Довжина кодових комбінацій і потужність коду у всіх трьох випадках однакова, бо для кожної зі згаданих ІКВ S_n=31. Максимальна кількість помилок, які підлягають виявленню або виправленню під час реалізації першого з вказаних кодів, дорівнює відповідно 9 і 4. Кожен з двох інших випадків застосування ІКВ дає змогу побудувати код, який виявляє до 15 і виправляє до 7 помилок.

Ми бачимо, що останні два результати значно кращі від попереднього. Отже,

Висновки

У принципі будь-яка ІКВ може стати основою для побудови завадостійкого коду. Однак найефективнішим відповідають коди, утворені за допомогою ІКВ, параметри яких зв'язані співвідношенням (14).Показана можливість спрощеної побудови за допомогою ІКВ моделей монолітного коду розширеного класу завадостійких кодів, створення ефективних алгоритмів кодування і декодування інформації.

УДК 004.934

Н. Андрейчук, І. Волошиновська
Національний університет “Львівська політехніка”

СЕМАНТИЧНИЙ РОЗПОДІЛ ХАРАКТЕРИСТИЧНИХ ДІЄСЛІВ ДЛЯ НАУКОВИХ РОБІТ ПРИКЛАДНОГО ТА ФУНДАМЕНТАЛЬНОГО НАПРЯМКІВ

Описано результати аналізу головних компонент, що був застосований до корпусу наукових текстів. Проаналізовано внесок характеристичних наборів дієслів, що описують стан наукового поступу. Виділено основні компоненти разом із відповідними наборами дієслів, що відповідають за прогрес у прикладних та фундаментальних науках. Розглянута можливість застосування характеристичних слів в інформаційному пошуку.

This paper reports the results of principal component analysis applied to the corpus of scientific texts. The contribution of the characteristic set of verbs, revealing the stage of scientific advance, is analyzed. The main components attributed to the progress in applied and fundamental science are extracted together with the respective verb sets. The possibility of characteristic word application in the information retrieval is discussed.

1. Вступ

Для організації системи інформаційного пошуку постає необхідність не лише розроблення механізмів опрацювання кількісних характеристик текстів, але й побудови алгоритмів виявлення латентних семантичних характеристик текстового корпусу. Опрацювання робіт попередників у напрямку запланованих та споріднених досліджень є необхідним за умов високого темпу розвитку сучасних технологій в природничих науках. Очевидним стає завдання правильного подання лінгвістичних знань для забезпечення успішної роботи комп’ютерних систем пошуку. Інформація, яку подає людина, є структурована, логічна. Ту саму інформацію системи розпізнавання сприймають на початкових етапах аналізу як набір незалежних змінних. Тому процедури: і) виділення основних компонент, що характеризують вхідні фрагменти корпусу текстів; іі) виявлення кореляції між основними характеристиками, треба розглядати як основні етапи відтворення початкової інформації. Наукова стаття, до деякої міри, є простою з погляду оброблення та пошуку інформації, оскільки вона має чітко окреслену структуру та характерні складові елементи [1]. Проте інформація, закладена у таких основних розділах, як назва, ключові слова, вступ, висновки, не завжди є достатньою для визначення відповідності інформації, викладеної в статті, до сформованого запиту. Підбір документів згідно із сформованим запитом можна подати як оцінку відповідності набору термінів запиту до характерних термінів документа. Однак для детальнішого

аналізу відібраних згідно із запитом документів необхідно виробляти критерії подальшої семантичної оцінки тексту. Таку оцінку запропоновано виконувати, ґрунтуючись на лексичних множинах, що характерні для ключових ознак текстів [2].

2. Формування семантичної векторно-просторової моделі текстового корпусу

В роботі [2] показано можливість визначення етапу довершеності наукових досліджень. Для розв'язання цієї задачі були сформовані характеристичні набори дієслів, на основі частотного аналізу лексики наукових статей, опублікованих в журналі "Physical Review B" (Тематика: Фізика твердого тіла та матеріалознавство) [3]. Сформовані набори дієслів ({*expect, ..., suggest*}{очікувати, ..., припускати}; {*achieve, ..., develop*}{досягати, ..., розробляти}) відображають характерні ознаки для початкових та завершальних етапів наукової роботи і використовуються у цій роботі як лексичний базис для побудови векторно-просторової семантичної моделі корпусу наукових робіт міжнародного дослідницького центру DESY (Deutsches Elektronen-Synchrotron, Німеччина) [4].

2.1. Аналіз головних компонент

Для встановлення латентно-семантичних зв'язків між елементами текстового корпусу та лексичного базису (набору характеристичних дієслів) сформовано двовимірну матрицю даних **A** розмірністю 63x28, елемент a_{ij} якої відображає частоту появи *j*-го характеристичного дієслова в *i*-му тексті. Сформована у такий спосіб матриця була використана як вхідні дані для аналізу головних компонент (principal component analysis, PCA) [5]. Під час такого аналізу матриця **A** подається сумою векторів $\mathbf{t}_i$, $\mathbf{p}_i$ та залишкової матриці **E**:

$$\mathbf{A} = \mathbf{t}_1\mathbf{p}^{\mathrm{T}}{}_1 + \mathbf{t}_2\mathbf{p}^{\mathrm{T}}{}_2 + ... + \mathbf{t}_k\mathbf{p}^{\mathrm{T}}{}_k + \mathbf{E}. \quad (1)$$

Значення k дорівнює або меншим найменшого значення розмірності матриці **A**. В нашому випадку $k \leq 28$. Вектори $\mathbf{t}_i$ відображають співвідношення зразків (в нашому випадку це є тексти), а вектори $\mathbf{p}_i$ – співвідношення змінних (слів). Вектори $\mathbf{p}_i$ знаходять як власні вектори коваріаційної матриці, тобто для кожного $\mathbf{p}_i$

$$\mathrm{cov}(\mathbf{A})\mathbf{p}_i = \lambda_i\mathbf{p}_i\,, \quad (2)$$

де λ_i є власними значеннями, що відповідають власним векторам $\mathbf{p}_i$**.** Коваріаційна матриця формується на основі матриці даних **A** за правилом:

$$\mathrm{cov}(\mathbf{A}) = (\mathbf{A}^{\mathrm{T}}\mathbf{A})/(m\text{-}1)\,, \quad (3)$$

де значення m визначається кількістю зразків (в нашому випадку кількість текстів m=63).

Для вхідної матриці даних **A** та пари векторів $\mathbf{t}_i$, $\mathbf{p}_i$ виконується умова:

$$\mathbf{A}\mathbf{p}_i = \mathbf{t}_i\,, \quad (4)$$

тобто вектори $\mathbf{t}_i$ є проекціями вхідних даних на напрямки $\mathbf{p}_i$.

Власні значення λ_i є прямо пропорційними до дисперсії (змісту) вхідних даних, що описуються відповідною парою векторів $\mathbf{t}_i$, $\mathbf{p}_i$. Послідовність λ_i є спадною і, здебільшого, вхідні дані описуються (без значної втрати інформації) кількома першими парами $\mathbf{t}_i$, $\mathbf{p}_i$, кількість яких є значно меншою від кількості вхідних змінних.

Аналіз головних компонент допомагає виділити приховані характеристики у великих масивах даних (текстах) та проаналізувати зв'язки (семантичні), наявні у досліджуваній системі (корпусі текстів). В результаті аналізу були виділені 24 дієслова (змінні), які найбільшою мірою відображали семантичні взаємозв'язки в межах текстового корпусу, формуючи лексичну основу для побудови семантичної моделі. Виявлені найвагоміші чотири головні компоненти (principal component, *PC*-1, ..., *PC*-4) описують близько 80% дисперсії (змісту) вхідних даних і були взяті до розгляду при побудові моделі. Головні компоненти, що описували лише незначну дисперсію вхідних даних, не брались до розгляду, щоб запобігти приросту похибки. Найінформативнішим виявилось відображення внеску змінних (дієслів) в першу і другу головні компоненти (рис. 1). Взагалі, для головної компоненти не притаманно мати чітко визначене семантичне значення. Здебільшого виявити приховане семантичне значення головної компоненти доволі складно через багатозначність множини змінних (слів), що дають найбільший внесок у формування відповідної головної компоненти. У розглянутому випадку однозначне семантичне значення можна чітко приписати

лише першим двом головним компонентам з чотирьох компонент, що формують модель. Перші дві основні компоненти охоплюють близько 50% дисперсії (змісту) вхідних даних.

2.2. Семантичний аналіз

Семантична сегментація змінних [6] була виконана для встановлення змісту (семантичного значення) першої та другої основних компонент. Змінні, що дають найвагоміший внесок у *PC*-1, є більше віддаленими ліворуч та праворуч від нуля на поданому графіку (рис. 1). Змінні з протилежними знаками формують антикореляційні набори в межах проаналізованого корпусу текстів. Проекції змінних на від'ємну сторону осі *PC*-1 формують послідовність дієслів {*propose, …, implement, develop*} {пропонувати, …, втілювати, розробляти}, що чітко відображає поступ у прикладних науках [2]. Проекції дієслів на додатну сторону осі *PC*-1 {*suggest, …, accomplish, confirm, expect, …, presume, elaborate, …, suspect*} {припускати, …, завершувати, очікувати, …, припускати, розробляти, …, допускати} не відображають чіткої послідовності розвитку. Проте ці дієслова описують передбачення та припущення, які є характерними для фундаментальних наук і можуть бути наявними на початкових та подальших етапах наукових робіт.

На графіку (рис. 1) можна чітко побачити розподіл змінних (дієслів) вздовж двох ортогональних векторів **a** та **f**. Змінні, що приписуються двом основним виділеним напрямкам, позначені на рис. 1 заповненими кружечками та трикутниками і відповідають за розвиток у прикладних та фундаментальних науках, відповідно. Змінні, що стосуються прикладної галузі науки, розташовані в межах другої та третьої чверті запропонованої координатної площини, тоді як змінні фундаментальної галузі охоплюють її першу та четверту четверті. Спільними змінними, що формують основу для розвитку фундаментальної та прикладної галузі, є *suggest* та *predict* (припускати, передбачати) – дієслова, що відображають початковий етап розвитку (позначені на графіку порожніми квадратами, рис. 1). На початковому етапі роботу неможливо віднести до фундаментальних або прикладних наук. Більшість праць беруть свій початок з фундаментальних наук. Тому спільна основа для фундаментальних та прикладних задач зсунута у третю четверть, де домінують змінні, що відповідають за фундаментальні напрямки.

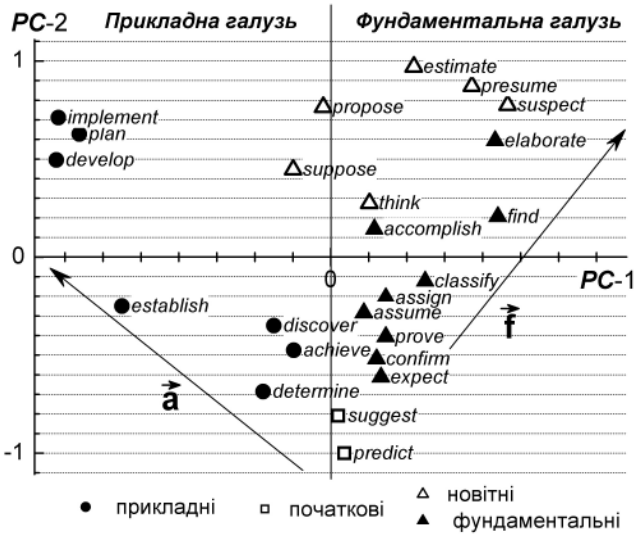

Рис. 1. Графік внеску змінних в головні компоненти PC-1 та PC-2

Відстань між змінними, що відповідають за прикладні та фундаментальні роботи, збільшується у першій та другій чвертях, що відображає етапи розвитку роботи. Отже, друга основна компонента (*PC*-2) описується змінними, що відповідають за стан розвитку наукового дослідження. Змінні, що відповідають за першу головну компоненту (*PC*-1), розподіляють наукові праці на прикладні та фундаментальні роботи. Виняток становлять змінні {*think, suppose, propose, suspect, presume, estimate*}, які не володіють чіткими семантичними зв'язками із станом розвитку наукової праці (позначені порожніми трикутниками на рис. 1). З іншого боку, дієслова цієї групи описують

такі процеси, як обдумування, припущення, сумніви, оцінку, які притаманні для подальшого розвитку фундаментальних наук. Як видно з графіка, для фундаментальних робіт на етапі завершення існує висока імовірність формування на їхній основі прикладних галузей. Група змінних {*think, suppose, propose*}{думати, припускати, пропонувати}, виявлена для фундаментальних досліджень на проміжному етапі, разом із групою слів {*classify, accomplish, find*}{класифікувати, завершувати, знаходити} формує кластер змінних, що розташовані вздовж вектора **a**. Отже, набір змінних {*classify, accomplish, find, think, suppose, propose*} можна віднести до початкового рівня нової прикладної роботи, що бере свій початок з проміжного етапу фундаментальних робіт.

У верхній частині першої чверті розташована група слів {*suspect, presume, estimate*}, що відповідає за найвищий рівень наукового розвитку. Ця група слів разом із словом {*elaborate*} відповідає за нові роботи прикладних наук, де можлива оцінка практичних досягнень. Отже, набір змінних{*think, suppose, propose, suspect, presume, estimate*}, що зображені прозорими трикутниками на рис. 1, формує семантичну групу характеристичних дієслів, які відповідають за описання подальшого розвитку у фундаментальних науках і є характерними для початкових стадій нових прикладних досліджень. Отже, побудовані вектори **a** та **b** відображають напрямки розвитку прикладних та фундаментальних наук, відповідно, у лексичному просторі.

3. Висновки

Застосування аналізу основних компонент для оброблення корпусу текстів наукового спрямування виявилось перспективним для формування характеристичних наборів слів, які зможуть полегшити процеси інформаційного пошуку.

Аналіз розподілу дієслів, що відповідають за етапи розвитку наукових досліджень, наведений у моделі чотирьох головних компонент, яка охоплює близько 80% змісту вхідних даних. Семантична сегментація змінних на графіку внеску змінних в головні компоненти показала, що перша головна компонента відповідає за розподіл робіт на прикладну та фундаментальну галузі. Друга головна компонента формується змінними, які описують етапи наукового поступу.

Досліджувані дієслова були прокласифіковані та чітко віднесені до: і) спільного початкового етапу для фундаментальних та прикладних наук; іі) прогресу прикладних наук; ііі) поступу у фундаментальних дослідженнях; іv) зародження нових прикладних розробок під час фундаментальних досліджень. Використану у роботі модель чотирьох головних компонент можна застосовувати для оцінки поступу наукових досліджень, а також їхньої приналежності до прикладної чи фундаментальної науки.

1. Hjorland B. Information retrieval, text composition and semantics // Knowledge Organization. – 1998. Vol. 25, № 1,2. – P. 16-31. 2. Андрейчук Н.І., Волошиновська І.А. Дієслова, що визначають глибину опрацювання предмета дослідження та їх статистичні характеристики у науково-технічних текстах // Лексикографічний бюлетень. – 2006. – № 13. – С. 170–174. 3. Physical Review B // <http://prb.aps.org>. – 2005. 4. Publications of the DESY // <http://msk.desy.de/Publikationen>. – 2006. 5. Jackson J.E. Principal Components and Factor Analysis: Part 1-Principal Components // Journal of Quality Technology. – 1981. Vol. 13. – P. 201–213. 6. Jackson P., Moulinier I. Text Retrieval, Extraction & Categorization // Natural Language Processing for Online Applications. – Amsterdam: John Benjamins Publishing Co., 2002. – 226 pp.

УДК 004.032.26:004.048

А. Дорошенко
Національний університет ”Львівська політехніка”,
кафедра автоматизованих систем управління

МЕТОДИ КЛАСИФІКАЦІЇ НА ОСНОВІ МОДЕЛІ ГЕОМЕТРИЧНИХ ПЕРЕТВОРЕНЬ ДЛЯ ЗАВДАНЬ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Проаналізовано особливості постановки та підходи до розв’язку задач класифікації для випадків великорозмірних завдань інтелектуального аналізу даних. Подано основи розроблених методів класифікації, вдосконалених завдяки використанню оптимізації відпалом металу.

The article analyses the features of the target setting and the approach to solving a problem of classification task for Data Mining tasks where data are high-dimensional. Essential principles of the methods of classification on the base of neural networks are proposed. This methods of classification are improved by simulated annealing algorithm.

Вступ

Розвиток сучасних інформаційних технологій надає можливість створювати сховища даних, що містять величезні обсяги даних з усіх сфер людської діяльності, від повсякденних та ділових (дані про транзакції в супермаркетах, записи про використання кредитних карток, інформація про телефонні дзвінки та урядову статистику) до наукових (такі, як зображення астрономічних тіл, база даних молекул чи медичних записів). Однак разом із збільшенням розміру сховищ даних ускладнюється можливість їх аналізу, який є головною метою накопичення даних. Для точного й оперативного аналізу даних виникли нові, відмінні від традиційно статистичних, методи, об’єднані під назвою видобуток даних, або інтелектуальний аналіз даних.

Постановка задачі

Більшість відомих методів класифікації не є адаптованими для задач інтелектуального аналізу даних, а саме до таких основних характерних особливостей завдань видобутку даних, як: об'ємність завдань та велика розмірність даних, виродженість задач, суперечливість, неповнота та неоднорідність даних.

Коротко розглянемо кожну з цих особливостей окремо:

1. Об'ємність завдань.

Якість інтелектуального аналізу даних певною мірою залежить від розмірів сховища даних, що обробляється. Відповідно для розв'язання задач видобутку даних із максимальною точністю обробляють вибірки великого обсягу та розмірності. Однак, це не лише підвищує якість видобутку даних, але й ускладнює його процес, робить його обчислювально складнішим.

2. Виродженість задач

Оскільки вхідні дані в задачах інтелектуального аналізу даних, як правило, є корельованими між собою, то в багатьох випадках такі задачі є виродженими.

Крім того, системи рівнянь, що описують залежність між вхідними та вихідними даними, часто є погано зумовленими (або надчутливими), тобто навіть незначна зміна коефіцієнтів задачі здатні призвести до великих змін її розв’язку. Відповідно, погано зумовлена задача не може бути розв’язана класичними методами (методом найменших квадратів тощо). Для розв’язання вироджених задач можливе застосування методу сингулярної декомпозиції, однак цей метод є дуже об’ємним та обчислювально складним. Наприклад, порівняно із QR-факторизацією, сингулярна декомпозиція