прийнятного для надійного копіювання траєкторій, а тривалість підготовки нової системи помітно скорочено порівняно з ручним налаштуванням.

Список джерел:

1. **DYNAMIXEL SDK: Overview** [Електронний ресурс]. — Режим доступу: https://emanual.robotis.com/docs/en/software/dynamixel/dynamixel_sdk/overview/
2. DYNAMIXEL Protocol 2.0 — Communication Manual [Електронний ресурс]. Режим доступу: https://emanual.robotis.com/
3. **ROBOTIS-GIT. DynamixelSDK** : репозиторій програмного забезпечення [Електронний ресурс]. — GitHub. — Режим доступу: https://github.com/ROBOTIS-GIT/DynamixelSDK
4. **Hugging Face. LeRobot — Documentation** [Електронний ресурс]. — Режим доступу: https://huggingface.co/docs/lerobot/en/index

АВТОМАТИЧНЕ ФОРМУВАННЯ ОНТОЛОГІЇ ТОВАРІВ НА ОСНОВІ АНАЛІЗУ ДАНИХ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

Жежерун О. П., Колесніков А.О.

Національний університет «Києво-Могилянська академія»

вул. Сковороди 2, м. Київ, 04070, Україна, anton.kolesnikov@ukma.edu.ua

The article presents a system for automatic generation of product ontology based on analysis of heterogeneous data from multiple e-commerce sources. The system architecture and algorithm for concept extraction from natural language texts without manual synonym dictionary creation are described. The system generated an ontology with 486 concepts and 1216 relationships with F1=95.2% extraction accuracy. The system uses a four-layer hybrid architecture with transformer embeddings (gte-small, 384-dimensional) and HNSW indexing (M=16, efSearch=16). Experimental deployment on 700,000 products from 34 sources in four languages showed F1=95.2% concept extraction accuracy at 13 products per second processing speed. Main advantages: no need for large labeled datasets, automatic multilingual processing without translation dictionaries, ability to supplement ontology with new concepts without retraining. The system can be adapted for other domains: medicine, finance, logistics.

Keywords: ontology engineering, knowledge base, product ontology, natural language processing, transformer embeddings, semantic matching, big data.

Після створення концепції Semantic Web онтологія стала синонімом рішення проблем розуміння природної мови комп'ютерами [1]. Проте ручне створення онтологій потребує значних інтелектуальних ресурсів та швидко застаріває. Щоб знайти рішення, з'явився напрям онтологічної інженерії, який вивчає шляхи автоматизації генерування знань з тексту [2,3].

У процесі роботи розглянуто задачу автоматизованої генерації онтології товарів з використанням гетерогенних даних з 34 джерел електронної комерції чотирма мовами. Побудовано систему, яка формує онтологію з 486 концептів без ручного створення словників.

Традиційні підходи потребують залучення експертів та ручного створення правил, що не масштабується при роботі з мільйонами товарів [3,4]. Сучасні методи можна поділити на три класи: словникові системи (потребують ручних синонімів), правилові системи (не масштабуються), системи на основі машинного навчання (потребують великих розмічених датасетів) [5].

МЕТОДОЛОГІЯ

Для виокремлення концептів використано трансформерні нейронні мережі, які генерують векторні представлення текстів у багатовимірному просторі. На відміну від класичних підходів (стемінг, лематизація), трансформери захоплюють семантичний зміст та автоматично виявляють близькість концептів у різних мовах без словників перекладу.

У роботі використано модель gte-small (384-вимірні вкладення) з прискоренням на GPU. Для швидкого пошуку застосовано алгоритм HNSW [6], який будує багаторівневий граф навігації зі складністю O(log N).

Онтологію побудовано у вигляді ієрархічної структури з класами: Brand (59 брендів з 847 варіаціями), ProductType (16 типів з ієрархією), Color (337 відтінків у 12 родинах), Material (47 матеріалів з 213 синонімами), Size (23 стандарти зі 156 зв'язками конвертації EU/US/UK/CM), Gender (4 категорії), Product (абстрактний клас з зв'язками до всіх класів).

Система має архітектуру конвеєра (рис. 1) з п'ятьма етапами. Модулі написано мовою Python з використанням sentence-transformers, hnswlib, MongoDB.

Етап 1. Збір даних з 34 джерел (JSON формат, 4 мови: англійська 60%, італійська 25%, іспанська 10%, французька 5%).

Етап 2. Попередня обробка – витягування та очищення тексту, формування запиту.

Етап 3. Генерація вкладень – завантаження моделі gte-small, генерація вкладень для канонічних концептів, побудова HNSW-індексів (M=16, efConstruction=200, efSearch=16, обсяг 4,2 МБ).

Етап 4. Гібридний пошук – чотиришарова архітектура:

• Шар 1: Правила (<1мс, 98% довіра, 45% товарів)

• Шар 2: Псевдоніми (~1мс, 95% довіра, 35% товарів)

• Шар 3: Семантика (~8мс, 70-90% довіра, 18% товарів)

• Шар 4: Валідація (~80мс, 60-80% довіра, 2% товарів)

Адаптивний оптимізатор пропускає шари за довірою: >0,98 → шари 3-4 пропускаються (80% товарів), >0,90 → шар 4 пропускається (95% товарів).

Етап 5. Доповнення онтології – додавання концептів, встановлення зв'язків, пошук синонімів, ієрархічні зв'язки. Для кольорів: текстова класифікація у 12 родин, за впевненості <0,7 – візуальна валідація через ΔE у LAB-просторі. Багатомовні еквіваленти через семантичну близькість (Blu↔Blue↔Azul↔Bleu).

РЕЗУЛЬТАТИ

Датасет: 700 000 товарів з 34 джерел, 4 мови. Обладнання: Apple M1 Max (10-ядерний CPU, 32-ядерний GPU, 32 ГБ RAM). Програмне забезпечення: Python 3.10, sentence-transformers, hnswlib, MongoDB.

Результат генерації. Система згенерувала онтологію: 59 брендів (847 варіацій), 16 типів (ієрархія до 3 рівнів), 337 кольорів (12 родин), 23 розміри (156 конвертацій), 4 статі, 47 матеріалів (213 синонімів). Загалом 486 концептів та 1216 зв'язків.

Для оцінки точності створено еталонну розмітку на 1000 товарів з усіх 34 джерел.

Концепт	Кількість	Зв'язків	Точність	Повнота	F1-міра
Brand	59	847	97,50%	96,80%	97,10%
ProductType	16	128	96,80%	95,40%	96,10%
Color	337 (12)	89	95,20%	93,80%	94,50%
Gender	4	64	95,70%	94,20%	94,90%
Size	23	156	94,30%	92,70%	93,50%
Material	47	213	91,80%	89,40%	90,60%
Всього	486	1216	95,20%	93,70%	94,50%

Таблиця 1. Результати автоматичної генерації онтології

Продуктивність: 581 товар за 44 секунди (13 тов/с). Пам'ять: HNSW 4,2 МБ, концепти 28 МБ, модель 133 МБ (165 МБ загалом).

Головна перевага системи – відсутність потреби у розмічених датасетах та можливість роботи з новими концептами без перенавчання. Система автоматично виявляє еквіваленти у різних мовах через семантичну близькість (cosine >0,95), що дозволяє інтегрувати товари від міжнародних постачальників без словників перекладу.

ВИСНОВКИ

Розроблена система автоматично генерує онтологію товарів без ручного створення правил та словників синонімів. Експериментальне розгортання на 700 000 товарів з 34 джерел чотирма мовами показало точність F1=95,2% при швидкості 13 товарів за секунду.

Система згенерувала онтологію з 486 концептів та 1216 зв'язків. Гібридна чотиришарова архітектура з адаптивним вибором забезпечує баланс між швидкістю та точністю.

Головні переваги: відсутність потреби у розмічених датасетах, автоматична багатомовність без словників перекладу, доповнення онтології без перенавчання. Систему можна адаптувати для інших галузей: медицина, фінанси, логістика.

ЛІТЕРАТУРА

1. Berners-Lee T. The Semantic Web. Scientific American. 2001. Vol. 284, no. 5. P. 34–43.
2. Biemann C. Ontology Learning from Text: A Survey of Methods. 2005. URL: https://www.researchgate.net/publication/200044378_Ontology_Learning_from_Text_A_Survey_of_Methods (дата звернення: 15.11.2024).
3. Жежерун О. П., Рєпкін М. С. Автоматична генерація онтологій на основі статей українською мовою. Наукові записки НаУКМА. Комп'ютерні науки. 2022. Том 5. С. 12–15. DOI: 10.18523/2617-3808.2022.5.12-15. URL: https://www.researchgate.net/publication/369350039_Automatic_Generation_of_Ontologies_Based_on_Articles_Written_in_Ukrainian_Language (дата звернення: 15.11.2024).
4. Chen Q., Lin J., Zhang Y. et al. Towards Knowledge-Based Personalized Product Description Generation in E-commerce. KDD 2019. URL: https://arxiv.org/abs/1903.12457 (дата звернення: 15.11.2024).
5. Papadakis G., Efthymiou V., Thanos E. et al. An analysis of one-to-one matching algorithms for entity resolution. The VLDB Journal. 2023. Vol. 32. P. 1369–1400. URL: https://doi.org/10.1007/s00778-023-00791-3 (дата звернення: 15.11.2024).
6. Malkov Y., Yashunin D. Efficient and robust approximate nearest neighbor search using HNSW graphs. IEEE TPAMI. 2020. Vol. 42, no. 4. P. 824–836. URL: https://doi.org/10.1109/TPAMI.2018.2889473