— сосредоточить внимание работников организации на проблемах развития бизнеса, а не на выполнении рутинных задач.

Выполнена программная реализация предложенного подхода, которая позволит вносить бизнес правила для управления ходом процессов разработки программных средств командой разработчиков.

6. Выводы

Предложен подход к адаптации процессов контроля команды разработчиков программных средств, основанный на использовании логических правил. Указанные правила регламентируют порядок создания программных средств при возникновении заданных ситуаций в процессах разработки, выполняющихся при использовании методологии SCRUM. Предложенный подход позволяет улучшить процесс мониторинга и управления командой разработчиков и, тем самым, сократить временные затраты на разработку.

Литература

1. Книберг, Х. Scrum и XP для тренеров [Текст] / Х. Книберг. — Вильямс, 2010. — 268 с.
2. Мартин, Р. Быстрая разработка программ. Принципы, примеры, практика. [Текст] / Р. Мартин. — Tennessy, 2011.— 264 с.
3. Ремси, А. Getting Real [Текст] / А. Ремси. — NY, 2011. — 191 с.
4. Субраманиам, В. Этюды на тему быстрой разработки программного обеспечения [Текст] / В. Субраманиам. — Москва, 2009. — 302 с.
5. Хамбл, Д. Непрерывное развертывание ПО. Автоматизация процессов сборки, тестирования и внедрения новых версий программ [Текст] / Д. Хамбл. — Вильямс, 2011. — 361 с.
6. Расмуссон, Дж. Гибкое управление IT-проектами. Руководство для настоящих самураев [Текст] / Дж. Расмуссон. — Вильямс, 2009. — 312 с.
7. Мартин, Р. Чистій код [Текст] / Р. Мартин. — Висконсин, 2010. — 201 с.
8. Поппедикс, К. Реализация разработки программного обеспечения: от концепции к деньгам [Текст] / К. Поппедикс. — Даллас, 2011. — 233 с.
9. Ларман, С. Масштабирование Agile разработки [Текст] / С. Ларман. — Чикаго, 2012. — 259 с.
10. Уэллс, Д. Производство без потерь. Канбан для рабочих [Текст] / Д. Уэллс. — NY, 2003. — 233 с.

АДАПТАЦІЯ ПРОЦЕСІВ КОНТРОЛЮ КОМАНДИ РОЗРОБНИКІВ ПРОГРАМ З ВИКОРИСТАННЯМ ЛОГІЧНИХ ПРАВИЛ

Дана робота описує основні принципи виробничого процесу у розробці програмних продуктів з використанням гнучких методологій, а також розкриває їх недоліки. Пропонується удосконалення SCRUM-методології шляхом використання логічних правил для автоматизованої адаптації процесів з метою підвищення ефективності управління процесом розробки. Рішення полягає у використанні двигунів бізнес-правила для управління обмеженнями бізнес-процесів.

Ключові слова: інформаційна технологія, методологія розробки, програмний продукт, ітерація, завдання, бізнес-правила.

Чалый Сергей Федорович, *доктор технических наук, профессор, кафедра информационных управляющих систем, Харьковский национальный университет радиоэлектроники, Украина.*
Цимбаленко Роман Николаевич, *кафедра информационных управляющих систем, Харьковский национальный университет радиоэлектроники, Украина,* ***e-mail: gurut.91@gmail.com.***
Буцукина Инна Борисовна, *кафедра экономической кибернетики, Харьковский национальный университет радиоэлектроники, Украина.*

Чалий Сергій Федорович, *доктор технічних наук, професор, кафедра інформаційних управляючих систем, Харківський національний університет радіоелектроніки, Україна.*
Цимбаленко Роман Миколайович, *кафедра інформаційних управляючих систем, Харківський національний університет радіоелектроніки, Україна.*
Буцукіна Інна Борисівна, *кафедра економічної кібернетики, Харківський національний університет радіоелектроніки, Україна.*

Chalyi Sergii, *Kharkiv National University of Radio Electronics, Ukraina.*
Tsymbalenko Roman, *Kharkiv National University of Radio Electronics, Ukraina,* ***e-mail: gurut.91@gmail.com.***
Butsukina Inna, *Kharkiv National University of Radio Electronics, Ukraina*

УДК 004.048

Ульяновська Ю. В.

МОДЕЛЮВАННЯ ПРОЦЕСІВ ПОШУКУ ТА КЛАСИФІКАЦІЇ СЛУЖБОВИХ ДОКУМЕНТІВ В АВТОМАТИЗОВАНИХ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ

У роботі розглядається вирішення важливої практичної задачі класифікації, пошуку та ранжування службових документів. Виявлена й обґрунтована можливість застосування для вирішення завдання методу латентно-семантичного аналізу та аналізу взаємозв'язків Google Page Rank. Проведено моделювання за зазначеними методами.

Ключові слова: *класифікація документів, латентно-семантичний аналіз*

1. Вступ

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. Аналіз сучасного стану проблеми дозволив переконатися в тому, що автоматизація документообігу в митних органах потребує вдосконалення [2].

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.

2. Мета роботи

Дана робота спрямована на 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.

3. Аналіз методів та алгоритмів вирішення завдання

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], у якій зазначено, що не зважаючи на важливість отримання інформації, інформаційні пошукові системи часто слабо відрізняються від перспектив взаємодії людини з комп'ютером. Саме тому необхідно оцінювати ефективність систем електронного документообігу за різними критеріями [9].

Однією з проблем, які постають перед автоматизованими системами описаного типу є проблема машинного розуміння природної мови. У сучасних документальних інформаційно-пошукових системах відповідність між запитом користувача та документом виконується комп'ютером, що практично виключає використання природної мови у якості основного засобу представлення інформації. Це пояснюється недоліками природної мови з точки зору машинної технології обробки інформації [6].

Перспективним інструментом розв'язання проблем у цьому напрямі є семантичний аналіз, який знаходить своє застосування у різних галузях [7, 8]. Під цим терміном мається на увазі підходи, методи, моделі комп'ютерної обробки текстів з метою виявлення значення цього тексту, виявлення тематики, пошуку схожих текстів.

4. Результати дослідження

Трудомісткість операції класифікації одного документу складається з витрат на:

1. Обчислення оцінки близькості документа до даної тематики (для кожної тематики):
   — Пошук необхідної інформації про кожен терм з опису тематики і документа.
   — Обчислення оцінки близькості двох термів.
   — Обчислення загальної оцінки близькості документу тематиці.
2. Вибору найбільш близької тематики.

Загальна трудомісткість класифікації одного документа складає порядку $O(|\Omega||D|_{avr}|C|W|k)$ операцій, де $|\Omega|$ — загальне число тематик, D_{avr} — середня кількість термів в документі, $|C|$ — середня кількість термів в описі тематики, $|W|$ — число термів в загальному словнику, k — розмірність простору гіпотез (число використовуваних сингулярних значень матриці терми-на-документи).

Таким чином хоча описаний підхід вимагає значних обчислювальних ресурсів на підготовчому етапі, обчислювальна трудомісткість етапу класифікації відносно невелика.

Для подальшого поліпшення якості класифікації будемо досліджувати ряд ідей.

1) Багаторівнева класифікація. Багато труднощів при класифікації викликано тим фактом, що деякі тематики значно ближче одна до одної, ніж в середньому. Як наслідок, загалом у тематичному просторі описи таких тематик дуже схожі один на одного, що погіршує результати класифікації.

Для вирішення цієї проблеми пропонується використовувати багаторівневий підхід [10]:

1. Виявлені групи дуже близьких тематик об'єднуються в мегатематики.
2. Проводиться класифікація по отриманій множині мегатематик.
3. Для кожної мегатематики проводиться додаткова класифікація документів, що потрапили в неї.

Попередні експерименти показують, що такий підхід дозволяє значно підвищити точність класифікації на етапі класифікації по мегатематикам.

2) Облік зворотного зв'язку. Перспективним методом поліпшення якості класифікації є облік коментарів користувачів системи для точнішого обчислення оцінок тематичної близькості. Такий підхід називається механізмом зворотного зв'язку (relevance feedback) [11].

2) Вибір документів для завдання тематики.

Набір документів, використовуваних для завдання тематики, в значній мірі визначає набір слів, які

використовуватимуться як опис даної тематики, а також побічно впливає на описи інших тематик. Весь набір документів, використовуваних для завдання тематик, також визначає загальний словник і функцію тематичної близькості.

Під час досліджень, що проводилися під час написання цієї роботи була виявлена можливість застосування моделі Google Page Rank також і для обробки документів на митну тематику.

Як відомо, документи мають чітку структуру. Крім того, документи митного спрямування не з'являються та не існують ізольовано від інших. Вони можуть бути створені на основі вже існуючого документу, можуть його доповнювати, модифікувати, припиняти дію.

Отже, можна виділити такі відношення між документами:

— Створений на основі.
— Відношення модифікації.
— Скасування дії документу.

Аналогічно підходу Google Page Rank, ці відношення можна використовувати для аналізу взаємозв'язків між документів.

Виявлену властивість можна використовувати для ранжування документів при пошуку, побудови ієрархії документів, класифікації або кластеризації документів. Крім того, відношення «скасування дії» та «модифікації» можуть використовуватися для актуалізації бази даних документів, тобто підтримання їх у стані, що відповідає чинному законодавству.

Розглянемо мережу з вершин (сторінки) і орієнтованих ребер (посилання). Моделюватимемо рух користувача по мережі таким чином: користувач стартує у випадковій вершині. З вірогідністю ε користувач переходить у випадкову вершину, а з вірогідністю 1 – ε він переходить по одному з випадкових вихідних ребер. На практиці припускають, що ε = 0,15.

Уявимо собі, що цей користувач рухається так нескінченно довго. Для кожного k можна $PR_k(i)$ як вірогідність опинитися у вершині i через k кроків. Хай користувач робить переміщення один раз в секунду. Тоді для кожної сторінки існує якась вірогідність, що користувач опиниться на ній через, наприклад, мільярд секунд. Тоді гранична вірогідність опинитися в i-й вершині і є PageRank:

$$PR(i) = \lim_{k\to\infty} PR_k(i). \quad (1)$$

Нехай $T_1,...,T_n$ — вершини, з яких йдуть ребра в i, $C(X)$ — позначення для вихідного ступеня вершини X. Оскільки ми стартуємо у випадковій вершині, то $PR_0(i) = \frac{1}{N}$ кількість всіх сторінок. За визначенням отримуємо наступне рекурентне рівняння:

$$PR_k(i) = \frac{\varepsilon}{N} + (1-\varepsilon)\sum_{j=1}^{n} \frac{PR_{k-1}(T_j)}{C(T_j)}. \quad (2)$$

Перейдемо до границі і отримаємо:

$$PR(i) = \frac{\varepsilon}{N} + (1-\varepsilon)\sum_{j=1}^{n} \frac{PR(T_j)}{C(T_j)}. \quad (3)$$

На практиці замість $PR(i)$ зазвичай використовують $PR_{50}(i)$, обчислене за ітеративною формулою.

Має місце достатньо важливий факт про те, що фактично PageRank власний вектор матриці всіх посилань. Визначимо матрицю L таким чином:

— якщо немає ребра з i в j, то $l_{ij} = \frac{\varepsilon}{N}$;

— якщо ребро є, то $l_{ij} = \frac{\varepsilon}{N} + (1-\varepsilon)\frac{1}{C(i)}$.

Введемо позначення:

$$\overline{PR_k} = (PR_k(1),...,PR_k(N))^T;$$
$$\overline{PR} = (PR(1),...,PR(N))^T. \quad (4)$$

Тоді виконуються співвідношення:

$$\overline{PR_k} = L^k \times \overline{PR_0};$$
$$\overline{PR} = L \times \overline{PR}. \quad (5)$$

Звідси робимо висновок, що $\overline{PR}$ власний вектор матриці всіх посилань L.

Для проведення дослідження поводження моделі Page Rank до роботи з документами будемо вважати, що відношення «створений на основі» — це посилання на основний документ, а відношення модифікації або скасування — як посилання на даний документ.

Розглянемо набір документів, наведений на рис. 1. Ранг жодного з документів невідомий. Тому кожному документу присвоїмо ранг 1.

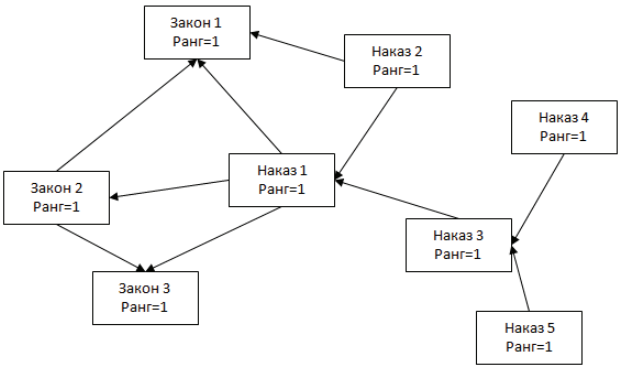

Рис. 1. Попереднє ранжування документів

Будемо застосовувати модель випадкового руху. Почнемо із Закону 1. З цього закону не на наші вершини ведуть зв'язки, тобто його вага залишається поки-що рівною 1. Візьмемо Закон 2. При переході на інші нормативно-правові акти він втрачає $\frac{1-\varepsilon}{n} = \frac{0,85}{2} = 0,425$, де n — кількість посилань з цього нормативно-правового акту. В результаті до рангу Закону 1 та Закону 3 додається по 0,425. Варто відмітити, що сума рангів на даному кроці по документам рахується в кінці кроку, що це не впливало на розрахунки на поточному кроці. Якщо цього не дотримуватися, результат роботи буде залежати від вибору порядку проходження.

Візьмемо Наказ 1. З нього йдуть три посилання. Тому всі три закони отримують до свого рангу ще по 0,283. Цей процес продовжується поки не будуть опрацьовані всі вершини (документи). Потім рахуємо по кожному документу отримані ранги (рис. 2).

На практиці достатньо 50 ітерацій для визначення рангів. В результаті роботи алгоритму ми маємо ранги

по кожному з взаємопов'язаних документів, які визначають ступінь важливості окремого документу.

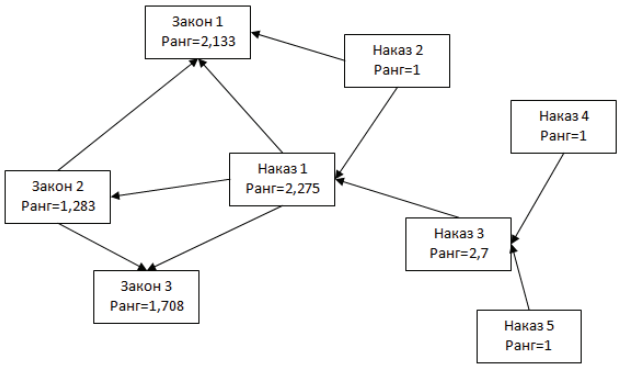

Рис. 2. Ранги після першої ітерації

Але алгоритм має суттєвий недолік: при розрахунку рангів по цьому алгоритму мають враховуватися документи рівнів «Закон України» або «Наказ ДМСУ» і т. д. Це зумовлено ієрархічним характером залежності між документами.

Як бачимо, після ранжування за даним алгоритмом, кожному з документів розраховано свій ранг, який визначає його важливість серед сукупності пов'язаних документів. Для застосування запропонованого способу при пошуку пропонується поєднати латентно-семантичний аналіз та аналіз взаємозв'язків між документами. Як можливий варіант можна розглядати такий підхід:

— Застосувавши формулу тематичної близькості двох термів $FSR(w_1,w_2)=\widehat{X}\widehat{X}^T[w_1,w_2]$ знайти всі документи, які відповідають запиту;

— Серед знайдених документів знайти взаємозв'язки;

— Застосувавши формулу (3) обчислити ранги для кожного документу й видати документи відповідно отриманим рангам.

5. Висновки

У відповідності до мети роботи, у статті досліджено застосування методу латентно-семантичного аналізу до пошуку та кластеризації документів на митну тематику; виявлена й обґрунтована можливість застосування методу аналізу взаємозв'язків Google Page Rank. Вироблені рекомендації щодо спільного застосування латентно-семантичного аналізу та Google Page Rank до пошуку документів на митну тематику. Робота є перспективною для подальших теоретичних та практичних досліджень. Такими перспективами можна вважати: програмну реалізацію методу латентно-семантичного аналізу для побудови документальної пошуково-інформаційної системи митних документів; побудова за допомогою Google PageRank ієрархічної моделі взаємозв'язків між нормативно-правовими актами, що характеризують їх важливість в даній тематиці; побудова експертами структури тематик, створення словників для них і класифікація документів по цим тематикам.

Література

1. Деркач, Л. Українська митниця: вчора, сьогодні, завтра [Текст] / Л. В. Деркач. — К.: Державна митна служба України, 2000. — 542 с.
2. Ульяновська, Ю. В. Автоматизація діловодства в митній справі [Текст] / 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 [Текст] / 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. В. Основи організації електронного документообігу. [Текст] / О. В. Матвієнко, М. Н. Цивін. — К.: Центр учбової літератури, 2008. — 112 с.
5. Belkin, N. Evaluating Interactive Information Retrieval Systems: Opportunities and Challenges N. Belkin, J. Scholtz, S. Dumais, R. Wilkinson [Text] / N. Belkin, J. Scholtz, S. Dumais, R. Wilkinson. — CHI 2004. — April 24–29, 2004. — Vienna, Austria.
6. Корнеев, В. В. Базы данных. Интеллектуальная обработка информации [Текст] / В. В. Корнеев, А. Ф. Гареев, С. В. Васютин, В. В. Райх. — М.: «Нолидж», 2000. — 352 с.
7. Marcus, A. Recovering documentation-to-source-code traceability links using latent semantic indexing [Text] / A. Marcus, J. I. Maletic // Software Engineering, 2003. Proceedings. 25th International Conference on. — Pp. 125—135.
8. Deerwester, S. Indexing by Latent Semantic Analysis [Text] / S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. A. Harshman // Journal of the American Society for Information Science. — 1990. — № 41. — Pp. 391—407.
9. Круковский, М. Ю. Критерии эффективности систем электронного документооборота [Текст] / М. Ю. Круковский // Системи підтримки прийняття рішень. Теорія і практика. — 2005. — С. 107—111.
10. Кураленок, И. Автоматическая классификация документов на основе латентно-семантического анализа [Текст] / И. Кураленок, И. Некрестьянов // Научные труды Донецкого национального технического университета. Серия: Информатика, кибернетика и вычислительная техника (ИКВТ-2006). — Вып. 25. — Донецк: ДонНТУ, 2006. — С. 324—335.
11. Callan, J. Learning while filtering documents. In Proc. of SIGIR'98 [Text] / J. Callan. — Melbourne, Australia, 1998. — Pp. 224—231.

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ ПОИСКА И КЛАССИФИКАЦИИ СЛУЖЕБНЫХ ДОКУМЕНТОВ В АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

В работе рассматривается решение важной практической задачи классификации, поиска и ранжирования служебных документов. Определена и обоснована возможность применения для решения задачи метода латентно-семантического анализа и анализа взаимосвязей Google Page Rank. Проведено моделирование указанными методами.

Ключевые слова: классификация документов, латентно-семантический анализ.

Ульяновська Юлія Вікторівна, *кандидат технічних наук, доцент, кафедра інформаційних систем та технологій, Академія митної служби України, Україна,* ***e-mail: uyv@rambler.ru.***

Ульяновская Юлия Викторовна, *кандидат технических наук, доцент, кафедра информационных систем и технологий, Академия таможенной службы Украины, Украина.*

Ul'yanovskaya Yulia, *Ukrainian Academy of Customs Service, Ukraina,* ***e-mail: uyv@rambler.ru.***