УДК 004.89:510.635

А.Ю. Дорошенко[1]

[1]НТУ «ХПІ», м. Харків, Україна, doroshenkoanastasiia@gmail.com

РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ФАКТОГРАФІЧНОЇ ІНФОРМАЦІЇ

З швидким розвитком технологій — інформацію розглядають як один з основних ресурсів розвитку спільноти, а інформаційні системи та технології – як знаряддя удосконалення продуктивності праці та ефективності роботи особового складу. У будь-яких соціально-економічних та організаційно-виробничих системах – опрацювання та переробка інформації – найважливіша функція, без якої неможлива цілеспрямована діяльність. Обсяги і швидкість інформаційних потоків постійно збільшуються, тому підприємства все частіше звертаються до інтелектуального аналізу як засобу, який дає змогу отримувати корисні для підприємства відомості з величезної кількості інформації, що зберігається в корпоративних базах даних. Інтелектуальний аналіз допомагає досягти розуміння взаємовідносин з клієнтами і партнерами, основних показників роботи підприємства, а також отримати комплексне уявлення про компанію на всіх рівнях. Головним завданням є підвищення ефективності роботи бізнесу і його прибутковості, розширення ринку, зростання і досягнення поставлених цілей.

ПОШУК ІНФОРМАЦІЇ, ІНФОРМАЦІЙНІ СИСТЕМИ, ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ, ОПРАЦЮВАННЯ ДАНИХ

Дорошенко А.Ю. Разработка информационной технологии интеллектуального анализа фактографической информации. С быстрым развитием технологий — информацию рассматривают как один из основных ресурсов развития общества, а информационные системы и технологии — как орудие совершенствования производительности труда и эффективности работы личного состава. В любых социально-экономических и организационно-производственных системах — разработка и переработка информации — важнейшая функция, без которой невозможна целенаправленная деятельность. Объемы и скорость информационных потоков постоянно увеличиваются, поэтому предприятия все чаще обращаются к интеллектуальному анализу как средству, которое позволяет получать полезные для предприятия сведения из огромного количества информации, хранящейся в корпоративных базах данных. Интеллектуальный анализ помогает достичь понимания во взаимоотношениях с клиентами и партнерами, основных показателей работы предприятия, а также получить комплексное представление о компании на всех уровнях. Главной задачей является повышение эффективности работы бизнеса и его прибыльности, расширение рынка, рост и достижения поставленных целей.

ПОИСК ИНФОРМАЦИИ, ИНФОРМАЦИОННЫЕ СИСТЕМЫ, ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ, ОБРАБОТКА ДАННЫХ

Doroshenko A.Y. Development of data mining information technology of factual information. With the rapid development of technology, information is considered as one of the main resources for the development of society, and information systems and technologies as an instrument for improving labor productivity and the efficiency of personnel. In any socio-economic and organizational-production systems — the development and processing of information is the most important function, without which purposeful activity is impossible. The volumes and speed of information flows are constantly increasing, therefore, enterprises are increasingly turning to intellectual analysis as a means of obtaining information useful for an enterprise from a huge amount of information stored in corporate databases. Intelligent analysis helps to achieve understanding in relations with customers and partners, the main indicators of the enterprise, as well as get a comprehensive picture of the company at all levels. The main task is to increase the efficiency of the business and its profitability, expand the market, grow and achieve its goals. Specific responsibilities often include offering their knowledge of development methodologies, project planning, analysis and management of new or upgraded information technology projects, including definition and evaluation of alternatives. This paper present a review about Intellectual Capital and Information Technology, showing relationship between these subject and outlines research method, based on a intellectual review approach. This work is a starting point for further research that aims to deepen the intellectual capital theory as a way to understand and share knowledge in IT projects. There are analyzed the problems of disparate text data processing. The problems of text data processing are highlighted. It is shown that the existing mathematical software and insufficient for the simultaneous solution of problems handling multiple text resources.

INFORMATION SEARCH, INFORMATION SYSTEMS, INFORMATION TECHNOLOGY, DATA PROCESSING

Вступ

Сучасне життя неможливе без використання електронних інформаційних ресурсів в усіх сферах. Бурхливий розвиток мережевих інформаційних технологій, в тому числі Інтернету, сприяють значному збільшенню доступних інформаційних ресурсів та обсягів переданої інформації. Найчастіше це різнорідна, слабо структурована і надлишкова інформація, що має високу динаміку оновлення. Необхідність ефективного використання цього колосального мінливого обсягу інформації обумовлює актуальність і значимість досліджень у галузі інтелектуальної обробки інформації.

Значна частина інформації подається у вигляді текстів природними мовами. У багатьох завданнях, наприклад, при обробці новин, або результатів

пошукової видачі наукових статей, кількість текстових документів, що вимагають обробки, є дуже велика. Тому велику значимість мають методи, що спрощують роботу з такими об'єктами.

Останнім часом значного поширення набувають нові технології та методи аналізу даних, зокрема методи інтелектуального аналізу, які використовують для виявлення прихованих закономірностей у великих масивах даних. Аналіз наукових робіт показав, що навіть при існуючому розмаїтті методів інформаційного пошуку проблема автоматизованого пошуку документів залишається недостатньо вирішеною.

1. Постановка задачі

Інформація, яка характеризує певний конкретний факт, фактографічну подію або їх сукупність, називається фактографічною. Безперервне накопичення текстових даних привело до необхідності розробки методів інтелектуального аналізу текстів для забезпечення ефективної роботи з великими корпусами текстів. Вся сукупність представлених на сьогоднішній день методів тематичного аналізу тексту можна розділити на дві великі групи: лінгвістичний аналіз та статистичний аналіз. Перший орієнтований на екстракцію смислу тексту за його семантичною структурою, другий – за частотним розподілом слів у тексті. Як правило, в реальних задачах обробки тексту використовується поєднання методик з обох груп з тим чи іншим акцентом.

Пропонована технологія орієнтована на аналіз документів жанру ділової прози, для якої характерні наступні особливості: обмеженість предметної області та мови документів, наявність суворої модельної ситуації (яка визначається характером автоматизації або призначенням Інформаційних технологій), чіткість функцій кожного повідомлення, що дозволяє сконцентрувати аналіз навколо найбільш значущих понять предметної області. Саме такі документи є найважливішими з точки зору комп'ютерної обробки для найрізноманітніших Інформаційних технологій.

Аналіз стану досліджень показав, що найбільш відомою технологією інтелектуальної обробки даних є інтелектуальний аналіз даних (Data Mining) [1]. Основна особливість Data Mining – об'єднання широкого математичного інструментарію і останніх досягнень в сфері інформаційних технологій, розроблених на основі штучного інтелекту, для організації процесу здобуття знань з потоку даних. До найбільш відомих підходів відносять системи на основі нейронних мереж [2], статистичних методів [3-8], нечіткої логіки, методів узагальнення за прикладами об'єктів [4-5] та ін., які забезпечують роботу в середовищах з різними типами даних і можуть працювати з експертом, які не є програмістом.

Реалізація ефективного пошуку фактографічної інформації вимагає вивчення структури предметної області, знаходження її специфічних семантичних ознак. Розробка теоретичних і методологічних аспектів розробки комп'ютерного лінгвістичного забезпечення почалася ще в 60-і роки. Проблеми лінгвістичних засобів для технологій автоматичного аналізу тексту розроблені в працях таких вчених як Шапот М. [2], Гаврилова Т.А. [3], Загоровский И.М. [4], Калинина Е.А.[5], Чубукова И.А. [6], Бондаренко М.Ф. [7], Н.В. Шаронова [8], М.Я. Дворкіна [11] та ін.

Серед зарубіжних фахівців, які присвятили себе дослідженнями в області автоматизованої обробки інформації, слід відзначити: O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, M.C. Suárez-Figueroa, Ju. Apresjan, I. M. Boguslavsky, *J. Makki, A.-M. .Alquier., V. Princ,* L. L. Iomdin, L. L. Tsinman, *Buitelaar P., Cimiano P., Magnini B.* та ін. [9-10, 16-21]. У роботах досліджено питання автоматизованого індексування текстової інформації. Особливо важливою і актуальною проблема автоматизації онтологічного інжинірінгу в системах вилучення знань із текстової інформації. У зв'язку з цим інтерес представляють дослідження таких авторів як O. Corcho, A. Gomez-Perez [9-10].

Автор Шапот М. у своїх дослідженнях наводить приклади використання методів інтелектуального аналізу даних у фінансових додатках і маркетинговому аналізі [2]. Автор Чубукова І.О. описує життєвий цикл онтологій, він збігається з життєвим циклом розробки будь-якої програмної систем, природно, з урахуванням різних способів класифікації етапів життєвого циклу і їх назви [6].

Тенденція сучасних досліджень спрямована саме на впровадження в інформаційних системах лексичних знань. Ряд робіт, як наприклад [16,18,20], присвячені визначенню конкретних лінгвістичних шаблонів (patterns), специфічних для певної предметної галузі і забезпечують якісне виявлення в тексті релевантних даних. Багато робіт цього напрямку присвячені дослідженням «нейронної парадигми». Нейромережевий підхід використовується у величезній кількості завдань — для кластеризації інформації з мережі Інтернет, автоматичної генерації локальних каталогів, уявлення образів (в рекурсивних нейронних мережах). Серед тем, що активно вивчаються останнім часом — неоднорідні нейронні моделі з відносинами подібності. (Heterogeneous Neural Networks with similarity relation) [17-20].

Таким чином, *метою даної роботи* є розробка принципів побудови інформаційної технології

інтелектуального аналізу текстової інформації, яка дозволяє автоматизувати процес обробки слабо структурованих фактографічних ресурсів та вдосконалити процес екстракції знань за рахунок визначення ознак та атрибутів предметних областей. Відмінною рисою такого підходу є орієнтація використовуваних лінгвістичних описів на конкретні предметні знання.

2. Результати

Розвиток інформаційних технологій забезпечив зростання кількості інформаційних ресурсів, більшість яких представлена у неструктурованому текстовому вигляді. Технологія, завдяки якій з'являється можливість формування анотованого опису інформаційних ресурсів, полягає в наступному:

– текст інформаційного ресурсу розділяється на окремі твердження, які описують ту чи іншу ознаку, ситуацію або дію;

– з онтології вибираються поняття, за допомогою яких проводиться опис змісту інформаційного ресурсу;

– обрані твердження перетворюються в триплети «об'єкт-атрибут-значення» з використанням концептів, описаних в онтології.

Будь-яка онтологія повинна грунтуватися на перевірених джерелах знань, а також передбачати повторне використання вже існуючих онтологій для того, щоб уникати дублювання інформації.

Позначимо $\sigma = \{\sigma_0, \sigma_2, ..., \sigma_6, \sigma_{num}, \sigma_{pers}, \sigma_{gend}\}$ — група іменника, де σ_0 – начальна форма (іменник в називному відміннику); $\sigma_2, ..., \sigma_6$ – форми слова в родовому, давальному, знахідному, орудному та місцевому відміннику; $\sigma_{num}, \sigma_{pers}, \sigma_{gend}$ – характеристики (число, особа і рід). Група дієслова (інфінітив, активний стан, пасивний стан, герундій, прийменник, характеристики – перехідне та неперехідне дієслово, а також союз) позначимо через $\mu = \{\mu_0, \mu_{act}, \mu_{pas}, \mu_{grd}, \mu_{prep}, \mu_{trans}, \mu_{nt}, \mu_{conj}\}$.

Розглянемо типові лінгвістичні шаблони, які зустрічаються у тексті, та можливі способи їх онтологічного подання. Наприклад, «літак приземлився», «вчений розробив прилад», «чашка розбилася», «знайти новини», «специфікація стандарту» та ін. Можна виділити наступні лінгвістичні шаблони фактографічного текстового ресурсу:

(1) (σ_s, μ_{act})
(2) $(\sigma_s, \mu_{act}, \sigma_o)$
(3) σ_s, μ_{pas}
(4) $(\sigma_s, \mu_i, \mu_{0j})$, $(\sigma_s, \mu_{0j}, \mu_i)$ або $(\mu_{0j}, \mu_i, \sigma_o)$
(5) $(\sigma_s, \mu_i, \sigma_o), \mu_i \in Serv$
(6) (μ_0, σ_o)
(7) (μ_{grd}, σ_o)
(8) $(\sigma_s, \mu_{prep}, \sigma_o)$
(9) $(\mu_{iact}, \mu_{j\,prep}, \sigma_o)$ або $(\mu_{i\,pas}, \mu_{j\,prep}, \sigma_o)$
(10) (σ_s, σ_o)
(11) $(\mu_{0i}, \mu_{j\,prep}, \sigma_o)$

Позначимо $\Omega = (\psi, \varepsilon, \varsigma, \phi)$ – це онтологія, де ψ – множина об'єктів, ε – відношень, ς – множина допустимих атрибутів (які задаються ім'ям та типом), та ϕ – правил виведення, ψ_s, де ψ_o – суб'єкт та об'єкт відношення, $\tilde{\varepsilon}$ – відношення успадкування. Тоді типові лінгвістичні шаблони матимуть відповідні онтологічні трактовки.

(1) 1. $\exists \varepsilon, \psi_s \in \Omega \,|\, \exists \varepsilon(\psi_s)$
    2. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_s : \exists \varepsilon(\psi_s) \wedge \varsigma_i \in \psi_s$
(2) $\exists \varepsilon, \psi_s, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_s, \psi_o)$
(3) 1. $\exists \boldsymbol{\varepsilon}, \boldsymbol{\psi_s} \in \boldsymbol{\Omega} \,|\, \exists \boldsymbol{\varepsilon}(\boldsymbol{\psi_s})$
    2. $\exists \varepsilon, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_o)$
    3. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_o : \exists \varepsilon(\psi_o) \wedge \varsigma_i \in \psi_o$
(4) 1. $\exists \varepsilon, \psi_s \in \Omega \,|\, \exists \varepsilon(\psi_s)$
    2. $\exists \varepsilon_i, \varepsilon_j, \psi_s \in \Omega \,|\, \exists \varepsilon_i(\psi_s) \wedge \exists \varepsilon_j(\psi_s)$
    3. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_s : \exists \varepsilon(\psi_s) \wedge \varsigma_i \in \psi_s$
(5) 1. $\exists \varepsilon, \psi_s, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_s, \psi_o)$
    2. $\exists \psi_i, \varsigma_j \in \Omega \,|\, \varsigma_j \in \psi_i$
(6) 1. $\exists \boldsymbol{\varepsilon}, \boldsymbol{\psi_O} \in \boldsymbol{\Omega} \,|\, \exists \boldsymbol{\varepsilon}(\boldsymbol{\psi_O})$
    2. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_o : \exists \varepsilon(\psi_o) \wedge \varsigma_i \in \psi_o$
(7) 1. $\exists \varepsilon, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_o)$
    2. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_o : \exists \varepsilon(\psi_o) \wedge \varsigma_i \in \psi_o$
(8) 1. $\exists \varepsilon, \psi_s, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_s, \psi_o)$
    2. $\exists \psi_i, \varsigma_j \in \Omega \,|\, \varsigma_j \in \psi_i$
    3. $\exists \varepsilon, \psi_i, \psi_j \in \Omega \,|\, \exists \varepsilon(\psi_i, \psi_j)$
(9) 1. $\exists \varepsilon, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_o)$
    2. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_o : \exists \varepsilon(\psi_o) \wedge \varsigma_i \in \psi_o$
(10) 1. $\exists \varepsilon, \psi_i, \psi_j \in \Omega \,|\, \exists \varepsilon(\psi_i, \psi_j)$
    2. $\exists \psi_i, \varsigma_j \in \Omega \,|\, \varsigma_j \in \psi_i$
    3. $\exists \varepsilon, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_o)$
(11) 1. $\exists \varepsilon, \psi_o \in \Omega \,|\, \exists \varepsilon(\psi_o)$
    2. $\exists \varepsilon, \varsigma_i \in \Omega \,|\, \exists \psi_o : \exists \varepsilon(\psi_o) \wedge \varsigma_i \in \psi_o$
    3. $\exists \varepsilon_i, \varepsilon_j, \psi_o \in \Omega \,|\, \exists \varepsilon_i(\psi_o) \wedge \exists \varepsilon_j(\psi_o)$

Особлива увага приділяється етапу перевірки онтології шляхом побудови семантичних дескриптів документів та аналізу протиріч, оскільки він є критичним для всієї процедури побудови онтології та представляє основну відмінність запропонованого підходу в порівнянні з відомими методами, при цьому він є не незалежним етапом, а постійним процесом автоматичної корекції та верифікації, що запускається після кожного з етапів.

Вводиться метрика коректності для синтаксичного $\rho_{syn}(w_i, w_j, \Omega_{syn})$ та семантичного зв'язку $\rho_{sem}(\alpha_i, \alpha_j, \Omega_{sem})$, яка показує, наскільки коректний побудований зв'язок між концептами α_i та α_j, та, відповідно, групами слів, які представляють їх в тексті, $\{w_i\} \to \alpha_i$ и $\{w_j\} \to \alpha_j$.

Синтаксична коректність:

$$\rho_{syn}(w_i, w_j, \Omega_{syn}) = \begin{cases} 1, & \exists \varepsilon_{syn}(w_i, w_j) \in \Omega_{syn}; \\ -1, & \bar{\exists} \varepsilon_{syn}(w_i, w_j) \in \Omega_{syn}. \end{cases}$$

Тоді ступень коректності слова w_l:

$$\rho_{syn}(w_l) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{N_{D_i}^{w_j}} \rho_{syn}(w_l, w_j, \Omega_{syn})}{N_{D_i}^{w_j}}, \forall j : w_j, w_l \in T_{D_i}$$

Загальна синтаксична коректність концепту онтології:

$$\rho_{syn}(\alpha_l) = \sum_i \rho_{syn}(w_i), \{w_i\} \to \alpha_l .$$

Концепт онтології виключається у випадку, коли

$$\rho_{syn}(\alpha_l) << \frac{\sum_{j=l}^{N\alpha_l} \rho_{syn}(w_j)}{N\alpha_l}, \quad \forall j \neq l ,$$

$N\alpha_l$ – число документів, у яких присутній концепт онтології. Коефіцієнти $\rho_{syn}(w_j)$ для решти термінів перераховуються без врахування $\{w_l\} \to \alpha_l$. У випадку декількох альтернатив термін α_i^m вважається надійним, якщо

$$\rho_{syn}(\alpha_1^m) >> \rho_{syn}(\alpha_1^i), \forall i \neq m, i = \overline{1, N\alpha_l} .$$

Аналогічно, семантична метрика концепту α_l на основі слів, які його представляють $\{w_l\}$, розраховується, як

$$\rho_{sem}(\alpha_l) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{N_{D_i}^{w_j}} \rho_{sem}(\alpha_l, \alpha_j, \Omega_{syn})}{N_{D_i}^{w_j}}, .$$

$$\forall j : \{w_j\} \to \alpha_j, w_j, w_l \in T_{D_i}$$

Вклад концепту в розуміння документу:

$$K_{sem}(\alpha_l, D_i) = \frac{\sum_{j=1}^{k} \rho_{sem}(\alpha_l, \alpha_j, \Omega)}{k}, \quad .$$

$$\forall j : w_j, w_l \in T_{D_i}, \{w_j\} \to \alpha_j, \{w_l\} \to \alpha_l$$

Термін розуміється незадовільно та погіршує загальне розуміння тексту у разі:

$$\sum_{i=1}^{N} K_{sem}(\alpha_l, D_i) << 0 .$$

Запропонований підхід до автоматизованої побудови онтології дозволяє домогтися наступних основних переваг в порівнянні з існуючими методами:

1. Не потрібна побудова початкової онтології предметної області людиною-експертом в якості базису для подальшої роботи.

2. Не потрібна попередня обробка людиною-експертом документів предметної області (включаючи стандартизацію шаблонів, перетворення форматів, попередню розмітку тексту, складання вручну словника термінів предметної області та ін.).

3. Процес побудови онтології повністю прозорий для користувача, обґрунтування усіх рішень, що приймаються, логіка та оцінка можуть бути простежені.

4. Процес побудови онтології не залежить від язика документу, за винятком підтримки синтаксичних онтологій для різних мов.

5. Процес побудови онтології ітеративний, завжди існує зворотний зв'язок з можливістю перевірити семантику онтології, яка згенерована автоматичним шляхом, коли вже побудована частина онтології сама є основою для аналізу семантичної коректності запропонованих змін та доповнень. При цьому процес саморегулювання автоматизовано й може обходитися без людини-експерта.

6. Аналіз та вичленення термінів з врахуванням їх семантики відбувається у рамках всього корпусу текстів, він не обмежується аналізом індивідуальних пропозицій.

7. Алгоритм може працювати як автономно, так и в інтерактивному режимі, причому користувач може вплинути на формування рішення на кожному з етапів роботи.

Технологія фактографичічного пошуку заснована на представленні змісту тексту у формі семантичної мережі. Семантична мережа містить значимі слова і словосполучення, які зв'язані одне з одним різними типами синтактико-семантичних зв'язків. Елементарна семантична мережа представляє результат синтаксичного аналізу та постсинтаксичних трансформацій дерева синтаксичних залежностей між словами у окремих реченнях. Повна семантична мережа тексту є сукупністю окремих семантичних мереж, які відповідають реченням.

Пошук факта – це пошук у семантичній мережі тексту такої підмережі, яка є ізоморфною до одного з шаблонів. Якщо підмережа знайдена, факт вважається встановленим, після чого здійснюється вилучення сутностей та їх маркування ролями, які задані у відповідних вузлах лінгвістичного опису.

Схему розробленої інформаційної технології представлено на рисунку.

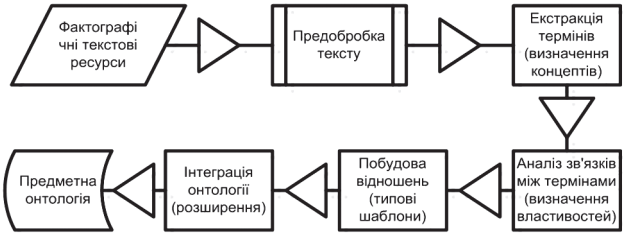

Схема інформаційної технології інтелектуального аналізу фактографічної інформації

Реалізація — створення онтології з допомогою деякого формального мови опису (OWL, OIL, RDF, SPARQL і т.д.) на основі концептуальної моделі. Ми пропонуємо використовувати мови сімейства OWL (OWL, OWL 2) як такі, що найбільш розвинену семантику.

Технологія METHONTOLOGY пропонує використовувати фреймових системи або дескриптивні логіки. їх пропозиція узгоджується з нашою рекомендацією по використання мови OWL, тому що дескриптивна логіка SHOIN (D) закладена в його основу.

Технологія On-To-Knowledge пропонує формалізувати концептуальну модель з використанням мови OIL, при цьому продовжувати процес виявлення знань (яким чином — не уточнюється) для збагачення онтології деталями і менш значущими концептами і відносинами.

У технології Enterprise Model Approach дається посилання на думку Грубера [18], з якого випливає, що стадії виявлення знань (в яку входить концептуалізація) і реалізації можна об'єднати. Однак сам автор вважає, що будь-яка методологія, яка об'єднує ці дві стадії повинна давати серйозне обґрунтування такого кроку.

У будь-який довільний проміжок часу будь-який кінцевий користувач може звернутися до створеної онтології. В результаті своїх дій він повинен отримати достовірні і актуальні дані. Основним дією цієї стадії є повноцінне і своєчасне зміна онтології відповідно до змінами реального світу. Онтологія повинна еволюціонувати, для чого процеси поновлення-видалення-вставки вимагають строгих правил, що є основою процесу супроводу.

Основою інформаційної технології семантичного пошуку є застосовування семантичних мереж, які характеризують значення слів людської природної мови та зв'язки між уявленнями, що окреслюються ними. Не досить опрацьовано і слабо виражено застосування підтримки української морфології. Семантична мережа словника мови включає близько 40 тисяч семантичних груп в базовому варіанті. Це дозволяє користувачу вводити запит природною мовою і система сама організовує пошук всіх документів, контекст яких збігається з контекстом користувацького запиту. Застосування семантики дозволяє враховувати загальний контекст документа.

Таким чином, інформаційна технологія інтелектуального аналізу фактографічної інформації удосконалює та доповнює існуючий підхід обробки текстових даних і не суперечить існуючій практиці, що свідчить про його практичну цінність та ефективність використання.

Висновки

У статті проаналізовано існуючі інформаційні технології, моделі та методи обробки фактографічних даних у слабо структурованих текстових ресурсах, сформульовано основні вимоги до розробки інформаційної технології інтелектуального аналізу фактографічних ресурсів. Враховано особливості інтелектуального аналізу фактографічної текстової інформації. Сформовано підхід до видобування фактографічних даних з текстових джерел на основі використання онтологічних специфікацій. Описано використання онтологій для опису процесів інтеграції фактографічної інформації. Запропоновано використання нового напівавтоматичного методу, основаного на принципах обробки природної мови, для розбудови та розширення базової предметної онтології. Запропонований підхід до автоматизованої побудови онтології дозволяє удосконалити та доповнити існуючий підхід обробки текстових даних, що свідчить про його практичну цінність та ефективність використання.

Доцільно використання отриманих результатів у розробці систем, які здатні забезпечити розв'язання задач поділу на частини, виділення ключових слів та опрацювання множини документів. Особливо це актуально для наукових установ та бібліотек.

Список літератури:

1. Оперативна аналітична обробка даних: концепції і технології [Електронний ресурс] / Іванівський держ. енергетичний ун-т. — Режим доступу URL: http://citforum.ru/seminars/cis99/sch_03.shtml — 2009. – Загл. з екрану. **2.** *Шапот М.* Інтелектуальний аналіз даних в системах підтримки прийняття рішень [Текст] / М. Шапот // Журн. відкриті системи. — 2008. — №1. — С. 30-35. **3.** *Гаврилова Т.А.* Бази знань інтелектуальних систем [Текст]: навч. / Т.А. Гаврилова, В.Ф. Хорошевський. — СПб: Пітер, 2000. — 384 с. **4.** *Загоровський І.М.* Вибір алгоритму навчання в системах придбання знань з даних [Текст]: матеріали 12-ої національ. конф. зі штучного інтелекту з міжнар. участю (КВІ 2010), — М .: Физматлит, Т. 1, 2005. — С. 131-135. **5.** *Калініна Е.А.* Застосування технології Data Mining для автоматизованої побудови баз знань інтегрованих експертних систем [Текст] / Е.А. Калініна, Г.В. Рибіна.: матеріали 8-ої національ. конф. зі штучного інтелекту з міжнар. участю (КВІ 2002), — М .: Физматлит, Т. 1, 2002.

— С. 119-127. **6.** Чубукова І.А. Data Mining [Текст] / І.А. Чубукова. — М.: БИНОМ. Лабораторія знань, Інтернет-університет інформаційних технологій — ІНТУІТ.ру, 2008. — 384 с. **7.** *Бондаренко, М. Ф.* Теорія інтелекту [Текст]: навч. / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. — Харків: Компанія СМІТ, 2006. — 576 с. **8.** *Шаронова, Н.В.* Автоматизовані інформаційні бібліотечні системи: завдання обробки інформації [Текст]: монографія, Нар. Укр. Акад. / Н.В. Шаронова, Н.Ф. Хайрова; [Каф. інформац. технологій і документознавства]. — Х., 2003 — 120 с. **9.** *Gomez-Perez A.* Ontological Engineering: what are ontologies and how can we build them [Текст] / O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, // In Cardoso (ed) Semantic Web: Theory, Tools and Applications. — IDEA Group. — 2007. — Pages 44-70. **10.** *Suárez-Figueroa,* How to write and use the Ontology Requirements Specification Document [Текст] / M.C. Suárez-Figueroa, A. Gómez-Pérez, Boris Villazón-Terrazas // Proceedings of the 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2009). – ISBN: 978-3-642-05150-0. LNCS 5871. — Volume: Part II. — 2009. — Pages: 966-982. **11.** *Дворкіна М. Я.* Бібліотечне обслуговування: нова реальність / М. Я. Дворкіна. — М.: МГУКИ, 2000. — 48 с. **12.** *Воронцов А. В.* Гібридні алгоритми лексико-граматичного аналізу тексту / А. В.Воронцов // Штучний інтелект. — 2006. — № 4. — С. 593-602. **13.** *Corcho O.* Methodologies, tools, and languages for building ontologies. Where is their meeting point? / *O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez* // Data & Knowledge Engineering, 46, 2003. **14.** *Єрмаков А.Е.* Автоматизація онтологічного інжинірингу в системах добування знань з тексту / А. Е. Єрмаков // праці Міжнародної конференції Діалог'2008. — Москва, Наука, 2008. — С. 136-140. **15.** *Каніщева О. В.* Використання методів Data Mining і Text Mining для обробки текстової інформації в інформаційних системах / О. В. Каніщева, Сайед Мохаммад Таухид Сіддікі, Н. В. Шаронова // Біоніка інтелекту. — Харків: ХНУРЕ, 2005. — № 2 (63). — С. 22-26. **16.** *Apresjan Ju.* Lexical Functions in NLP: Possible Uses / Ju. Apresjan, I. M. Boguslavsky, L. L. Iomdin, L. L. Tsinman // Computational Linguistics for the New Millenium: Divergence or Synergy? : proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg. – Frankfurt am Main, 2002. – P. 55-72. **17.** *Buitelaar P.* Ontology Learning from Texts: An Overview. / Buitelaar P., Cimiano P., Magnini B. In Ontology Learning from Text: Methods, Evaluation and Applications, 2005, Vol. 123, Eds. IOS Press. P. 634.-265. **18.** *Simperl E.* Achieving Maturity: the State of Practice in Ontology Engineering / E. Simperl, M. Mocho // In International Journal of Computer Science and Applications, Technomathematics Research Foundation Vol. 7 No. 1, pp. 45-65, 2010. **19.** *Makki J.* Semi Automatic Ontology Instantiation in the domain of Risk Management / J. Makki, A.-M. .Alquier., V. Prince // In IFIP, Advances in Information and Communication Technology. 2008. Volume 288. p. 254. **20.** *Builеaar P.* Topic extraction from scientific literature for competency management [Текст]: / Builеaar P., Eigner T. In The 7th International Semantic Web Conference PICKME 2008, 27 octobre Karlsruhe, Germany, p. 55-67. **21.** *Zhou, L.* Ontology Learning: State of the Art and Open Issues/ Zhou, L. Information Technology and Management, 2007, 8(3), p. 241-252.

Надійшла до редколегії 07.06.2018