Báo cáo tuần 2

Tìm hiểu Github Recommenders và Chạy thử một số Model trên bộ dữ liệu MIND

Người phụ trách: Anh Ngô Văn Vĩ

Người thực hiện: Nguyễn Đình Hiếu

Mục lục

Sau khi tìm hiểu repo github em đã clone repo về máy và tải dataset MIND về để chạy thử được kết quả như sau:

Phần 1: Dataset MIND

Bộ dữ liệu Microsoft News Dataset (MIND) là một bộ dữ liệu quy mô lớn phục vụ cho nghiên cứu hệ thống gợi ý tin tức. Bộ dữ liệu này được thu thập từ các nhật ký hành vi đã được ẩn danh của người dùng trên trang web Microsoft News.

MIND bao gồm khoảng 160.000 bài báo tiếng Anh và hơn 15 triệu bản ghi lượt hiển thị được tạo ra bởi 1 triệu người dùng. Mỗi bài báo chứa nội dung văn bản phong phú bao gồm tiêu đề, tóm tắt, nội dung chính, chuyên mục và các thực thể liên quan. Mỗi bản ghi lượt hiển thị bao gồm các sự kiện nhấp chuột, các sự kiện không nhấp chuột và lịch sử các bài báo đã được người dùng đó nhấp vào trước thời điểm hiển thị. Để bảo vệ quyền riêng tư của người dùng, mỗi người dùng đã được tách khỏi hệ thống sản xuất và được mã hóa thành ID ẩn danh một cách an toàn.

Phần 2: Chạy thử một số model

1, Deep Knowledge-Aware Network (DKN)

DKN là một mô hình học sâu sử dụng thông tin từ đồ thị tri thức (knowledge graph) để cải thiện hệ thống gợi ý. Cụ thể, DKN sử dụng phương pháp TransX để học biểu diễn từ đồ thị tri thức, sau đó áp dụng một khung CNN có tên là KCNN để kết hợp embedding của thực thể (entity embedding) với embedding của từ (word embedding), từ đó tạo ra vector embedding cuối cùng cho một bài báo. Việc dự đoán tỷ lệ click (CTR) được thực hiện thông qua một bộ chấm điểm neuron dựa trên cơ chế attention.

Đặc điểm của DKN:

- DKN là một mô hình học sâu dựa trên nội dung dùng để dự đoán CTR, thay vì dựa trên phương pháp lọc cộng tác truyền thống dựa trên ID.
- Mô hình khai thác các thực thể tri thức và kiến thức thông thường trong nội dung tin tức thông qua việc học kết hợp giữa biểu diễn ở cấp độ ngữ nghĩa và cấp độ tri thức của bài báo.

- DKN sử dụng một module attention để tính toán linh hoạt biểu diễn lịch sử tổng hợp của người dùng.

2, Long- and Short-term User Representations (LSTUR)

LSTUR là một phương pháp gợi ý tin tức nhằm nắm bắt cả sở thích dài hạn và ngắn hạn của người dùng. Cốt lõi của LSTUR gồm hai thành phần: bộ mã hóa tin tức và bộ mã hóa người dùng. Trong bộ mã hóa tin tức, mô hình học biểu diễn của bài báo từ tiêu đề của chúng. Trong bộ mã hóa người dùng, chúng tôi đề xuất học biểu diễn dài hạn của người dùng từ embedding của ID người dùng. Ngoài ra, chúng tôi đề xuất học biểu diễn ngắn hạn của người dùng từ các bài báo họ vừa xem gần đây thông qua mạng GRU. Bên cạnh đó, chúng tôi đề xuất hai phương pháp để kết hợp biểu diễn dài hạn và ngắn hạn của người dùng. Phương pháp đầu tiên là sử dụng biểu diễn dài hạn của người dùng để khởi tạo trạng thái ẩn của mạng GRU trong phần biểu diễn ngắn hạn. Phương pháp thứ hai là nối biểu diễn dài hạn và ngắn hạn lại thành một vector biểu diễn người dùng hợp nhất.

Đặc điểm của LSTUR:

- LSTUR nắm bắt cả sở thích dài hạn và ngắn hạn của người dùng.
- Mô hình sử dụng embedding của ID người dùng để học biểu diễn dài hạn.
- Mô hình sử dụng tin tức mà người dùng vừa đọc gần đây thông qua mạng GRU để học biểu diễn ngắn hạn.

3, Neural News Recommendation with Personalized Attention (NPA)

NPA là một mô hình gợi ý tin tức sử dụng cơ chế attention được cá nhân hóa. Cốt lõi của NPA bao gồm mô hình biểu diễn tin tức và mô hình biểu diễn người dùng. Trong mô hình biểu diễn tin tức, một mạng CNN được sử dụng để học biểu diễn ẩn của các bài báo dựa trên tiêu đề của chúng. Trong mô hình biểu diễn người dùng, mô hình học biểu diễn người dùng dựa trên biểu diễn của các bài báo mà họ đã nhấp vào. Ngoài ra, attention cá nhân hóa ở cấp từ và cấp bài báo được sử dụng để nắm bắt mức độ quan trọng khác nhau đối với từng người dùng.

Đặc điểm của NPA:

- NPA là một phương pháp gợi ý tin tức dựa trên nội dung.
- Mô hình sử dụng mạng CNN để học biểu diễn tin tức, và học biểu diễn người dùng từ các bài báo họ đã nhấp vào.
- Attention cá nhân hóa ở cấp từ giúp NPA tập trung vào những từ quan trọng đối với từng người dùng khác nhau.
- Attention cá nhân hóa ở cấp bài báo giúp NPA tập trung vào những bài báo đã nhấp có mức độ quan trọng khác nhau đối với từng người dùng.

4, Neural News Recommendation with Multi-Head Self-Attention (NRMS)

NRMS là một phương pháp gợi ý tin tức sử dụng mạng nơ-ron với cơ chế multi-head self-attention. Cốt lõi của NRMS gồm bộ mã hóa tin tức và bộ mã hóa người dùng. Trong bộ mã hóa tin tức, mô hình sử dụng multi-head self-attention để học biểu diễn tin tức từ tiêu đề bằng cách mô hình hóa sự tương tác giữa các từ. Trong bộ mã hóa người dùng, mô hình học biểu diễn người dùng từ các bài báo họ đã đọc và sử dụng multi-head self-attention để nắm bắt mối liên hệ giữa các bài báo này. Ngoài ra, NRMS còn áp dụng attention cộng (additive attention) để chọn lọc các từ và bài báo quan trọng, từ đó tạo ra biểu diễn tin tức và người dùng giàu thông tin hơn.

Đặc điểm của NRMS:

- NRMS là một phương pháp gợi ý tin tức sử dụng mạng nơ-ron dựa trên nội dung.
- Mô hình sử dụng multi-head self-attention để học biểu diễn tin tức bằng cách mô hình hóa tương tác giữa các từ, và học biểu diễn người dùng bằng cách nắm bắt mối quan hệ giữa các bài báo họ đã đọc.
- NRMS sử dụng attention cộng (additive attention) để học biểu diễn tin tức và người dùng giàu thông tin hơn thông qua việc lựa chọn các từ và bài báo quan trọng.

Nguồn tham khảo:

[1] *Recommenders*

https://github.com/recommenders-team/recommenders