УДК 004.94:159.95

О.В. Бісікало, к.т.н., **І.О. Назаров**

ОБРАЗНИЙ АНАЛІЗ ТЕКСТОВОЇ ІНФОРМАЦІЇ З WIKIPEDIA

Вінницький національний технічний університет, Вінниця, e-mail: obisikalo@gmail.com

У роботі розглянуто побудову інформаційної технології образного аналізу текстів з використанням бази знань Wikipedia. В основу технології покладено підхід до моделювання асоціативного мислення людини, поняття мовного образу, інфологічної системи та образного сенсу природномовного контенту.

***Ключові слова**: образний аналіз тексту, мовний образ, Wikipedia, Java-Based Wikipedia Library.*

Вступ. В умовах швидкого збільшення обсягів природномовного контенту (ПМК) високу актуальність набувають методи машинного аналізу й обробки текстової інформації. При розв'язанні даних задач важливим є комп'ютерне моделювання асоціативного образного мислення людини. Поряд з іншими підходами в роботі [1] запропонований інфологічний, який доцільно використати для побудови технології образного аналізу текстової інформації.

Побудова інформаційної технології. Розглянемо інформаційну технологію з урахуванням таких формальних обмежень на об'єкт дослідження – процеси асоціативного образного мислення людини (рис. 1):

Рис. 1. Взаємозв'язок системи обмежень та результатів формалізації об'єкту дослідження

a) на рівні концептуальної ідеї дослідження – формулювання понять онтогенетичного принципу, мовного образу, інфологічної системи (ІС) та образного сенсу електронного контенту (ЕК); розмежування в останньому сенсу-властивості та сенсу-параметра;
b) на рівні формальної метамови – закладення онтогенетичного принципу в основу функціонування ІС; представлення сенсу-властивості у вигляді семантичної мережі асоціативної мережі образів (АМО), а сенсу-параметра на основі нечітких відношення та простору образного сенсу;
c) на рівні теорії образного сенсу природної мови – формалізація ІС як конструктивної моделі об'єкта дослідження; аксіоматика прикладної теорії першого порядку та аксіоматика простору з нечіткою мірою;
d) на рівні моделі асоціативного образного мислення – механізм онтогенезу ІС на основі кібернетичної інтерпретації нейропсихологічних даних; склад блоків, функцій та образних потоків моделі ІС; представлення нескінченних множин скінченними; вибір основ та сигнатури алгебраїчної системи;
e) на рівні комплексу інфологічного моделювання образного мислення (КІМОМ) – обмеження функцій КІМОМ класифікацією образного пошуку та функціональними вимогами до ІС; послідовна побудова графових моделей для базових типів образного пошуку; обмеження моделей прикладних задач образного мислення класифікацією асоціативних відношень.

Розглянемо можливості побудови інформаційної технології на основі отриманих 9 результатів формалізації об'єкту дослідження, позначених на рис.1 цифрами та вище окресленою 5-рівневою системою обмежень. Будемо виходити з того, що вхідна інформація для формування бази знань ІС має надходити безпосередньо з ПМК, а саме з текстів, попередньо відібраних експертом з предметної області. У межах інформаційної технології базовим носієм знань запропоновано вважати модель асоціативної пам'яті людини у вигляді АМО. Можливими шляхами накопичення такого типу знань принципово можуть бути:

- використання існуючих асоціативних словників, отриманих в результаті широкомасштабного вільного асоціативного експерименту;
- «вирощування» власної асоціативно-вербальної мережі на основі запропонованої онтогенетичної концепції моделювання образного мислення.

Пропонується напівавтоматична методика внесення інформації з ЕК в АМО ІС:

1. Користувачу надається можливість ввести текст речення в систему.
2. Кожному слову речення ставиться у відповідність один з мовних образів:
   a) автоматично – у випадку, коли саме таке слово вже існує в пам'яті системи;
   b) напівавтоматично – шляхом вибору користувачем з списку існуючих образів;
   c) шляхом ручного введення користувачем нового образу – у випадку відсутності потрібного образу в системі.

З метою прив'язки вибраного слова пари до образу у випадку 2б для кожного слова речення складається ранжований список найбільш схожих зовні вербальних позначень з образів *Image*, що вже існують в словнику. При цьому користувач може вибрати в меню існуючу опцію зі списку, при необхідності відкоректувати відповідну статтю словника образів або ввести абсолютно нову.

Для введення асоціативних пар синтагми в систему користувачеві надається можливість задати кількість пар, а потім вибрати кожне слово пари в меню, складеного із слів поточного речення *Event*.

Питальний займенник між образами пари задається двома способами:

- спочатку можна вибрати тип зв'язку з 7 кортежів відношення *Link* (визначення, присудок, підмет, обставина місця, обставина часу, обставина, доповнення), тоді вибраний тип стає фільтром, і кількість можливих займенників *Inter-Pronoun* зменшується;
- якщо спочатку вибрано питання як кортеж з *Inter-Pronoun*, то для контролю користувачеві автоматично демонструється відповідний йому тип зв'язку з *Link*, оскільки однакові займенники можуть відноситися до різних типів, наприклад питання «що?» ставиться як до підмету, так і до доповнення.

Розглянута методика напівавтоматичного введення природномовної інформації позначена як блок 3 на структурній схемі інформаційної технології (рис. 2).

Через блок (3) експерти з предметної області знань (1) вносять до бази знань системи на основі АМО (5) окремі речення з попередньо відібраного природно-мовного ЕК предметної області (2). Прискорення цієї процедури забезпечується блоком автоматичного введення ПМК (4) як альтернативним шляхом досягнення значного обсягу семантичної мережі АМО. Для останнього блоку згадане обмеження на рівні практичної реалізації інформаційної технології набуває критичного характеру – якість обслуговування користувачів технології (7) ІС безпосередньо залежить від якості відібраного природно-мовного контенту.

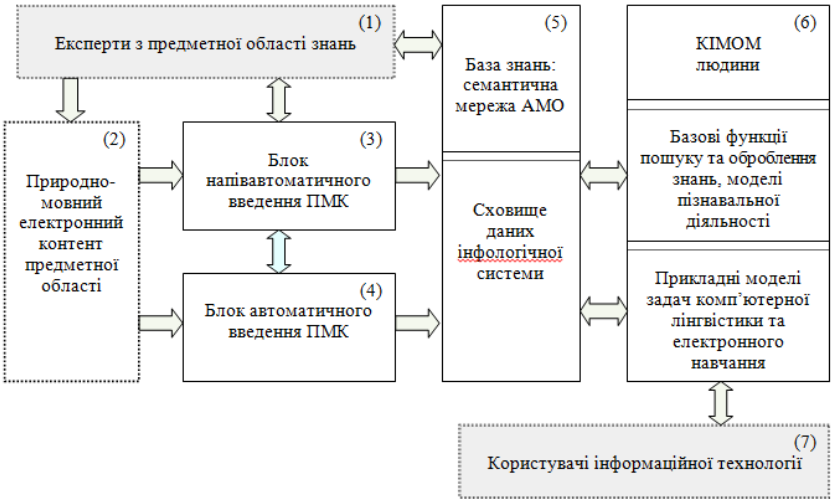

Рис. 2. Структурна схема інформаційної технології

На відміну від розглянутої методики побудови блоку (3) автоматичне введення ЕК до ІС, яка представлена на схемі блоками (5) та КІМОМ людини (6), у значно більшій мірі залежить від мови контенту. Для підвищення рівня автоматизації в системі потрібно послідовно застосувати алгоритми, представлені в табл. 1.

Такий стан речей пов'язаний з тим, що ключовим елементом у розглянутому процесі автоматизації є парсер – синтаксичний аналізатор. Загалом невисока якість роботи парсерів природних мов на даний час помітно вища, наприклад, для англійської мови, ніж для афіксних мов – української або російської. Блок бази знань (5) в якості інформаційної основи використовує сховище даних ІС. Користувачі інформаційної технології (7) мають можливість отримувати результати як базових, так і прикладних моделей для розв'язання відповідних задач комп'ютерної лінгвістики та електронного навчання.

З метою повної функціональної реалізації задач потрібно додатково побудувати такі прикладні моделі: отримання формальних ознак результатів пізнавальної діяльності, конструювання образу-

рішення проблемної ситуації, генерація множини повідомлень щодо стану та потреб ІС, побудова різних типів відповідей на питання.

Таблиця 1

Склад алгоритмів блоку автоматичного введення ПМК до ІС

№ з/п	Назва алгоритма (модуля)	Вхід	Вихід
1.	Відділення синтагми з тексту як речення або автономної частини речення	Текст	Список синтагм
2.	Перетворення синтагми у список слів, що ідентифікуються модулем словника	Окрема синтагма	Список відомих системі слів
3.	Побудова словника мовних образів обраної мови	Окреме слово	Ієрархія: слово → словоформа → лексема → мовний образ
4.	Використання парсеру обраної мови для отримання дерева підлеглості (залежностей)	Окрема синтагма	Дерево орієнтованого графу, вершини якого – МО, а ребра – питальні займенники
5.	Образна індексація предметної області	Корпус відібраних текстів предметної області	Семантична мережа АМО з образним індексом корпусу

З метою практичної реалізації описаної інформаційної технології пропонується застосування зручного інструментарію для доступу до бази знань Wikipedia на мові програмування Java – Java Wikipedia Library.

Java-Based Wikipedia Library. Wikipedia – це багатомовна, вільно доступна енциклопедія, розроблена спільними зусиллями добровольців. Її об'єм збільшується швидкими темпами, і з приблизно 7,5 мільйонами статей більш ніж 250 мовами світу вона стала найбільшою колекцією вільно доступних знань. Статті у Wikipedia формують сильно зв'язану базу знань, збагачену новою системою категорій зі спільними тегами, яка представляє собою тезаурус. Таким чином, Wikipedia містить багату лексичну семантичну інформацію, аспекти якої докладно описані в [2]. Ця інформація включає в себе знання про назви сутностей, область конкретних термінів і сенсів слів. Крім того, система статей Wikipedia може бути використана як словник синонімів, варіантів написання і скорочень.

Wikipedia є одним з видів спільних баз знань (іншими прикладами можуть слугувати dmoz і Citizendium). Основні відмінності властивостей спільних та лінгвістичних баз знань наведені в таблиці 2:

Таблиця 2

Порівняння спільних та лінгвістичних баз знань

	Лінгвістичні бази знань	Спільні бази знань
Розробники	Лінгвісти	В основному непрофесійні добровольці
Конструктивний підхід	Прихильники теоретичної моделі або зібрання доведень	Прихильники необов'язкових директив
Витрати на розробку	Значні	Відсутні
Оновленість	Швидко стають застарілими	В основному оновлюються
Розмір	Обмежений вартістю розробки	Великі або швидко зростаючі
Доступні мови	Основні мови	Багато мов

Як слідує з наведеної таблиці, основними перевагами спільних баз знань є відсутність витрат на розробку, швидке оновлення даних, великі обсяги інформації та доступність великої кількості різних мов.

Оскільки Wikipedia використовується для задач обробки природної мови великої розмірності, необхідним є ефективний програмний доступ до знань у ній. В технічному університеті німецького міста Дармштадт був розроблений інтерфейс програмування додатків на мові Java - Java-Based Wikipedia Library (JWPL). JWPL наразі є вільно доступним для дослідницьких цілей.

Оригінальна структура бази даних Wikipedia оптимізована для пошуку статей за ключовими словами, які проводяться мільйонами користувачів кожного дня. Тим не менше, програми, призначені для досліджень обробки природної мови, повинні підтримувати широкий діапазон шляхів доступу, в тому числі ітерації по всіх статтях, синтаксис запиту, а також ефективний доступ до такої інформації, як посилання, категорії та ін. Таким чином, JWPL оперує оптимізованими базами даних

(як показано на рисунку 3), які створені зі сховищ баз даних, доступних з Wikimedia Foundation. Переваги архітектури даної системи наступні: обчислювальна ефективність дозволяє розв'язання задач обробки природної мови великої розмірності; відтворювані результати дослідження; зручне використання об'єктно-орієнтованого програмування.

Достовірні результати експериментів є прямим наслідком використання фіксованих сховищ баз даних на відміну від онлайн Wikipedia, які з високою ймовірністю будуть змінені між двома запусками певних експериментальних налаштувань.

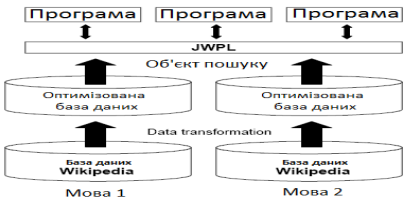

Рис. 3. Архітектура системи JWPL

Ефективність обчислень також є наслідком доступу до бази даних з використанням механізмів швидкого пошуку. Дані з бази даних безпосередньо відображаються в Java-обєктах за допомогою об'єктно-реляційного відображення. Це також означає, що JWPL не обмежується використанням певної бази даних і може працювати на верхньому рівні найбільш поширених систем управління базами даних.

Розробка об'єктно-орієнтованого програмного інтерфейсу сконцентрована навколо об'єктів: Wikipedia, Page і Category. Об'єкт Wikipedia використовується для встановлення з'єднання з базою даних і для одержання об'єктів Page і Category. JWPL підтримує пошук за ключовими словами або за допомогою інтерфейсу запитів, який дозволяє одержання підмножини статей чи категорій в залежності від таких параметрів, як кількість лексем у статті або кількість вхідних посилань. Об'єкт Wikipedia також дозволяє перебір статей, категорій, посилань і неоднозначних сторінок.

Об'єкт Page представляє собою звичайну статтю Wikipedia, посилання на статтю або неоднозначну сторінку. Кожен об'єкт Page забезпечує доступ до тексту статті (з розміткою інформації або як простий текст), переданих категорій, вхідні і вихідні посилання статті, а також усі перенаправлення, які посилаються на цю статтю.

Об'єкт Category представляє категорії Wikipedia і дозволяє доступ до статей у категоріях. Так як категорії у Wikipedia сформовані в тезаурус, об'єкт Category також надає засоби для одержання батьківських і дочірніх категорій, а також усіх рекурсивно зібраних нащадків. JWPL також забезпечує об'єкт CategoryGraph, який, наприклад, дозволяє знайти найкоротший шлях між двома заданими категоріями.

Останні версії JWPL містять парсер для розмітки мови Wikipedia. Парсер легко дозволяє визначити і отримати більш «тонку» інформацію зі статей Wikipedia, наприклад, розділи, параграфи, шаблони, посилання, текстові посилання, контекстні посилання, списки і таблиці. Рисунок 4 показує структуру статті Wikipedia «Natural language processing», яка проаналізована парсером.

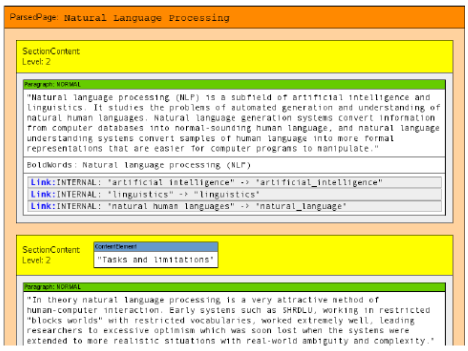

Рис. 4. Візуалізація парсером структури статті Wikipedia

Образний аналіз текстової інформації в першу чергу передбачає розв'язання задач семантичного аналізу ПМК [3], серед яких основними є наступні: виявлення мовних образів; вирішення лексичної багатозначності; побудова семантичного графу; реалізація алгоритмів на семантичному графі.

Виявлення мовних образів. З метою формалізації тексту його подають в адаптованому вигляді, у якому відсутні одиниці мови (слова), які фактично не викликають в уяві жодного образу. Для кожного значущого слова вибирається набір статей Wikipedia, які потенційно можуть описувати його значення. Програма дозволяє визначити кількість посилань на сторінку, кількість посилань з даної сторінки на іншу, кількість категорій, які відносяться до сторінки.

Вирішення лексичної багатозначності. З метою уникнення мовного явища полісемії кожному слову ставиться у відповідність сукупність мовних образів – його можливих сенсів. За ключовим словом встановлюються наступні характеристики: неоднозначність сторінки, наявність перенаправлення на іншу сторінку. На даному етапі забезпечується знаходження перенаправлень, категорій, вхідних і вихідних посилань сторінки.

Побудова семантичного графу. В роботі [3] описана процедура побудови семантичного графу на основі попередньої формальної обробки текстового матеріалу. Така обробка полягає в приведенні тексту до адаптованого вигляду, яка проводиться на першому етапі

семантичного аналізу. Величина сили семантичного зв'язку визначається кількістю поєднань пари образів в тестовому прикладі. Таким чином утворюється асоціативна мережа образів.

При появі невідомого системі слова знаходиться відповідна категорія Wikipedia, аналізуються усі сторінки категорії і таким чином відбувається поповнення знань системи про нові образи. Текст програми реалізації вибору шуканих сторінок наведений в лістингу 1. Подібний підхід дозволяє формувати тезаурус мовних образів, побудова якого визнається пріоритетним напрямком подальшої роботи.

Лістинг 1

```
package de.tudarmstadt.ukp.wikipedia.api.tutorial;
import de.tudarmstadt.ukp.wikipedia.api.DatabaseConfiguration;
import de.tudarmstadt.ukp.wikipedia.api.Page;
import de.tudarmstadt.ukp.wikipedia.api.WikiConstants;
import de.tudarmstadt.ukp.wikipedia.api.Wikipedia;
import de.tudarmstadt.ukp.wikipedia.api.exception.WikiApiException;
import de.tudarmstadt.ukp.wikipedia.api.exception.WikiPageNotFoundException;
public class T2_PageInfo implements WikiConstants {
public static void main(String[] args) throws WikiApiException {
    DatabaseConfiguration dbConfig = new DatabaseConfiguration();
    dbConfig.setHost("SERVER_URL");
    dbConfig.setDatabase("DATABASE");
    dbConfig.setUser("USER");
    dbConfig.setPassword("PASSWORD");
    dbConfig.setLanguage(Language.german);
    Wikipedia wiki = new Wikipedia(dbConfig);
    String title = "Hello world";
    Page page;
    try {
       page = wiki.getPage(title);
    } catch (WikiPageNotFoundException e) {
       throw new WikiApiException("Page " + title + " does not exist");  }
    System.out.println("Queried string      : " + title);
    System.out.println("Title               : " + page.getTitle());
    System.out.println("IsDisambiguationPage : " + page.isDisambiguation());
    System.out.println("redirect page query  : " + page.isRedirect())
    System.out.println("# of ingoing links   : " + page.getNumberOfInlinks());
    System.out.println("# of outgoing links  : " + page.getNumberOfOutlinks());
    System.out.println("# of categories      : " + page.getNumberOfCategories()); }}
```

Реалізація алгоритмів на графі. В роботі [3] подана формальна постановка задачі визначення сенсу як виділення певного підграфу в даному семантичному графі. У якості перспективного напрямку дослідження в даній області пропонується розглядати алгоритми пошуку на графах з їх удосконаленням та врахуванням вказаних властивостей шуканого розв'язку, що дозволить розв'язання поставленої задачі.

Висновки. Отже, кінцеві результати запропонованої інформаційної технології повністю визначає система формальних обмежень, що сформульована на 6-ти рівнях – від концептуальної ідеї та формальної метамови до практичної реалізації з урахуванням всіх отриманих наукових результатів. Розроблена схема інформаційної технології образного аналізу текстів на основі моделювання асоціативного образного мислення людини забезпечує внесення попередньо відібраного природномовного матеріалу в напівавтоматичному або автоматичному режимах, що імітує природний шлях накопичення знань людиною, побудову бази знань у вигляді семантичної мережі АМО та розв'язок ряду актуальних задач комп'ютерної лінгвістики та електронного навчання на основі базових та прикладних функцій КІМОМ. Для програмної реалізації запропоновано використання інструментарію JWPL, який дозволяє ефективний доступ до бази знань Wikipedia.

Список літературних джерел

1. Бісікало О.В. Інфологічний підхід до моделювання образного мислення людини / О.В. Бісікало // Вісник СумДУ (Серія "Технічні науки"). – 2009. – № 2. – С. 15–20.

2. Zesch T. Comparing Wikipedia and German WordNet by evaluating semantic relatedness on multiple datasets / T. Zesch, I. Gurevych // The Annual Conference of North American Chapter of the Association for Computational Linguistics, 2007. – pp. 205-208.

3. Квєтний Р.Н. Визначення сенсу текстової інформації на основі моделі розповсюдження обмежень / Р.Н. Квєтний, О.В. Бісікало, І.О. Назаров // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2012. – № 1. – С. 93-96.