8. *Тарасов В.А., Герасимов Б.М., Левин И.А., Корнейчук В.А.* Интеллектуальные системы поддержки принятия решений: Теория, синтез, эффективность. – К.: МАКНС,2007. – 336

Поступила 14.9.2015р.

УДК 004.451.36:681.5

В. І.Сабат, Українська академія друкарства, м.Львів

ВИКОРИСТАННЯ СЕМАНТИЧНИХ МЕРЕЖ ДЛЯ АВТОМАТИЗОВАНИХ СИСТЕМ ДОКУМЕНТООБІГУ

Анотація. У статті проаналізовано методи семантичного аналізу тексту документів в автоматизованих системах документообігу (АСДО) при використанні мережних технологій.

Ключові слова: системи документообігу, глибинний аналіз тексту, семантика.

Вступ. Перехід від паперових до електронних документів стимулює подальший розвиток нових технологічних рішень та напрямків для автоматизації процесу створення, опрацювання та використання документів і, відповідно, аналізу інформації, яка в них міститься. З розвитком інформаційних технологій та засобів мережного зв'язку з'являються нові автоматизовані системи документообігу (АСДО), які облегшують не тільки сам процес обігу документів на їхньому життєвому циклі, — від проектанта до виконавця чи замовника, але й відкривають нові можливості для аналізу великих масивів тексту документів, їх захисту та контролю виконання управляючих дій, які містяться в документах.

Основна частина. На сьогодні для глибинного аналізу тексту використовується концепція контент-аналізу (Text Mining). Контент-аналіз — це якісно-кількісний метод вивчення документів, яке характеризується об'єктивністю висновків і строгістю процедури і складається з обробки, квантифікації тексту, з подальшою інтерпретацією результатів. [1] Отже, контент-аналіз в рамках дослідження електронних інформаційних масивів — це відносно новий напрям, який передбачає аналіз безлічі текстових документів.

Якщо розглянути *кількісний контент-аналіз*, то він у більшості випадків зводиться до пошуку частоти появи в документах певних характеристик змісту (ключових слів, фраз, понять). Показники можуть відрізнятися або, навпаки, бути близькі за абсолютним значенням, яке буде враховуватися при інтерпретації результатів обробки. Завдання можна ускладнити поставивши в якості попередньої умови виділення всіх змістовних у смисловому відно-

шенні одиниць відповідних текстів, а потім підрахувавши відносну значущість даного виразу в порівнянні з іншими.

Якісний контент-аналіз націлений на поглиблене змістовне вивчення текстового матеріалу, в тому числі з точки зору контексту, в якому представлені виділені категорії. Результати такого аналізу формуються з врахуванням взаємозв'язків змістовних елементів і їх відносної значущості (рангу) у структурі тексту.

Відповідно до вже сформованої методології до основних елементів Text Mining відносяться: класифікація (classification), кластеризація (clustering), виділення фактів, понять (feature extraction), сумаризація (summarization), відповідь на запити (question answering), тематичне індексування (thematic indexing), пошук за ключовими словами (keyword searching) та побудова семантичних мереж. Також в деяких випадках набір доповнюють засоби підтримки і створення таксономій (oftaxonomies) і тезаурусів (thesauri).

При *класифікації* текстів використовуються статистичні кореляції для побудови правил розміщення документів в певні категорії. Завдання класифікації — це класичне завдання розпізнавання, де по деякій контрольній вибірці система відносить новий об'єкт до тієї або іншої категорії. Особливістю системи Text Mining є те, що кількість об'єктів і їх атрибутів може бути дуже великою, тому необхідно заздалегідь передбачити інтелектуальні механізми оптимізації процесу класифікації.

Кластеризація базується на ознаках документів, які використовує лінгвістичні і математичні методи без використання певних категорій. Результат кластеризації — це таксономія або візуальна карта, що забезпечує ефективний обсяг великих об'ємів даних. Кластеризація в Text Mining розглядається як процес виділення компактних підгруп об'єктів з близькими властивостями. Система повинна самостійно знайти ознаки і розділити об'єкти по підгрупах. Кластеризація, як правило, випереджує класифікацію, оскільки дозволяє визначити групи об'єктів. Розрізняють два основних типи кластеризації — ієрархічну і бінарну.

Виділення фактів, призначене для зчитування деяких фактів з тексту з метою поліпшення класифікації, пошуку і кластеризації.

Автоматичне реферування (Automatic Text Summarization) — це складання коротких викладів матеріалів, анотацій або дайджестів, тобто зчитування найбільш важливих відомостей з одного або декількох документів і генерація на їх основі лаконічних і інформаційно-насичених звітів.

Семантичні методи формування рефератів-викладів допускають два основні підходи: метод синтаксичного розбору пропозицій, і методи, що базуються на розумінні природної мови. Цей підхід ґрунтується на системах штучного інтелекту, в яких також на етапі аналізу виконується синтаксичний розбір тексту, але синтаксичні дерева не породжуються.

Семантичні мережі — це нова концепція розвитку всесвітньої мережі Інтернет на основі розробок RDF (resource description framework), онтологій та інтелектуальних агентів, прийнята Консорціумом Всесвітньої павутини

(World wide web consortium, W3C).

За допомогою семантичних веб-технологій користувачі автоматизованих систем документообігу можуть виділяти корисну інформацію з великих масивів даних, які містяться в документах. При здійсненні семантичного аналізу документів стає можливим не тільки досягати автоматизованої інтерпретації інформації на рівні ключових слів, але й визначати зміст досліджуваних документів. Отже, семантичні технології та мережі — це ефективний спосіб представлення інформації документів в Інтернет. Інформація подається за допомогою онтологій із забезпеченням аргументації, зв'язків, правил, понять, логіки та умов, визначених в онтологіях.

Онтології — створюються на основі всеохоплюючої та детально формалізованої певної області знань за допомогою концептуальної схеми. Зазвичай така схема складається із ієрархічної структури даних, які містять усі релевантні класи об'єктів, їх зв'язки та правила виводу, прийняті в цій області. Онтології призначені для опису предметної області в термінах відносин між сутністю об'єктів та їх обмеженнями, що необхідно для комп'ютерів для сприйняття семантики з метою аналізу, порівняння, співставлення даних, а також виводу нових знань із вже існуючих понять. [2]

У найпростішому випадку, — для АСДО, в основі семантичного аналізу текстів закладено побудову семантичного словника досліджуваної предметної області та логічних правил виводу, поєднання слів у фрази в контексті понятійних груп. На відміну від традиційних тлумачних чи орфографічних словників, семантичний словник містить змістову інтерпретацію слів предметної області та фраз а також прописані правила виводу і взаємозв'язки між словами, фразами, смисловими поняттями. При цьому основні компоненти досліджень семантики — це семантична значущість, семантична повнота, семантична суперечність тощо за допомогою яких визначаються семантичні характеристики документа в АСДО. [3]

До основних семантичних веб-технологій можна віднести такі.

Стандартний синтаксис опису даних (RDF, Resource Description Framework), де RDF — це специфікація, що визначає модель представлення даних і синтаксис для обміну даними. RDF по суті забезпечує спосіб опису і роботи з будь-яким інтернет-ресурсом: від текстових сторінок і графіків до аудіо та відеоінформації. Він закладає синтаксичні можливості взаємодії мереж і формує базовий шар для створення семантичної мережі. Принцип побудови відносин між мережевими ресурсами в специфікації RDF передбачає наявність трьох компонент: об'єкта, атрибута і його значення (аналогічних класичній схемі «підмет — присудок — доповнення»). Модель даних RDF є лише синтаксичною основою семантичного аналізу. Для того, щоб опис мав сенс, необхідно скористатися словниками термінів і понять, які задаються за допомогою технології – схема RDF, що відіграє для RDF таку ж роль, що і схема для XML.

Стандартні способи опису властивостей даних (Schema RDF, RDF-S) — це семантичне розширення RDF, що забезпечує механізми опису пов'язаних

ресурсів і самих зв'язків між ними. RDF — це самий низькорівневий з існуючих мов опис метаданих, оскільки оперує лише поняттями зв'язків примітивної суті, наприклад, «об'єкт А володіє суб'єктом Б».

Стандартні способи опису зв'язків між об'єктами даних (онтологія, яка визначається за допомогою онтологічної веб-мови (Ontology Web Language, OWL)). OWL використовується, для щоб явно представляти значення термінів і відношення між цими термінами в словниках. OWL має більше засобів для виразу значення і семантики, ніж XML, RDF, і RDF-S, і, таким чином, OWL перевершує усі веб-мови в здатності представити контент мережі, що піддаються машинній обробці. [4]

Висновок. Побудова семантичної мережі дозволяє значно підвищити ступінь автоматизації систем документообігу і сприяє розвитку динамічних інформаційних систем, що характеризуються гнучкістю, масштабованістю, платформною незалежністю. Використання семантичної мережі для АСДО успішно вирішує такі важливі проблеми пошуку та опрацювання інформації в документах: забезпечує семантичний аналіз інформації з можливістю багатокритерійного пошуку та зчитування релевантної інформації за допомогою програм-агентів; реалізує семантичну інтеграцію інформації в електронних документах; створює основи для використання інтелектуальних веб-сервісів роботи з великими масивами документів, що в подальшому може призвести до нових розробок універсальної онтологічної мета-мови формального представлення інформаційних одиниць синтезу і семантики.

1. *Костенко Н.* Досвід контент-аналізу: моделі та практики: монографія / Н. Костенко, Іванов В. — К. : Центр вільної преси, 2003. — 62 с.
2. *Ландэ Д. В.* Поиск знаний в Internet / Д. В. Ландэ. — М. : Диалектика-Вильямс, 2005. — 272 с.
3. *Дурняк Б. В.* Семантичний захист інформації в системах документообігу: монографія / Б. В. Дурняк, В. І Сабат. — Л. : УАД, 2010. — 160 с.
4. *Кальченко Д.* Интеллектуальные агенты семантического Web'а / Д. Кальченко. — М. : «Компьютер-Пресс», 2004. — №10. — С. 45–48.

Поступила 28.9.2015р.