УДК 004.891

В.М. ДЖУЛІЙ, О.М. ГОРБАТЮК
Хмельницький національний університет

МЕТОДИ ТА АЛГОРИТМИ КЛАСТЕРИЗАЦІЇ ПРИ КОМПЛЕКСНОМУ АНАЛІЗІ ДАНИХ

В статті виділено особливу роль вирішенні задачі кластеризації при комплексному аналізі даних - про властивості яких на момент початку аналізу нічого невідомо. Висвітлено відомі методи розв'язання задачі кластеризації. Виявлено особливості та недоліки існуючих методів кластеризації. Дані недоліки виникають з припущень в відношенні досліджуваних даних: кластери мають задану форму, в кожному з кластерів є вузлова точка, по відношенню до якиих і будується розбиття. Такі допущення не завжди коректні. Крім того, при побудові кластерів не приймаються до розрахунку відношення між елементами множини даних.

Ключові слова: *інтелектуальна обробка даних, програмні системи, алгоритми, база даних і знань, інформаційна система.*

V.M. DZULIY, A.M. GORBATYUK
Khmelnitsky national university

METHODS AND CLUSTERING ALGORITHMS IN COMPLEX DATA ANALYSIS

The article indicated the special role of solving the problem of clustering in complex data analysis - the properties of which at the time the test is not known. The article known methods of solving the problem of clustering. The features and shortcomings of existing clustering methods. These deficiencies arise from assumptions concerning the survey data, clusters are specified shape, each cluster node is in relation towhich and built a partition. These assumptions are not always correct. In addition, the construction of clusters are not taken into account the relationship between the elements of the set of data.

Keywords: *data mining, software systems, algorithms, database and knowledge information system.*

Вступ. В сучасних умовах, багато компаній прагнуть впровадити особливу корпоративну культуру, корпоративні цінності, особливий стиль вирішення виробничих завдань і т.п. З цих причин корпоративне навчання є необхідним для будь-якої компанії, а іноді приймає стратегічне значення. У 2011 р компанія «Амплуа» провела комплексне дослідження систем корпоративного навчання і розвитку в Росії та Україні під назвою Trainings INDEX. У цьому дослідженні виділено ряд тенденцій, що дозволяють говорити про те, що компанії проявляють все більший інтерес до корпоративного навчання і готові вкладати в цю сферу значні інвестиції. Крім іншого, в дослідженні була оцінена тенденція використання технологій електронного навчання та систем управління навчанням. Слід визнати, що на даний момент вкладення в системи електронного навчання мізерно малі в порівнянні з традиційними видами навчання. Тому актуальною є проблема підвищення ефективності електронних засобів, що застосовуються для корпоративного навчання.

Як правило, системи корпоративного навчання створюються для вирішення наступних завдань: навчання та оцінка персоналу; впровадження корпоративної культури; побудова та розвиток кар'єри; управління ефективністю діяльності співробітників; управління знаннями. Перші два завдання легко вирішуються як традиційними, так і електронними засобами навчання. Третє і четверте завдання вимагають особистісного підходу. *П'ята задача, пов'язана з проблемами здобуття знань з неформалізованих джерел і представлення знань в інформаційних системах.*

Постановка завдання. Основним джерелом корисної, з точки зору навчання, інформації є тексти на природній мові, збережені у файловому архіві підприємства. Ефективна обробка та зберігання інформації не можливі без урахування семантики. Більшість сучасних сховищ даних не оперують даними на смисловому рівні. Результатом цього є надмірність і перетин збережених даних, складність організації ефективного пошуку та доступу до даних користувача.

Логічний підхід до семантичного аналізу природних мов інтенсивно досліджується в роботах лінгвістів і логіків і має великий потенціал при ефективній реалізації засобами сучасних програмних інтелектуальних систем. Інтелектуальний аналіз даних - область знань, що відноситься до обробки даних, вивчає пошук і опис прихованих, нетривіальних і практично корисних закономірностей в досліджуваних даних. Методи інтелектуального аналізу даних отримали особливий розвиток внаслідок явища, що отримало назву інформаційного вибуху. Інформаційний вибух - лавиноподібне зростання кількості інформації, що накопичується в різних областях людської діяльності, і пов'язане з розвитком засобів обчислювальної техніки, особливо, засобів зберігання даних. Одним з наслідків інформаційного вибуху стало неможливість обробки всієї накопичуваної інформації за допомогою наявних методів (в основному статистичних). Важлива відмінність інтелектуального аналізу даних від традиційних методик пов'язана з виявленням з його допомогою прихованих закономірностей в даних, в той час як інші методики займаються параметричної оцінкою вже відомих закономірностей. Способи вирішення завдань інтелектуального аналізу даних можна розділити на «навчання з учителем» і «навчання без учителя». Дані терміни пов'язані з наданням («учителем») або ненаданням додаткової інформації про дані, що аналізуються. Таке рішення задачі класифікацією відноситься до навчання з учителем, оскільки завжди передбачає наявність інформації про класову приналежність кожного з елементів досліджуваної множини. Задача кластеризації може бути віднесена до навчання без учителя, оскільки в ідеальному випадку не вимагає надання додаткової інформації. Крім того, навчання з учителем часто пов'язано з етапом уточнення рішення, якість якого забезпечується експертом або спеціальною адаптивною процедурою.

В задачах регресії та класифікації потрібно визначити значення залежної змінної об'єкта на підставі значень інших змінних, що характеризують даний об'єкт. Якщо дана кінцева множина об'єктів $I = \{i_1, i_2, ..., i_j, ... i_n\}$, кожен з яких характеризується деяким описовими ознаками

$(x_1, x_2, ..., x_k, ..., x_m, x_{m+1})$, де $x_k \in X_k$ і X_k - допустима множина значень ознаки. Якщо значення ознак $(x_1, x_2, ..., x_k, ..., x_m)$, відомі, тоді задача полягає у визначенні за відомими ознаками невідомої ознаки x_{m+1}.

Якщо множина X_{m+1} значень ознаки x_{m+1} кінцева, то завдання називають класифікацією. Якщо X_{m+1} лічильно або має потужність континууму, то задача регресії.

Методи інтелектуального аналізу даних істотно відрізняються як по використовуваному математичному апарату, так і по алгоритмічній реалізації. Математична статистика, як теоретична основа, раніше інших методів знайшла своє застосування в аналізі даних. З математичною статистикою пов'язане поняття про регресійний аналіз. Методи математичної статистики знаходять своє застосування при вирішенні задач класифікації, регресії, аналізу відхилень. Апарат нейронних мереж знайшов найбільш широке застосування при рішенні задач класифікації (наприклад, розпізнавання символів, зображень, геофізичних даних. Важливою властивістю нейронних мереж є їх здатність до навчання, що реалізується, наприклад, на основі алгоритму зворотного поширення помилки. Алгоритмічне рішення формалізованої задачі інтелектуального аналізу даних пов'язане з різними завданнями пошуку: екстремуму цільової функції, виду цільової залежності. Часто такий алгоритм зводиться до повного перебору можливих варіантів рішення, але в ряді завдань можливі більш ефективні методи. Число методів і алгоритмів розбиття множини на кластери досить велике. Всі методи можна поділити на ієрархічні та неієрархічні. У неієрархічних алгоритмах характер їх роботи і умову зупинки часто необхідно заздалегідь регламентувати досить великим числом параметрів, що іноді важко. Але в таких алгоритмах досягається велика гнучкість у варіюванні кластеризації. В ієрархічних алгоритмах фактично відмовляються від певного числа кластерів, будуючи повне дерево вкладених кластерів.

Існують різні варіанти вирішення проблеми інформаційного пошуку. Перший варіант - реалізація локальних пошукових підсистем в кожному інформаційному модулі. Перевага такого підходу - простота реалізації, недоліки - неможливість пошуку по декількох джерел з єдиної точки доступу, складність модифікації і оновлення системи, висока вартість адміністрування та налаштування. Необхідна реалізація системи, що дозволяє організувати доступ до гетерогенних територіально розподілених джерел інформації. Серед сучасних технологій побудови корпоративних інформаційних систем з подібними характеристиками, варто виділити сервісну і агентну архітектуру. Агентна архітектура має ряд переваг при вирішенні подібного класу задач, тому для створення розподіленої системи пошуку інформації в масштабах підприємства має сенс використовувати саме агентний підхід.

Ефективне використання знань, що містяться в текстах, потребує нових стратегій обробки інформації, відмінних від традиційних підходів. Такі стратегії повинні враховувати семантичні закони природної мови.

Вивчення семантики речення тісно пов'язане з мисленням. Тому історично перші спроби формалізації методів роботи з семантикою робилися в рамках логіки. Логічний підхід до формалізації семантичного аналізу донедавна мав поширення лише в середовищі лінгвістів і логіків. Перспективним є застосування логічного підходу до проблеми автоматизації обробки природно-мовних текстів за допомогою ЕОМ. Виділимо основні плюси такого підходу.

Просте розпаралелювання єдиного завдання. Символьні формули досить просто дробляться на підформули. Достовірність окремої підформули найчастіше можна перевіряти незалежно від інших. Ця властивість особливо важливо з огляду розвитку сучасної обчислювальної техніки в сторону багатоядерності мікропроцесорів і розподіленої обробки даних.

Множинність вирішуваних завдань на єдиному наборі знань. Логічне подання дозволяє на єдиному наборі даних/знань виконувати різні операції. Так, на одній базі можуть функціонувати пошукові системи, питально-відповідні системи, системи розпізнавання образів і т.д.

Природне пояснення результатів операцій. Логіка формалізує правила мислення, тому результати набору логічних операцій при достатній дружності інтерфейсу користувача можуть легко адаптуватися під користувача. У логічній пошуковій системі можна показати ланцюжок умовиводів, на основі якої пропонований текст віднесений до релевантних результатів пошуку. Більшість існуючих методів не дають такої можливості пояснення результату, оскільки набори векторних і статистичних даних, з якими вони працюють, набагато складніше представити в доступному для непідготовленого користувача вигляді. Це досить важлива обставина, оскільки досвід використання систем підтримки прийняття рішень показав, що користувач найчастіше відкидає результат роботи програми, якщо не може усвідомити, яким чином такий результат був отриманий.

Розвиток логічного підходу до вивчення семантики можна простежити на основі існуючих семантик некласичних логік: семантики сенсу і денотата Г. Фреге, теорії об'єктів і пропозицій Б. Рассела, теорії істини А. Тарского, семантики можливих світів С. Кріпке, логіки сенсу і денотата А. Черча. Найбільш перспективним з точки зору застосовності в автоматичному семантичному аналізі представляється підхід, запропонований Річардом Монтегю. Він сформував цілий напрям, що отримав назву «формальна семантика». Основна ідея робіт Монтегю виражена в назві однієї з його основних праць «English as a formal language». Будь яка природна мова (зокрема англійська) пропонується розуміти як формальна логічний мова, яка є більш складною по відношенню до існуючих формальних мов. Отже, при описі природної мови можна використовувати ті ж поняття і конструкції, що і для інших логічних мов.

Основна частина. Центральний принцип формальної семантики полягає в композиційному відношенні між синтаксисом і семантикою. Принцип композиційності можна виразити таким чином: значення виразу є функція його частин і способу їх синтаксичної комбінації. При цьому істинність визначається не абсолютно, а в межах деякої моделі.

Можливі два підходи до вивчення семантико-синтаксичних зв'язків: описати синтаксис природної мови (у нашому випадку української) та інтерпретувати вирази мови в моделях; використовувати проміжну логічну мову, для чого описати синтаксис і семантику досить близької до природної логічної мови; при цьому опис семантики природної мови зводиться до подання тексту в сконструйованій логічній мові. Монтегю використовував обидва підходи. Другий підхід є більш перспективним, оскільки дозволяє

реалізувати описані вище плюси логічного представлення мов.

У інтелектуальному агентно-орієнтованому комплексі передбачається наявність таких інтелектуальних можливостей, як періодичне тестування за знаннями, накопиченими в компанії; підбір інформації з конкретних питань. Для реалізації цих можливостей необхідна структурована база знань, з якої відповідні інтелектуальні агенти могли б отримувати необхідну інформацію.

З точки зору реалізації технічної системи можна виділити наступні фрагменти обробки текстової інформації засобами формальної семантики (рис. 1).

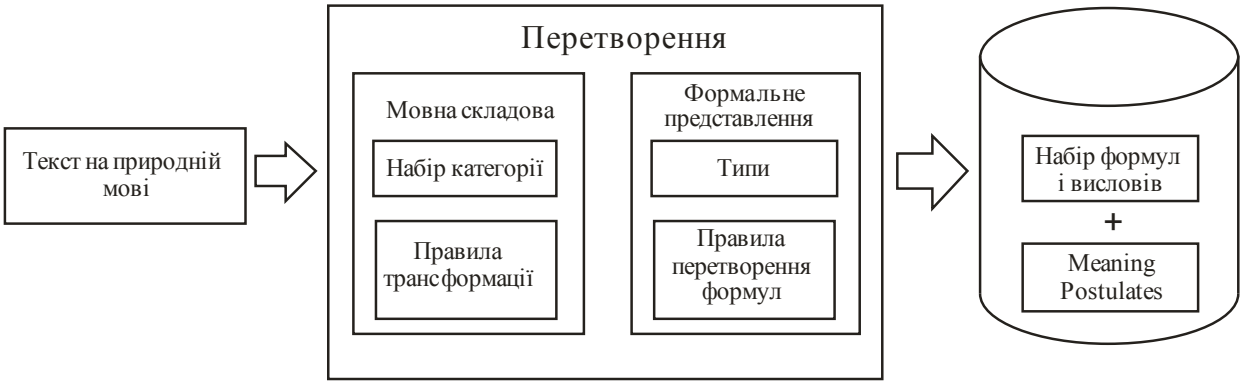

Рис. 1. Обробка текстової інформації у формальній семантиці на основі логіки Монтегю

Схема наочно показує, як відбувається функціональний розподіл частин єдиної системи під назвою «логіка Монтегю». У блоці «мовна складова» об'єднані елементи, специфічні для конкретного природної мови. Формальне подання не залежить від конкретної мови і є єдиним для багатьох реалізацій мовних складових. Зараз основні дослідження зосереджені в сфері природно-мовної складової (з внесенням необхідних змін до постулатів значень). На мовну складову вводяться обмеження таким чином, щоб вона представляла підмножину природної мови, мінімально необхідну для представлення простих мовних фраз.

На даний момент вивчення формальної семантики зводиться саме до визначення способів переходу від природно-мовного представлення до формалізованого логічного. Можливі різні способи таких переходів.

Вхідні дані для системи - це текст на природній мові. На основі лінгвістичної обробки тексту будується набір категорій інтенсіональної логіки для подальшого застосування правил трансформації синтаксичних конструкцій в елементи єдиної формули, що відображає зміст висловлювання. Формальне подання не залежить від конкретної природної мови і являє собою набір типів та операцій над формулами. Результатом обробки є формалізоване представлення змісту тексту у вигляді набору формул, відображаючих сенс речень і множин постулатів значень, які представляють фонові знання про світ. Метод полягає в застосуванні алгоритмів формалізації сенсу природно-мовних текстів, заповнені бази знань та інтерпретації на ній запитів користувачів чи інтелектуальних агентів.

Висновки. Виділена особлива роль вирішення задачі кластеризації при комплексному аналізі даних - кластеризація є необхідним первинним етапом при комплексному аналізу даних, про властивості яких на момент початку аналізу нічого невідомо. Висвітлено відомі методи розв'язання задачі кластеризації. Виявлено особливості та недоліки існуючих методів кластеризації. Дані недоліки виникають з припущень в відношенні досліджуваних даних: кластери мають задану форму, в кожному з кластерів є вузлова точка, по відношенню до яким і будується розбиття. Такі допущення не завжди коректні. Крім того, при побудові кластерів не приймаються до розрахунку відношення між елементами множини даних. В рамках представленої концепції можуть бути побудовані програмні інтелектуальні агенти, які здатні витягувати нові знання з природно-мовних текстів і формувати багаторівневі бази знань по взаємопов'язаних предметних областях, що дозволить забезпечити накопичення та ефективну передачу корпоративних знань.

Література

1. Швецов А. Н. Распределенные интеллектуальные информационные системы / А. Н. Швецов, С. А. Яковлев – СПб.: СПбГТУ «ЛЭТИ», 2003 – 318 с.

2. Микиртумов И. Б. Теория смысла и интенсиональная логика. – Спб.: Изд-во С.-Петерб. ун-та, 2006. – 351 с.

3. Герасимова И. А. Формальная грамматика и интенсиональная логика - М.: 2000. – 156 с.

4. Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-ое издание [Текст] / Барсегян А. А., Куприянов М. С, Степаненко В. В., Холод И. И. - СПб.: БХВ-Петербург, 2007. 384 с.

References

1. A. Shvetsov Raspredelennыe yntellektualnыe Clearing system / AN Shvetsov, S. Yakovlev - SPb .: SPbHTU "LETI", 2003 - 318 p.

2. Mykyrtumov I. B. Theory and Meaning yntensyonalnaya logic. - SPb .: Publishing House of St. Peterborough. University Press, 2006. - 351 p.

3. Gerasimova IA Formalnaya Grammatika and yntensyonalnaya Logic - M .: 2000. - 156 p.

4. Barseghyan AA Technologies analysis of data: Data Mining, Visual Mining, Text Mining, OLAP. 2-th edition [Text] / A. Barseghyan, Kupryyanov M. C. V. Stepanenko, Cold I. I. - SPb .: BHV-Petersburg, 2007. 384 pp

Рецензія/Peer review : 7.11.2014 р. Надрукована/Printed :2.1.2015 р.