УДК 004.91:004.415
DOI: 10.15587/2312-8372.2015.56824

Кунченко-Харченко В. І.

МЕХАНІЗМ ТЕРМІНОЛОГІЧНОГО АНАЛІЗУ ПОКАЗНИКІВ ФУНКЦІОНУВАННЯ ІНТЕГРОВАНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ

Проаналізовано проблему автоматичного виявлення в текстовому потоці елементів часового ряду для машинночитаних документів та виділення статистичних та лінгвістичних критеріїв, що дозволило конкретизувати задачі дослідження: задачі для блоку синтаксично-семантичного аналізу та формування контекстних правил. При апробації розробленого на базі семантичних мереж програмного забезпечення з'ясовано, що механізм термінологічного аналізу показників функціонування інтегрованої інформаційної системи виконує покладені на нього завдання.

Ключові слова: *семантична мережа, термінологічний аналіз, машинночитані документи, статистичні критерії, інтегрована інформаційна система.*

1. Вступ

Проблема автоматичного виявлення в текстовому потоці інформаційної системи слів і словосполучень, що називають поняття визначеної проблемної області вивчається з точки зору різних галузей застосувань, наприклад, для виділення часового ряду з різних машиночитаних документів [1].

Актуальність роботи пояснюється тим, що запропонований механізм термінологічних аналіз дає змогу отримати саме такі показники часового ряду системи, що дозволяють без застосування додаткових методів виділення пошукового шуму застосовувати їх для прогнозування поведінки системи в цілому.

2. Аналіз літературних даних та постановка проблеми

Загальноприйнятим підходом при формулюванні механізму термінологічного аналізу статистичних показників для інтегрованої інформаційної системи є застосування часткового синтаксичного аналізу тексту і використання для розпізнаванні термінів лінгвістичних та статистичних критерії [2, 3].

Статистичні критерії так або інакше засновані на частоті вживання термінів і дають достатню для великого кола задач точність та повноту розпізнання, але тільки в корпусах текстів [4].

Лінгвістичні критерії враховують в першу чергу типову структуру іменних термінологічних словосполучень, та в більшості проаналізованих з теми дослідження наукових праць [1–6] для якісного та кількісного аналізу документів. Існуючи методи термінологічного аналізу мало підходять для виділення незашумленого часового ряду, придатного для подальшого опрацювання.

3. Об'єкт, мета та задачі дослідження

Об'єктом дослідження є інтегрована інформаційна система.

Метою дослідження є формування механізму термінологічного аналізу показників функціонування інтегрованої інформаційної системи для отримання часового ряду показників системи.

Для досягнення поставленої мети необхідно виконати такі задачі:

1. Сформувати механізм виділення термінів.
2. Сформувати блок синтаксично-семантичного аналізу та окреслити його функції.
3. Проаналізувати результати роботи розробленого на базі створеного механізму термінологічного аналізу програмного забезпечення.

4. Механізм виділення термінів

Термінологічний аналіз має на меті отримати синонімічні перетворення, розшифровку скорочень, виділення термінів [4]. Для цього використовуються фрагменти наступного виду:

TERMIN(<результ.слово>,<слово1>,<слово2>) або
TERMIN(<результ.слово>,<слово1>,<слово2>,
<слово3>),

де <слово1>,... — це може бути окреме слово, ознака, код «АБО» реквізити документу. Фрагменти «АБО» описуються оператором STR_OR(...), де перераховуються факультативні слова або їх ознаки. Фрагменти типу «І» описуються оператором STR_AND(...), де пропонується обов'язковість слів з вказуванням ознак.

Наприклад, TERMIN(НЕЛІКВІДНИЙ,НЕ,ЛІКВІДНИЙ) вказує на перетворення: НЕ ЛІКВІДНИЙ –> НЕЛІКВІДНИЙ.

Для термінів може бути заданий допустимий контекст-слово або їх ознаки, що стоять ліворуч або праворуч. Може бути також недопустимим контент-слово або його ознаки, які не повинні бути праворуч або ліворуч. У результаті вдається виділити терміни і словосполучення, цифрові показники, значення яких залежить від контексту [5].

Для синонімів використовуються багато місцевих фрагментів. SYNON (<результ.слово>,<поч.слово> ... <поч.слово>).

Наприклад, SYNONIM (БАЛАНС, ФІНРЕЗУЛЬТАТ) — слово «ФІНЗРЕЗУЛЬТАТ» повинно бути замінено на слово «БАЛАНС».

Багато з синонімів носить умовний характер. Для них вказується допустимий або недопустимий контекст. Наприклад, у наведеному вище випадку недопустимі заміни для особових назв: прізвищ, назв вулиць, міст тощо.

Формування блоків синтаксично-семантичного аналізу:

Блок синтаксично-семантичного аналізу виконує такі функції:

1) за ознаками і контекстом виділяють значимі об'єкти (ПІБ людей, організації та ін.);

2) для кожного виявленого значущого об'єкту знаходить у документі зв'язану інформацію.

Для цього використовуються «контексті правила», які контексті правила необхідні для подальшого синтаксично-семантичного аналізу.

Синтаксично-семантичний аналіз необхідний для виділення адрес, номерів справ, організацій, термінів, грифів тощо. Зазвичай, це набори слів, які граматично ніяк не узгоджені. Їх виділення може здійснюватися за виключно формальними ознаками. Наприклад, адреса може розглядатися як набір літер, слів з великої літери і чисел або окремих цифр. Кожен такий набір може мати свої межі і недопустимі компоненти.

Наприклад, в адресі не може міститися ПІБ, дієслів, дієприкметників. Виділення таких наборів слів (описів об'єктів) засновано на використанні контекстних правил такого виду:

CONTEXT (<слово1>,<слово2>,...,<словоN>) –> <результ. фрагмент>

де <слово1>,... — це може бути окреме слово, ознака, а також І-АБО графи коди документа , а також окреме слово, ознака. А також І-АБО графи.

Для цих правил вказується, з якої позиції почати застосування. Правило також повинно обумовлювати допустимість того чи іншого контенту. Окрім цього, в правилах може вказуватися, які слова з тими чи іншими ознаками не повинні стояти на тій чи іншій позиції. Такий підхід забезпечує диференційоване застосування правил.

Такі правила виділяють з тексту групи слів за їх ознаками. Ці слова описують деякий об'єкт і замінюють їх на одне слово, з яким надалі пов'язується відповідний фрагмент семантичної мережі. Наприклад, фінансовий термін.

Синтактико-семантичний аналіз документів з виділенням словосполучень і аналізом форм відбувається на основі контекстних правил, які застосовуються у визначеній послідовності [7].

Спочатку виділяються об'єкти, потім їх ознаки, словосполучення і, наприкінці, — цифровий показник, що необхідний для формування часового статистичного ряду для подальшої побудови прогнозу. Надалі, по мірі застосування таких правил будується семантична мережа — змістовний портрет документа з інтегрованої системи.

Наприклад, розглянемо наступне контекстне правило GG~1:

MUSTBE(GG~1,1) STR_OR(ADJ,PRON/2+)
CONTEXT(2-,NOUN/GG~1)

P_P(GG~1,3+) WORD_C(1,2/3-) 3-(2,MORF)
NOTBE(GG~1,2,LETT)

Це правило здійснює перетворення: ПРИКМЕТНИК ІМЕННИК –> <комбінація слів> і ЦИФРА ІМЕННИК –> <комбінація слів>.

Фрагмент MUSTBE вказує на те, що правило GG~1 потрібно застосовувати з 1-ої позиції, тобто шукати слова з ознаками ПРИКМЕТНИК (ADJ) й іменник (PRON). Фрагмент P_P відділяє ліву частину від правої (–>), а WORD_C — вказує, що слова на 1-й і 2-ой позиції лексичної аксіоми повинні «склеюватися» в комбінацію слів, яке в подальшому буде розглядатися як одне слово з морфологічними ознаками 2 шуканого слова чи числової послідовності (DIGIT). Фрагмент NOTBE вказує, що на 2-ій позиції можуть бути відмінні від шуканих слова (ознака LETT). Це приклад найбільш простого правила. До таких правил додаються фрагменти, що вказують на контекст, на можливість використання деяких символів. Спеціальні правила здійснюють ідентифікацію об'єктів, наприклад, на основі іменників і коротких описів і багато іншого [5, 8].

Кожне контекстне правило це семантична мережа [8]. Усі лінгвістичні знання, що оброблюються ЛП записуються у вигляді семантичної мережі. Над ними працюють продукції програми, що застосовують ці правила і грають роль пустої лінгвістичної оболонки, що підтримують записи лінгвістичних знань семантичної мережі [8].

Зрозуміло, що таку оболонку можна через інтерфейс користувача налагоджувати на різні алфавітні аксіоми і, відповідно, будувати різні лінгвістичні процесори.

Результати застосування семантико-орієнтованих лінгвістичних процесорів для обробки семантичних конструкцій документів на основі опису машинно-зчитуваної граматики DTD і XML [9, 10].

На базі сформованих вище правил і алгоритмів термінологічного аналізу побудоване програмне забезпечення. При проведенні процедури розпізнавання та тестування в науково технічному тесті виділено ряд цільових термів.

Результат розпізнавання термінів за їх морфосинтаксичним зразком називається термінами — кандидатами. Така назва пов'язана з тим, що в числі з термінів — кандидатів з досить великою ймовірністю можуть виявитися загальнонаукові словосполучення, які надалі виявляться рішенням задачі.

Формування термінів — кандидатів відбувається на 1 кроці роботи алгоритму.

На вхід кожній процедурі за винятком процедури виявлення лексико-синтаксичних варіантів, надходить аналізований текст і відповідна група шаблонів. На першому етапі роботи процедури виконується пошук всіх фрагментів тексту, що представляють собою шукані терміновживання, а на другому — підраховується частота вживання кожного виявленого в них терміна. Поділ цих двох етапів необхідний для коректного підрахунку частоти вживання у випадках повного вкладення одного терміна в іншій (адреса — логічна адреса), оскільки частота повинна вживань терміну повинна бути визначена без урахування таких вкладень.

Таким чином, на виході зазначених процедур виходить список фрагментів-терміновживань і список виділених термінів з частотою їх вживання в тексті.

Розроблене програмне забезпечення (ПЗ) було протестоване за методикою, викладеною в [4]. Отримані внаслідок роботи ПЗ результати, наведені в табл. 1.

Таблиця 1

Результати аналізу повноти і точності виділення показників термінологічного аналізу

Процедура	Виділення термінів		Виділення терміновживань	
	Повнота	Точність	Повнота	Точність
Терміни-кандидати	72	41	70	49
Авторські терміни	76	94	82	96
Словарні терміни	92	96	94	97
Зв'язки	76	41	—	—
Числові показники	98	52	99	96

Результати роботи ПЗ є в цілому такими, що задовольняють постановці. Отримані результати є значно кращими від аналогічних, описаних в [4].

Найгірші результати дала процедура виявлення термінів-кандидатів, що спирається на незначний обсяг лінгвістичної інформації. Вона давала багато неприйнятних результатів (великий розмір, аналогічний результат). В той же час не були розпізнані поєднання, чия морфосинтаксична структура не врахована в наборі шаблонів (наприклад, індексна інформація, html-теги і зворотний зв'язок за релевантністю).

Для словникових термінів нерозпізнаними виявилися переважно їх терміновживання всередині сполук, а деякі розпізнані терміни виявилися частиною знайдених термінів (наприклад, термін ряд — частиною загальнонаукових виразів). Що стосується процедур виявлення з'єднань, на деякі причини невисоких значень повнота і точність такі ж, як для групи термінів кандидатів.

При розпізнаванні та виділенні в ряді випадків, за рядом авторських термінів, синонімів основна втрата повноти була через обмеженість морфосинтаксичних зразків термінів, так і через причини неврахування деяких конструкцій визначення термінів та їх синонімів.

5. Обговорення результатів дослідження механізму термінологічного аналізу

В результаті проведених досліджень сформовано механізм термінологічного аналізу, який дозволяє отримати вектор часового ряду. Даний масив значень максимально вільний від шуму для аналізу показників інтегрованої системи. Надалі, отриманий часовий ряж може бути використаний для прогнозування поведінки інтегрованої інформаційної системи.

6. Висновки

В результаті проведених досліджень були розв'язані такі задачі, як формування завдань для блоку синтаксичного аналізу та формування контекстних правил. Це дало змогу сформувати блок синтаксичного аналізу для виділення часових рядів, контекстних правил. Розроблене програмне забезпечення підтвердило ефективність розробленого механізму, який у порівнянні з існуючими більше підходять для виділення незашумленого часового ряду, придатного для подальшого опрацювання.

Внаслідок аналізу результатів роботи прикладного програмного забезпечення було встановлено, що при формуванні числових показників точність слововживання в переважній більшості випадків досягає 99 %, що і є основним шуканим результатом і дозволяє формувати часовий ряд з мінімальним шумом. Тобто можна стверджувати, що механізм термінологічного аналізу показників функціонування інтегрованої інформаційної системи виконує покладені на нього завдання.

Література

1. Salton, G. Term-weighting approaches in automatic text retrieval [Text] / G. Salton, C. Buckley // Information Processing & Management. — 1988. — Vol. 24, № 5. — P. 513–523. doi:10.1016/0306-4573(88)90021-0
2. Jacquemin, C. Term extraction and automatic indexing [Text] / C. Jacquemin, D. Bourigault; by ed. R. Mitkov // Handbook of Computational Linguistics. — Oxford University Press, 2003. — P. 599–615. doi:10.1093/oxfordhb/9780199276349.013.0033
3. Добров, Б. В. Формирование базы терминологических словосочетаний по текстам предметной области [Текст] / Б. В. Добров, Н. В. Лукашевич, С. В. Сыромятников // Труды пятой конференции всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». — 2003. — С. 201–210.
4. Ефремова, Н. Э. Терминологический анализ текста на основе лексико-синтаксических шаблонов. Библиотеки [Электронный ресурс] / Н. Э. Ефремова, Е. И. Большакова, А. А. Носков, В. Ю. Антонов. — 2010. — Режим доступа: \www/URL: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/20.pdf. — 23.10.2015.
5. Beuster, G. MIC — A System for Classification of Structured and Unstructured Texts [Electronic resource]: Master's thesis / G. Beuster. — University Koblenz, 2001. — Available at: \www/URL: http://www/gb/papers/thesismic/mic.pdf. — 10.10.2015.
6. Методика работы с источниками информации. Раздел 2 [Электронный ресурс]. — Режим доступа: \www/URL: http://edu.dvgups.ru/METDOC/EKMEN/ETEOR/ORGANIZ_ISSLED_D/METOD/SIMONENKO/UP/frame/frame_tema5.htm. — 27.10.2015.
7. Леонтьева, Н. Н. К теории автоматического понимания текста [Текст]. Ч. 3. Семантический компонент. Локальный семантический анализ / Н. Н. Леонтьева. — М.: Изд. Моск. ун-та, 2002. — 49 с.
8. Kuznetsov, I. P. Linguistic Processor «Semantix» for Knowledge extraction from natural texts in Russian and English [Text] / I. P. Kuznetsov, E. B. Kozerenko // Proceeding of International Conference on Machine Learning, ISAT-2008, 14–18 July, 2008. — Las Vegas, USA CSREA Press, 2008. — P. 835–841.
9. XML DTD — An Introduction to XML Document Type Definitions [Electronic resource]. — Available at: \www/URL: http://www.xmlfiles.com/dtd/. — 01.11.2015.
10. Boyer, J. Canonical XML Version 1.0 [Electronic resource]: Report / J. Boyer. — March 2001. — Available at: \www/URL: http://dx.doi.org/10.17487/rfc3076

МЕХАНИЗМ ТЕРМИНОЛОГИЧЕСКОГО АНАЛИЗА ПОКАЗАТЕЛЕЙ ФУНКЦИОНИРОВАНИЯ ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Проанализирована проблема автоматического обнаружения в текстовом потоке элементов временного ряда для машинных документов и выделения статистических и лингвистических критериев, что позволило конкретизировать задачи исследования: задачи для блока синтаксически-семантического анализа и формирования контекстных правил. При апробации разработанного на базе семантических сетей программного обеспечения установлено, что механизм терминологического анализа показателей функционирования интегрированной информационной системы выполняет возложенные на него задачи.

ISSN 2226-3780

Ключевые слова: семантическая сеть, терминологический анализ, машинные документы, статистические критерии, интегрированная информационная система.

Кунченко-Харченко Валентина Іванівна, *доктор технічних наук, професор, кафедра інформатики і інформаційної безпеки, Черкаський державний технологічний університет, Україна,* ***e-mail: valentine.kun@ukr.net****.*

Кунченко-Харченко Валентина Ивановна, *доктор технических наук, профессор, кафедра информатики и информационной безопасности, Черкасский государственный технологический институт, Украина.*

Kunchenko-Kcharchenko Valentina, *Cherkassky State Technological University, Ukraine,* ***e-mail: valentine.kun@ukr.net***

УДК 004.08:005.8
DOI: 10.15587/2312-8372.2015.56825

Нестеренко С. А.,
Становський А. О.,
Оборотова О. О.

РОЗПІЗНАВАННЯ СТАНУ БЕЗДРОТОВИХ КОМП'ЮТЕРНИХ МЕРЕЖ ЗА ДОПОМОГОЮ ТРИВИМІРНОГО ПОЛЯ НАПРЯМКІВ

Показано, що особливості зорового відображення бездротових комп'ютерних мереж з частково недоступними моніторингу в експлуатації елементами не дозволяють використовувати для розпізнавання їх стану відомі інтелектуальні методи обробки нерухомих зображень. Розроблений і впроваджений метод такого розпізнавання за допомогою тривимірного поля напрямків. Наведено приклад використання цього підходу в реальній практиці Збройних Сил України.

Ключові слова: *бездротова комп'ютерна мережа, розпізнавання стану, зоровий образ, тривимірне поле напрямків.*

1. Вступ

Сучасні розповсюджені складні системи, до яких, в першу чергу, відносяться бездротові комп'ютерні мережі (БКМ), потребують постійного моніторингу своєї працездатності, особливо у випадках, коли ці системи експлуатуються в небезпечних для їхнього стану умовах, наприклад, в умовах бойових дій. Ця небезпека для самого факту існування деяких елементів БКМ та зв'язків між ними багаторазово посилюється неможливістю діагностувати їхній поточний стан з-за віддаленості та недоступності для безпосереднього тестування.

Натомість існують методи інтелектуального (побудованого на знаннях) розпізнавання стану частково недоступних для моніторингу технічних систем [1], коли дослідникові вдається отримати ймовірнісну оцінку стану недоступних елементів (а, отже, і всієї БКМ) по сигналах від доступних. На жаль, такі методи відрізняються низькою швидкодією з-за своєї складності [2], або невеликою точністю [3].

Останнім часом з'явилися пропозиції по інтелектуальному розпізнаванню стану БКМ за допомогою побудови проміжного зорового образу (зображення) цього стану [4]. Однак вибір методу фінішної обробки такого образу з метою отримання можливостей його порівняння з базою даних і прийняття діагностичних рішень не зв'язаний із геометричними особливостями саме зорових образів БКМ, що суттєво знижує їхню ефективність.

Тому розробка швидкодіючого методу розпізнавання стану бездротових комп'ютерних мереж за допомогою вибору та вдосконалення методу обробки проміжного зорового зображення цього стану є головним завданням даної роботи.

2. Аналіз літературних даних і постановка проблеми

Широке використання БКМ в військовій практиці призводить до зростання вимог щодо їхньої надійності. При проектуванні та експлуатації таких відповідальних БКМ важливо вміти оцінювати стан їхньої структури. Адже на відміну від «звичайних» дротових мереж, бездротові позбавлені можливості постійного внутрішнього структурного моніторингу [5, 6]. Справа ускладнюється також тим, що елементи бездротових мереж не мають сталих «сусідів» для взаємотестування, оскільки вони часто-густо переміщуються в просторі, постійно змінюючи перелік найближчих серверів та вузлів іншого призначення.

Зазначені проблеми обумовлюють високі часові витрати на пошук несправності, а також звужують діапазон суб'єктів, що забезпечують коректне розв'язання завдання пошуку структурної несправності, що, в свою чергу, призводить до високої трудомісткості і складності розв'язання даної проблеми.

В той же час, на справних серверах бездротової мережі під час її роботи накопичується багато інформації, яка може взагалі не використовуватися для основної роботи, але яка, в той же час, містить на прихованому рівні важливі знання про структуру мережі та її «історію» від початку експлуатації до поточного часу. Важливо, що з виходом з ладу окремих структурних одиниць системи,

