Поняття взаємозв'язків між сутностями при видобуванні інформації з текстових даних

Антоніна Заяць[1], Андрій Романюк[2]

Кафедра систем автоматизованого проектування, Національний університет «Львівська політехніка», УКРАЇНА, м. Львів, вул. С. Бандери, 12;
e-mail: 1. tonya-zayats@yandex.ru, 2. anrom@polynet.lviv.ua

Abstract – the rapid development of hardware and information technologies requires new methods and means of improving large data amounts processing. A major breakthrough in this direction is information extraction. This paper is devoted to relations types during extraction from the text data.

Ключові слова – видобування інформації, взаємозв'язки, відношення, сутність, декартів добуток.

I. Вступ

Стрімкий розвиток технічних засобів та галузі інформаційних технологій вимагає нових методів та засобів покращення опрацювання великих об'ємів даних. Значну увагу у цьому напрямку сьогодні приділяють екстракції даних.

II. Основна частина

Видобування інформації – процес обробки та аналізу неструктурованих об'ємів даних для виділення окремих сутностей. Поряд з цим питанням стоїть встановлення семантичних взаємозв'язків між виокремленими одиницями, що допомагає краще проаналізувати та зрозуміти структуру отриманих даних.

Говорячи про семантичні зв'язки потрібно розрізняти їхній обсяг та внутрішній зміст. Обсягом або розмірністю зв'язку вважають множину впорядкованих сутностей, що задовольняють цей зв'язок. Наприклад, для відношення частина – ціле членами цієї множини будуть *<професор, факультет>*, але не *<факультет, професор>*. Зміст цього зв'язку полягає у тому, що, власне, означає зв'язок «професор є частиною факультету». В загальному випадку відношення можна представити у вигляді декартового добутку множин $S_1 \ldots\ldots S_n$, де кожна множина належить конкретному аргементу відношення. Якщо дві сутності *x* та *y* перебувають у бінарному зв'язку *R*, це можна описати як *– xRy* або *R (x, y)*.

Вперше визначення та класифікація зв'язків розглядалось як окреме завдання в галузі екстракції інформації у 1998 році на UMC – 7 (Message Understanding Conferences). Тоді виділяли три типи зв'язків, що

стосувалися організацій: розташований в (location of), працівник в (employee of) та вироблено в (product of) [1, 11-12].

На сьогодні вчені розрізняють понад 7 типів зв'язків та приблизно 24 підтипи. Основні з них наведено у Табл.1 [2, 837].

ТАБЛИЦЯ 1

ТИПИ ТА ПІДТИПИ ЗВ'ЯЗКІВ

Тип зв'язку	**Підтип зв'язку**
Фізичний	Розташування, частина-ціле, близькість
Особистісно-соціальний	Бізнес, сім'я
Працевлаштування/членство/ієрархічні зв'язки	Працівник, член групи, учасник, партнер
Агент-артифакт	Власник/користувач, винахідник/виробник,
Інші зв'язки між особами чи організаціями	Етнічний, ідеологічний
Інші зв'язки між геополітичними сутностями	Мешканець чи житель, заснований у
Дискурсивний	

Висновок

Сьогоднішня наукова та технологічна ситуації зумовили значне зацікавлення сферою екстракції інформації. Створення системи встановлення взаємозв'язків між отриманими сутностями буде вагомим грунтом для подальшого розвитку цього напряму.

Література

1. S. Katrenko. A Closer Look at Learning Relations from Text. PhD thesis, University of Amsterdam. – 2009, – 222 p.
2. Doddington G., Mitchell A. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. Proceedings of the LREC 2004, Portugal. European Language Resources Association, – 2004, – P. 837 – 840.