ОСНОВНІ АСПЕКТИ МОДЕЛЮВАННЯ ТА ВПРОВАДЖЕННЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ ОБРОБКИ ТА КЛАСИФІКАЦІЇ ДОКУМЕНТІВ В МИТНІЙ СПРАВІ

УДК 519.633:536.24

ПЕТРЕНКО Юлія Миколаївна
старший інспектор м/п «Смільниця» Львівської митниці.
Наукові інтереси: методи вдосконалення митних процедур, автоматизовані системи обробки документів.

ВСТУП

Необхідною умовою ефективного функціонування як митних підрозділів, так і Державної митної служби в цілому є організація оперативного автоматизованого інформаційного обміну як у межах Державної митної служби України, так і з іншими міністерствами та відомствами України. Розвиток України як незалежної держави обумовлює поширення її економічних стосунків з іншими державами. Ці та інші фактори обумовлюють появу нових та вдосконалення вже існуючих нормативно-правових документів, що регулюють організацію митної справи. Працівники митних органів для прийняття кваліфікованих і своєчасних управлінських рішень щоденно використовують значну кількість зазначених документів. Управлінське рішення приймається не тільки на підставі розпорядницького документа, але й припускає аналіз всієї можливої інформації, що стосується суті питання. Чим ширшим є коло такої інформації, доступної керівникові в реальному режимі часу, тим вищою є ймовірність успішного вирішення проблеми.

Сучасним шляхом вирішення цієї проблеми є створення автоматизованої системи документообігу, яка дозволила організувати не тільки контроль за внутрішньою документацією, але й забезпечувала б у реальному режимі часу доступ до нормативно-правових документів, пов'язаних з поточним документом. Слід зазначити, що питання про автоматизацію обігу, облік, обробку та зберігання документів актуальне і для інших державних структур. Це пояснюється тим, що за останні декілька десятиліть обсяги документів значно зросли, проте використовувані методи роботи з ними виявляються неефективними. Це проявляється, перш за все, в зберіганні, оперативному пошуку й обміні документів. Тому актуальним стає передачі левової частки цієї роботи на електронно-обчислювальні машини.

АНАЛІЗ ПОПЕРЕДНІХ ДОСЛІДЖЕНЬ

Аналіз робіт показує, що для підвищення ефективності управління необхідно покладатися не тільки на оптимізацію традиційних методів роботи з документами, але й на впровадження сучасних інформаційних технологій. Серед них можна виділити, з одного боку, технології створення, обробки, пошуку й зберігання документів за допомогою персональних комп'ютерів (частіше поєднуваних у локальну обчислювальну мережу), а з іншого боку – засоби телекомунікації, що дозволяють швидко й надійно переміщувати великі масиви інформації на будь-які відстані. Перший аспект інформаційних технологій стосовно до завдань діловодства втілився в автоматизовані системи документаційного забезпечення управління (АСДЗУ), які все частіше знаходять застосування в органах державного керування й підвідомчих їм організаціях. Другий аспект проявляється у використанні систем електронної пошти

й інформаційних ресурсів, підключених до глобальної мережі Інтернет.

Сучасне поняття «документаційне забезпечення управління» (ДЗУ) охоплює всі стадії життєвого циклу документа: його створення (включаючи керування рухом проектів), обіг або власне документообіг (як по каналах зв'язку, так і всередині організації), забезпечення обліку й контролю виконання, зберігання (оперативне – у відділі організації, відомче – в архіві цієї ж організації, організацію експертизи цінності й передачу частини документів на постійне зберігання в державний архів).

Головна особливість ДЗУ в порівнянні із традиційним діловодством полягає в тому, що воно припускає використання більш ефективних технологій обробки інформації. У системі ДЗУ, як правило, задіяні локальні й корпоративні обчислювальні мережі, документи створюються й обробляються за допомогою комп'ютера. Поняття ДЗУ охоплює не тільки організаційно-розпорядницькі документи (з якими в основному має справу традиційне діловодство), але всю документацію, що використовується в організації. Поширюється воно й на принципово нові види документації, обумовлені як «інформаційні ресурси» – реєстри, регістри, бази й банки даних, Інтернет-сайти й т.ін. У сучасних умовах порядок обігу з усіма цими видами документації й умови доступу до них повинні бути єдиними.

Необхідність впровадження АСДЗУ особливо гостро відчувається органами державного керування, які зіштовхуються з тенденцією до різкого росту документообігу при відносно незмінній чисельності працівників апарату. Митна служба на даному етапі – це специфічна організація, в якій працюють тисячі людей і структура якої охоплює велике число митниць, митних постів, відділів, департаментів. Цілком природно, що в ній також рухається велика кількість різноманітної документації. При цьому відбувається постійне накопичення даних, які згодом стають незручними в роботі. Тому відчувається гостра потреба в АСДЗУ для митних органів.

За останній час в напрямку створення програмних комплексів для автоматизації роботи з документами зроблено багато [1-3]. Особливо ця тенденція відстежується в країнах СНД. З конкретних розробок можна виділити такі програми як «Ліга», QDPro, «Діло». Але ці програми є в основному нормативною базою й не можуть задовольнити більшість потреб митних органів в автоматизації.

ПОСТАНОВКА ЗАВДАННЯ

Мета статті – моделювання процесу документообігу в митній службі, визначення основних характеристик та вимог до автоматизованої системи митного документообігу, огляд алгоритмів класифікації як складової частини інформаційної технології обробки митних документів.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

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

На митниці вхідний документ обробляється згідно такої схеми: Загальний відділ (реєстрація, облік, попередній розгляд) → Керівник → Загальний відділ → Структурний підрозділ → Керівники структурного підрозділу. Для вихідних документів ця схема виконується в зворотному порядку. На сьогоднішній день в митних органах автоматизовано за допомогою ПІК ЄАІС ДМСУ лише документація ВМД, а всі інші види документації майже не автоматизовані [4]. Схема документообігу в митній службі наведена на рис. 1.

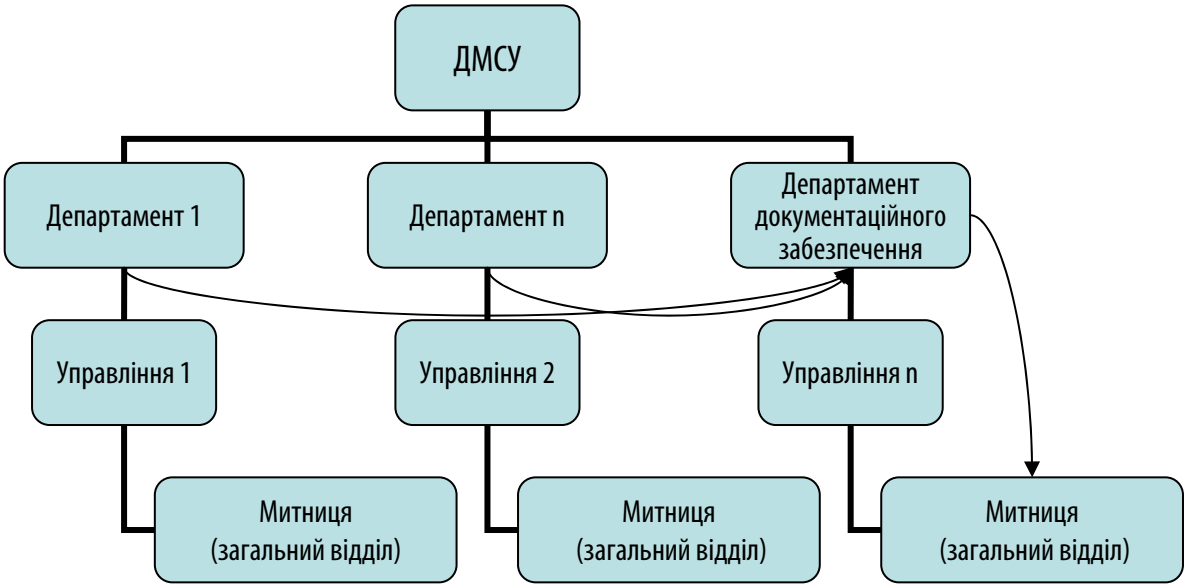

Рисунок 1 – Схема документообігу

На даному етапі аналізу предметної області можна сформулювати такі 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 по визначеним реквізитам та можливість працювати з супровідними документами (доповненнями, уточненнями);
- забезпечення зв'язку із супровідними документами (додатками та ін.);
- тривале зберігання документів;
- архівація та захист даних;
- забезпечення мережевої роботи.

Основні типові рішення:

- 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 – сімейство Win32;
- використання мови розмітки xml, або створення спеціального типу файлу для обміну файлами між робочою станцією та сервером;
- зберігання документів у вигляді окремих файлів.

Розглянемо завдання класифікації митних документів по заданому набору тематик Ω для автоматизованої системи, що пропонується. Завдання полягає у визначенні для кожного документа, що надходить в систему, однієї (або декількох) тематик до яких цей документ відноситься. Відзначимо, що на відміну від завдання фільтрації документів, тут мається на увазі, що в систему не надходить «сміття», тобто, що кожен з даних документів насправді відноситься хоч би до однієї із заданих тематик.

Всі методи класифікації використовують один і той же узагальнений алгоритм, який складається з наступних етапів: задання/побудови описів для всіх тематик, побудови опису даного документа, обчислення оцінок близькості між описами тематик і описом документа і вибору найбільш близьких тематик

Відмінності ж між методами визначаються реалізацією цих етапів.

Описи тематик і документів. Пропонується підхід, заснований на припущенні, що тематика документа визначається його словниковим запасом. Ми виключили з розгляду так звані стоп-слова, тобто найбільш споживані слова, які можуть використовуватися в документах будь-якої тематики, такі як прийменники, займенники і т. п. Будемо вважати, що різні синтаксичні форми одного і того ж слова не відбиваються на

загальній тематиці документа і, отже, можуть представлятися єдиною базовою словоформою (термом).

Як опис документа використовується вся множина термів, що зустрічаються в документі, за винятком загальновживаних.

Тематики також представляються в системі наборами термів, проте ці набори містять не всі вживані в даній тематиці слова, а тільки невелика їх підмножина, яка вибирається автоматично.

Побудова описів тематик. Тематика задається відносно невеликою множиною документів, що відносяться до неї. За результатами аналізу цієї множини документів, а також множини документів, що задають решту даних тематик, автоматично будується опис тематики у вигляді набору термів.

Метою аналізу є виявлення відмінностей цієї тематики від інших і вибір термів, що найкращим чином підкреслюють особливості цієї тематики.

Вибір слів для опису кожною з тематик проводиться за допомогою наступного алгоритму [5]:

Побудова загального словника термів *W*: У цей словник включаються всі терми, які використовуються хоч би в одному з документів, що задають тематику.

Обчислення імовірнісних оцінок: Для кожного терму $\omega \in W$ обчислюється оцінка вірогідності його використання в документах *d* даної тематики *C*:

$$P(w \mid C) = \frac{\left|\{d : d \in C, d \supset w\}\right|}{\left|C\right|} \qquad (1)$$

Для кожної тематики *C* будується «тематичний» словник. У цей словник потрапляють терми, вірогідність використання яких в цій тематиці перевершує вірогідність їх використання в будь-якій іншій тематиці $C_i \in \Omega$, тобто:

$$P(w \mid C) \geq \frac{\sum_{C_i \in \Omega} P(w \mid C_i)}{|\Omega|} \qquad (2)$$

Для кожного з відібраних термів обчислюється його значущість в рамках даної тематики згідно наступній емпіричній формулі:

$$TermValue(w \mid C) = \frac{P^3(w \mid C)}{\sum_{C_i \in \Omega} P(w \mid C_i)^2} \qquad (3)$$

Відбір термів для опису. Значущість термів, отримана на попередньому етапі, задає відношення порядку на кожному з «тематичних» словників. Використовуючи це відношення з «тематичного» словника тематики, вибирається декілька термів для використання як опис цієї тематики.

Оптимальна кількість термів для включення в опис залежить від конкретного завдання. Експерименти показали, що із зростанням числа термів якість класифікації спочатку поліпшується, а потім починає погіршуватися. При цьому оптимум досягається при невеликому розмірі опису – від 10 до 30 термів.

Обчислення оцінок близькості. Як вже було сказано вище, описуваний підхід ґрунтується на припущенні, що тематика документа визначається його словарним запасом. Ми визначаємо функцію *FSR*, яка зіставляє кожній парі термів деяку оцінку їх тематичної близькості. Оцінка тематичної близькості документа і тематики визначається тематичною близькістю термів що входять в їх описи.

У експериментах було розглянуто декілька варіантів обчислення оцінок близькості документа і тематики. Найбільш ефективним виявилося обчислення оцінки, як середнього арифметичного попарних оцінок тематичної близькості термів з описів документа *d* і тематики $C_i \in \Omega$:

$$Goodness(d, C) = \frac{\sum_{w^d_i \in d} \sum_{w^d_j \in C} (FSR(w_i^d, w_j^C))}{|C| \times |D|} \qquad (4)$$

Після оцінки документу з погляду всіх тематик можна вибрати одну або декілька найбільш високих оцінок і класифікувати документ в одну або декілька тематик.

Тематична близькість пари термів характеризує наскільки часто ці терми використовуються в документах однієї і тієї ж тематики. Відзначимо, що це не має на увазі обов'язкового використання цих термів в одних і тих же документах. Наприклад, тематична близькість термів фільтрація і класифікація може бути досить велика, навіть якщо вони ніколи не зустрічаються в одному документі.

Обчислення оцінок тематичної близькості термів, і, як наслідок, завдання функції *FSR*, відбувається за результатами аналізу використання термів в множині документів, якими описуються тематики.

По використовуваному для навчання системи набору документів будується матриця терми-на-документи *X*, рядки якої відображають розподіл термів по документах. Як оцінка тематичної близькості двох термів використовується скалярний добуток відповідних рядків цієї матриці. Таким чином для обчислення оцінок близькості між всіма парами термів досить обчислити матрицю XX^T.

Такий підхід аналогічний класичним методам пошуку інформації заснованих на векторному представленні опису документа. Тому йому властиві ті ж недоліки:

- метод не виявляє залежності між термами, які часто використовуються в документах однієї і тієї ж тематики, але рідко зустрічаються разом;
- випадкові залежності і помилки правопису роблять істотний вплив на отримувані оцінки і негативно позначаються на точності методу;
- розмір матриці терми-на-документи дуже великий навіть для невеликого (з погляду статистики) числа документів і тому використання цієї матриці доволі ресурсоємне.

Подальшим розвитком такого підходу є використання латентно-семантичного аналізу.

По матриці XX^T будується її апроксимація $\hat{X}\hat{X}^T$, де $\hat{X}$ – це апроксимація *X*, отримана методом латентно-семантичного аналізу.

Таким чином [5]:

$$\hat{X}\hat{X}^T = U_{lsa}^T \Sigma_{lsa} V_{lsa} (U_{lsa}^T \Sigma_{lsa} V_{lsa})^T \Leftrightarrow \hat{X}\hat{X}^T = U_{lsa}^T \Sigma_{lsa} V_{lsa} V^T{}_{lsa} \Sigma_{lsa} U_{lsa} \quad (5)$$

Функція тематичної близькості двох термів $FSR(w_1,w_2)$ однозначно задається матрицею $\hat{X}\hat{X}^T$:

$$FSR(w_1, w_2) = \hat{X}\hat{X}^T[w_1, w_2] \quad (6)$$

Відзначимо, що діагональна матриця Σ має розмірність *k*, де *k* – це вибрана при апроксимації бажана розмірність простору гіпотез. Таким чином при такому підході трудомісткість обчислення тематичної близькості двох термів при обчислених матрицях U_{lsa} і Σ_{lsa} складає $O(k)$, тобто вона не залежить від кількості аналізованих документів і розміру загального словника.

ВИСНОВКИ

Використання передових інформаційних технологій з метою забезпечення оперативного і кваліфікованого реагування на події – це основи захисту інтересів держави. Аналіз сучасного стану проблеми дозволив переконатися в тому, що автоматизація документообігу в митних органах потребує вдосконалення. Сучасним вирішенням проблеми є розробка та впровадження автоматизованої системи оперативного інформаційного обміну. У зв'язку з цим та у відповідності до мети роботи, у статті промодельована схема оперативного автоматизованого інформаційного обміну та визначено основні вимоги до автоматизованої системи документообігу, розглянуті алгоритми класифікації як складової частини інформаційної технології обробки митних документів. Детально розглянуті етапи класифікації та визначені недоліки та шляхи вдосконалення запропонованих алгоритмів. Розглянуті підходи можуть бути використані в автоматизованих системах класифікації та обробки документів.

ЛІТЕРАТУРА:

1. 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.
2. Матвієнко О., Цивін М. Основи організації електронного документообігу. – К.: Центр учбової літератури, 2008. – 112 с.
3. Круковский М.Ю. Критерии эффективности систем электронного документооборота /М.Ю. Круковский //Системи підтримки прийняття рішень. Теорія і практика. – 2005. – С.107-111.
4. Ульяновська Ю., Яковенко В., Ганжа В. Автоматизація діловодства в митній справі //Вісник Академії митної служби України. – 2006. – №1 (29). – С.77-80.
5. Кураленок И., Некрестьянов И. Автоматическая классификация документов на основе латентно-семантического анализа //Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника. – 2006. – Вып. 25. – С.324-335.