УДК 004.82 (045)

А.І. ВАВІЛЕНКОВА, канд. техн. наук, доц., НАУ, Київ

АНАЛІЗ МЕТОДІВ ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ

У статті надано характеристику основних методів *Data Mining,* виділено їх переваги та недоліки. В результаті аналізу виявлено, що жоден з описаних методів не здатен вилучати знання з інформації. Продемонстровано роботу методу резолюцій Робінсона для порівняння двох простих речень. Запропоновано алгоритм порівняння логіко-лінгвістичних моделей текстової інформації за змістом. Бібліогр.: 8 назв.

Ключові слова: методи *Data Mining*, вилучення знань, метод резолюцій, логіко-лінгвістичні моделі, текстова інформація.

Постановка проблеми і аналіз літератури. Засоби сучасних ЕОМ, що використовуються для обробки електронних текстів, дозволяють задавати різні обмеження на шукані комбінації слів в тексті, визначаючи обов'язковість або необов'язковість, допустиму відстань між словами та порядок їх знаходження в тексті. Це дає можливість проводити аналіз слова у всіх граматичних формах, точно і повно описуючи можливі способи представлення необхідного змісту в тексті. Для підвищення точності аналізу текстів розробляються методи попередньої лінгвістичної обробки, що вимагає, по-перше, значних обчислювальних витрат для лінгвістичного аналізу індексованої колекції текстів, по-друге, розробки спеціалізованої пошукової машини.

Автоматизоване вилучення знань з тексту є однією з основних задач штучного інтелекту і безпосередньо пов'язане з розумінням текстів на природній мові.

Задачу автоматизованої аналітичної обробки текстової інформації намагаються вирішити багато іноземних та вітчизняних вчених. Зокрема, ще у 1979 році Кузін Н.Т. [1] описав методи частотної оброки текстової інформації, які згодом були удосконалені в роботах Бродера А. [2] та Ланде Д.В. [3]. У своєму навчальному посібнику Барсегян А.А. та Купріянов М.С. [4] узагальнили дані щодо сучасних методів автоматичного аналізу *Data Mining* і *Text Mining*. Проте жоден з описаних методів не забезпечує вилучення з текстової інформації знань.

Мета статті – проаналізувати основні методи автоматичного аналізу текстової інформації, виявити їх переваги та недоліки для задачі вилучення знань з природної мови; здійснити порівняльний аналіз простих речень природної мови за допомогою методу резолюцій

© А.І. Вавіленкова, 2013

Робінсона; описати основні кроки алгоритму порівняльного аналізу текстової інформації, представленої у вигляді логіко-лінгвістичних моделей.

Засоби обробки текстової інформації. Для вилучення знань з текстової інформації використовуються різноманітні методи автоматичного аналізу *Data Mining*. Такі методи використовують алгоритми та засоби штучного інтелекту для дослідження і вилучення з великих об'ємів інформації знань, які будуть практично корисні та доступні для інтерпретації людиною [5]. До основних методів *Data Mining* належать класифікація, кластеризація, регресія, пошук асоціативних правил, анотування та автореферування.

Задача ***класифікації*** зводиться до визначення класу об'єкту за його характеристиками, при чому множина класів задається завчасно. ***Класифікація*** – використовує статистичні кореляції для побудови правил розміщення документів у наперед заданій категорії; задача класифікації – це задача розпізнавання, коли система відносить новий об'єкт до тієї чи іншої категорії. В *Data Mining* задачу класифікації розглядають як визначення значення одного з параметрів об'єкту на основі значення інших параметрів [4].

Задача ***регресії*** подібна до задачі класифікації і дозволяє визначити за відомими характеристиками об'єкту значення деякого його параметру. Тут значенням параметру є не кінцева множина класів, а множина дійсних чисел.

Класифікація та регресія передбачають здійснення двох обов'язкових етапів. Перший етап – виділення набору об'єктів, для яких відомі значення залежних і незалежних змінних. На основі отриманого набору будується модель визначення значення залежної змінної (функція класифікації або регресії). На другому етапі побудовану модель застосовують до об'єктів, які аналізуються. Недоліком класифікації та регресії є те, що розробник системи повинен фіксувати кількість класів та характеристик, за якими буде проводитись дослідження. Це означає, що якщо система не виявить ознаки або класу, до якого можна віднести, наприклад, текстовий, документ, він не буде коректно оброблений.

Анотування – це процес створення коротких повідомлень про електронний текст, які дозволяють робити висновки щодо доцільності його докладного вивчення [6]. Сучасні системи аналітичної обробки текстової інформації володіють засобами автоматичного складання анотацій, при цьому існує два підходи до вирішення цієї проблеми.

У першому підході програма-анотатор вилучає з першоджерела невелику кількість фрагментів, у яких найбільш повно представлено

зміст документу. При другому підході анотація представляє собою синтезований документ у вигляді короткого змісту. Анотація, сформована відповідно з першим підходом, якісно поступається анотації, одержаній при синтезі. Для підвищення якості анотування необхідно вирішити проблему орієнтування на вузьку предметну область. Тоді у такому процесі необхідна участь людини.

Метод анотування тексту довільної структури передбачає:

1. Формування множини анотованих фрагментів, які є цілими реченнями даного тексту, містять у своєму складі дієслово або короткий прикметник, і не є питальним чи окличним реченням.

2. Створення таблиці всіх можливих пар основних тематичних вузлів (тут використовується система продукцій для встановлення характеристик структурних одиниць тексту, описана раніше).

3. Відбір таких речень, які містять декілька різних тематичних вузлів, що не зустрічалися раніше у тексті.

Здійснення автоматичної анотації є прикладною задачею, що вирішується перед тим, як інформація із заданого тексту потрапить до пошукового серверу.

Автоматичне реферування – представляє собою створення коротких викладів матеріалів, анотацій, дайджестів, тобто вилучення найбільш важливих відомостей з одного або декількох документів і генерація на їх основі лаконічних та інформаційно-ємних звітів. На сьогодні існує два основних напрямки автореферування: квазіреферування (засноване на екстрагуванні фрагментів документів, тобто виділенні найбільш інформативних фраз і формування з них квазірефератів) і коротке викладення змісту первинних документів (дайджести) [3].

Автоматичне реферування та анотування використовуються в основному для економії часу користувачам, створення каталогів інформаційних ресурсів, використання словників-тезаурусів загального та спеціального призначення. Застосовується автоматичне реферування та анотування в корпоративних системах документообігу, пошукових машинах та каталогах ресурсів Інтернет, автоматизованих інформаційно-бібліотечних системах, каналах зв'язку, службах розсилки новин і т.д.

Пошук асоціативних правил представляє собою метод пошуку часткових залежностей між об'єктами та суб'єктами. Знайдені залежності представляються у вигляді правил та використовуються для кращого розуміння природи даних, що аналізуються. Тобто з великої кількості наборів об'єктів визначаються такі набори, що найбільш часто зустрічаються. При виявленні закономірностей можна з певною ймовірністю передбачити появу подій у майбутньому, що дозволяє

приймати рішення. Така задача є різновидом задачі пошуку асоціативних правил і називається сиквенційним аналізом.

Кластеризація – це розбиття множини документів на кластери (групи документів зі спільними ознаками), які представляють собою підмножини, смислові параметри яких заздалегідь невідомі. Числові методи кластеризації базуються на визначенні кластера як множини документів. Кластеризація може застосовуватися в довільній області, де необхідне дослідження експериментальних та статистичних даних [4].

Для задачі кластеризації характерний пошук груп найбільш схожих об'єктів. Після визначення кластерів використовуються інші методи *Data Mining*. Кластерний аналіз дозволяє розглядати великий об'єм інформації та скорочувати, стискати великі масиви інформації. Результат кластеризації залежить від природи даних об'єктів та від представлення кластерів. Кластеризація відрізняється від класифікації тим, що для проведення аналізу не потрібно мати виділену залежну змінну. Задача кластеризації вирішується на початкових етапах дослідження, а її розв'язок допомагає краще зрозуміти дані.

Всі описані вище методи автоматичного аналізу *Data Mining* забезпечують певну структуризацію текстової інформації, її узагальнення або анотування. Проте для вилучення знань з електронних текстів, зокрема, порівняння текстів та виявлення в них збігів, необхідні засоби автоматичного лінгвістичного аналізу.

Основний метод, що використовується сьогодні для логічного порівняння текстової інформації є метод резолюцій Робінсона. Наприклад, нехай є два простих речення, для кожного з яких побудовано логічну модель.

Перше речення: "Експерт відповідає на запитання слухачів":

$$P(x_1, x_2[x_3]), \tag{1}$$

$$Відповідає(експерт, запитання[слухачів]). \tag{2}$$

Друге речення: "Експерт аналізує запитання спеціалістів":

$$P'(x_1, x_2[x_3']), \tag{3}$$

$$Аналізує(експерт, запитання[спеціалістів]). \tag{4}$$

За алгоритмом уніфікації шукаємо підстановку $Q = \{P / P'\}, x_3 / x_3'$. Якщо здійснити підстановку у вираз (3), будемо мати множини, що містять літерали з однаковими предикатами. Після цього, застосовуючи метод резолюцій до виразів (1) та (3), отримаємо резольвенту, що не дорівнює пустій множині, що свідчить про те, що речення однакові.

Аналізуючи зміст заданих речень можна зробити висновки про те, що а даному випадку в алгоритмі уніфікації не можна було застосовувати підстановку P/P', так як предикати різні за змістом і не є синонімами. Метод резолюцій не дає змогу визначити це в процесі заміни, через те, що не аналізує зміст слів, що входять до речення природної мови [7].

Це означає, що для здійснення коректного порівняння текстових документів за змістом необхідні нові алгоритми лінгвістичного аналізу, які забезпечать змістовну обробку текстової інформації. Одним із таких алгоритмів може бути алгоритм порівняння логіко-лінгвістичних моделей речень природної мови, що включає в себе наступні етапи.

1. Побудова логіко-лінгвістичних моделей [8]. На цьому етапі кожному реченню природної мови ставиться у відповідність логічна формула, що представляє собою одновимірний масив слів, з яких складаються речення, упорядковані у відповідності до того, яку синтаксичну роль вони виконують.

2. Ідентифікація. Відбувається почерговий перегляд елементів всіх логіко-лінгвістичних моделей: предикатів, предикатних змінних (суб'єктів), предикатних змінних (аргументів), предикатних констант. Серед складових логіко-лінгвістичних моделей шукаються спільнокореневі слова, синоніми, активні та пасивні форми спільнокореневих дієслів.

3. Заміна тотожних змінних. Якщо на етапі ідентифікації знайдено тотожні змінні, у всіх логіко-лінгвістичних моделях відбувається їх перепозначення, завдяки чому одні й ті самі слова (навіть якщо вони мають різні граматичні рамки) будуть позначатися однаково, і, відповідно, мати ідентичний зміст.

4. Логічний вивід. Після ідентифікації та заміни тотожних змінних застосовується система продукцій, що містить правила порівняння логіко-лінгвістичних моделей. Такі правила дозволяють через встановлені зв'язки між словами переходити до представлення значень слів у вигляді комбінацій елементарних компонентів змісту.

Висновки. Методи автоматичного аналізу *Data Mining* базуються на використанні певних статистичних закономірностей (класифікація, регресія), пошуку ключових слів, проте не використовують алгоритмів лінгвістичної обробки текстів. Таким чином, автоматичний аналіз текстової інформації, який здійснюється сучасними засобами аналітичної обробки, не здатен опрацьовувати зміст текстів. Для порівняння двох простих речень за змістом було використано метод резолюцій. Як показали дослідження, при застосуванні алгоритму уніфікації, зміст речень не враховується. Тому як вирішення проблеми порівняльного

аналізу текстової інформації за змістом запропоновано новий алгоритм роботи з логіко-лінгвістичними моделями.

Список літератури: 1. *Кузин Л.П.* Основы кибернетики: в 2-х т. Т.2. Основы кибернетических моделей. – М.: Энергия, 1979. – 584 с. **2.** *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. – P. 1-10. **3.** *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа / *Д.В. Ландэ*. – М.: ООО "Вильямс", 2005. – 272 с. **4.** *Барсегян А.А.* Технологии анализа данных: Data Mining, Viual Mining, Text Mining, OLAP / *А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод.* – Спб.: БХВ-Петербург, 2007. – 384. **5.** Data Mining and Knowledge Discovery – 1996 to 2005: Overcoming the Hype and moving from "University" to "Business" and "Analytics", Gregory Piatetsky-Shapiro, Data Mining and Knowledge Discovery journal. – 2007. – 365 с. **6.** *Башмаков А.И.* Интеллектуальные информационные технологии: Учеб.пособие / *А.И. Башмаков, И.А. Башмаков*. – М.: Изд-во МГТУ им. Баумана, 2005. – 304 с. **7.** *Вагин В.Н.* Дедукция и обобщение в системе принятия решений / *В.Н. Вагин*. – М.: Наука, 1988. – 384 с. **8.** *Вавіленкова А.І.* Логіко-лінгвістична модель як засіб відображення синтаксичних особливостей текстової інформації / *А.І. Вавіленкова* // Математичні машини та системи. – 2010. – № 2. – С. 134–137.

Надійшла до редакції 25.03.2013

Статтю представив заступник директора Інституту проблем реєстрації Інформації НАН України д-р техн. наук Додонов О.Г.

УДК 004.82 (045)

Анализ методов обработки текстовой информации / Вавиленкова А.И. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2013. – № 39 (1012). – С. 35 – 40.

В статье представлена характеристика основных методов Data Mining, выделены их преимущества и недостатки. В результате анализа выявлено, что ни один из описанных методов не способен извлекать знания из информации. Продемонстрирована работа метода резолюций Робинсона для сравнения двух простых предложений. Предложен алгоритм сравнения логико-лингвистических моделей текстовой информации по смыслу. Библиогр.: 8 назв.

Ключевые слова: методы Data Mining, извлечение знаний, метод резолюций, логико-лингвистические модели, текстовая информация.

UDC 004.82 (045)

Analysis of methods for processing textual information / Vavilenkova A.I. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling.– Kharkov: NTU "KhPI". – 2013. – № 39 (1012). – P. 35 – 40.

The article presents a description of the main methods of Data Mining, highlighted its advantages and disadvantages. The analysis reveals that none of the above methods are capable to extract knowledge from data. The study demonstrates the operation of Robinson's resolution method for comparing two simple sentences. It is proposed an algorithm of comparing the logic-linguistic models of textual information within the meaning. Refs.: 8 titles.

Keywords: methods of Data Mining, extraction of knowledge, the method of resolutions, the logic-linguistic models, textual information.