УДК 681.3:355

О.С. Андрощук

Національна академія Державної прикордонної служби України імені Богдана Хмельницького, Хмельницький

МЕТОДИКА ВИЗНАЧЕННЯ КЛАСІВ ТЕКСТОВИХ ДОКУМЕНТІВ ЩОДО ДІЯЛЬНОСТІ ДЕРЖАВНОЇ ПРИКОРДОННОЇ СЛУЖБИ УКРАЇНИ

У статті подано методику визначення класів описів та прецедентів рішень щодо подій, надзвичайних та кризових ситуацій на основі природно-мовних описів, які виникають у діяльності Державної прикордонної служби України, із застосуванням інтелектуального аналізу текстових даних. Застосовуються статистичні методи та методи на основі штучних нейронних мереж для кластерного аналізу, що надає можливість більш точно класифікувати текстові документи з подальшим застосуванням результатів у підтримці прийняття рішень.

Ключові слова: *клас, кластеризація, прецедент, опис, документ, база знань, особлива ситуація.*

Вступ

Постановка проблеми. Забезпечення ефективності діяльності підрозділів, органів та управлінь Державної прикордонної служби України (ДПСУ) значною мірою визначається засобами формування рішень. Особливо нагально це питання постає у випадку подій, надзвичайних та кризових ситуацій (далі – особливих ситуацій). На теперішній час підготовка цих рішень здійснюється на підставі аналізу керівних, інструктивних та додаткових документів тощо (далі – документів), що подаються у письмовому, друкованому або електронному вигляді. Цей аналіз здійснюється керівниками (експертами) без засобів автоматизації. Застосування інтегрованої інформаційно-телекомунікаційної системи "Гарт" ДПСУ сприяє тому, що більшість документів подається саме в електронному текстовому форматі, що призводить до значного збільшення їх кількості. Так, за 2010 рік кількість документів зросла у 5 разів порівняно з 2009 роком. У середньому щодобово до підрозділів, органів та управлінь надходило 200 нових документів. У даних умовах проблема підтримки прийняття рішення (ППР) керівників, у тому числі в особливих ситуаціях (ОС), представляється як подальша автоматизація аналізу природно-мовних електронних текстових даних на предмет їх розподілу за класами (підгрупами або видами) стосовно описів ОС для подальшого вибору класу рішення у конкретній ОС.

Аналіз останніх досліджень і публікацій. Класифікація і кластеризація текстових документів із метою забезпечення ефективного пошуку інформації є предметом досліджень багатьох учених 70-х років XX століття до теперішнього часу. Так, Ван Рійсберген у роботі [1] розглянув застосування формальних методів ієрархічної кластеризації, Г. Селтон та інші дослідники виконали велику роботу з розробки методів кластеризації термінів [2]. У роботах [3, 4] здійснюється застосування статистичних методів. Їх аналіз свідчить, що вони не дають прийнятного рівня якості класифікації текстових документів через специфіку завдання, яке вирішується, і початкових даних, що використовуються.

Не зважаючи на наявність загальної постановки завдання кластеризації, спроби врахувати при його вирішенні ту чи іншу специфіку оброблюваних даних призводять до різних обмежень і свідомо позбавляють універсальності методи вирішення, які пропонуються. У цій роботі пропонується здійснювати визначення класів описів та рішень стосовно ОС у діяльності ДПСУ на підставі неформалізованих описів за допомогою інтелектуальних методів аналізу текстових даних – Text Mining.

Мета статті – подати методику визначення класів опису та прецедентів рішень щодо ОС, які виникають у діяльності ДПСУ на підставі інтелектуального аналізу текстових даних.

Виклад основного матеріалу

Інтелектуальний аналіз текстових даних (ІАТД) визначають як процес аналітичного дослідження великих масивів інформації з метою виявлення певних закономірностей і систематичних взаємозв'язків між змінними, які потім можна застосувати до нових сукупностей даних [5]. Його застосування до опису прецедентів ОС проводиться з метою встановлення відповідності між класами прецедентів ОС і класами рішень щодо управління. Отже, стає можливою побудова моделей типових варіантів процесу розвитку ОС і прийняття рішень керівниками та збереження їх у базі знань (БЗ) у складі інформаційно-телекомунікаційних систем (ІТС) ДПСУ.

Необхідно відзначити, що ідея кластеризації документів на основі термінів, що в них містяться, успішно використовується для вирішення завдань інформаційного пошуку в різних системах. Проте

типова сукупність документів-описів ОС і рішень у діяльності ДПСУ, як правило, має низку особливостей, у силу яких отримання задовільного результату в застосуванні методів кластерного аналізу уявляється проблематичним. Можна виділити такі типи проблем, що виникають при аналізі описів таких ОС:

забезпечення репрезентативної вибірки, що пов'язане з певними труднощами, обумовленими, по-перше, специфікою ОС як рідкісних явищ, і, по-друге, не відлагодженістю механізмів їх (автоматизованої) реєстрації;

малий обсяг вибірки обумовлює низьку частоту появи термінів, які є важливими для ідентифікації ОС;

надмірна стисліть описів і, відповідно, невелика кількість присутніх термінів не надає можливості ідентифікувати ОС достатньо точно;

формулювання описів у неформальному, недокументальному стилі ускладнює виокремлення термінів і розуміння природи ОС;

наявність в описах чисел, дат, скорочень, позначень іноземними мовами ускладнює семантичний аналіз текстів;

певну трудність при проведенні кластеризації становить "вдалий" вибір метрики і міри відстані, який надав би можливість одержати більш-менш однорідні і зпівставні за потужністю класи об'єктів.

Отже, ураховуючі дані особливості, необхідно розбити початкову множину описів ОС на класи так, щоб одержати семантично близькі групи описів для організації швидкого пошуку необхідного рішення і запуску на виконання відповідного сценарію управління в умовах ОС, які виникають.

На рис. 1 подано основні етапи розробленої методики визначення класів описів та рішень прецедентів особливих ситуацій із застосуванням ІАТД.

Запропонований підхід засновано на відображенні множини можливих описів ОС на класи рішень щодо управління і сценарії (плани) їх реалізації. Це означає, що необхідно виконати інтелектуальний аналіз описів ОС із метою визначення множини дескрипторів їх описів у контексті процесу управління і підтримки прийняття рішень.

З'ясовано, що найбільш складним завданням у вирішенні проблем на основі прецедентів є вибір інформативних ознак для розробки індексної структури БЗ. Індексування множини ОР описів ОС дає можливість подати кожний окремий його елемент op_i у вигляді документа Doc_i за допомогою одержаного набору індексних термінів, тобто $Doc_i = (t_1, t_2,\ldots,t_j,\ldots,t_N)$. Для цього випадку є характерним якісне представлення ознак, що містяться в описі ОС на природній мові. Автором описано, яким чином формується множина індексних термінів словника предметної сфери (ПС) $T= \{t_1,...,t_n\}$, яка є необхідним набором елементів для індексування текстових описів ОС.

Рис. 1. Основні етапи методики визначення класів описів-прецедентів особливих ситуацій

У результаті повного індексування всієї множини описів ОР одержуємо деяку матрицю "термін – опис ОС" $TD = \{td_{ij}\}$, i = 1,..,M; j = 1,…,N, M – кількість термінів, M – кількість документів. Її елементами можуть бути як ваги ознак-термінів, у цьому випадку td_{ij} – вага j-го терміну в описі i-ї ОС, так і значення двійкових змінних, що відображають факт присутності/відсутності термінів у відповідних описах, тобто $td_{ij} \in \{0;1\}$.

Одержана матриця TD є, як правило, надмірною, оскільки деякі набуті ознаки (індексних термінів) можуть бути залежними, і тому неінформативними з погляду класифікації. Виокремлення найбільш значущих термінів традиційними методами прикладної лінгвістики (наприклад, виокремлення термінів за частотою зустрічі) не завжди є ефективним, оскільки більшість термінів словника має низьку частоту зустрічі в короткому повідомленні про ОС. Тому додатково здійснюється редукція простору ознак за допомогою методів кластерного аналізу. Виокремлені найбільш інформативні ознаки стають термінами-дескрипторами. Одержана таким чином матриця "термін – ситуація" є початковою для проведення кластеризації ОС. У результаті кластеризації виявляються класи ОС, які потім узгоджуються з експертом. Кожному з класів ставиться у відповідність типове рішення за висновком з ОС, тим самим формується множина класів рішень.

Наступний крок – застосування штучної нейронної мережі (ШНМ) щодо класифікації ОС. Для навчання ШНМ формується таблиця незалежних спостережень ОС із різних класів $\{x'_1,...,x'_n, OP^D_m)^L_i$, де $(x'_1,...,x'_n)$ – вектор термінів-дескрипторів прецедентів; OP^D_m – вказівка експерта про дійсну приналежність рішення, що міститься в прецеденті, до одного з m класів рішень на множині рішень OP^D; L – кількість спостережень ОС. Отже, як вхідні дані ШНМ виступають вектори термінів-дескрипторів описів прецедентів, а виходами є класи типових рішень.

Для того, щоб навчити ШНМ коректно розпізнавати класи, необхідно мати достатньо велику вибірку достовірних навчальних прикладів, що на порядок перевищує кількість ознак класифікації, які використовуються. Якщо вдається налаштувати вагові коефіцієнти мережі, досягнувши прийнятного рівня помилки навчання, то ШНМ здатна класифікувати з дуже великою точністю, ураховуючи особливості прикладів, що не формалізуються.

Наведемо приклад використання запропонованої методики для класифікації прецедентів ОС, що виникають у процесі управління оперативно-службовою діяльністю відділу прикордонної служби (підрозділу) ДПСУ. Розглядається випадок повеней на державному кордоні.

Спочатку на основі результатів моделювання процесів управління й автоматизованого лінгвістичного аналізу сукупності 70 текстових описів прецедентів випадків підтоплення (повені) на державному кордоні за допомогою програми Text Analyst було сформовано множину індексних термінів, що включає 68 одиниць. На рис. 2 подано фрагмент словника зазначеної ПС.

Потім було одержано матрицю $DT=\{dt_{ij}\}$, елементами якої є значення бінарної змінної, що відображає факт присутності/відсутності термінів у відповідних описах.

Далі було проведено кластеризацію ознак за допомогою ієрархічних агломеративних методів. Для оцінки ефективності використання методів кластерного аналізу стосовно описів ОС було проведено низку експериментів.

Обчислювальні процедури кластеризації реалізовувалися за допомогою статистичного пакету SPSS v13.0.

Було проаналізовано декілька різних комбінацій метрик і мір відстаней, які давали прийнятні результати.

У табл. 1 наведено метрики для визначення відстаней між векторами термінів (позначеними символами X і Y), які найбільш часто використовуються для кластеризації текстових документів. Вибір міри схожості визначається структурою початкових даних ПС, яка розглядається.

№	Термін	№	Термін	№	Термін	№	Термін
1	АРтаІНР	11	продукти	21	інсп_пост	31	моб_прик_зас
2	авто (УАЗ)	12	ДПС	22	інспектор	32	наказ
3	загроза_підтоп	13	місц_нес_сл	23	керівник	33	нас_пункт
4	беріг	14	евак_група	24	комісія	34	населення
5	блокування	15	евакуація	25	кордон	35	нач_відд
6	будинок	16	евак_пункт	26	ліквідація	36	нач_заг
7	взаємодія	17	контроль	27	майно	37	обстановка
8	ВПС	18	зброя	28	маршрут	38	оповіщення
9	відслідковування	19	зв'язок	29	мед_зах	39	охорона_кордону
10	вода	20	злива	30	персонал	40	патрулювання

№	Термін	№	Термін	№	Термін
41	переправа	51	пункт_проп	61	територія
42	персонал	52	район	62	транспорт
43	підприємства	53	резерв	63	формування
44	підр_МЧС	54	рівень_води	64	черг_сили
45	підрозділ	55	розрах_запр	65	ЦУС
46	підтоплення	56	сит_група	66	черговий
47	повінь	57	ситуація	67	човен
48	прик_заг	58	служби	68	штаб
49	прик_нар	59	сім'ї		
50	приміщення	60	стик		

Рис. 2. Фрагмент словника предметної сфери управління прикордонними підрозділами у випадку повені

Таблиця 1

Обчислення схожості для ознак із символьними значеннями

Найменування функції Sim(*x,y*)	Двійковий вектор термінів	Вектор зважених термінів
Евклідова відстань	$\lvert X \cap Y \rvert$	$\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Коефіцієнт Міньковського	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\left(\sum_i \lvert x_i - y_i \rvert^r\right)^{1/r}$
Коефіцієнт Дейка	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_i (x_i \cdot y_i)}{\sum_i x_i^2 + \sum_i y_i^2}$
Коефіцієнт Рассела і Рао	$\frac{\lvert X \cap Y \rvert}{n}$	$\frac{\sum_i (x_i \cdot y_i)}{n}$
Коефіцієнт косінуса	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert}\sqrt{\lvert Y \rvert}}$	$\frac{\sum x_i \cdot y}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$

На рис. 3 подано фрагмент дендрограми, яка ілюструє результати кластеризації методом Варда з використанням квадратичної Евклідової відстані у якості метрики. Метод Варда надає можливість об'єднувати ті кластери, які дають найменший внесок у функцію якості $\sum_k \sum_{tj} \sum_n (X_{ij} - \bar{X}_{ij})^2$, де k – кількість кластерів; tj – кількість об'єктів у кластері $\tilde{J}$; i – індекс ознаки; j – номер об'єкту в кластері $\tilde{J}$; n – кількість ознак; X_{ij} – значення ознаки i для об'єкту $\tilde{j}$; $\bar{X}_{ij}$ – середнє значення ознаки i для кластера $\tilde{J}$. Метод оптимізує критерій для кожного рівня групування. У результаті об'єднання об'єктів у кластери сформувалося 17 груп термінів. Відповідна структура є оптимальною з точки зору потужності кластерів, їх однорідності і ступеня деталізації.

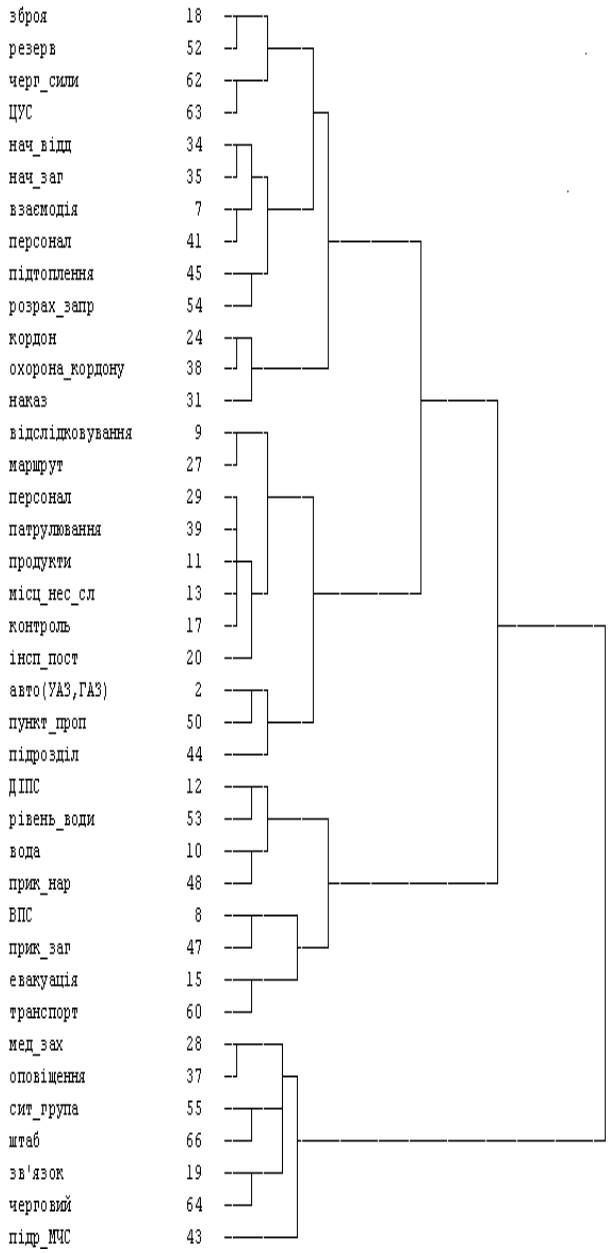

Рис. 3. Фрагмент дендрограми кластеризації індексних термінів прецедентів особливих ситуацій методом Варда

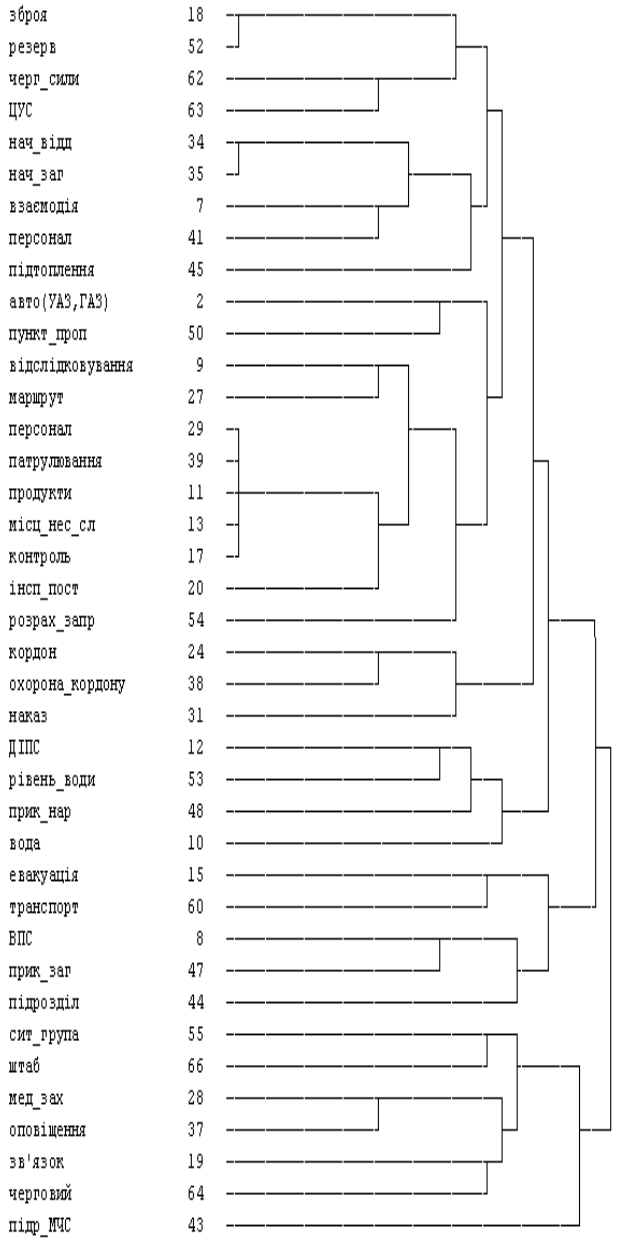

Рис. 4. Фрагмент дендрограми кластеризації індексних термінів прецедентів особливих ситуацій методом "Зв'язок між групами"

На рис. 4 подано фрагмент дендрограми, яка ілюструє результати кластеризації методом "Зв'язок між групами" з використанням у якості метрики коефіцієнта Міньковського.

У результаті об'єднання кластерів утворилося 15 груп термінів, склад яких дещо відрізняється від структури, одержаної за допомогою методу Варда.

Проте, у результаті проведеного аналізу було виявлено 11 стійких груп термінів, серед яких було виокремлено дискримінанти і, крім того, виключено малоінформативні терміни.

Отже, на основі результатів кластерного аналізу з урахуванням змістовних міркувань було сформовано множину дескрипторів, що включає 39 термінів.

Результатом проведеної класифікації термінів є породження груп асоційованих термінів, придатних до включення в тезаурус [5]. Групи при цьому можуть бути непов'язаними одна з одною або, навпаки, між ними можуть бути визначені відносини. Якщо за своєю природою відносини між групами термінів належать до родового типу, то одержують ієрархії термінів; в інших випадках групи можуть бути впорядковані у вигляді двовимірної семантичної мережі. Кластер семантичних відносин фактично формує структуру спрямованого ациклічного графа, що є фрагментом тезауруса.

Для попереднього виявлення класифікаційної структури в описах про прецеденти ОС було проведено їх кластеризацію за допомогою ієрархічних

агломеративних методів. На рис. 5 подано фрагмент відповідної дендрограми кластеризації, що відображає результати за методом Варда з використанням квадратичної Евклідової відстані в якості метрики.

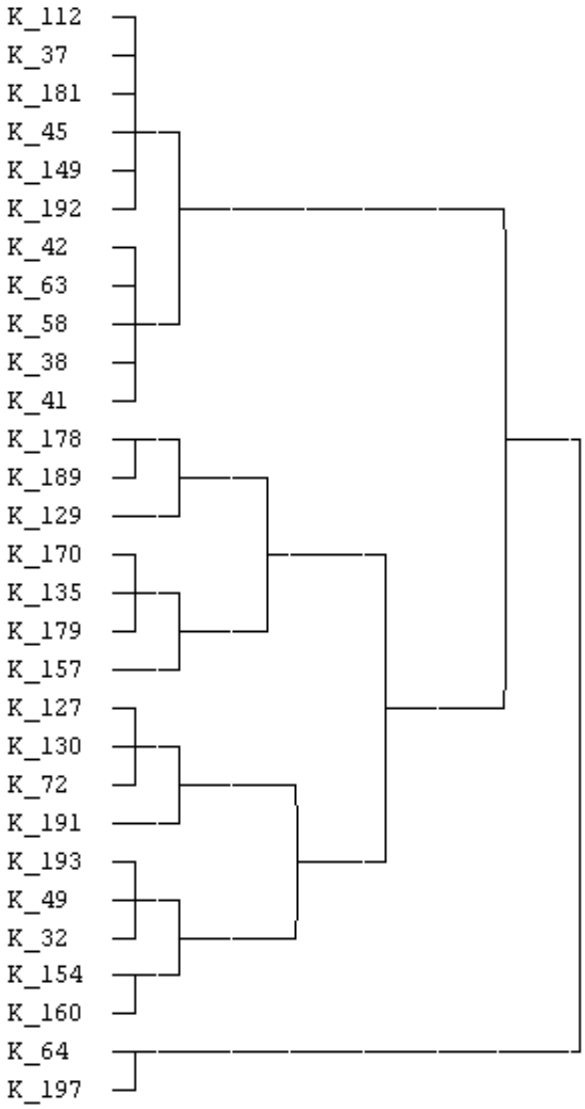

Рис. 5. Фрагмент дендрограми кластеризації описів прецедентів особливих ситуацій

На рис. 6 подано фрагмент матриці зазначених описів прецедентів ОС. Аналіз отриманих результатів надав можливість виокремити 7 укрупнених класів ОС.

Отже, використання методів кластерного аналізу в умовах, обумовлених особливостями описів ОС, надає можливість виокремити стійку класифікаційну структуру в даних про прецеденти і сформувати інформаційний простір для пошуку рішень у БЗ.

Далі за допомогою експертів було сформовано вибірку описів ОС, що включає класи типових рішень, які відповідають виокремленим категоріям ОС.

Наступна частина експериментів щодо класифікації ОС відповідно до рішень з управління підрозділами проводилася з використанням статистичних класифікаторів, класифікаторів, які засновані на функціях подібності, ШНМ типу багатошаровий перцептрон (MLP) та використанням ШНМ Кохонена (SOM).

Документ	ОС	Повінь	Евакуація	Пожежа	Оповіщення	Резерв	Прик_заг
154	0	0	0	0	0	1	1
160	0	0	0	0	0	1	1
32	0	0	0	0	0	0	1
193	0	0	0	0	0	0	1
72	1	0	0	0	0	0	1
127	1	0	0	0	0	0	1
130	1	0	0	0	0	0	1
49	0	0	0	0	0	0	1
129	0	0	0	1	1	0	1
157	0	0	0	1	0	0	1
135	1	0	0	1	0	0	1
178	1	0	0	1	0	0	1
191	1	0	0	1	0	0	1
189	1	0	0	1	1	0	1
64	1	1	1	0	1	0	0
197	1	1	1	1	1	0	0
58	1	0	0	0	0	0	0
38	1	0	0	0	0	0	0
41	0	0	0	1	0	0	0
42	0	0	0	1	0	0	0
63	0	0	0	1	0	0	0
181	0	0	0	1	0	1	0
192	0	0	0	1	0	1	0
45	0	0	0	1	0	1	0
149	0	0	0	1	0	1	0
112	0	0	0	1	0	1	0
37	0	0	0	1	0	1	0

Рис. 6. Фрагмент матриці описів прецедентів особливих ситуацій

Результати застосування різних класифікаторів подано у табл. 2.

Таблиця 2

Експериментальне дослідження класифікаторів

Показник	1	2	Нейромережеві класіфікатори	
			3	4
Точність (V)	0,25	0,43	0,68	0,89
Повнота (U)	0,34	0,45	0,75	0,91

Примітка: 1 – статистичні класифікатори, 2 – класифікатори, які засновані на функціях подібності, 3 – багатошаровий перцептрон, 4 – мережа Кохонена.

Практичну реалізацію наведених підходів здійснено у вигляді програмних модулів (підсистем), що

отримують та обробляють первинну (вихідну) інформацію і входять до складу інформаційно-телекомунікаційних систем ДПСУ "Гарт", наприклад, у систему "Реєстрація подій/надзвичайних(кризових) ситуацій".

Отже, у результаті застосування методики у розглянутому прикладі на підставі аналізу природномовних текстових описів особливих ситуацій було визначено їх клас та відповідні їм класи рішень у випадках повені на державному кордоні.

Висновки

Отже, у результаті роботи описаної методики, яка реалізує інтелектуальний аналіз текстових даних щодо описів прецедентів ОС, стає можливим визначити типові рішення за висновком з ОС на основі природно-мовних описів, які можуть поступати, наприклад, з органів (підрозділів) охорони державного кордону у вигляді письмових (з подальшим переводом до електронних), електронних донесень, інструкцій і наказів тощо.

Використання методів кластерного аналізу інтелектуального аналізу текстових даних в умовах, обумовлених особливостями описів ОС у діяльності Державної прикордонної служби України, надає можливість виокремити відносно стійку класифікаційну структуру в даних про прецеденти і сформувати інформаційний простір для пошуку рішень у базах знань у складі інформаційно-телекомунікаційних систем Державної прикордонної служби України.

Застосування штучних нейронних мереж для вирішення завдання класифікації ОС на основі їх описів дають достатньо добрі результати і дозволяють достатньо точно відтворити розумову діяльність керівника (експерта).

Слід урахувати, що завдання доводиться вирішувати в межах певних обмежень, пов'язаних у першу чергу з необхідністю мати великий зумовлений набір якісних навчальних прикладів.

Комплексне застосування методів інтелектуального аналізу текстових даних надає можливість більш точно класифікувати документи з діяльності Державної прикордонної служби України.

Напрямком подальших досліджень слід вважати синтез алгоритмів підтримки прийняття рішення з використанням баз знань, які містять прецеденти.

Список літератури

1. Rijsbergen C.J. Information Retrieval / C.J. van Rijsbergen; Second Edition. – Butterworths, Glasgou, 1979.

2. Сэлтон Г. Автоматическая обработка, хранение и поиск информации / Г. Сэлтон; [пер. с англ., под ред. А.И. Китова]. – М. : Сов. радио, 1973. – 560 с.

3. Кочетков П.А. Краткий курс теории вероятностей и математической статистике: учебное пособие / П.А. Кочетков. – М. : МГИУ, 1999. – 51 с.

4. Елисеева И.И. Общая теория статистики: учебник; 5-е изд., перераб. и доп. / И.И. Елисеева, М.М. Юзбашев. – М. : Финансы и статистика, 2004. – 656 с.

5. Дюк В. Data mining: учебный курс / В. Дюк, А. Самойленко. – СПб : Питер, 2001. – 368 с.

6. Дж. Вэн Райзин. Классификация и кластер / Дж. Вэн Райзин; пер. с англ. – М. : Мир, 1980. – 389 с.

Надійшла до редколегії 8.10.2010

Рецензент: д-р техн. наук, проф. М.С. Вертузаєв, Інститут Служби зовнішньої розвідки, Київ.

МЕТОДИКА ОПРЕДЕЛЕНИЯ КЛАССОВ ТЕКСТОВЫХ ДОКУМЕНТОВ В ДЕЯТЕЛЬНОСТИ ГОСУДАРСТВЕННОЙ ПОГРАНИЧНОЙ СЛУЖБЫ УКРАИНЫ

О.С. Андрощук

В статье подана методика определения классов описаний и прецедентов решений относительно событий, чрезвычайных и кризисных ситуаций на основе естественно-языковых описаний, которые возникают в деятельности Государственной пограничной службы Украины, с применением интеллектуального анализа текстовых данных. Применяются статистические методы и методы на основе искусственных нейронных сетей для кластерного анализа, что предоставляет возможность более точно классифицировать текстовые документы с последующим применением результатов в поддержке принятия решений.

Ключевые слова: *класс, кластеризация, прецедент, описание, документ, база знаний, особенная ситуация.*

METHOD OF DETERMINATION OF CLASSES OF TEXTS DOCUMENTS IN RELATION TO ACTIVITY OF GOVERNMENT BOUNDARY SERVICE OF UKRAINE

O.S. Androshchyk

In the article the method of determination of classes of descriptions and precedents of decisions is given in relation to events, extraordinary and crises situations on the basis of naturally-linguistic descriptions which arise up in activity of Government boundary service of Ukraine, with application of intellectual texts data analysis. Statistical methods and methods are used on the basis of artificial neurons networks for the cluster analysis that gives possibility more exactly to classify texts documents with subsequent application of results in support of decision-making.

Keywords: *class, clusterization, precedent, description, document, knowledges base, special situation.*