

Streszczenie

Systemy rekomendacyjne w e-commerce w znacznej mierze przejęły techniki modelowania sekwencyjnego z przetwarzania języka naturalnego, traktując historie interakcji użytkowników jako uporządkowane ciągi tokenów. Analogia ta jest jednak błędna: w odróżnieniu od słów w zdaniu, kolejność zakupów i kliknięć wynika z układu strony, harmonogramu promocji czy względów logistycznych, a nie z rzeczywistej dynamiki preferencji. Reprezentacje punktowe pogłębiają ten problem: uśrednianie wektorów produktów do pojedynczego wektora sprawia, że wielomodalne preferencje zwijają się do centroidu pozbawionego użytecznej informacji. W efekcie struktura behawioralna jest systematycznie tracona już na poziomie reprezentacji, zanim jakkolwiek architektura modelu zdąży to skompensować.

Niniejsza rozprawa proponuje alternatywę dystrybucyjną opartą na estymacji gęstości na rozmaitościach reprezentacji wektorowych produktów. Pierwszym wkładem jest Efficient Manifold Density Estimator (EMDE), metoda wykorzystująca haszowanie wrażliwe na lokalność z podziałem zależnym od gęstości danych do reprezentowania historii behawioralnych jako rozkładów prawdopodobieństwa. Metoda ta generuje szkice o stałym rozmiarze, które zachowują wielomodalną strukturę preferencji i umożliwiają integrację heterogenicznych modalności danych poprzez prostą konkatenację. Drugim wkładem jest BaseModel, samonadzorowany model fundacyjny, który na podstawie przeszłych rozkładów behawioralnych przewiduje przyszłe. Pojedynczy wstępnie wytrenowany rdzeń obsługuje zadania rekomendacji, klasyfikacji i regresji bez etykiet specyficznych dla danego zadania, co przenosi koszt obliczeniowy z trenowania specyficznego dla każdego zadania na jednorazowy etap pretrenowania.

Aby zweryfikować te wkłady, sformułowano i przetestowano hipotezy badawcze, a cele badawcze zdefiniowano i zrealizowano w trzech uzupełniających się kontekstach. Na benchmarkach akademickich BaseModel uzyskuje ponad 100% poprawy względem HSTU firmy Meta oraz najlepsze wyniki w 10 z 12 zadań RelBench. W międzynarodowych konkursach metody konsekwentnie plasowały się w czołowej trójce, obok zespołów z DeepMind, Baidu i NVIDIA, a rozwiązanie zaprezentowane na WSDM WebTour Challenge zdobyło nagrodę za najlepszą pracę (Best Paper Award). W środowisku produkcyjnym system przetwarza 42 miliardy zdarzeń miesięcznie w ponad 200 organizacjach, dostarczając mierzalnych wzrostów zaangażowania i konwersji. Uzyskane wyniki pokazują, że wybór reprezentacji ma większe znaczenie niż wybór architektury: proste sieci jednokierunkowe operujące na szkicach dystrybucyjnych przewyższają złożone architektury sekwencyjne i grafowe, które odrzucają tę samą strukturę, którą próbują modelować.

Słowa kluczowe: systemy rekomendacyjne, estymacja gęstości, reprezentacja wielomodalna, modele fundacyjne, haszowanie wrażliwe na lokalność, modelowanie behawioralne