УДК 004.5:81

Андреєв С.С. ,
Джуренко Т.С.

Методи видалення об’єктів і фактів з текста

Національний університет харчових технологій

Анотація*:*

У даній роботі представлено метод здобуття сутності (наприклад, імен людей і географічні назви) з текстів і запитів. А також видобування фактів, тобто зв'язків між об'єктами.

Ключові слова*:* Text Mining, факти, об’єкти

Abstract:

The article presents the method of obtaining essence (for example, names of people and geographical names) with texts and requests, and also the extracting facts, namely connections between objects.

Keywords: Text Mining, facts, objects

На сьогодні в особистих ПК, локальних і глобальних мережах накопичено величезну кількість інформації і її обсяг стрімко збільшується. Пошук в гігантських масивах текстових даних і аналіз об'ємних текстів є малоефективними, тому стають потрібними технології, які спроможні обробляти неструктуровані або слабкоструктуровані тексти.

Вилучення об'єктів і фактів з текстів - це частина NLP (Natural Language Processing - автоматична обробка природної мови). Кінцева мета - навчити машину повноцінно розуміти звичайний людський текст.

Сьогодні NLP успішно застосовується для декількох цілей:

- текстовий пошук;
- вилучення фактів;
- діалогові системи і Question Answering;
- синтез і розпізнавання мови;
- оцінка тональності відгуків;
- кластеризація і класифікація текстів.

Вилучення структурованої інформації з не структурованого тексту називається Text Mining. Основна частина цього процесу присвячена визначенню об'єктів, їхніх зв'язків і властивостей в тестах.

Його робота полягає в тому, що на вхід подається документ, описаний природною мовою, а на виході користувач отримує запитану інформацію в структурованому вигляді. Структури можуть являти собою як прості сутності (персони, організації, географічні назви), так і складні (факти, що містять якась подія, його учасників, дату, фінансові параметри та ін.).

Аналізувати подібні набори даних, безумовно, простіше і швидше, ніж результати роботи пошуковика. Однак постає проблема по інтеграції засобів Text Mining з джерелами документів, пошуковиком і аналітичними інструментами.

Основною проблемою використання цих технологій є складність налаштування і підтримки таких інструментів.

Тим не менш користувач вже позбавлений від необхідності вручну переглядати тисячі документів і підбирати ключові слова. За нього це робить система. З'являються додаткові можливості автоматичної класифікації і порівняння подібних документів. Крім того, програма здатна сама розпізнавати змістовні елементи тексту, наприклад факти, події, і передавати їх на подальшу обробку.

Реалізація описаної ідеї описується чотирма процедурами.

1. Information Extraction (витяг інформації):

- Feature Extraction - видобуток слів чи груп слів, які, з точки зору користувача є важливими для опису змісту документа. Це можуть бути згадки персон, організацій, географічних місць, термінів предметної області та інших слів або словосполучень. Видобуті сутності також можуть бути найбільш значущими словосполученнями, що характеризують документ по його основній темі;
- Feature Association Extraction - простежуються різного роду зв'язки між видобутими сутностями;
- Relationship, Event and Fact Extraction - найскладніший варіант видобутку інформації (Information Extraction), що містить видобуток сутностей, розпізнавання фактів і подій, а також витягнення інформації з цих фактів.

Технологія Text Mining, у відповідності з заданими обмеженнями відрізняє факти, що відносяться до необхідної теми від тих, що з нею не пов'язані.

2. Summarization (автоматичне реферування, анотування) – побудова короткого змісту документа за його повним текстом.

3. Categorization (категоризація, класифікація)

Віднесення документа або його частини до однієї або кількох категорій. Категорії можуть визначати "спрямованість" тексту - тематичну, жанрову, емоційну, оцінкову.

4. Clusterization (кластеризація) - об'єднання документів в групи за принципом їх схожості.

Проблеми застосування таких технологій очевидні і пов'язані з багатокомпонентністтю рішення.

Тому виникає необхідність побудування та інсталяції пошукової системи, модулів індексації, інструментів отримання даних з тексту, засобів аналізу, а крім того, зробити всю супутню інтеграцію.

Зважаючи на істотну складність і об'ємність завдань інтелектуалізації при побудуванні пошукової систем, їх розробка та розвиток повинні бути здійснені поетапно, шляхом послідовного нарощування їх функціональності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ландэ Д.В. Поиск знаний в Интернет. Издательский дом "Диалектика- Вильямс", 2005. 272 с.

2. http://poiskbook.kiev.ua/dialog.html

Андреєв Сергій Сергійович, студент групи АКС 4-5, факультет Автоматизації та комп'ютерних систем, Національний університет харчових технологій, м. Київ.

Джуренко Тетяна Сергіївна, асис. Національного університету харчових технологій, м. Київ, email: tania-dzhurenko@mail.ru

Sergey Andreev - student of Automation and Computer Systems, National University of Food Technologies, Kyiv

Dzhurenko Tetyana - assistant of the National University of Food Technologies, Kyiv, email: tania-dzhurenko@mail.ru