УДК 004.827

О.А. ЩЕРБИНА
Дніпропетровська митниця

АНАЛІЗ ПРОБЛЕМИ АВТОМАТИЗАЦІЇ ДОКУМЕНТООБІГУ В ІНФОРМАЦІЙНИХ СИСТЕМАХ ПЕРЕДАЧІ ДАНИХ МИТНОЇ СЛУЖБИ

У роботі розглядається важливе практичне завдання автоматизації роботи з митними документами. Виділенні для розгляду питання пошуку та класифікації службових документів з метою створення автоматизованої системи документообігу. Розглянуті основні етапи автоматизованої системи з врахуванням її використання в митній системі. Приділена увага основним вимогам до системи та етапам, з яких складається процес класифікації документів.

Ключові слова: документообіг, митна служба, методи пошуку та класифікації документів.

OLENA ANATOLIEVNA SCHERBINA
Dnipropetrovsk Customs

PROBLEM ANALYSIS DOCUMENT CIRCULATION AUTOMATION IN DATA TRANSMISSION INFORMATION SYSTEMS OF CUSTOMS SERVICE

Abstract. The problem of automated document for the customs service. Effectiveness depends on the decision promptly obtain the necessary information. Therefore, the task set is of practical value. The problems that are not solved in automated document management systems. For the study highlighted the problem of finding and classifying documents. The purpose of this paper is to examine the methods of solving problems identified in relation to official customs documents. Analyzed the text of the record and show what characteristics it possesses. Species listed documents which exist in customs and circuit processing. System requirements document are showed.

Keywords document management, customs service, methods of detection and identification documents.

Дослідження процесів класифікації та пошуку даних в системах обробки та передачі інформації є актуальним питанням для органів державної служби, які виконують фіскальні, контролюючі функції і робота яких спрямована на запобігання порушенню законодавства. Використання передових інформаційних технологій з метою забезпечення оперативного і кваліфікованого реагування на події – це основи захисту інтересів держави. Ефективність прийняття управлінського рішення безпосередньо залежить від оперативності і своєчасності отримання інформації, тобто від якості інформаційного пошуку. Сучасним вирішенням зазначеної проблеми є розробка та впровадження автоматизованої системи оперативного інформаційного обміну, яка дозволить організувати не тільки контроль за внутрішньою документацією, але й забезпечити у реальному режимі часу доступ до нормативно-правових документів, пов'язаних з поточним документом. В останні роки спостерігається зростання обсягів і номенклатури митної інформації [1]. Проте використовувані методи роботи з нею виявляються неефективними. Це проявляється, перш за все, в зберіганні, оперативному пошуку й обміні документів.

В системах автоматизованої обробки інформації в митній системі ще на даний момент не вирішені такі проблеми:

- Проблема пошуку документів – існуючі досі системи пошуку не завжди дають задовільний результат.
- Класифікація документів – кожен документ має відношення до якогось аспекту діяльності митних органів і їх можна класифікувати на цій основі, але існуючі системи не проводять автоматичної класифікації.
- Визначення важливості та старіння інформації – при пошуку документів необхідно враховувати критерії інформації;
- Автоматична актуалізація баз даних – в силу постійних змін законодавства автоматично проводити зміни у базі даних документів на основі нових нормативно-правових актів.

У даній статті розглядаються методи вирішення перших двох задач при створені автоматизованої системи документообігу в митній службі.

Аналіз попередніх досліджень.

На сьогоднішній день розроблено достатньо спеціальних програмно - апаратних засобів, які беруть на себе основні аспекти роботи по зберіганню, обробці, пересиланню документів [3, 4]. Отримання та пошук інформації відіграє важливу роль у широкому діапазоні задач управління інформацією та задач електронної комерції. Перспективам розробки інтерактивних пошукових систем присвячена робота [5].

Складовою частиною цієї проблеми є задача класифікації документів. класичне завдання класифікації документів полягає у їх класифікації по заданому набору тематик Ω, тобто 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.

Необхідно відмітити, що всі методи класифікації використовують один і той же узагальнений алгоритм, який складається з наступних етапів:

- 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..

Постановка завдання. *Мета роботи* – Дана робота спрямована на аналіз проблеми розробки автоматизованої системи документообігу в митній структурі, виділення основних проблем, дослідження можливості моделювання процесів пошуку та класифікації службових документів методами латентно-семантичного аналізу з метою застосування цієї моделі для подальшої розробки інформаційної технології пошуку текстів, їх автоматичної класифікації та виявлення пов'язаних документів.

Рішення задачі. 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) **à** Керівник **à** Загальний відділ **à** Структурний підрозділ **à** Керівники структурного підрозділу. Для вихідних документів ця схема виконується в зворотному порядку.

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 виділені у вступі до розгляду проблеми більш детально. В даний час існує декілька підходів до представлення інформації в базах даних для забезпечення подальшого пошуку цієї інформації. Найбільш популярними до пошуку і представлення текстової інформації, що динамічно надходить в бази даних інформаційно-пошукових систем є наступні підходи. Перший з них базується на теорії множин, другій на векторній алгебрі, а третій на теорії вірогідності. Всі ці підходи досить ефективні на практиці, проте в канонічному вигляді у всіх них є загальний недолік, який витікає з припущення, що полягає в тому, що контент документа, його основний зміст визначається безліччю ключових слів – термінів і понять, які входять в нього. Звичайно ж, такий підхід частково веде до втрати змістовних відтінків документів, проте дозволяє виконувати швидкий пошук і групування документів по формальних ознаках. Сьогодні названі підходи найпопулярніші. Слід зазначити, що існують також інші методи, наприклад, семантичні, в рамках яких робляться спроби виявити зміст за рахунок аналізу граматики тексту, використання баз знань і тезаурусів, які відображають семантичні зв'язки між окремими словами і їх групами. Очевидно, що такі підходи вимагають істотних витрат на підтримку баз знань і тезаурусів для кожної мови, тематики і виду документів, область їх застосування – професійні аналітичні системи.

Усі документи, які функціонують в митній системі сформульовані на природній мові з використанням специфічних термінів. Як відомо, природна мова (ПМ) є універсальною знаковою системою, що служить для обміну інформацією між людьми. Оскільки документи на вході документально-пошукових інформаційних систем (ДПІС), записані на природній мові, справедливо було б задатися питанням, а чи не можна використовувати природну мову як основний засіб представлення інформації під час всього циклу функціонування документально-інформаційних пошукових систем? Відповідь буде позитивною, якщо мова йде про ті інформаційно-пошукових систем, в яких відповідність між запитом і документом встановлює людина. Проте в сучасних ДПІС ця операція виконується комп'ютером, що практично виключає застосування природної мови як основного засобу представлення інформації. Це пояснюється істотними

недоліками природної мови з погляду машинної технології обробки інформації, основні з яких розглянуті нижче.

1. Різноманіття засобів передачі сенсу. Не дивлячись на те, що основним засобом передачі сенсу повідомлення є лексика природної мови, в повідомленнях на природній мові функцію передачі сенсу виконує і ряд інших елементів:

- контекст;
- парадигматичні відносини між словами;
- текстуальні відносини між словами;
- посилання на слова (словосполучення, фрази і так далі), раніше згадувані в тексті повідомлення.

Семантична неоднозначність. Повідомлення, записані на природній мові, можуть бути семантично неоднозначними. Семантична неоднозначність виникає в основному із-за синонімії і багатозначності слів природної мови.

Синонімія є тотожністю або близькістю за значенням слів, що виражають одне і те ж поняття, які відрізняються одне від іншого або відтінками значень, або стилістичним забарвленням, або одночасно обома названими ознаками. Синонімами природної мови є як окремі слова, так і словосполучення.

Багатозначність характеризує можливість неоднозначного розуміння сенсу окремих слів природної мови. Багатозначність слів представлена двома різновидами: полісемією і омонімією. Полісемія – це збіг назв різних предметів, що мають між собою які-небудь загальні властивості або ознаки. До типових загальних властивостей слід віднести схожість предметів, їх суміжність (просторову, тимчасову і так далі), а також однакове функціональне призначення. Прикладами полісемії є: "команда" (військовий підрозділ) - "команда" (екіпаж судна) - "команда" (спортивна). *Омонімія* – це збіг назв різних предметів, що не мають між собою яких-небудь загальних властивостей. Наприклад: "коса" (дівоча) - "коса" (сільськогосподарський інструмент); "ключ" (джерело) - "ключ" (дверний). Омонімічні слова, співпадаючі між собою як за написанням, так і за звучанням, слід відрізняти від омографів – слів, що позначають різні предмети, однакові за написанням, але різні за звучанням, наприклад: "замок" (дверний) - "замок" (палац). Проте, оскільки ДПІС оперують з повідомленнями на природній мові, представленими у письмовій формі, унаслідок чого фонетика мови не робить вирішального впливу на сенс таких повідомлень, омографи можуть бути прирівняні до омонімічних слів.

Еліпсність. У багатьох повідомленнях на природної мови зустрічаються еліпси або пропуски слів, що маються на увазі. Еліпсність повідомлення часто грає негативну роль при безпосередній роботі з ним людини. Очевидно, що вона тим більше негативно позначиться в тому випадку, якщо повідомлення на природній мові оброблятимуться комп'ютером.

Проблеми розуміння природної мови в повному обсязі стосуються й текстів митного спрямування. Як приклад, візьмемо витяг наказу ДМСУ № 314 від 20 квітня 2005 «Про затвердження Порядку здійснення митного контролю й митного оформлення товарів із застосуванням вантажної митної декларації»:

«2. Начальникам регіональних митниць, митниць:

2.1. Забезпечити інформування суб’єктів зовнішньоекономічної діяльності про вимоги цього наказу й Порядку».

Якщо взяти окремо цей текст, то незрозуміло, що за Порядок мається на увазі, хто дає наказ і на основі чого, тобто існує еліпсність.

Тому для організації систем зберігання і обробки документів, а точніше, для організації пошуку, класифікації документів потрібен етап попереднього аналізу текстів документів. Одним із перспективних напрямів у системах пошуку і класифікації документів є семантичний аналіз.

Розглянемо завдання класифікації митних документів за заданим набором 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 з 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 обирається 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 відповідних алгоритмів. Прикладом одного з алгоритмів може бути процедура, яка розглянута в роботі [6], і є аналогічний класичним методам пошуку інформації, заснованим 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.

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. В роботі проаналізовано проблему організації оперативного автоматизованого інформаційного обміну та визначено основні вимоги до автоматизованої системи документообігу, а також розглянуті питання пошуку та класифікації документів як складової частини інформаційної технології обробки митних документів. Подальшим розвитком роботи є аналіз застосування методів латентно-семантичного аналізу до митних документів. Проведений в роботі аналіз має практичну цінність при розробці автоматизованих систем класифікації та обробки документів в митній службі.

Литература

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

2. Ульяновська Ю. В. Автоматизація діловодства в митній справі / Ю.В. Ульяновська, В.О. Яковенко, В.М. Ганжа // Вісник Академії митної служби України. – 2006. – № 1(29). – С. 77–80.

3. 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.

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

5. Belkin N., Scholtz J., Dumais S., Wilkinson R. Evaluating Interactive Information Retrieval Systems: Opportunities and Challenges, CHI 2004, April 24–29, 2004, Vienna, Austria.

6. 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). – Донецк : ДонНТУ, 2006. – Вып. 25. – С. 324–335.

References

1. Derkach, L. (2000). Ukrainian Customs: yesterday, today, tomorrow. *State Customs Service of Ukraine*, 2000, 542.

2. Ulianovska, Yu., Yakovenko, V., Ganzha, V. (2006). Office-work automation in customs business. *The bulletin of Ukrainian Academy of Customs*, №1(29), 77–80

3. Velichkevich M.B., Mitrophan N.V., Kunanec N.E. (2010). Electronic document circulation, tendencies and prospects. *The bulletin of National University "Lviv Polytechnic".Information systems and networks*, № 689, 44–54.

4. Matvienko O., Cyvin M. (2008). Bases of the organization of electronic document circulation. *The centre of the educational literature*, 112.

5. N. Belkin, J. Scholtz, S. Dumais, R. Wilkinson, (2004). Evaluating Interactive Information Retrieval Systems: Opportunities and Challenges, *CHI* 2004, April 24–29, Vienna, Austria.

6. Kuralenok I., Nekrestianov I. (2006). Automatic classification of documents on the basis of the latentno-semantic analysis. Proceedings of Donetsk national technical university. A series: Computer science, cybernetics and computer facilities, №. 25, 324–335.

Рецензія/Peer review : 25.7.2013 р. Надрукована/Printed :29.9.2013 р.
Рецензент: д.т.н., проф. Яковенко В.О.,
кафедра інформаційних систем та технологій Академії митної служби України