї або кількох тематик, до яких цей документ належить. Зазначимо, що, на відміну від завдання фільтрації нормативів, у систему не надходить "сміття", тобто кожен документ насправді належить хоча б до однієї із заданих тематик.

Для класифікації об'єкта слід зазначити номер (чи назву) класу, до якого він належить. Усі методи класифікації здійснюються за однаковим узагальненим алгоритмом, який складається з таких етапів: задання/побудова описів для всіх тематик, побудова опису даного документа, оцінка близькості між описами тематик і документа, вибір подібних тематик.

У математичній статистиці задачі класифікації називають також задачами дискретного аналізу. У машинному навчанні завдання класифікації виконується, як правило, за допомогою методів штучної нейронної мережі під час експерименту – навчання з учителем.

Існують також інші способи постановки експерименту (навчання без учителя), але вони використовуються для виконання іншого завдання – кластеризації, або таксономії. У цих завданнях поділ об'єктів навчальної вибірки на класи не задається, і потрібно класифікувати об'єкти тільки на основі їх подібності. У деяких прикладних галузях, і навіть у самій математичній статистиці, через близькість завдань часто не відрізняють кластеризацію від класифікації.

Деякі алгоритми для класифікації комбінують навчання з учителем і без учителя. Наприклад, одна з версій нейронних мереж Кохонена – мережі векторного квантування, які вивчають за допомогою навчання з учителем.

Відмінності ж між методами визначаються реалізацією цих етапів.

Сформулюємо завдання:

Нехай X – множина описів об'єктів, Y – множина номерів (чи назв) класів. Існує невідома цільова залежність, відображення $y^* : X \to Y$, значення якої відомі лише на елементах кінцевої навчальної вибірки $X^m = \{(x_1, y_1),\ldots,(x_m, y_m)\}$. Потрібно побудувати алгоритм $a : X \to Y$, здатний класифікувати довільний об'єкт $x \in X$.

Описи тематик і документів. Пропонується підхід, який ґрунтується на припущенні, що тематика документа визначається його словниковим запасом. Ми виключили з розгляду так звані стоп-слова, тобто найуживаніші лексеми, котрі можуть використовуватися в документах будь-якої тематики, такі як прийменники, займенники тощо. Вважатимемо, що різні синтаксичні форми одного й того ж слова не впливають на загальну тематику документа, тому їх узагальнюємо єдиною базовою словоформою. Як опис документа використовується вся множина словоформ документа, за винятком загальновживаних. Тематики також подано в системі як набори термінів, містять не всі вживані в даній тематиці слова, а тільки невелику їх підмножину, що обирається автоматично.

Побудова описів тематик. Тематика задається відносно невеликою кількістю необхідних документів. За результатами аналізу цієї множини документів, а також множини документів, що задають решту даних тематик, автоматично будується опис тематики у вигляді набору словоформ.

Мета аналізу – виявити відмінності цієї тематики від інших та обрати словоформи, які найкраще підкреслюють особливості цієї тематики.

Кожна тематика за допомогою відповідних алгоритмів обирає свій набір слів для опису. Прикладом одного з алгоритмів може бути процедура, запропонована в праці [4], вона аналогічна класичним методам пошуку інформації, що базуються на векторному описі документа. Тому такий алгоритм має ті ж недоліки:

- метод не виявляє залежності між словоформами, які часто використовуються в документах тієї ж тематики, але рідко зустрічаються разом;
- випадкові залежності та помилки правопису значно впливають на його оцінки й точність методу;
- розмір матриці словоформ дуже великий навіть для невеликої кількості документів.

Подальшим розвитком такого підходу є використання латентно-семантичного аналізу.

Латентно-семантичний аналіз – це метод обробки інформації природною мовою для аналізу взаємозв'язку між документами і термінами, які в них зустрічаються. Він зіставляє деякі фактори (теми) документів зі словоформами.

Висновки. Використання передових інформаційних технологій для забезпечення оперативного і кваліфікованого реагування на події – це основи захисту інтересів держави. Сучасний стан проблеми свідчить, що автоматизація документообігу в митних органах потребує вдосконалення. Для цього слід розробити та впровадити автоматизовану систему оперативного інформаційного обміну. Ми проаналізували проблему організації оперативного автоматизованого інформаційного обміну та визначили основні вимоги до автоматизованої системи документообігу, а також розглянули питання пошуку та класифікації документів як складової частини інформаційної технології обробки митних документів. З'ясували застосування методів латентно-семантичного аналізу до митних документів. Проведений аналіз має практичну цінність під час розробки автоматизованих систем класифікації та обробки документів у митній справі.

Література

1. Деркач Л. В. Українська митниця: вчора, сьогодні, завтра / Деркач Л. В. – К. : Державна митна служба України, 2000. – 542 с.

2. Величкевич М. Б. Електронний документообіг, тенденції та перспективи / М. Б. Величкевич, Н. В. Мітрофан, Н. Е. Кунанець // Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі. – 2010. – № 689. – С. 44–54.

3. Матвієнко О. Основи організації електронного документообігу / О. Матвієнко, М. Цивін. – К. : Центр учбової літератури, 2008. – 112 с.

4. Кураленок И. Автоматическая классификация документов на основе латентно-семантического анализа / И. Кураленок, И. Некрестьянов // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2006). – 2006. – Вып. 25. – С. 324–335.