РОЗДІЛ V. ІНФОРМАЦІЙНО-КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 004.052:004.275

В.В. Литвинов, д-р техн. наук
О.П. Мойсеєнко, аспірант
Чернігівський державний технологічний університет, м. Чернігів, Україна

ПРИЙНЯТТЯ РІШЕНЬ ПРИ СЕМАНТИЧНОМУ РОЗБОРІ ТЕКТОВОГО ДОКУМЕНТА

У статті розглядаються питання, пов'язані з пошуком можливості вдосконалення існуючого методу кластеризації текстових документів (ЛСА) до рівня семантичного розбору. Проведено огляд систем обробки текстових документів на природній мові. Визначені необхідні умови для вдосконалення методу. Зроблено підтверджувальний висновок про можливість реалізації методу семантичного розбору, взявши за основу метод кластеризації.

Ключові слова: *автоматична обробка тексту, аналіз тексту, зміст тексту, семантичний розбір.*

В статье рассматриваются вопросы, связаные с поиском возможности улучшения существующего метода кластеризации текстовых документов (ЛСА) до уровня метода семантического разбора. Проведен обзор систем обработки текстовых документов на естественном языке. Определены нужные условия для улучшения метода. Дан утвердительный вывод в пользу возможности реализации метода семантического разбора, взяв за основу метод кластеризации.

Ключевые слова: *автоматическая обработка текста, анализ текста, содержание текста, семантический разбор.*

In this article the questions that are associated with the search for opportunities to improve the existing method of clustering text documents (LSA) to the level of the Semantic analysis. A review systems handle text documents in natural language . Determine the proper conditions for the improvement of the method. Is affirmative conclusion in favor of the feasibility use of the method of semantic analysis, based on the clustering method.

Key words: *automatic text processing, text analysis, text contents, semantic analysis.*

Вступ. Завдання аналізу та прийняття рішень щодо авторської частки текстової інформації в матеріалах наукових робіт потребують обробки великого об'єму текстових даних, які в наш час масово представляються у вигляді електронних документів. Прикладні системи підтримки прийняття рішень (СППР) мають можливість використовувати знання, що містяться в текстових документах. Однак спершу потрібно в автоматичному режимі вилучити знання із тексту, підтримуючи високий рівень продуктивності операцій аналізу текстових документів та процедур пошуку знань, а також провести їх верифікацію по базі знань чи іншим джерелам. Наприклад, мережа Інтернет є найбільшим сховищем електронної текстової інформації. Відомо, що автоматичний аналіз тексту має декілька типів, зорієнтованих на текстові рівні – фонетичний, морфологічний, семантичний. Як стверджує Ю.М. Марчук [1], наразі момент семантико-зорієнтований підхід лежить в основі більшості сучасних технологій аналізу тексту. Різнорідна структурованість текстових документів з науковим змістом та часті посилання на інші джерела заважають достовірно оцінити автентичність та змістовність таких авторських робіт. Виникає ситуація, коли людина, що приймає рішення про авторську приналежність наукової роботи, серед масиву доступних їй текстових документів не в змозі якісно і за короткий час їх опрацювати з мінімальними похибками без використання спеціалізованих програмних засобів. Структурно-семантичні засоби аналізу логічної зв'язності традиційно є суб'єктивною характеристикою тексту. Основна складність полягає у виявленні математичних залежностей характерних для змісту текстового документа в цілому чи окремих його структурних елементів. Саме тому виділення змістовних характеристик з електронного тексту на природній мові є складним завданням. Система підтримки прийняття рішень на основі семантичного розбору текстових документів могла б більш ефективно, ніж у ручному режимі, вирішити цю проблему. Нижче розглянуто декілька систем автоматичної обробки тексту, які вирішують згадану проблематику, однак у своїй роботі вони викори-

стовують словники та тезауруси, що на порядок знижують швидкість їх роботи та не гарантують безпомилкових результатів. У цій статті, розглядається можливість реалізації безсловникового підходу виявлення унікальності змісту текстового документа та фіксації змістовних співпадінь з тексовими елементами бази знань. Такий підхід, в змозі значно підвищити швидкість систем автоматичної обробки тексту, залишаючись на прийнятному рівні допустимих похибок аналізу.

Аналіз останніх досліджень та публікацій. Аналіз останніх публікацій показав, що сьогодні на практиці існує невелика частка програмно реалізованих продуктів аналізу та семантичного розбору для природно-мовних текстів. Мала кількість семантичних систем пов'язана, в першу чергу, зі складністю їх реалізації. Існують проекти, роботи над якими тривають більше десятка років. Нижче наведено список систем автоматичної обробки текстових даних на природній мові з коротким описом їх можливостей. Варто додати, що подібні системи розробляються під потреби замовника та є ефективними при роботі з електронними тексовими документами конкретного направлення та специфіки.

1. Інформаційно-аналітична система "Аріон" [2].

Система дозволяє працювати як зі структурованими (xml-документи), так і з неструктурованими (тексти природною мовою) джерелами інформації. На вході лінгвістичний процесор системи отримує текстовий документ. Результатом його праці є массив фактографічної інформації, що в подальшому використовується для виявлення схожих і співпадаючих об'єктів. Виділення фактографічної інформації здійснюється за допомогою спеціалізованих правил та словників. Ефективний для опрацювання звітних паперів.

2. Система автоматичного аналізу тексту TextAnalyst [3].

Це інструмент для аналізу змісту текстів з автоматичним формуванням семантичної мережі (змістовний портрет тектового документа в термінах основних понять та їх змістовних зв'язків у вигляді тематичного дерева з гіперпосиланнями) та пошуку інформації з урахуванням виявлених змістовних зв'язків слів запиту зі словами в тексті. На відміну від попередньої системи, це не комерційний продукт. Ефективний для автоматичного конспектування великих текстів до менших, не втрачаючи загального сенсу.

3. Інститут лінгвістики РГГУ та робоча група АОТ [4].

Розробляє програмне забезпечення у сфері автоматичної обробки текстів. На їх електронному ресурсі є можливість безкоштовно ознайомитися з вихідним кодом програм та словників, в основному спрямованих на опрацювання російськомовних текстових документів.

Постановка завдання. Дослідити можливості адаптації методу латентно-семантичного аналізу (ЛСА) [5; 6] для створення алгоритму семантичного розбору електронних текстих документів з науковим змістом, що надасть можливість виокремлення авторського змісту при порівнянні з іншими документами бази знань.

Розв'язання завдання. Для створення складової системи прийняття рішень при семантичному розборі текстових документів, без використання словників чи тезаурусів, та подальшої реалізації алгоритму виявлення змістовних співпадінь у документах з бази знань важливо відзначити, що серед етапів роботи такого алгоритму мають бути такі, які він не повинен виключати:

– завантаження та редагування текстових документів;
– обробка тексту, відштовхуючись від початкових можливостей медоду ЛСА;
– побудова семантичної мережі;
– допустимий час роботи.

Відповідно, на вході повинен поступати тексовий документ, а на виході повинна будуватися семантична мережа у вигляді орієнтованого графа. Відмінність від існую-

чих алгоритмів – у можливості виключення функцій створення, підключення та роботи зі словниками та тезаурусами.

Узагальнену структуру семантичних систем можна побачити на рисунку.

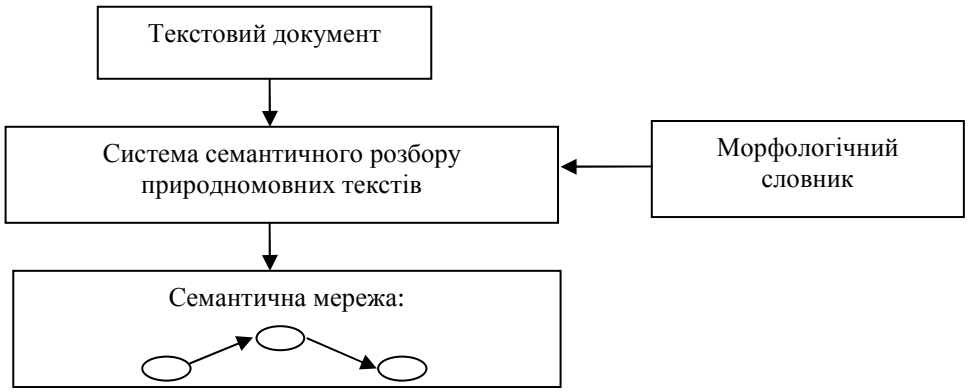

Рис. Узагальнена структура систем семантичного розбору

Загальні принципи систем автоматизованої обробки текстів являють собою компоненти, що складають структуру систем аналізу текстів, а саме – лінгвістичні процесори, які послідовно опрацьовують вхідний текст [7]. Вхід одного процесора є виходом іншого.

Виділяють такі компоненти:

– графематичний аналіз – виокремлення цифро-буквенних сполучень, формул;

– морфологічний аналіз – побудова морфологічної інтерпретації слів вхідного тексту;

– синтаксичний аналіз – побудова дерева залежностей конкретного речення;

– семантичний аналіз – побудова семантичного графа тексту.

Метод латентно-семантичного аналізу вже адаптовувався для SEO оптимізації інтернет-сайтів та кластеризації текстових документів за тематикою [5; 8].

ЛСА дозволяє виявляти значення слів з урахуванням контексту їх використання шляхом обробки великого набору текстів. Головний принцип дії методу полягає в думці, що порівняння множини всіх контекстів, у яких слова чи групи слів вживаються і контектів, у яких вони не вживаються, дозволяє зробити висновок про ступінь схожості змістовних понять цих самих слів чи їх групи [5].

Як практичний метод, що характеризує значення слова, ЛСА дозволяє вимірювати кореляції типу «слово-слово», «слово-фрагмент» та «фрагмент-фрагмент». Можна сказати, що ЛСА представляє значення слова як середнє значення уривків, у яких воно зустрічається, а значення уривка – як середнє значення всіх слів, що входять до нього.

З точки зору латентного семантичного аналізу, документи розглядаються як набори ключових слів, що зустрічаються у цих документах і називаються термами. Терм – це просте слово, семантика якого допомагає описати основний зміст документа.

Метод відображає документи й окремі слова (терми) в «семантичний простір», в якому і відбуваються подальші порівняння без використання тезаурусів. Текстовий документ визначається як набір слів, порядок яких ігнорується. Важлива лише кількість появи конкретного слова в документі. Припускається, що кожне слово має лише одне значення. Похибка через омоніми під час обробки великих документів є дуже мізерною.

Метод ЛСА можливо адаптувати до використання при пошуку дублікатів текстової інформації шляхом реалізації таких можливостей:

– виключення стоп-символів;

– виокремлення текстових блоків (парсинг);

– виокремлення кореня слів (стемінг);

– виключення одиничних слів;

– виокремлення індексованих слів (слова, що залишились після попередніх дій).

Необхідно скласти частотну матрицю з індексованих слів, у якій стрічки відповідатимуть проіндексованим словам, а стовпці – текстовим документам. На їх перетині відповідно буде вказана кількість повторів конкретного слова в конкретному документі.

Після отримання матриці, понижується її ранг, так-як високий ранг ускладнює обчислення та початкова матриця неодмінно містить «шуми» – випадкові потрапляння слів, що не відповідають змісту фрагмента чи текста в цілому. Пониження рангу дозволить певною мірою позбавитися від «шумів». У результаті перетворення розмірність розкладу слів за документами зменшиться.

Вибір нового рангу матриці після перетворення задається в автоматичному режимі, шукаючи найбільш оптимальний варіант (адаптивний метод, що не виключає самонавчання), або ж в ручному режимі, покладаючись на людський досвід.

Використавши розкладання матриці за сингулярним значенням (SVD), величезна вихідна матриця розкладається в безліч з k, наприклад, від 20 до 160, ортогональних матриць, лінійна комбінація яких є наближенням вихідної матриці.

Більш формально, відповідно до теореми про сингулярне розкладання, будь-яка дійсна прямокутна матриця M може бути розкладена в добуток трьох матриць:

$$M = UWV^t,$$

де U та V^t – ортогональні матриці, а W – діагональна матриця, значення на діагоналях якої називаються сингулярними значеннями матриці M.

Таке розкладання має цікаву особливість: якщо в W залишити тільки k найбільших сингулярних значень, а в матрицях U і V тільки відповідні до цих значень стовпці, то добуток матриць, що вийшли, і буде найкращим наближенням вихідної M матриці матрицею k рангу:

$$M \approx \hat{M} = U_L W_L V_L.$$

Кожен терм і документ представляються за допомогою векторів у загальному просторі розмірності k. Близькість між будь-якою комбінацією термів і/або документів може бути обчислена за допомогою скалярного добутку векторів.

Матриця аналізується на предмет кореляцій, використовуючи, наприклад, алгоритм рангової кореляції Спірмена.

На практиці розрахунок коефіцієнта рангової кореляції складається з таких кроків:

– співставлення кожному окремому із признаків порядкового номера;

– оцінка різниці рангів кожної пари співставлених значень;

– привести в квадрат кожну різницю та скласти отримані результати.

Алгоритм Спірмана дещо уступає параметричному коефіцієнту кореляції, однак саме цей алгоритм доцільніше використовувати в умовах, коли в наявності є невелика кількість явищ.

Таким чином, результатом роботи методу ЛСА з метою виявлення змістовних співпадінь може бути кореляційна матриця, що відображає змістовні зв'язки між виокремленими словами та набором текстових документів. Реалізувавши вищезгадані пункти в роботі методу, можна отримати повноцінний інструмент семантичного розбору для виявлення змістовних співпадінь в україномовних текстових документах.

Висновки. Створення систем автоматичної обробки природномовних текстів є одним з актуальних завдань у комп'ютерній лінгвістиці. Принципи роботи таких систем входять до кола інтересів систем підтримки прийняття рішень. Створення системи, що опрацьовує тексти будь-якої складності українською мовою, не можливо. Однак можливо за допустимої похибки реалізувати семантичний розбір та пошук текстових даних, схожих за змістом, взявши за основу метод ЛСА, що дозволить досягти високого рівня формалізації мовних структур у різноманітних прикладних цілях.

Список використаних джерел

1 Марчук Ю. Н. Компьютерная лингвистика / Ю. Н. Марчук. – М.: АСТ: Восток-Запад, 2007. – С. 60-70.

2 SyTech – разработка программного обеспечения: аналитические системы, электронный документооборот, корпоративные системы, информационные порталы [Электронный ресурс]. – Режим доступа: http://www.sytech.ru.

3 TextAnalyst 2.0 – персональная система автоматического анализа текста [Электронный ресурс]. – Режим доступа: http://www.analyst.ru.

4 Автоматическая обработка текста [Электронный ресурс]. – Режим доступа: http://www.aot.ru.

5 Landauer, T. K., Foltz, P., and Laham, D. 1998. An Introduction to Latent Semantic Analysis. DiscourseProcesses, 25: 259-284.

6 Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41: 391-407.

7 Нгуен Ба Нгок. Обзор подходов семантического поиска [Электронный ресурс] / Нгуен Ба Нгок, А. Ф. Тузовский // Доклады Томского государственного университета систем управления и радиоэлектроники: периодический научный журнал. – 2010. – № 2 (22). Ч. 2. – С. 234-237. – Режим доступа: http://www.tusur.ru/filearchive/reports-magazine/2010-2-2/234.pdf.

8 Косинов Д. Локальные параметры текстов и проблема определения почти-дубликатов [Электронный ресурс] / Д. Косинов // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2008. – Т. 1. – С. 83-85. – Режим доступа: http://www.vestnik.vsu.ru/pdf/analiz/2008/01/kosinov.pdf.

УДК 004.4

А.М. Акименко, канд. фіз.-мат. наук
Чернігівський державний технологічний університет, м. Чернігів, Україна

ДЕЯКІ АСПЕКТИ ЗАСТОСУВАННЯ UML ПРИ РОЗРОБЛЕННІ СКЛАДНИХ ПРОГРАМНИХ СИСТЕМ

Розглянуто проблеми, пов'язані з формуванням завдання на розробку коду програми після завершення стадій аналізу та проектування складної програмної системи. Запропоновано використання специфікацій сумісного типу для організації роботи груп експертів та постановки завдань на виконання.

Ключові слова: *розробка програм, сумісний тип, специфікація, діаграма варіантів використання, технічне завдання.*

Рассмотрены проблемы, возникающие в процессе формирования задачи на разработку кода программы после завершения стадий анализа и проектирования сложной программной системы. Предложено использование спецификации совместного типа для организации работы групп экспертов и постановки задачи на исполнение.

Ключевые слова: *разработка программ, совместный тип, спецификация, диаграмма вариантов использования, техническое задание.*

The problems associated with the formation of a task to develop the application code after the stages of analysis and design of sophisticated software systems. Proposed the use of the specifications for the type of joint organization of the expert group and set goals for performance.

Key words: *development programs, joint type, specification, use cases diagram, technical task.*

Постановка проблеми. Традиційно розробка складної програмної системи складається з таких етапів [1]:

1. Постановка та аналіз завдання, визначення вимог до проекту.
2. Проектування.
3. Розробка, кодування.
4. Тестування та оцінювання якості.
5. Впровадження та супровід.

Як бачимо, розроблення програмного забезпечення – це доволі складний процес. Кожна зі стадій розроблення має визначені методики її виконання та перелік документів, які повинні бути сформовані для переходу на наступний етап. Однак існує кілька