УДК 004.93.14

І.Г. Оксанич, Д.М. Піскунов, Д.П. Черниш

Кременчуцький національний університет імені Михайла Остроградського, Кременчук

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ МАСИВУ ТЕКСТОВИХ ДОКУМЕНТІВ НА ОСНОВІ ТЕХНОЛОГІЇ TEXT MINING

Запропоновано підхід до оцінки тематичної близькості документів з використанням редукції простору ознак і на його основі розроблено алгоритм формування інформаційно-пошукових образів документів, що дозволяє підвищити якість і швидкість виконання автоматичної кластеризації документів. Пропонується модель для програмної системи, що дозволить виконати інтелектуальний аналіз текстового масиву авторефератів дисертацій, використовуючи методи технології Text Mining. Особливістю розробленої програмної системи є аналіз великих масивів текстових документів за рахунок вилучення прихованих нетривіальних знань.

Ключові слова: *інформаційно-пошукові образи, кластеризація, Text Mining, карти Кохонена.*

Вступ

Постановка проблеми. Інтелектуальний аналіз даних – область знань, яка відноситься до обробки даних, що вивчає пошук і опис прихованих, нетривіальних і практично корисних закономірностей у досліджуваних даних. До задач інтелектуального аналізу даних відноситься множина напрямків, такі як пошук документів в локальних і глобальних мережах, сортування, класифікація і кластеризація документів, автоматичне анотування та реферування, побудова тезаурусів і онтологій, системи автоматичного контролю, діалогові системи, системи, які навчаються, модифікація і поповнення баз знань, експертні системи і машинний переклад. Data Mining – дослідження і виявлення "машиною" (алгоритмами, засобами штучного інтелекту) в сирих даних прихованих знань, які раніше не були відомі, і є нетривіальними. практично корисними, доступними для інтерпретації людиною [1].

Розглянемо більш докладно властивості знань, що виявляються при застосуванні технології Data Mining. Знання повинні бути нові, оскільки зусилля витрачені на відкриття знань, які вже відомі користувачеві, не окупаються. Тому цінність представляють саме нові, раніше невідомі знання. Знання повинні бути нетривіальні, так як результати аналізу повинні відображати неочевидні, несподівані закономірності в даних, так звані приховані знання. Результати, які могли б бути отримані більш простими способами (наприклад, візуальним переглядом), не виправдовують застосування потужних методів Data Mining. Також, знання повинні бути практично корисними, бо повинні застосовуватися, в тому числі і на нових даних, з досить високим ступенем достовірності. Корисність полягає в тому, щоб ці знання могли принести певну вигоду при їх застосуванні і знання повинні бути доступні для розуміння людиною. Знайдені закономірності повинні бути логічно зрозумілі, в іншому випадку існує ймовірність, що вони є випадковими. Крім того, виявлені знання мають бути представлені в зрозумілому для людини вигляді [2].

Text Mining представляє собою множину методів обробки тексту, в результаті застосування яких з'являються нові, раніше не виявлені знання. Сьогодні це міждисциплінарна область, у якій використовуються базові технології Data Mining в поєднанні з техніками інших дослідницьких областей, таких як пошук інформації – Information Retrieval, вилучення інформації – Information Extraction, математична лінгвістика, класифікація – Classification, кластеризація – Clustering, створення онтологій – Ontology engineering і т. і.

Незважаючи на те, що в кожній з цих областей вирішуються свої специфічні прикладні задачі, часто буває досить складно провести чітку межу між Text Mining та іншою областю досліджень, оскільки всі вони мають справу з текстами а, отже, - загальні проблеми і підходи до їх вирішення. Так з Information Retrieval в Text Mining запозичені деякі алгоритми та методи обробки тексту. Різниця між цими областями полягає в кінцевій меті. В інформаційному пошуку метою є знайти документи, які хоча б частково співпадали з пошуковим запитом і, серед знайдених, відібрати ті, для яких збіг є найбільш повним. Методи Text Mining спрямовані на виявлення невідомих «фактів» і прихованих взаємозв'язків, які можуть виявитися по семантичним, лексичним і статистичним ознакам в множині текстів.

Інша область, з якої запозичені методи Text Mining – це вилучення інформації. Information Extraction відрізняється від Text Mining тим, що в цій області розглядаються способи вилучення специфічної інформації, структурованих даних, такі як імена людей, географічні назви, заголовки книг по заздалегідь заданим відносинам. У Text Mining наперед невідомо, яка саме інформація може бути виявлена. Методи Text Mining можна ефективно застосовувати при створенні баз даних та знань.

За допомогою методів Text Mining можна багато чого зробити для вирішення задач інтеграції даних. Вони дозволяють виявити загальні властивості в текстах, узятих з різних джерел. Критерії подібності в Text Mining можуть враховувати синтаксичну та семантич-

© І.Г. Оксанич, Д.М. Піскунов, Д.П. Черниш

ну інформацію, і застосовуватися не тільки до слів, але і до фраз, і до більш великих фрагментів тексту. Важливою проблемою в цій області залишається вибір найкращого критерію оцінки відстані між текстами - поняття, яке вперше було запроваджено в Text Mining.

Область досліджень, яка може значно виграти від використання Text Mining – це інформаційний пошук, тому що виконання запитів, вимагає перевірки семантичних відношень між текстами. Застосування Text Mining покращує точність інформаційних пошукових систем і зменшує кількість документів, що повертаються по одному запиту. Класифікація текстів також є частиною Text Mining, тому що наявність структурованого підходу полегшує пошук, перегляд та маркування документів.

Аналіз сучасного стану проблеми, що досліджується. Бурхливе зростання кількості електронних документів, що спостерігається в даний час, наочно показує, що традиційні механізми обробки електронних документів не спроможні впоратись з потребами користувачів. Ця тенденція помітна як в мережі Інтернет, так і у корпоративних мережах.

Таким чином, можна виділити основні проблеми, пов'язані зі збільшенням кількості інформації:

- швидкий ріст обсягу інформації, що міститься в Інтернеті, є причиною все більш і більш зростаючих труднощів пошуку необхідних документів та організації їх у вигляді структурованих за змістом сховищ;

- більшість технологій роботи з текстовими документами орієнтовані на організацію зручної роботи з інформацією для людини, але практично відсутні можливості для передачі смислового змісту тексту, тобто відсутнє семантичне індексування;

- для ефективного вирішення завдання пошуку необхідно розширити поняття традиційного документа: з документом необхідно пов'язати знання, що дозволяють інтерпретувати й обробляти дані, які зберігаються в цьому документі;

- неструктурована інформація становить значну частину сучасних електронних текстових документів.

При сучасному темпі зростання обсягів інформаційних масивів, неважко уявити, якими надмірно трудомісткими процесами будуть, як класифікація всього фонду електронних текстових документів вручну, так і його кластеризація. Допомогти у вирішенні даної проблеми здатні програмні засоби, які автоматично виконують інтелектуальну обробку даних. Останнім часом стало можливим втілення ідеї автоматичної класифікації або кластеризації документів по ряду причин. По-перше, мова йде про текстові документи, які можуть бути представлені у вигляді, придатному для автоматичного аналізу за допомогою програмних засобів. По-друге, на цей момент в науковому співтоваристві накопичився досить великий досвід дослідження і розробки таких систем. Причому інтерес до даної проблеми, не тільки не згасає, але в останні два десятиліття є підвищеним. Це в першу чергу викликано стрибком у розвитку програмно-апаратної бази, яка стала придатною для тестування розроблених раніше математичних методів інтелектуальної обробки текстів.

Кластерний аналіз займає одне з центральних місць серед методів аналізу даних і являє собою сукупність методів, підходів і процедур, розроблених для вирішення проблеми формування однорідних класів (кластерів) у довільній проблемній області. Ми проаналізуємо можливі методи для інтелектуальної обробки інформації з певної теми в різних електронних джерелах. Проблемна область, яка розглядається в даній роботі, являє собою великий масив авторефератів дисертацій, що робить неможливим його кластеризацію за допомогою експертів. Крім того, експертна розбивка авторефератів дисертацій на кластери може бути суб'єктивною і відображати лише думку конкретного експерта, що робить актуальною задачу розробки програмної системи для інтелектуальної обробки великих масивів текстових документів.

Матеріал і результати дослідження

Під автоматичною кластеризацією текстових документів розуміють процес класифікації колекції текстових документів, який базується тільки на аналізі та виявленні внутрішньої тематичної структури колекції без наявності апріорної інформації про неї, тобто при відсутності визначеного рубрикатора і множини документів-зразків. Класифікація документів з використанням алгоритмів кластеризації призводить до розбиття множини документів на однорідні, у відповідному розумінні, групи або кластери, шляхом автоматичного аналізу тематичної близькості між ними. Кластеризація текстів базується на гіпотезі: тісно пов'язані за змістом документи намагаються бути релевантними одним і тим же запитам, тобто документи, релевантні запиту, віддалені від тих, які не релевантні цьому запиту.

Задачу автоматичної кластеризації текстових документів у загальному вигляді можна сформулювати наступним чином: дано множину текстів на природній мові – колекція текстових документів. Передбачається, що існує множина тематичних груп (кластерів) , на які можна розбити дану колекцію документів. Тоді задача автоматичної кластеризації колекції текстових документів зводиться до пошуку невідомої множини таким чином, щоб підсумкова множина була оптимальною у відповідності з деяким критерієм якості розбиття документів.

Слід зазначити, що вихідними даними задачі кластеризації є не самі тексти на природній мові, а їх інформаційно-пошукові образи.

Інформаційно-пошуковий образ документа представляє собою багатовимірний вектор в евклідовому просторі ознак документа, що характеризує смисловий зміст вихідного документа. Ознаки документів автоматично витягуються з текстів відповідно до обраного способу представлення тексту, і в самому поширеному випадку є словами. Ознаки документів всієї колекції об'єднуються в загальну множину. Вектор ознак кожного документа має роз-

мірність N_p. Процес кластеризації базується на аналізі тематичної близькості документів, визначення якої полягає в припущенні, що геометрична близькість векторів документів в просторі ознак документів всієї колекції означає дійсну подібність предметних областей даних документів.

Оцінка тематичної близькості документів заснована на обчисленні деякої міри близькості. Часто використовуваними мірами близькості між векторами текстових документів у просторі їх ознак є косинусна міра, яка обчислює значення косинуса між двома векторами документів і міри близькості, засновані на вимірюванні відстані між векторами документів в багатовимірному просторі ознак документів [3].

Вихідними даними задачі автоматичної кластеризації документів є отриманий в результаті її вирішення набір кластерів, структура яких залежить від вибору алгоритму кластеризації та може належати до одного з наступних типів:

- плоский набір кластерів – множина кластерів, що відображає деяке число незалежних один від одного груп документів;

- ієрархічний набір кластерів – множина кластерів, елементам якої співставлені зв'язки ієрархічного типу, які відображають деревоподібну структуру груп документів, при якій кожен вузол дерева представляє групу, що містить всі документи її групнащадків;

- набір кластерів у вигляді графа – множина кластерів, елементам якого співставлені зв'язки довільного типу, що відображають деякі відносини між групами документів.

Кластеризація масиву авторефератів дисертацій буде складатися з наступних основних етапів обробки даних:

- формування інформаційно-пошукових образів текстових документів;

- формування множини кластерів інформаційно-пошукових образів.

Розроблюваний алгоритм формування образів документів заснований на статистичному підході до аналізу текстів на природній мові. Образ кожного документа пропонується формувати у вигляді багатовимірного вектора нормалізованих і зважених одиночних слів (ознак), що зустрічаються в тексті даного документа. Розмірність такого вектора буде дорівнювати кількості унікальних ознак у колекції документів.

Запропонований спосіб формування образів Φ_D складається з таких основних етапів:

$$\Phi_D = < \Phi_P, \Phi_{DP}, \Phi_R >,$$

де Φ_P – спосіб вилучення ознак із текстів документів; Φ_{DP} – спосіб відображення документів у простір їх ознак; Φ_R – алгоритм редукції простору ознак документів.

Спосіб вилучення ознак Φ_P полягає в послідовному виконанні наступних операцій: лексичний аналіз тексту (видалення розмітки, пунктуації, цифр, перетворення всіх букв у прописні і т. і.), видалення стопслів, тобто широковживаних слів, які не несуть самостійного сенсу, наприклад, прийменників, сполучників, часток і займенників; морфологічний аналіз.

Нами пропонується використання такого підходу до морфологічного аналізу, як виділення псевдоосновслів. В результаті даного аналізу слова з тексту приводяться до спеціального виду, і в подальшому, слова, що мають однаковий спеціальний вид (псевдооснову) розглядаються як одна ознака. В результаті вилучення ознак способом Φ_P вдається отримати N_P - розмірну множину ознак (псевдоосновслів) колекції документів P, яке також називають загальним словником ознак колекції документів.

Спосіб відображення документів у простір їх ознак Φ_{DP} заснований на процедурі зважування ознак. Зважування ознак документів пропонується виконувати за допомогою традиційної техніки tf*idf, яка є незалежною від наявності навчальної множини, враховує частоту входження терму, як в окремий документ, так і у всю колекцію в цілому.

Необхідність розробки алгоритму редукції простору ознак документів Φ_R обумовлена тим, що високорозмірні та розріджені вектори документів мають у просторі ознак недостатньо виразну орієнтацію для того, щоб автоматичні методи шляхом обчислення відстані між ними могли б зробити однозначний висновок про їхню спорідненість або відмінність. Для вирішення даної проблеми в сучасних інформаційно-пошукових системах застосовується примусове скорочення простору ознак за критерієм DF. Алгоритм примусової редукції за критерієм DF видаляє із загального словника ознак P колекції документів всі ті ознаки, документна частота яких вище порогового значення τ_{max}^{DF} та нижче порогового значення τ_{min}^{DF}.

Кластеризацію інформаційно-пошукових образів документів (авторефератів дисертацій) запропоновано виконувати у відповідності з підходом, заснованим на самоорганізуючих картах Кохонена [4]. Штучна нейронна мережа Кохонена або самоорганізуюча карта ознак (SOM) представляє собою двошарову мережу. Кожний нейрон першого (розподільного) слою з'єднаний з усіма нейронами другого (вихідного) слою, які розташовані у вигляді двовимірної решітки. Нейрони вихідного слою називаються кластерними елементами, їх кількість визначає максимальну кількість сегментів, на які система може розділити вихідні дані. Збільшуючи кількість нейронів другого слою можна збільшувати деталізацію результатів процесу кластеризації.

Система працює за принципом змагання – нейрони другого слою змагаються один з одним, перемагає той елемент-нейрон, чий вектор ваги ближче всього до вихідного вектора сигналів. За міру близькості двох векторів зазвичай береться евклідова відстань між ними. Таким чином, кожен вихідний вектор відноситься до деякого кластерного елементу.

Перш ніж мережа почне працювати її необхідно навчити на множині даних, яка буде піддана кластеризації. На кожному кроці навчання з вихідного набору даних випадково вибирається один вектор. Потім проводиться пошук нейрона вихідного слою, для якого відстань між його вектором ваги і вихідним вектором – мінімальна. За певним правилом здійснюється корегування ваги для нейрона-переможця і нейронів з його околу, яка задається відповідною функцією околу $h(t, j, m)$, де m – нейрон-переможець; j – нейрон вихідного слою, для якого обчислюється значення функції околу; t – параметр часу. Радіус околу повинен зменшуватися із збільшенням параметра часу. Цю проблему можна вирішити використанням функції Гаусса.

Кластером буде група векторів, відстань між якими всередині цієї групи менше, ніж відстань до сусідніх груп. Структура кластерів при використанні алгоритму самоорганізуючих карт Кохонена може бути відображена шляхом візуалізації відстані між опорними векторами (ваговими коефіцієнтами нейронів). При використанні цього методу найчастіше використовується уніфікована матриця відстаней (u-matrix), тобто обчислюється відстань між вектором ваги нейрона в сітці і його найближчими сусідами. Потім ці значення використовуються для визначення кольору, яким цей вузол буде відображатися. Зазвичай використовують градації сірого, причому, чим більше відстань, тим темніше відображається вузол. При такому використанні, вузлам з найбільшою відстанню між ними та сусідами відповідає чорний колір, а навколишнім вузлам – білий. Таким чином, розташовані поблизу кластери зі схожими кольорами утворюють більш глобальні кластери. Зазвичай в них розташовані близькі за ознаками документи.

На вхід нейронної мережі подається множина векторів документів у вигляді матриці розміром N на M, де N – кількість документів, що кластеризуються, а M – кількість унікальних термів у колекції документів, які кластеризуються. На перетині стовпчиків і рядків розташовуються ваги j-того терму в i-тому документі, обчислені за методом tf * idf.

Базовий алгоритм навчання мережі Кохонена виглядає наступним чином:

Крок 1. Ініціалізувати матрицю ваги малими випадковими значеннями (на відрізку [0, 1]).

Крок 2. Випадковим чином вибрати вектор з вихідної множини.

Крок 3. Для кожного вихідного нейрона j обчислити відстань між його вектором ваги w_i та вихідним вектором x:

$$d_j = \sqrt{\sum_{i=1}^{n} (w_{ij} - x_j)^2}$$

Крок 4. Знайти вихідний нейрон-переможець j_{min} з мінімальною відстанню $\min(d_j)$.

Крок 5. Для вихідного нейрона-переможця j_{min} та для його сусідів з околу оновлюються вектори ваги за правилом:

$$w_{ij}(t+1) = w_{ij}(t) + e(t) * h(t, j, m) * (x_i - w_{ij}(t)),$$

де $w_{ij}(t)$ – значення вагового коефіцієнта зв'язку вхідного нейрона i та вихідного нейрона j у момент часу t; $h(t, j, m)$ – значення функції околу з центральним нейроном вихідного слою m для нейрона вихідного слою j у момент часу t; $e(t)$ – коефіцієнт швидкості навчання в момент часу t; x_i – вихід нейрона першого слою з номером i.

Крок 6. Повторити кроки з кроку 2 для всіх елементів вихідної множини.

Цикл навчання триває до досягнення системою потрібного стану. В якості критеріїв зупинки процесу навчання можна використовувати наступні:

- топологічну впорядкованість карти ознак (матриці ваги);
- зміни ваги стають незначними;
- вихід мережі стабілізується, тобто вихідні вектори не переходять між кластерними елементами;
- досягнуто граничне значення помилки на карті;
- пройдено задану кількість епох.

Висновки

Таким чином, запропоновано підхід до оцінки тематичної близькості документів з використанням методу редукції простору ознак, які складають інформаційно-пошукові образи.

Запропоновано метод автоматичної кластеризації текстового масиву авторефератів дисертацій, який містить наступні кроки:

– формування інформаційно-пошукових образів документів, що дозволяє будувати образи текстових документів у вигляді векторів у просторі їх ознак;

– редукцію простору ознак документів, яка дозволяє підвищити представницьку здатність сформованих на передньому кроці образів документів;

– алгоритм кластеризації заснований на самоорганізуючих картах Кохонена.

Список літератури

1. Self organization of a massive document collection / T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela // IEEE Transactions on neural networks. – 2000. – Vol. 11, No. 3. – P. 574 – 585.

2. Freeman R.T. Adaptive topological tree structure for document organisation and visualisation / R.T. Freeman, H. Yin // Neural Networks. Elsevier Science Ltd. – 2004. – Vol. 17. – P. 1255–1271.

3. Губин М.В. Модели и методы представления текстового документа в системах информационного поиска / М.В. Губин // Научно-техническая информация. Сер. 1. – 2004. – №12. – С. 12-24.

4. Кохонен Т. Самоорганизующиеся карты. Пер. с англ. / Т. Кохонен; В.Н. Агеева/ − М.: БИНОМ. Лаборатория знаний, 2008. – 655 с.

Надійшла до редколегії 17.01.2013

Рецензент: д-р техн. наук, проф. В.Ф. Шостак, Кременчуцький національний університет імені Михайла Остроградського, Кременчук.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МАССИВА ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ ТЕХНОЛОГИИ TEXT MINING

И.Г. Оксанич, Д.М. Пискунов, Д.П. Черныш

Предложен подход оценки тематической близости документов с использованием редукции пространства признаков и на его основе разработан алгоритм формирования информационно-поисковых образов документов, что позволило повысить качество и скорость выполнения автоматической кластеризации документов. Предлагается модель для программной системы, которая позволит выполнить интеллектуальный анализ текстового массива авторефератов диссертаций, используя методы технологии Text Mining. Особенностью разработанной программной системы является анализ больших массивов текстовых документов за счет извлечения скрытых нетривиальных знаний.

Ключевые слова: *информационно-поисковые образы, кластеризация, Text Mining, карты Кохонена.*

INTELLIGENT ANALYSIS OF AN ARRAY OF TEXT DOCUMENTS BASED ON TEXT MINING TECHNOLOGIES

I.G. Oksanych, D.M. Piskunov, D.P. Chernysh

An approach to evaluating content close documents using reduction feature space, and on its basis algorithm is the information retrieval of images of documents that can improve the quality and speed of automatic clustering of documents. The model for software systems that allow me to perform intelligent analysis of text array dissertation abstracts using the methods of technology Text Mining. The peculiarity of the developed software system is to analyze large volumes of text documents by extracting hidden trivial knowledge.

Keywords*: information retrieval images, clustering, Text Mining, Kohonen maps.*