УДК 004.78

В.Я. ЛЯШКЕВИЧ, О.П. МІРОШ

Чернівецький національний університет ім. Ю. Федьковича, Україна

ОЦІНКА ЕФЕКТИВНОСТІ РОБОТИ МЕТОДІВ КЛАСИФІКАЦІЇ ТЕКСТІВ

Представлено аналіз алгоритмів класифікації текстів, методів машинного навчання класифікаторів текстів, структуру словника інвертованого файлу, реалізацію тематико-зорієнтованої інформаційно-пошукової системи та результати дослідження методів класифікації текстів за показником залежності їх від розмірності простору ознак і обсягу навчальної множини документів.

пошук інформації, тематико-зорієнтовані інформаційно-пошукові системи, класифікація текстів, методи класифікації текстів

Актуальність

Бурхливий розвиток мережевих технологій призводить до значного збільшення об’ємів доступної інформації, яка є різнорідною, слабоструктурованою й надмірною. Необхідність ефективного використання цього величезного об’єму інформації обумовлює актуальність і значущість досліджень в області інформаційного пошуку.

Парадокс у розвитку мережевих пошукових систем полягає в тому, що їхнє технічне вдосконалювання в рамках традиційної парадигми неминуче призводить до лавиноподібного зростання баз даних, і відповідно, обсягів релевантних вибірок, які кінцевий користувач, у підсумку, не в змозі опрацювати. Сучасні технології дозволяють здійснювати витончені операції над даними, але чим ефективніше вони застосовуються, тим менш “придатним” виявляється результат. Схоже, технічний прогрес у цьому випадку не поліпшує, а погіршує ситуацію.

Постановка задачі

Сучасні інформаційно-пошукові системи (ІПС) первісно проектувалися для забезпечення релевантності вибірки в поєднанні з вимогою повноти пошуку, але саме в цьому і полягає їхній головний недолік. Неконтрольований рівень пертинентності вибірки при цьому різко знижує імовірність одержання користувачем саме тієї інформації, яка йому потрібна.

Причини надлишковості результатів стандартного інформаційного пошуку можуть бути розділені на дві якісно різні категорії: дублювання інформації та інформаційна невідповідність. Істотним є те, що приналежність документа до числа дублів носить цілком об’єктивний характер і може визначатися автоматично на підставі формальних критеріїв. Навпаки, інформаційна невідповідність породжує проблеми суб’єктивного характеру, тому що машина не в змозі визначити, чи відповідає зміст даного документа інформаційним потребам даного користувача.

Звідси, пошукові технології повинні бути розширені за рахунок застосування додаткових засобів, що дозволяють скоротити розрив між рівнями релевантності й пертинентності. Одним з перспективних з існуючих сьогодні напрямків, безсумнівно, є автоматичне групування результатів пошуку, тобто розбивка релевантної вибірки документів на класи. Вона вирішує проблему шляхом скорочення обсягів, представляючи користувачу документи тільки заданої теми.

Відбір документів за допомогою тематичної фільтрації дозволяє суттєво скоротити кількість документів, отриманих за допомогою пошуку за ключовими словами. Це призводить до зменшення часових затрат на пошук необхідної інформації. Ряд

досліджень [1, 2] показали, що класифікація результатів пошуку дозволяє істотно скоротити час пошуку потрібної інформації. Таким чином, введення додаткової класифікації на отримувані користувачем документи дозволяє підвищити зручність використання пошукової системи і дозволяє швидше орієнтуватися в отриманих результатах.

Процес пошуку розділяється на два етапи:

1. Відбір документів, відповідних запиту за ключовими словами. Даний етап повинен забезпечити високу повноту пошуку.

2. Уточнення результатів пошуку за допомогою класифікації результатів по темах. Цей етап дозволяє забезпечити високу точність результатів пошуку.

Для підвищення ефективності процесу пошуку інформації необхідно вирішити такі задачі:

1. Провести аналіз сучасних алгоритмів класифікації текстів та оцінку ефективності їх застосування.

2. Розробити архітектуру тематико-зорієнтованої ІПС для пошуку інформації за ключовими словами.

3. Запропонувати метод рішення задачі тематичного пошуку, який базується на комбінації пошуку за ключовими словами й тематичною фільтрацією з використанням класифікаторів текстів.

Аналіз сучасних алгоритмів класифікації текстів та методів машинного навчання класифікаторів тексту

Класифікацію текстів на природній мові називають рубрикацією, тому в подальшому ці терміни приймаються ідентичними. Класифікатори текстів підрозділяються на три основні класи: плоскі, ієрархічні й мережеві. Плоскі класифікатори складаються з двох рівнів. На першому рівні розміщується коренева рубрика, а на другому – дочірня. Як показано в [3], ієрархічні й мережеві класифікатори можуть бути представлені у вигляді композиції декількох плоских. Завдання класифікації визначається таким чином. Є множина об'єктів $T=\{t_i\}$, не обов'язково скінченна, і множина $C=\{c_i\}$ $i=1..N_c$, що складається з N_c класів об'єктів. Кожен клас c_i представлений деяким описом F_i, що має деяку внутрішню структуру. Процедура класифікації f об'єктів $t \in T$ полягає у виконанні перетворень над ними, після чого робиться висновок про відповідність t одній із структур F_i, що означає віднесення t до класу c_i або висновок про неможливість класифікації t. Стосовно текстової інформації, елементами множини T є електронні версії текстових документів.

Загальна модель плоского текстового рубрикатора може бути представлена алгебраїчною системою наступного вигляду

$$R=<T,C,F,R_c,f>,$$

де T – множина текстів, які необхідно класифікувати, C – множина класів-рубрик, F – множина описів, R_c – відношення на $C \times F$, f – операція рубрикації виду $T \to C$. Відношення R_c має властивість

$$\forall c_i \in C \exists! F_i \in F : (c_i, F_i) \in R_c,$$

тобто класу відповідає єдиний опис. Зворотна умова необов'язкова. Відображення f не має ніяких обмежень, так що можливі ситуації, коли $\exists t \in T : f(t) = C_t \subset C \wedge |C_t| > 1$, тобто деякий текст може бути віднесений до декількох класів одночасно.

Крім сформульованого завдання класифікації розглядають навчання класифікатора, під яким розуміється часткове або повне формування C, F, R_c і f на основі деяких апріорних даних [4].

На сьогодні існує безліч систем для забезпечення пошуку інформації, які містять об'ємні бази даних і знань. В основу таких систем, покладено композицію вербального й контекстного підходу до пошуку інформації. Вербальний пошук передбачає отримання набору документів, що відповідають заданому запиту за ключовими словами, але основний інтерес являє контекстна частина таких

систем, що реалізується за допомогою алгоритмів класифікації [5].

Класифікація текстових документів розглядається як один з можливих варіантів вирішення проблеми використання інформаційних ресурсів тому, що отримати повну інформацію по конкретній тематиці з накопичених баз даних і знань не так легко. Дослідження ж конкретної тематичної області вимагає затрат на безпосередній пошук і аналіз інформації, тому рішення приймаються на основі неповного уявлення про проблему [6].

Для вирішення завдання класифікації запропоновано багато методів із використанням автоматичних процедур, які можна розділити на два принципово різних класи: методи машинного навчання та методи, що базуються на знаннях (“інженерний підхід”). Принципова різниця між двома групами методів полягає в тому, що методи машинного навчання використовують математичні методи для виділення знань з навчальної колекції текстів, тоді як “інженерний підхід” використовує знання експерта, що грунтується на надбаному досвіді, зокрема, на великій колекції прочитаних раніше текстів. На жаль, проблемою інженерного підходу є висока трудомісткість створення системи автоматичної класифікації [7]. Класифікатори текстів можуть бути розділені залежно від способу представлення описів класів (внутрішня структура елементів множини F), а також від організації процедури класифікації f. На сьогодні, практичне застосування отримали такі:

Статистичні класифікатори, на основі імовірнісних методів. Найбільш відомим в даній групі є сімейство Байєсових. Загальною рисою для таких систем є процедура f, в основі якої лежить формула Байєса для умовної імовірності.

Класифікатори, що використовують нейромережні методи. Даний вид класифікаторів добре зарекомендував себе для розпізнавання зображень. В даній роботі досліджені можливості їх використання для опрацювання текстів.

Класифікатори, побудовані на функціях подібності. Характерною рисою даного методу є універсальність описів F, які з одного боку використовуються для представлення змісту рубрик, а з іншого – зміст аналізованих текстів.

Важливим етапом при вирішенні задачі класифікації текстів є вибір методу машинного навчання. Розглянемо найпоширеніші методи машинного навчання для задач класифікації текстів.

Метод Байєса. Даний метод базується на аналізі сумісних розподілів ознак документу й рубрик [8]. Апостеріорна імовірність приналежності документа рубриці обчислюється за формулою Байєса, що пов’язує апріорну імовірність з апостеріорною. На практиці існує два підходи до використання методу Байєса для класифікації текстів:

1. Для кожної рубрики окремо приймається рішення про приналежність документу до неї.

2. Приймається рішення про приналежність документа для всіх рубрик, а вибираються ті, для яких ця ймовірність буде максимальною.

Метод k-найближчих сусідів (k-nearest neighbours, k-NN), на відміну від інших, не вимагає фази навчання. Для того, щоб знайти рубрики, релевантні документу d, цей документ порівнюється зі всіма документами з навчальної вибірки. Для кожного документа e з навчальної вибірки, знаходиться відстань - косинус кута між векторами ознак.

Класифікатор Роше (Rocchio classifier). Згідно даного методу для кожної рубрики обчислюється зважений центроїд. Особливість метода – зважені центроїди можна швидко перерахувати при додаванні нових класифікованих прикладів. Ця особливість корисна при використанні адаптивної фільтрації, коли користувач поступово вказує системі, які документи вибрані правильно, а які ні.

Метод PrTFIDF (Probabilistic TF-IDF). Даний метод використовує відмінні від метода Байєса способи апроксимації ймовірності $P(c_i|d)$. Робиться припущення про те, що ознака w несе вичерпну

інформацію про документ *d* в цілому, так що приналежність документа до тієї або іншої рубрики не вносить додаткової інформації про документ.

Метод опорних векторів (Support Vector Machines, SVM) розроблений В. Вапником на основі принципу структурної мінімізації ризику – одночасного контролю кількості помилок класифікації на навчальній множині й "ступеня узагальнення" виявлених залежностей.

Метод із використанням імовірнісних нейронних мереж (ІНМ). Використання ІНМ вперше було представлено у двох публікаціях: "Probabilistic Neural Networks for Classification", 1988 рік та "Mapping or Associative Memory and Probabilistic Neural Networks", 1990 р. Згідно методу можна оцінити густину імовірності для кожного класу, порівняти між собою ймовірності приналежності до різних класів та обрати модель з параметрами, при яких густина ймовірності буде найбільшою. [9]

Реалізація тематико-зорієнтованої ІПС. Для дослідження методів класифікації текстів розроблено тематико-орієнтовану ІПС [10] (рис. 1). ІПС складається з діалогового компоненту (ДК), модуля морфологічного аналізу (МА), модуля формування запитів (ФЗ) та пошукового агента (ПА). Пошуковий агент в свою чергу складається з системи пошуку інформації (СПІ), бази даних (БД) знайдених документів, модуля індексації (МІ) та підсистеми класифікації (ПКл), до якої входять словник ознак, модуль формування вектору ознак тексту (МФВО) та модуль класифікації текстів.

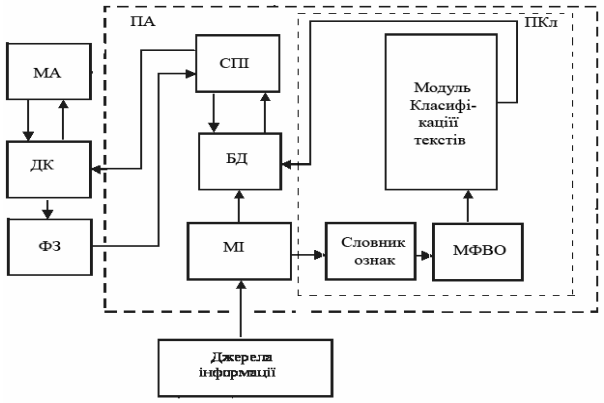

Рис. 1. Архітектура тематично-орієнтованої ІПС

Система працює наступним чином: користувач вводить через ДК ключові слова та назву тематичної області. Далі проводиться морфологічний аналіз ключових слів і формуються запити, котрі відправляються до ПА відповідної тематичної області. В ПА запит подається на СПІ, котрий взаємодіючи з базами даних знайдених документів формує і відправляє користувачу відповідь на запит.

МІ та ПКл заповнють та поновлюють БД знайдених документів. Це відбувається наступним чином: МІ знаходить всі слова в джерелі інформації і заносить їх в БД. При цьому, взаємодіючи з словником ознак, МФВО тексту формує вектор ознак даного джерела інформації й подає його на модуль класифікації текстів, який приписує джерелу відомості, котрі зберігає у БД знайдених документів.

Для функціонування розробленої ІПС удосконалено структуру даних: словника й таблиці документів, що дозволяє швидко опрацьовувати запити, а також ефективно використовувати пам'ять комп'ютера. В ІПС використовується структура даних, що розташовується в оперативній пам'яті комп'ютера. Ці дані по своїй суті є інвертованим файлом та складаються з декількох частин. Словник містить всі слова, що зустрілися в переглянутих документах, інформацію про те в яких саме документах вони зустрічаються, та позиції в них, тобто дуже багато інформації, яку помістити статичній пам'яті не можливо. Найбільш придатною для цього є динамічна пам'ять тому, що не можна наперед визначити скільки разів і в якій кількості документів будуть зустрічатися відповідні слова. Саме з цих причин масиви використовувати для організації словника не можливо. Найбільш придатними для цього є зв'язні списки, які відносяться до динамічних структур даних. Особливістю зв'язних списків є те, що для отримання доступу до елемента, потрібно переглянути всі попередні елементи. Оскільки словник містить велику кількість інформації, пошук в ньому необхідної інформації займе досить великий проміжок часу. Для вирішення цієї проблеми використано

метод прискорення доступу до даних – хешування. Для цього створюється хеш-таблиця. В цій таблиці є велике число (в даній ІПС 100000) хеш-ключів.

Для кожного слова вираховується свій хеш-ключ. Хеш-ключ не є унікальним для кожного слова, тобто один той самий хеш-ключ може належати багатьом словам. Тому хеш-ключ вказує на перше слово, яке вказує на друге, яке в свою чергу на наступне. Так, поки слово буде мати вказівник на NULL, тобто більше слів яким відповідає даний хеш-ключ немає. Вибір виду хеш-функції було здійснено емпіричним шляхом. Кожне слово, крім вказівника на наступне слово, має вказівник на структуру, що містить номер документа в таблиці опрацьованих гіпертекстових сторінок, в якому воно зустрічається.

Ця структура містить і вагу слова в документі, а також має два вказівника: на структуру наступного документа, в якому міститься дане слово та на позицію даного слова в цьому документі. Позиція вказує на наступну позицію даного слова на цій сторінці. За рахунок хешування скорочується область і час пошуку (рис. 2).

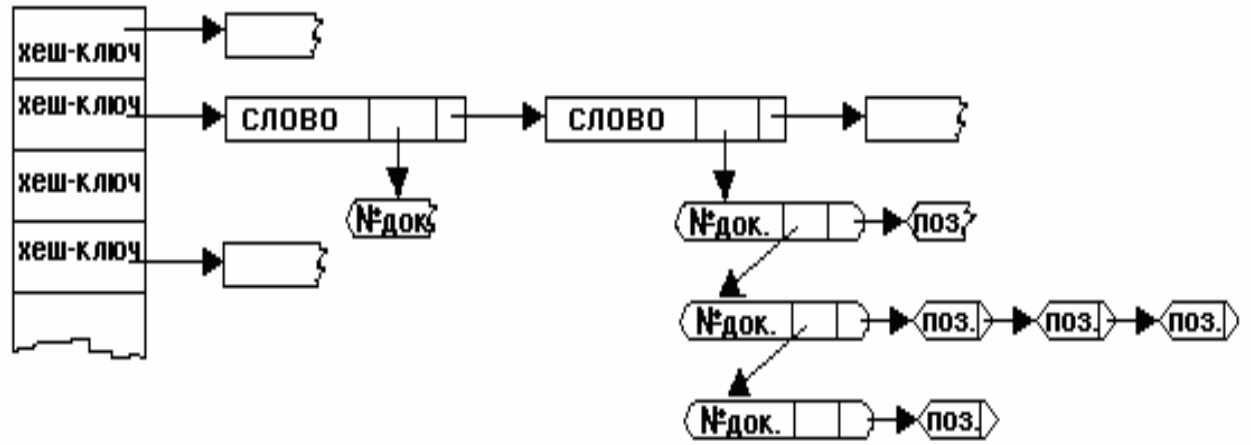

Рис. 2. Структура словника інвертованого файлу

Оскільки в одному документі є велика кількість слів, то немає сенсу кожному слову в словнику приписувати його назву та шлях (текстові змінні займають велику частину пам'яті).

Доцільніше назву та шлях записувати в іншу структуру даних (таблиця гіпертекстових сторінок), а кожному слову приписати його номер запису в цій таблиці.

Таблиця гіпертекстових сторінок – це масив записів (в даній ІПС 10000 елементів, цим і обмежується кількість документів, яка може бути опрацьована).

Кожен запис має 5 полів:

а) текстове – містить адресу та назву документу;

б) цілочисельне – містить загальну вагу документу;

в) цілочисельне – містить номер рубрики, якій приписаний документ;

г) логічне – вказує чи існує документ який був записаний раніше (якщо документ не існує, то на його місце можна записувати інший);

д) логічне – вказує чи опрацьовано документ при поточному пошуку.

Розроблена ІПС складається з кількох вузлів, схему взаємодії яких зображено на рис. 3.

Для покращення характиристик класифікації текстів за допомогою ІНМ запропонований такий шлях:

1) згрупувати нейрони прошарку прикладів по класах,

2) підібрати для кожного класу своє значення параметру σ.

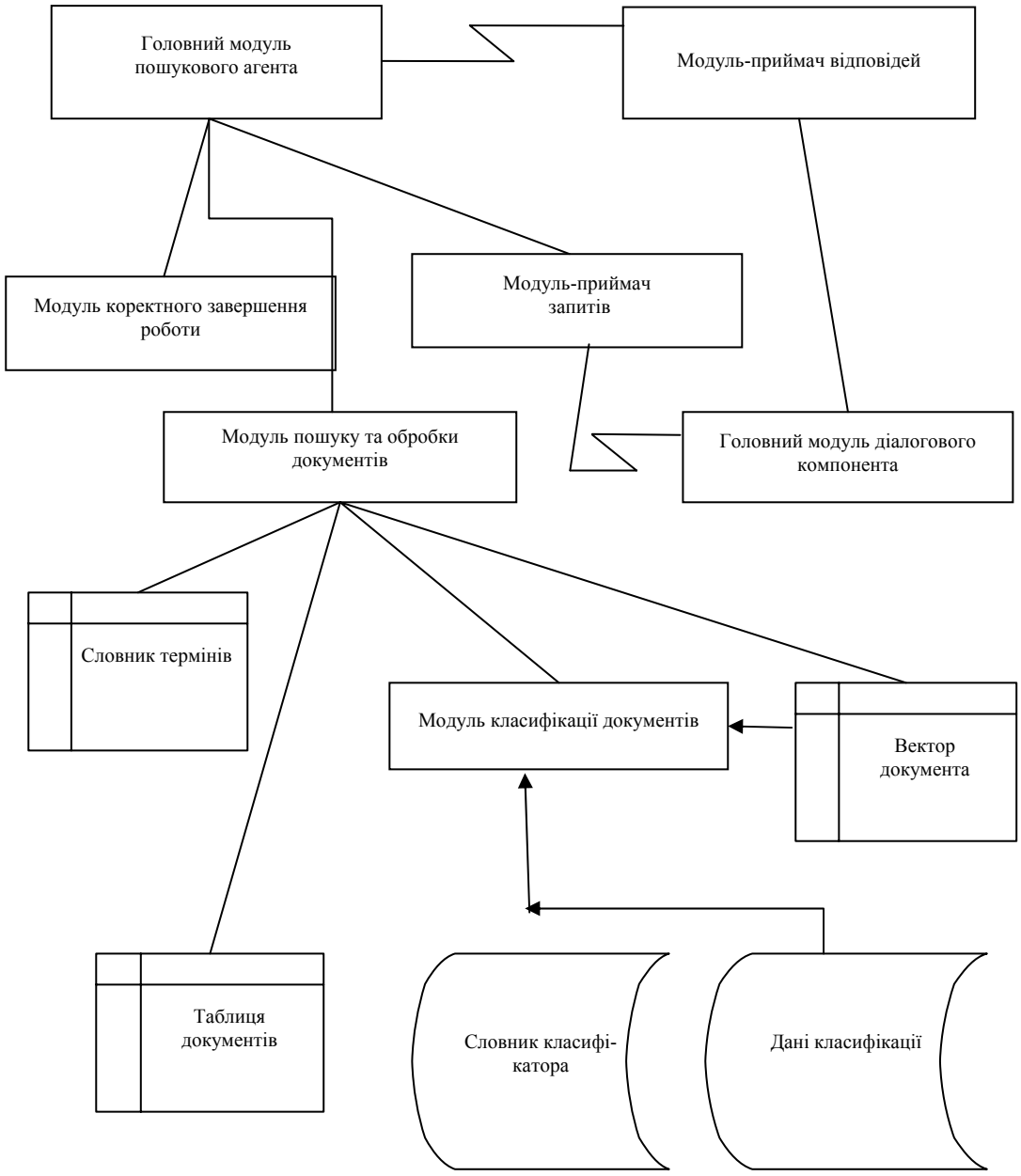

Рис. 3. Схема взаємодії програм ІПС

Наступні набори значень параметра σ дали найкращі результати: σ1=0,302; σ2=0,299; σ3=0,3; σ4=0,303; σ5=0,294 та σ1=0,302; σ2=0,3; σ3=0,301; σ4=0,303; σ5=0,294.

На рис. 4 ці набори позначені як набір1 та набір2 відповідно.

Для порівняння на рисунку зображено і характеристики з підібраним в дослідженнях σ = 0,3.

Як видно з рис. 4 значення σ1 = 0,302; σ2 = 0,3; σ3 = 0,301; σ4 = 0,303; σ5 = 0,294 дозволяють покращити характеристики класифікатора більш ніж на 2%. Дані значення і були використані при функціюванні ІНМ.

За допомогою розробленої ІПС були досліджені різні методи машинного навчання класифікатора текстів.

Результати дослідів, які представлені на рис. 5, 6, дозволили зробити наступні висновки:

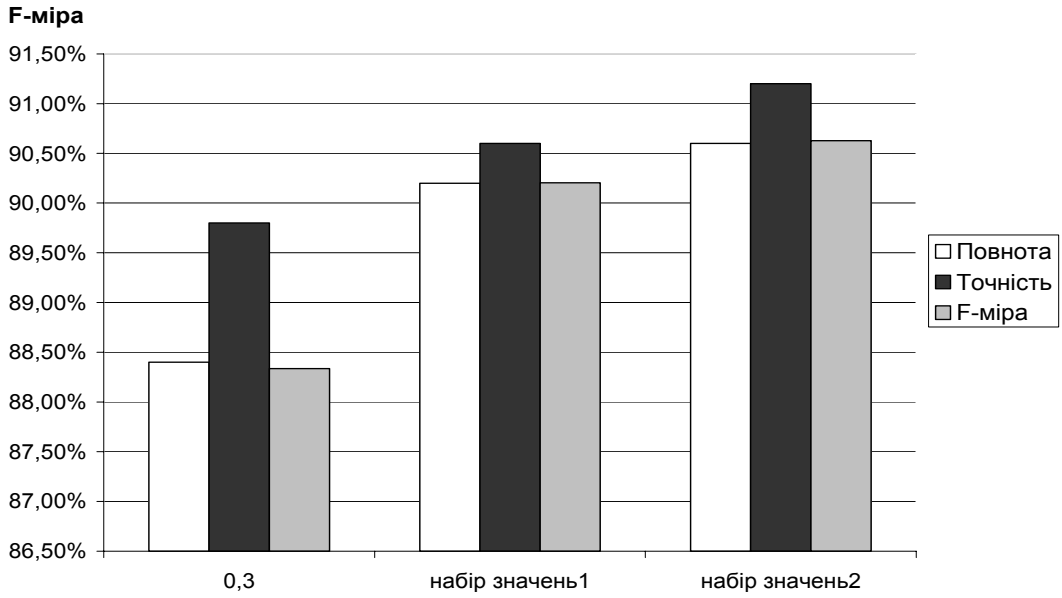

Рис. 4. Характеристики ІНМ з різними наборами значень σ

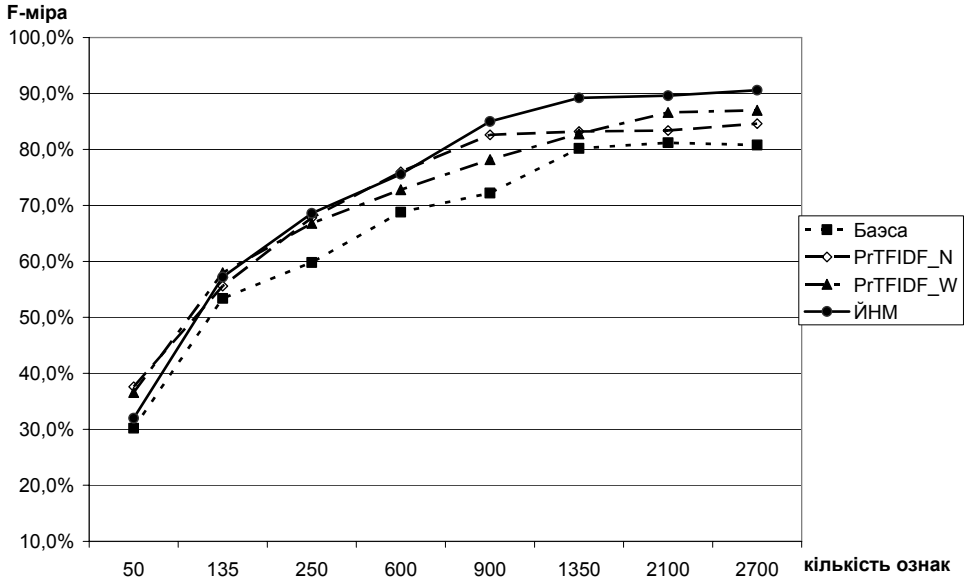

Рис. 5. Залежність класифікатора від розмірності простору ознак

Найгіршим методом в дослідах на залежність різних методів від розмірності простору ознак можна визнати метод класифікації Байєса. Він при всіх випробуваннях показав найгірші результати. Найкращим методом в цих дослідах можна визнати класифікатор на основі ІНМ.

Найстабільнішим при зменшені кількості прикладів виявився також класифікатор на основі ІНМ. Для всіх досліджених обсягів навчальної множини документів цей метод показав найкращі результати. У експериментах на якість класифікації документів з різних рубрик для вираження якості класифікації одним числом використовується F-міра. Ця оцінка для різних методів зображена на рис. 7. Класифікатор Байєса для всіх рубрик показав найгірші результати. Безперечно найкращим класифікатором текстів для всіх рубрик в цих дослідах є класифікатор на основі ІНМ.

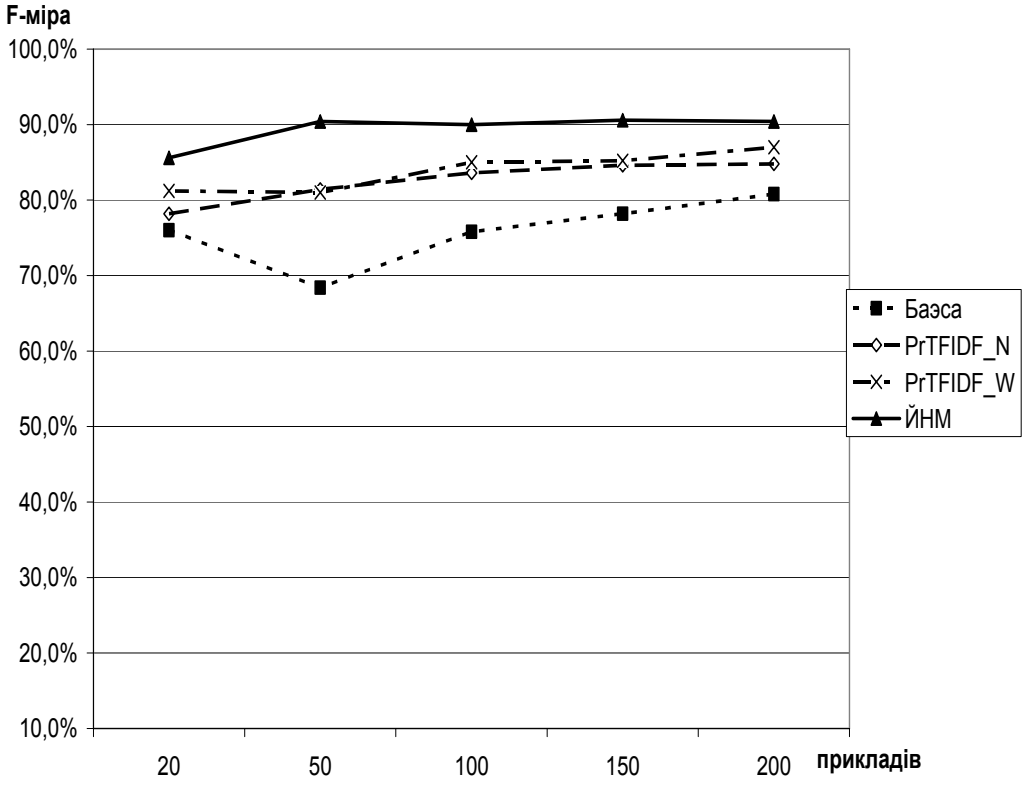

Рис. 6. Залежність класифікатора від обсягу навчальної множини документів

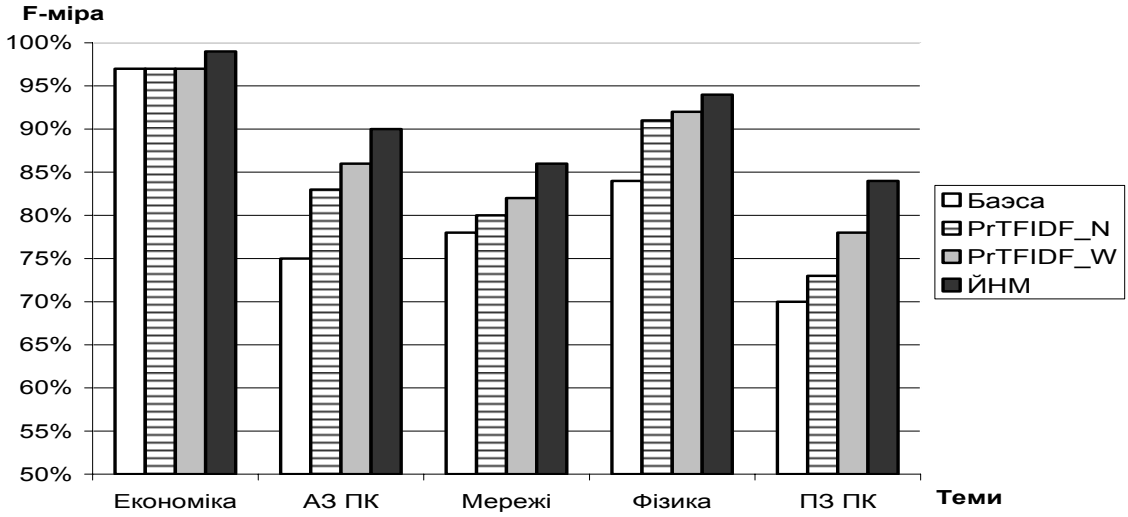

Рис. 7. Залежність класифікатора від обсягу навчальної множини документів

Висновки

1. Проведено аналіз сучасних алгоритмів класифікації текстів та методів машинного навчання класифікаторів тексту.

2. Розроблено архітектуру та реалізовано тематико-зорієнтовану ІПС, яка складається з двох частин: пошукового агента, що реалізує функційні можливості ІПС, та діалогового компоненту, що забезпечує діалог з користувачем.

3. Для покращення швидкодії методів класифікації було проведено зниження розмірності словника, початковий об’єм якого складав 193257 слів. Після морфологічного аналізу, виділення найбільш

корисних для класифікації термінів і групування цих термінів, був отриманий словник, що складається з 2700 груп слів й містить 62948 словоформ на двох мовах.

4. Використано представлення тексту у векторному вигляді, що дозволили значно оптимізувати ІНМ щодо використання ресурсів комп'ютера, що й полегшило реалізацію класифікатора текстів, а також значно збільшило швидкодію його роботи.

5. Для навчання класифікатора було використано 1000 прикладів документів з п'яти тем. Після навчання на вхід системи були подані контрольні приклади, які не брали участь в навчанні, третина від числа прикладів, використаних для навчання.

6. Програмно реалізовані й досліджені чотири методи автоматичної класифікації текстів. Серед них найкращий класифікатор текстів для всіх рубрик – на основі ймовірнісних нейронних мереж

Література

1. Chen H., Dumais S. Bringing Order to the Web: Automatically Categorizing Search // Proc. of ACM CHI 2000 Conference on Human Factors in Computing Systems. – 2000. – V. 1. – P. 145-152.

2. Driori O., Aron N. Using Documents Classification for Displaying Search Results List. – Technical Report No. 34/02 of the Leibniz Center for Research in Computer Science, Hebrew University of Jerusalem, Jerusalem, 2002. – P. 45-52.

3. Андреев А.М., Березкин Д.В., Сюзев В.В., Шабанов В.И. Модели и методы автоматической классификации текстовых документов // Вестник МГТУ. Сер. Приборостроение. – М.: МГТУ, 2003. – № 3. – С. 56-70.

4. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Труды V Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – С-Пб., 2003. – С. 25-36.

5. Максаков А.В. Исследование способов уменьшения набора характеристик в алгоритмах классификации текстов // Труды Всерос. науч. конф. "Методы и средства обработки информации". – М.: ВМиК МГУ, 2003. – С.234-240.

6. Кураленок И.Е., Некрестьянов И.С. Автоматическая классификация документов с использованием семантического анализа // Программирование. – 2000. – № 4. – С. 31-41.

7. Агеев М.С. Методы автоматической рубрикации текстов, основанных на машинном обучении и знаниях экспертов: Дисс. на соиск. уч. степ. к.ф.-м.н. – М.: МГУ, 2004. – 286 с.

8. Yang Y., Liu X. A re-examination of text categorization methods // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99). – 1999. – P. 42-49.

9. Ляшкевич В.Я., Мірош О.П. Дослідження можливостей апарату імовірнісних нейронних мереж для класифікації текстів // Вісник Хмельн. нац. університ. – 2007. – № 6. – С.99-105.

10. Lyashkevych V., Mirosh O. Subject-oriented informational retrieval system for investigation text classifier methods // Proc. of the 5th Int. Conf. "Microelectronics and Computer Science". – Sept. 19-21, 2007. – V. 1 – P. 454-457.

Надійшла до редакції 15.01.2008

Рецензент: д-р техн. наук, проф. В.С. Харченко, Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ», Харків.