УДК 004.822

Н.Ф. Хайрова

Національний технічний університет «Харківський політехнічний інститут», Харків

ІНТЕЛЕКТУАЛЬНІ ЗАСОБИ ВКЛЮЧЕННЯ НЕСТРУКТУРОВАНОЇ ІНФОРМАЦІЇ У ПРОЦЕС ІНФОРМАЦІЙНО-АНАЛІТИЧНОГО ЗАБЕЗПЕЧЕННЯ УПРАВЛІННЯ ВІЙСЬКОВОЮ ЧАСТИНОЮ

У статті розглядається необхідність включення неструктурованої інформації у процес інформаційно-аналітичного забезпечення управління військовими частинами, який об'єднує технології Data Mining та Text Mining. Показана необхідність використання технології OLAP-кубів для представлення багатовимірного простору об'єктивних та достовірних знань колекції документів. Запропоновано використання методу компараторної ідентифікації для виділення семантичних ідентифікаторів ієрархії понять колекції. Розроблена технологія створення атрибуту інформаційних понять OLAP-куба колекції.

Ключові слова: *неструктурована інформація, інформаційно-аналітичне управління, військова частина, Text Mining, OLAP-куби, видобування знань, метод компараторної ідентифікації.*

Вступ

Глобальна проблема, яка існує сьогодні при обробці інформації, полягає не тільки у необхідності її відбору, як із традиційних комп'ютерних систем, так і з різнорідних джерел, але й у перетворенні її на знання, які потім використовуються для ефективного управління бізнесом зокрема та суспільством взагалі. Це завдання ускладнюється ще й тим, що близько 85% інформації зберігається не у СУБД, а у текстових документах та файлах: Web-сторінках, електронних листах та аналогічних документах. Стрімке зростання об'єму повнотекстової інформації і далі продовжується не тільки в глобальних мережах (де кожні 18 місяців відбувається подвоєння об'єму неструктурованої інформації), але й у таких корпоративних інформаційних системах, якими є сучасні військові частини.

Високі темпи зростання об'єму інформаційного простору за умов домінування у його складі неструктурованих даних (як за абсолютними показниками, так и за більш високими темпами приросту) створюють загрозу перетворення сховищ інформації на «кладовища інформації» [1].

До останнього часу у галузі інтелектуальної обробки інформації існували близькі області Data Mining та Text Mining, які не перетиналися між собою. Засоби першого напрямку Data Mining, в основному, обробляли структуровану інформацію – тобто дані, а напрямок Text Mining займався обробкою слабкоструктурованої (текстової) інформації. Але сьогодні, у зв'язку з тим, що загальна картина інформаційного простору корпорації, її «інформаційний ландшафт», містить як структуровані дані, так і неструктуровану інформацію, виникає необхідність у конвергенції цих двох напрямків.

Мета статті полягає в обґрунтуванні спеціалізованої інтелектуальної технології щодо представлення неструктурованої природно-мовної інформації у інформаційно-аналітичній системі управління військовою частиною.

Загальна постановка задачі

Традиційно включення неструктурованої інформації у процес інформаційно-аналітичного забезпечення управління військовою частиною потребує її попереднього інтелектуального осмислення та змістової розмітки метаданими, яка у більшості своїй здійснюється вручну. В роботі пропонується метод динамічної побудови багатовимірного представлення досліджуваної колекції документів. Під документом, в даному контексті, ми будемо розуміти будь-які об'єкти, які являють собою «сірі» тексти, вони надходять до корпорації з різноманітних джерел, и не мають заздалегідь встановленої цінності. Група документів, за якими здійснюється обробка, являє собою колекцію або масив текстів.

Запропоноване багатовимірне представлення інформації колекції документів дозволить використати навігацію по OLAP-кубу для відбору необхідних документів або фактів, вилучених з даних документів. Аналітик може зосереджуватися на елементах різних вимірів (наприклад, адміністративно-географічні області публікації документа), проглядати документи у комірках з необхідними значеннями частот та ін.

Додатково можуть використовуватися загальні методи аналізу та прогнозування даних.

Незалежні виміри гіперкуба у моделі, яка розглядається, являють собою багатовимірний простір об'єктивних та достовірних знань досліджуваної колекції документів. Атрибутами, які характеризують якісні значення вимірів (тобто осями OLAP-кубу), є метадані документа (рубрики, автори, дата публікації, джерела та ін.) та семантико-факто-графічна інформація, вилучена з документів. Кількісним показ-

ником, розміщеним у комірці куба, на перетині вимірів, є агрегатна функція COUNT().

Одна з головних переваг OLAP технології, як відомо, полягає у використанні операції стискання гіперкуба та побудови його перетинів, яка з'являється завдяки використанню значень атрибутів-вимірів більш високих рівнів ієрархії та відповідного агрегування значень показників обчислювальних полів. У кубах, які створюються, така ієрархія будується за датами публікації, географічним розташуванням місць видання, входженням інформаційних понять в документ.

Але більш цікавою є запропонована у даному дослідженні класифікація усіх інформаційних понять масиву документів за деякими змістовими ознаками, яка здійснюється аналогічно інтелекту людини, з виділенням семантичних ідентифікаторів ієрархії. Для чого пропонується моделювати одну з найвищих форм інтелектуальної діяльності людини – розуміння та віднесення лексичних одиниць мови до одного класу за деякими змістовими ознаками.

Мета дослідження. Розробити метод побудови багатовимірного подання інформації колекції документів, у вигляді OLAP-кубів, які містять атрибут інформаційних понять колекції документа, класифікованих за деякими змістовними ознаками (рис. 1). Для виділення семантичних ідентифікаторів ієрархії понять колекції документів моделювати інтелектуальне розуміння та класифікацію лексичних одиниць за деякими змістовними ознаками.

- Вимірювання куба:
  - Метадані:
    - дата публікації;
    - країна публікації.
  - понятійні одиниці.
- Обчислювані поля даних:
  - count () – частота.

Інтелектуальна інформаційна технологія, яка пропонується

Для формування вимірів інформаційних понять колекції документів з можливістю побудови ієрархічних відношень пропонується використовувати технологію, яка поєднує аналіз окремих документів колекції та компараторну ідентифікацію [2] документів та лексикона інформаційних понять колекції документів.

На першому етапі здійснюється автоматичний аналіз окремих документів колекції. На цьому етапі використовується дещо змінена технологія Information Extraction напряму Text Mining, яка полягає у виділенні з документів, написаних різними мовами, інформаційних понять, які з визначеним ступенем достовірності відображають тематичну спрямованість документа.

На етапі передпроцесорної обробки визначається мова документа та його формат, відповідно до якого обирається декодер. На наступному етапі лінгвістичної обробки виділяємо токени (лексеми) – екземпляри послідовності символів у документі, поєднані у семантичну одиницю для обробки. Для чого при обробці російськомовних та україномовних текстів текст розбивається відповідно до пробілів та відкидаються знаки пунктуації. При обробці утворення присвійних прикметників та скорочень у англомовних текстах використовується додатковий алгоритм аналізу використання апострофу.

Система виділення токенів, включає підсистему обробки лексем, які мають графемне оформлення. Для виявлення семантично значущих змістовних понять документа, які мають графемне оформлення, комбінуються два методи: словниковий (словник містить лексеми, графемне оформлення яких важко формалізувати), та алгоритмічний, який використовує формальні правила універсальних випадків.

На морфологічному етапі обробки для приведення лексем до канонічної форми використовується стеммінг. Під стеммінгом розуміється наближений евристичний процес, в ході якого від слів відкидаються квазізакінчення.

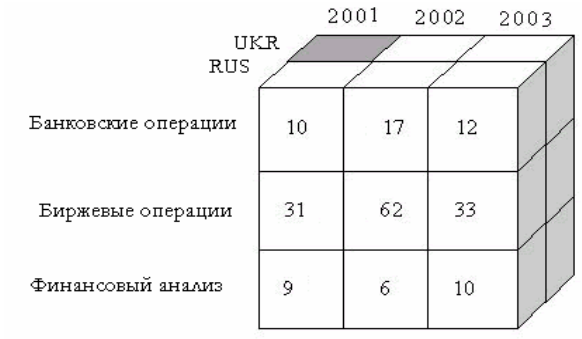

Рис. 1. Багатовимірне представлення простору знань масиву текстів

На етапі контекстного аналізу для виділення термінологічних словосполучень використовується наступний алгоритм: виділяються словосполучення за типом керування або узгодження, потім здійснюється пошук подібних словосполучень в усьому документі. При наявності більше трьох подібних словосполучень, дане словосполучення вважатиметься інформаційним поняттям документа.

Для остаточного формування словника документа будується гіперболічна залежність Ципфа рангу терміна від частоти. У центральній зоні гіперболи знаходяться слова, які максимально характеризують даний текст та виражають його специфічність, які за своїм визначенням є інформаційними поняттями даного тексту.

Таким чином, на першому етапі технології, в результаті аналізу усіх документів колекції формується лексикон інформаційних понять текстової колекції сховища документів, яка аналізується. Цей лексикон буде одним з вимірів створюваного гіперкуба OLAP.

Використання методу компараторної ідентифікації для визначення ієрархії семантичних ідентифікаторів

Для поєднання інформаційних понять у класи еквівалентності з використанням змістовної класифікації лексики, пропонується використовувати один з методів штучного інтелекту – метод компараторної ідентифікації. Використовуючи семантичний трикутник знаку, запропонований Фреге [3], на множині лексикона інформаційних понять $R= \{r_1, r_2, ..., r_n\}$ та множині концептів, які розглядаються у колекції документів, $\Re = \{\rho_1, \rho_2, \ldots, \rho_m\}$ введемо функцію розуміння інформаційного поняття

$$\rho = f(r),$$

де ρ – концепт інформаційного поняття. Під концептом ми будемо розуміти інформацію, яку несе r про можливі денотати [4], тобто сукупність суджень про якийсь об'єкт, яка відображає його сутність та відносить його за загальними та специфічними ознаками до предметів певного класу.

Розуміючи денотат поняття, який виражається ім'ям поняття r, класифікатор співвідносить його з певним концептом, змістом, десигнатом ρ. Функція розуміння інформаційного поняття описує процес встановлення класифікатором тотожності між ім'ям інформаційного поняття та концептом, знаком якого він є. Якщо класифікатор розглядає множину інформаційних понять лексикона колекції документів, то функція f відображає множину понять лексикона на множину усіх значень функції, тобто сукупність усіх змістів, породжуваних інформаційними поняттями з множини R.

Для виділення рівнів ієрархії вимірів інформаційних понять OLAP куба необхідно виділити класи рівнозначних або семантично близьких інформаційних понять даної предметної області. Такі інформаційні поняття відповідають одним і тим самим або близьким за змістом концептам, які зазвичай відносяться до одного дескриптора, і, як показують дослідження, часто розглядаються в одному зв'язному тексті ділової документації корпорації, яка характеризується однією тематикою.

Аналізуючи текст документа з колекції документів, яка розглядається, і розуміючи його, класифікатор, зазвичай, формує у своїй свідомості певний зміст, який є основним значенням документа. Зміст документа однозначно визначається текстом документа, який його породив (рис. 2) [5].

Відповідно до визначення тлумачного словника [6]: зміст – це ідеальна концептуальна сутність, ідея, призначення, кінцева мета (цінність) чого-небудь, сенс якогось висловлювання, який не може бути зведений до значень його складових частин та елементів, через те, що він сам визначає ці значення. Розуміння класифікатором тексту документа визначає компонент його мислення, психологічний стан, який визначає вірне сприйняття або інтерпретацію даного документу, тобто встановлення зв'язку виявлених нових властивостей об'єкта пізнання з вже відомими.

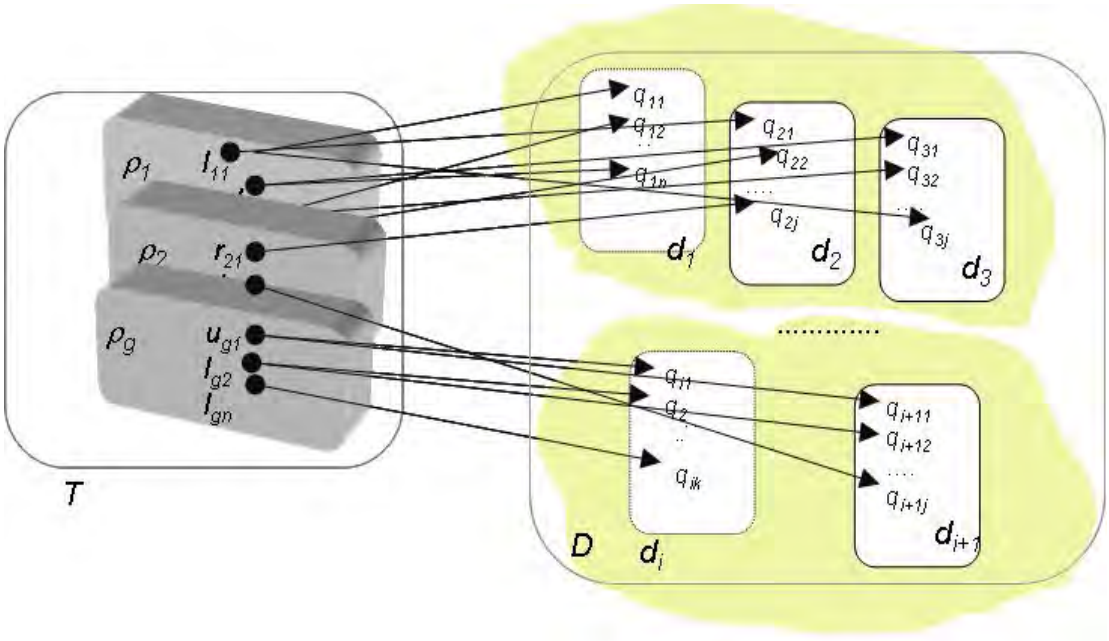

Рис. 2. Реалізація методу компараторної ідентифікації для формування класів еквівалентності лексики колекції документів.

Функція розуміння тексту ω = g(d) встановлює залежність змісту документа від імені (знаку) документа, де d ∈ D, D – множина документів колекції. Розуміючи зміст документа, класифікатор однозначно встановлює відповідність або невідповідність між певним концептом, розглядуваного у тексті денотата, та змістом тексту, визначаючи предикат.

Використовуючи метод компараторної ідентифікації можна перейти від суб'єктивного розуміння змістів документів та концептів денотатів до об'єк-

тивного відношення між елементами лексикона інформаційних понять колекції документів та текстами документів колекції, яке визначається предикатом P (d, r). Предикат аналітико-синтетичної обробки документа $\varepsilon = Z(\omega, \rho)$, який реалізує відношення предмета розгляду документа та концепту інформаційного поняття лексикона, пов'язаний з предикатом P (d, r) відношенням [7]:

$$P(d, r) = Z(\,(\,g(d), f(r)\,) = Z(\omega, \rho)\,.$$

Використовуючи предикат аналітико-синтетичної обробки можна отримати предикати еквівалентності, які однозначно розбивають лексикон інформаційних понять досліджуваної колекції документів на шари розбиття, у яких усі інформаційні поняття, що відносяться до одного шару, будуть відноситися до одного класу семантично близьких змістових одиниць, які поєднуватимуться на осі вимірів куба OLAP:

$$\Psi b(r) = \forall d \in D\,(P(d, r) \sim P(d, b)).$$

Висновки

Таким чином, використання метода компараторної ідентифікації для об'єднання інформаційних понять у класи еквівалентності з використанням змістовної класифікації лексики, дозволяє виділити головний атрибут багатовимірного куба OLAP тематичної колекції документів. Використання такої технології дозволяє «на льоту» будувати багатовимірне представлення масиву слабкоструктурованих документів, у яких в якості вимірів виступатимуть як метадані документа, так и семантично значущі інформаційні поняття колекції документів, що циркулює у середовищі інформаційно-аналітичної системи управління військовою частиною.

Список літератури

1. Владимир Рубанов. Между стандартами управления и информационной стихией. Ежеквартальный журнал Российское издание. Технологический прогноз. БОЛЬШИЕ ДАННЫЕ: как извлечь из них информацию. – 2010. – Вып. 3. – С. 7-14.

2. Бондаренко М.Ф. Теория интеллекта. Учебник / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Х.. Изд-во СМИТ, 2007. – 576 с.

3. Фреге, Г. Смысл и значение / Г. Фреге // Избранные работы – М.: Дом интеллектуальной книги, 1997. – 128 с.

4. Поспелов, Д.А. Введение в прикладную семиотику [Текст] / Д.А. Поспелов, Г.С. Осипов // Новости искусственного интеллекта. – М.: Изд-во РАИИ, 2002. – № 6. – С. 28-35.

5. Nina Khairova. Multidimensional Representation of the Documents Collection / Computer Science and Information Technologies: Materials of the VI th international Scientific and Technical Conference CSIT 2011. – Lviv: Publishing House Vezha&Co, 2011, P. 164-165.

6. Советский энциклопедический словарь / [Гл. ред. А.М. Прохоров]. – 4-е изд. – М.: Сов. энциклопедия, 1988. – 1600 с.

7. Хайрова Н.Ф. Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов корпоративной информационной системы / Н.Ф. Хайрова, В.А. Тарловский // Вісник Національного технічного університету «ХПІ». Збірник наукових праць. Тематичний випуск «Системний аналіз, управління та інформаційні технології». – Х.: НТУ «ХПІ». – 2010. – № 67. – С. 132-138.

Надійшла до редколегії 28.11.2012

Рецензент: д-р техн. наук, проф. И.В. Шостак, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

ИНТЕЛЛЕКТУАЛЬНЫЕ СРЕДСТВА ВКЛЮЧЕНИЯ НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ В ПРОЦЕСС ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОГО УПРАВЛЕНИЯ ВОЕННОЙ ЧАСТЬЮ

Н.Ф. Хайрова

В статье рассматривается необходимость включения неструктурированной информации в процесс аналитического обеспечения корпоративного управления, который объединяет технологии Data Mining и Text Mining. Показано необходимость технологии OLAP-кубов для представления многомерного пространства объективных и достоверных знаний коллекции документов. Предложено использование метода компараторной идентификации для выделения семантических идентификаторов иерархии понятий коллекции. Разработана технология создания атрибута информационных понятий OLAP-куба коллекции.

Ключевые слова: *неструктурированная информация, информационно-аналитическое управление, военная часть, Text Mining, OLAP-кубы, извлечение знаний, метод компараторной идентификации.*

INTELLECTUAL TECHNIQUES FOR UNSTRUCTURED INFORMATION INCLUSION INTO INFORMATION-ANALYTICAL MANAGEMENT OF MILITARY UNIT

N.F. Khairova

In the article it is considered necessity of unstructured information inclusion into analytical provision of corporative management which includes Data Mining and Text Mining. It is shown the necessity of OLAP-cube technology to present multidimensional area of objective and true knowledge from documents. It is suggested to use comparative identification method to extract semantic identifiers of notion hierarchy. It is developed technology for development of attribute of information notions in OLAP-cube collection.

Keywords: *unstructured information, information-analytical management, military unit, Text Mining, OLAP-cubes, knowledge extraction, comparative identification method.*