Литвин В. Метод класифікації текстових документів із використанням онтологічного підходу / Литвин В., Бобик І., Мельник А. // Вісник ТНТУ. — 2011. — Том 17. — № 2. — С.208-215. — (математичне моделювання. математика. фізика).

УДК 004.65

В. Литвин, канд. техн. наук; І. Бобик, канд. фіз.-мат. наук; А. Мельник

Національний університет „Львівська політехніка”

МЕТОД КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ ІЗ ВИКОРИСТАННЯМ ОНТОЛОГІЧНОГО ПІДХОДУ

***Резюме**. Розглянуто підхід до класифікації текстових документів із використанням онтологічного підходу. Розроблено метод рубрикування електронних текстових документів, що ґрунтується на метриці, яка, у свою чергу, використовує специфіку онтології рубрик.*

***Ключові слова:** текстовий документ, онтологія, рубрикування, прецедент, метрика.*

V. Lytvyn, I. Bobyk, A. Meljnyk

CLASSIFICATION METHODS OF TEXT DOCUMENTS USING ONTOLOGY BASED APPROACH

***The summary.** This article discusses an approach to classification of text documents using ontological approach. The method of text documents categorization based on metrics, which uses the rubric ontology specificity, is built.*

***Key words:** text document, ontology, rubrication, precedent, metrics.*

Умовні позначення

ТД – текстовий документ;
Pr – прецедент;
O – онтологія;
C – термін.

Постановка проблеми у загальному вигляді. Класифікація текстових документів (ТД) розглядається як один із можливих варіантів вирішення проблеми використання інформаційних ресурсів. Коротко вона характеризується таким чином. Різними сховищами знань (у тому числі бібліотеками) накопичені величезні інформаційні масиви. Проблема полягає в складності орієнтуватись у цих масивах через їх значний розмір. Тим самим втрачається можливість отримувати найактуальнішу і найповнішу інформацію з різних тем, що цікавить користувача, бо відсутність класифікації робить некорисною більшу частину накопичених ресурсів. Оскільки дослідження деякої конкретної задачі вимагає значних трудовитрат на пошук і аналіз інформації з теми, до якої відноситься ця задача, тому багато рішень приймають на основі неповного уявлення про проблему. Задача класифікації ТД також виникає, коли необхідно навести порядок у банку даних електронних ресурсів. Наприклад, на персональному комп’ютері необхідно автоматично порозносити множину файлів за різними каталогами.

Класифікацію природомовних ТД називають рубрикуванням. Використання рубрикаторів дозволяє скоротити витрати на пошук потрібної інформації, представленої електронними текстами [1]. Застосування семантичного підходу (онтологій) дозволяє підвищити ефективність процесу рубрикування.

Математично задача класифікації ТД визначається таким чином. Існує множина текстів $T = \{T_1, T_2, ..., T_M\}$, множина N рубрик, які ми будемо розглядати як прецеденти

$\Pr = \{\Pr_1, \Pr_2, ..., \Pr_N\}$. Кожна рубрика подається деяким описом, яка має певну внутрішню структуру [2]. Процедура класифікації f текстів $T_i \in T$ полягає у виконанні певних процедур, на основі яких робиться висновок про відповідність T_i одній зі структур $\Pr_j$, що означає віднесення T_i до прецедента $\Pr_j$, або висновок про неможливість класифікації T_i. Тобто у множину прецедентів необхідно додати порожній прецедент $\Pr_0$, якому будемо ставити у відповідність не прокласифіковані тексти. У нашому випадку елементами множини T є електронні версії ТД. Отже загальну модель класифікації ТД запишемо у вигляді

$$f : T \to \Pr . \quad (1)$$

Аналіз останніх досліджень і публікацій. Для розв'язування задачі (1) використовуються різні види класифікаторів:

1. Статистичні класифікатори на основі ймовірнісних методів. Найвідомішими серед них є сімейство Байесівських класифікаторів. Загальною рисою таких методів є процедура f, в основі якої лежить формула Байеса для умовної ймовірності. Аналізований текст T_i представляється у вигляді послідовності термінів $\left(C_{i_1}, C_{i_2}, ..., C_{i_{Ni}}\right)$. Кожна рубрика $\Pr_j$ характеризується безумовною ймовірністю її вибору $p\left(\Pr_j\right)$ у процесі класифікації деякого документа T_i (сукупність таких подій для всіх рубрик утворюють систему гіпотез, де $\sum_{j=1}^{N} p\left(\Pr_j\right) = 1$), та умовною ймовірністю $p\left(C \mid \Pr_j\right)$ – зустріти термін C у документі T_i за умови вибору рубрики $\Pr_j$. Ці величини утворюють елементи V_j множини $V = \{V_1, V_2, ..., V_N\}$ описів рубрик і використовуються при розрахунку ймовірностей $p_{ij} = p\left(T_i \mid \Pr_j\right)$ того, що текст T_i буде класифікований за умови вибору рубрики $\Pr_j$. При розрахунку p_{ij} враховується представлення T_i у вигляді послідовності термінів $\left(C_{i_1}, C_{i_2}, ..., C_{i_{Ni}}\right)$. Підставивши ці величини у формулу Байеса, отримуємо ймовірність

$$\tilde{p}_{ji} = p\left(\Pr_j \mid T_i\right) = \frac{p\left(\Pr_j\right) \cdot p\left(T_i \mid \Pr_j\right)}{\sum_{k=1}^{N} p\left(\Pr_k\right) \cdot p\left(T_i \mid \Pr_k\right)}$$

того, що документ T_i буде віднесений до рубрики $\Pr_j$. Процедура f зводиться до підрахунку $\tilde{p}_{ji}$ для всіх рубрик $\Pr_j$ і вибору тієї, для якої ця величина максимальна. Навчання рубрикатора зводиться до складання словника $\left(C_1, C_2, ..., C_K\right)$ та визначення для кожної рубрики величин $p\left(\Pr_j\right)$ і $p\left(C \mid \Pr_j\right)$, де $C \in \left(C_1, C_2, ..., C_K\right)$.

2. Класифікатори, що використовують методи на основі штучних нейронних мереж. Даний вид класифікаторів добре зарекомендував себе в задачах розпізнавання зображень. Опис прецедентів $\Pr$, як правило, являє собою багатовимірні вектори дійсних чисел синаптичних ваг штучних нейронів, а процедура класифікації f характеризується способом перетворення аналізованого тексту T до такого вектора, що залежить від функції активації нейронів, а також топологій мережі. Процес навчання класифікатора в даному випадку співпадає з процедурою навчання нейронної мережі та залежить від обраної топології [3].

3. Класифікатори, засновані на функціях подібності. Характерною рисою даного методу є універсальність описів V, які, з одного боку використовуються для представлення змісту рубрик, а з іншого – змісту аналізованих текстів. Процедура класифікації f використовує міру подібності вигляду $d: V \times V \to [0,1]$, що дозволяє кількісно оцінювати тематичну близькість описів $V_T \in V$ і $V_i \in V$, де опис V_T представляє зміст аналізованого тексту, а V_i – зміст *i*-ої рубрики. Процедура класифікації f зводиться до перетворення аналізованого тексту T у представлення V_T та знаходження оцінки подібності опису V_T з описами рубрик V_i. Тобто обчислюється відстань між описами $d(V_T, V_i)$. На основі цієї відстані здійснюється висновок про належність тексту рубриці. По суті пропонований підхід є частинним випадком цих класифікаторів. У якості описів V ми пропонуємо використовувати онтології рубрик O, а для обчислення відстані ми побудували метрику, яка базується на онтології.

4. Класифікатори на основі кластерного аналізу. Основна мета цього аналізу – виділення у вихідних даних однорідних груп. Процедури кластерного аналізу здійснюють виділення структури даних на основі попарного порівняння елементів вихідного масиву. Самі об'єкти, що підлягають структурній класифікації (кластеризації), розташовуються в просторі, вимірами якого є ознаки. Якщо ознака n, то простір n-вимірний, але у зв'язку з можливою кореляцією ознак розмірність простору може бути зменшена. Не зважаючи на широке застосування кластерного аналізу, загальноприйнятого визначення кластера не існує. Є лише інтуїтивне уявлення про те, що елементи одного кластеру ближчі один до одного, ніж до інших елементів, які знаходяться поза цим кластером. Також не існує точної постановки задачі кластерного аналізу. На основі загального поняття про кластери задача може бути сформована таким чином: необхідно множину текстів $T = \{T_1, T_2, ..., T_M\}$ віднести до деякого елемента множини Pr, використовуючи деякий критерій подібності елементів. Подібність елементів множини T зумовлюється наявністю множини ознак $X = \{x_1, x_2, ..., x_m\}$, які характеризують елементи множини T. Виміряні значення ознак для елемента T_i утворюють вектор $U_i = (u_{i1}, u_{i2}, ..., u_{im})$. Ввівши для множини векторів $U = \{U_1, U_2, ..., U_M\}$ деяку міру їх подібності, можна виконати різні процедури групування (кластеризації) документів.

Формування цілей. Розробити метод рубрикування ТД, який базуватиметься на онтології рубрик. Апробувати його шляхом побудови інтелектуального агента класифікації ТД, в основі якого використати розроблений метод.

Основний матеріал. З метою побудови метрики для рубрикування ТД на основі онтологій розширимо класичне поняття онтології шляхом введення в її структуру скалярних величин. Під класичною моделлю онтології O розуміють структуру вигляду

$$O = \langle C, R, F \rangle,$$

де C – поняття; R – відношення між поняттями; F – інтерпретація понять та відношень (аксіоми). Аксіоми встановлюють семантичні обмеження для системи понять та відношень. Ми пропонуємо зважувати поняття та зв'язки, оскільки для певних рубрик важливість зустрічі окремих понять у ТД різна. Коефіцієнт важливості поняття (зв'язку) – це чисельна міра, котра характеризує значущість певного поняття (зв'язку) для рубрики. Таким чином, отримаємо модель онтології

$$O = \langle C, R, F, W, L \rangle,$$

де W – важливість понять C; L – важливість відношень R.

Така онтологія однозначно представляється у вигляді зваженого концептуального графа (КГ) [4]. Тому метрику будемо будувати, використовуючи зважені КГ.

Ми пропонуємо визначати відстань між прецедентом і ТД як суму відстаней між „найважливішими" поняттями прецедента та ТД [5]. Наприклад, в авторефераті чи десартаційній роботі завжди вказується об'єкт досліджень, який і є „найважливішим" поняттям. Оскільки онтологія подається у вигляді зваженого КГ, то таке поняття є центром ваг відповідного зваженого КГ. Таких „важливих" понять може бути одне, два; однак якщо їх є більше-дорівнює трьом, то пропонуємо вибирати перші три. Ця кількість визначена на основі опитувань експертів різних предметних областей і ми її вважаємо оптимальною. Таким чином, маємо три центри ваг *i*-го прецедента pr_i^1, pr_i^2, pr_i^3 і три центри ваг ТД s^1, s^2, s^3. Тоді існує дев'ять різних відстаней $d\left(pr_i^j, s^k\right)$, $j = 1,2,3$; $k = 1,2,3$. Вибираємо три з них d^1, d^2, d^3 таким чином, щоб їх сума була найменшою і кожне із шести понять, які є центрами ваг, брали участь в обчисленні відстані. Приклад такої комбінації трьох співвідношень між центрами ваг для обчислення відстаней наведено на рис. 1.

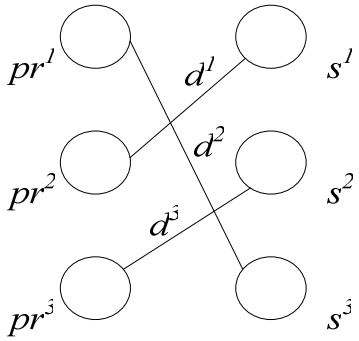


Рисунок 1. Приклад комбінації центрів ваг КГ прецедента та ТД для обчислення відстані

Отримана таким чином сума й буде відстанню між прецедентом та ТД. З математичної точки зору центром ваг КГ є поняття, середня відстань від якого до всіх інших понять є найменшою. Очевидно, що визначена таким чином відстань залежатиме від того, як ми визначимо відстань між двома суміжними вершинами КГ. Для цього пропонується визначати відстані між вершинами, що з'єднані зв'язком як

$$d_{ij} = \frac{Q}{L_{ij}\left(W_i + W_j\right)}, \tag{2}$$

де W_i та W_j – коефіцієнти важливості вершин C_i та C_j відповідно; L_{ij} – коефіцієнт важливості зв'язку між вершинами; Q – константа, яка залежить від конкретної онтології. Приймемо, що $L_{ii} = \infty$, тоді $d_{ii} = 0$.

Далі знаходимо центри ваг КГ. Це перші три вершини, для яких середня відстань $\overline{d}_i$ є найменшою:

$$\overline{d}_{i*} = \min_i \overline{d}_i . \tag{3}$$

Середню відстань $\overline{d}_i$ для вершини C_i обчислюємо згідно з формулою:

$$\overline{d}_i = \frac{\sum_{j=1, j\neq i}^{n} d_{ij}^*}{n-1}, \tag{4}$$

де n – кількість вершин графа; d_{ij}^{*} – найкоротший шлях між вершинами C_i та C_j, який обчислюємо за допомогою відомих алгоритмів, наприклад, Форда, Дейкстри, Флойда-Уоршалла [6].

Далі згідно з КГ, що задає онтологію прецеденту, шукаємо відстань від даного прецедента до ТД. Якщо поняття ТД не входять у КГ, то його відсутність зумовлює зростання відстані до безмежності, що означає не близькість прецеденту із ТД.

Накладаємо два КГ, які відповідають прецеденту та ТД. Можливі два випадки:

а) якщо вони мають спільні дуги, то відстань між вершинами, що з'єднані такими дугами, визначаємо як середню відстань двох графів:

$$\overline{d}^{12} = \frac{\overline{d}^{1} + \overline{d}^{2}}{2}; \tag{5}$$

б) якщо дуги не є спільними, то відстань між вершинами береться з відповідного графа.

Обчислюємо найкоротший шлях між трьома центрами ваг двох КГ

$$d(\mathrm{Pr}_i, T) = \sum_{j=1}^{3} d^{j}, \tag{6}$$

де $d^{j} = \overline{d}^{st}$; C^{s} – центр ваги 1-го графа; C^{t} – центр ваги 2-го графа.

Найкоротший шлях між вершинами обчислюємо за допомогою алгоритму Дейкстри.

Очевидно, що залежно від прецеденту ваги понять різні. Тобто насправді W – вектор вимірності кількості прецедентів $W = (W_1, W_2, ..., W_N)$. Для рубрикування ТД ваги коефіцієнтів важливості понять прецедентів онтології ми обчислювали на основі статистичного аналізу наявності понять у ТД, для яких відомо, до якої рубрики вони належать. Тобто при кожному входженні деякого поняття C_i у ТД, який належить до прецеденту Pr_j, вага цього поняття збільшувалася на одиницю $W_{ji} = W_{ji} + 1$. Очевидно, що на початку всі ваги $W_{ji} = 0$. Детальніше присвоєння ваг важливості поняттям і відношенням онтології описано в [7, 8].

Для побудови КГ ТД використано готові програмні засоби, які опрацьовують речення та розпізнають у них граматичні зв'язки [9].

Розроблений інтелектуальний агент (ІА) рубрикування ТД написаний мовою програмування Python, онтологія розроблена в редакторі Protégé-OWL. В якості формату словників обрано формат, який використовується для словників Hunspell. Для кожної мови використано кілька файлів, а саме, словник, який містить слова, файл афіксів, який визначає значення спеціальних позначок у словнику, файл стоп-слів, які фільтруються при визначенні термінів, та файл біграм, який використовується під час використання N-грамних моделей для визначення мови. Для автоматичного визначення кодування ми використали метод розподілу символів. Система підтримує формати даних doc, docx, docm, pdf, rtf, txt, html, htm.

Для прикладу функціонування ІА обрано такі рубрики: психологія, релігія, логістика, філософія, безпека життєдіяльності, охорона праці, цивільна оборона, екологія, соціологія, культура, комп'ютерні мережі. Для визначення ваг термінів використано Інтернет-джерела, зокрема реферати з порталу http://ua.textreferat.com. На рис. 2 наведено терміни, які мають найбільшу вагу в онтології, залежно від рубрики.

Для підтримки прийняття рішень щодо рубрикування, отримані відстані $d_{ij} = d\left(\mathrm{Pr}_i, T_j\right)$ перетворимо у ймовірнісні величини p_{ij} (ймовірність того, що текстовий документ T_j належить до прецедента Pr_i)

$$p_{ij} = \frac{d_{ij}}{\sum_{k=1}^{N} d_{kj}}. \qquad (7)$$

Наведемо результати роботи розробленого ІА. Документ «Будизм» містить інформацію на релігійну тематику, показує історію виникнення та різновиди. Система присвоїла тексту тематику релігієзнавство та частково зачепила філософію і культуру (рис. 3а). Насправді текст зачіпає ці питання, проте основною тематикою все ж є релігія та її історія, отже вибір можна вважати вірним.

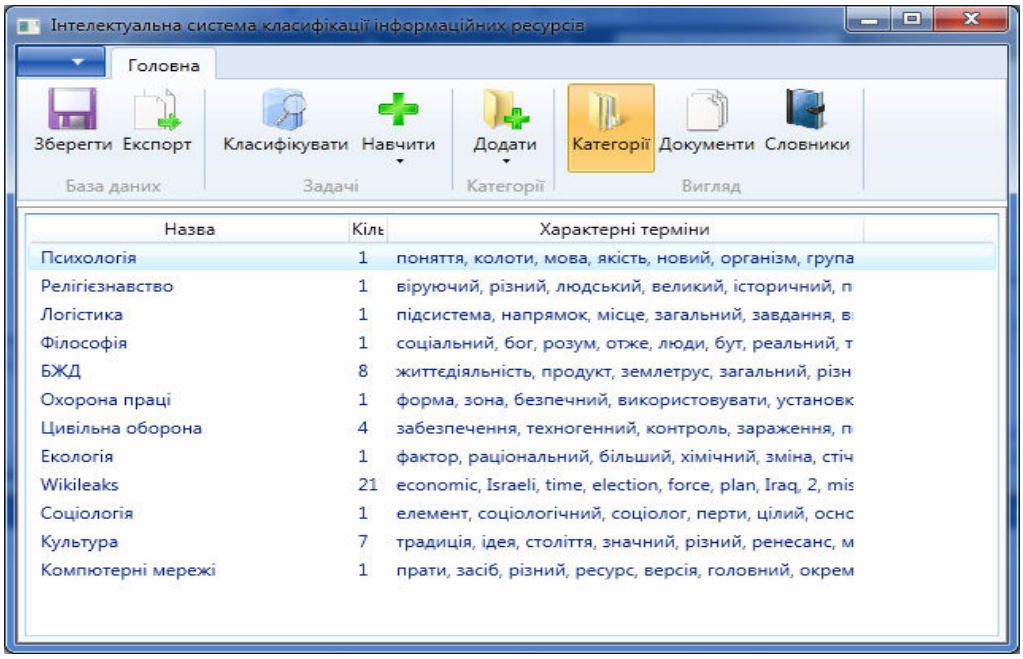
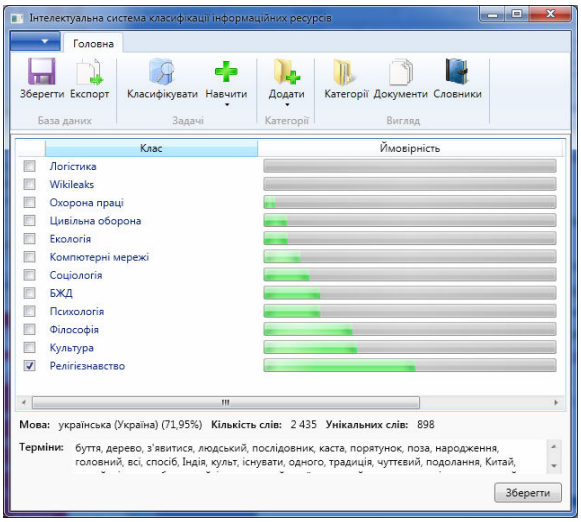
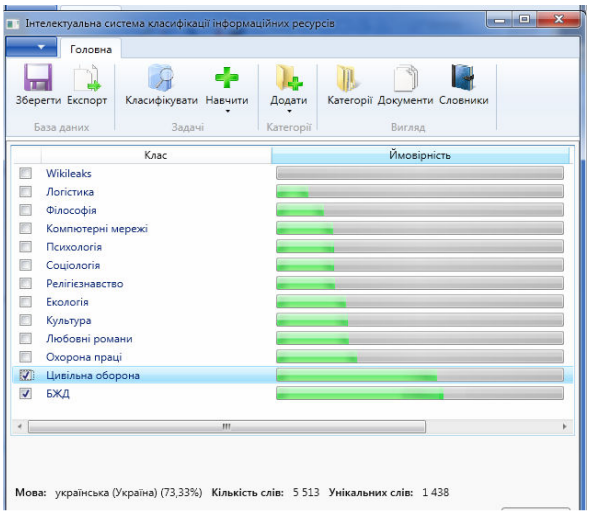


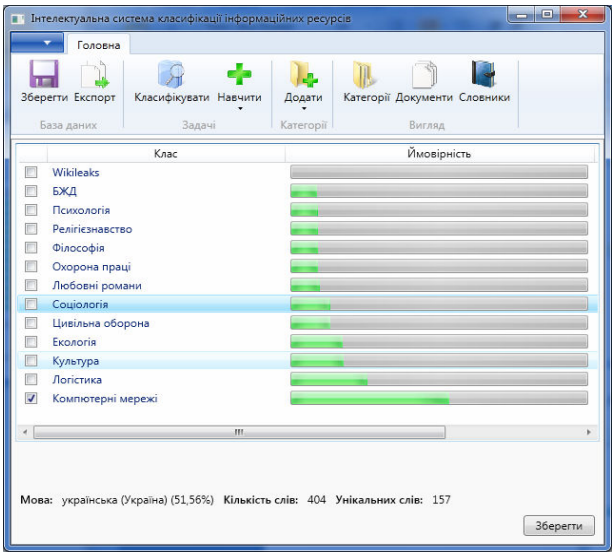
Рисунок 2. Характерні терміни рубрик



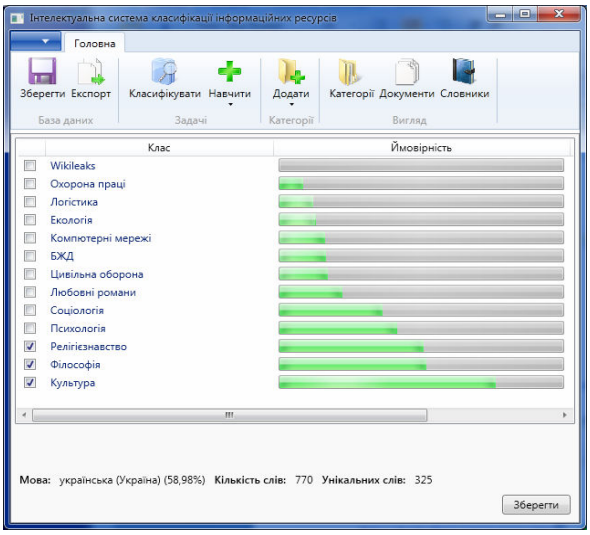
а)



б)



в)



г)

Рисунок 3. Приклади рубрикування документів

Текст «Чорнобиль – глобальна екологічна катастрофа» містить історію аварії на Чорнобильский АЕС. У тексті детально описано події з вказанням заходів, що були забезпечені, доз опромінення, як проводилася евакуація тощо. Таким чином, він належить до тематик цивільної оборони, безпеки життєдіяльності та екології. Система віднесла лише до перших двох (рис. 3б). Це пояснюється спрямованістю та акцентуацією тексту в цьому напрямку. Тематика екології, за допомогою якої створювалася програма, є більш загального характеру.

Документ «Навіщо потрібні комп'ютерні мережі», як бачимо з назви, містить текст про комп'ютерні мережі, її становлення та обґрунтування її необхідності. Система вірно віднесла її до тематики, проте частково зачепила тематику логістики (рис. 3в). Це обґрунтовується наявністю в тексті опису проблем встановлення та прокладання мережі.

Документ «Передумови епохи Відродження» містить інформацію про філософські концепції в епосі відродження. Джерело, з якого було взято цей текст, стверджує, що він належить до тематики філософії. Проте все ж у ньому розглянуто культуру, філософію та вплив релігії у цей період. Тобто текст містить інформацію на різну тематику, не обмежену лише рамками філософії. Система зробила вірний вибір, однак, можливо, було б вірно обрати психологію, оскільки ця проблема частково розглядається (рис. 3г). Проте це питання є спірним, і тому психологія знаходиться на межі вибору.

Результати тестування показали, що розроблений ІА класифікації ТД на основі онтології 87% файлів прокласифікував правильно. Однак не всі тексти можна чітко віднести до якоїсь рубрики. Система не може визначити точну тематику таких текстів, тому пропонує віднести до кількох рубрик.

Висновки. Розроблено метод рубрикування текстових документів на основі онтологій. Цей метод базується на метриці. Для побудови такої метрики використано

розширену класичну структуру онтології. З цією метою у загальноприйнятий трьохелементний кортеж, який задає онтологію (множина понять, відношень та їх інтерпретація), ми додали дві скалярні величини (важливість понять і відношень), які використовуються для обчислення відстаней. Побудовано інтелектуальний агент, який здійснює рубрикування на основі розробленого методу. Розглянуто приклад функціонування такого агента. Отримані результати показують ефективність запропонованого методу.

Література

1. Андреев, А.М. Модели и методы автоматической классификации текстовых документов [Текст] / А.М. Андреев, Д.В. Березкин, В.В. Сюзев, В.И. Шабанов // Вестн. МГТУ. Сер. Приборостроение. – 2003. – №3. – С. 45–51.
2. Литвин, В.В. Мультиагентні системи підтримки прийняття рішень, що базуються на прецедентах та використовують адаптивні онтології [Текст] / В.В. Литвин // Радіоелектроніка, інформатика, управління. – 2009. – №2(21). – С. 120–126.
3. Круглов, В.В. Искусственные нейронные сети. Теория и практика [Текст] / В.В. Круглов, В.В. Борисов. – М.: Горячая линия – Телеком, 2001. – 256с.
4. Sowa, J. Conceptual graphs for a database interface / J.Sowa // IBM Journal of Research and Development. – Vol. 20. – № 4. – 1976. – P. 336–357.
5. Даревич, Р.Р. Оцінка подібності текстових документів на основі визначення інформаційної ваги елементів бази знань [Текст] / Р.Р. Даревич, Д.Г. Досин, В.В. Литвин, З.Т. Назарчук // Искусственный интеллект. – 2006. – № 3. – С. 500–509.
6. Свами, М. Графы, сети и алгоритмы [Текст] / М. Свами, К. Тхуласираман. – М.: Наука, 1984. – 512с.
7. Проектування інтелектуальних агентів прийняття рішень в просторі ознак з використанням онтологічного підходу [Текст] / В.В. Литвин, Р.Р. Даревич, Д.Г. Досин, Н.В. Шкутяк // Штучний інтелект. – 2010. – Т.2. – С. 100–104.
8. Інтелектуальні системи, базовані на онтологіях [Текст] / Д.Г. Досин, В.В. Литвин, Ю.В. Нікольський, В.В. Пасічник. – Львів: Цивілізація, 2009. – 414с.
9. Link Grammar Homepage [Електронний ресурс]. – Режим доступу: http://www.link.cs.cmu.edu/link/

Отримано 11.04.2011