Tiêu đề:

**Xây dựng hệ thống gợi ý dịch vụ công trực tuyến sử dụng LightGCN**

**Ngụy Đình Thành, Lê Trung Hiếu và Phan Minh Tân**

Cục Cảnh sát quản lý hành chính về trật tự xã hội, Bộ Công an, 47 Phạm Văn Đồng, Mai Dịch, Cầu Giấy, Hà Nội

**Tóm tắt:**

Báo cáo trình bày việc xây dựng hệ thống gợi ý dịch vụ công trực tuyến nhằm cá nhân hóa trải nghiệm người dùng dựa trên lịch sử tương tác giữa người dùng và các dịch vụ công. Sử dụng thuật toán LightGCN, hệ thống khai thác dữ liệu giả lập gồm 1 triệu bản ghi tương tác từ 50.000 người dùng đối với 4.417 dịch vụ. Kết quả cho thấy LightGCN mang lại độ chính xác cao trong việc đề xuất so với một số mô hình truyền thống.

**Từ khóa:**

LightGCN, hệ thống gợi ý, dịch vụ công trực tuyến, khai thác dữ liệu, cá nhân hóa.

**1. Giới thiệu**

Trong bối cảnh chuyển đổi số và sự gia tăng đào sâu việc ứng dụng công nghệ trong dịch vụ công, việc phát triển các hệ thống gợi ý cá nhân hóa đang trở thành xu hướng quan trọng. Báo cáo này nhấn mạnh vai trò của công nghệ khuyến nghị trong việc tăng cường trải nghiệm người dùng và hiệu quả cung cấp dịch vụ công.

Các nghiên cứu trước đây đã chứng minh tác dụng của mô hình khuyên nghị như Collaborative Filtering (CF) và Matrix Factorization (MF) trong việc dự đoán sở thích người dùng. Tuy nhiên, các mô hình truyền thống này đều có hạn chế như không khai thác đủ thông tin kết nối đồ thị. LightGCN (Light Graph Convolution Network) đã được chứng minh là nâng cao hiệu quả gợi ý nhờ khai thác thông tin cấu trúc đồ thị một cách tối ưu. LightGCN loại bỏ các thành phần phức tạp không cần thiết trong GCN truyền thống và tập trung vào việc lan truyền và tổng hợp thông tin từ các nút lân cận trên đồ thị, giúp cải thiện hiệu quả và giảm chi phí tính toán.

LightGCN khác biệt so với các mô hình truyền thống ở việc tập trung hoàn toàn vào các bước lan truyền (propagation) và loại bỏ các thành phần như nonlinear activation và transformation matrix, vốn được sử dụng trong GCN tiêu chuẩn. Điều này làm cho LightGCN trở nên nhẹ hơn nhưng vẫn đạt hiệu quả cao trong các bài toán gợi ý.

**2. Phương pháp nghiên cứu**

**2.1. Dữ liệu**

Dữ liệu sử dụng là bộ giả lập gồm 1 triệu bản ghi lịch sử tương tác từ 50.000 người dùng đối với 4.417 dịch vụ công. Mỗi bản ghi bao gồm các thuộc tính:

* **Người dùng (userID):** danh tính duy nhất của người sử dụng.
* **Dịch vụ công (itemID):** danh tính duy nhất của dịch vụ.
* **Thời điểm (timestamp):** thời gian tương tác.

**2.2. Quy trình nghiên cứu**

Quy trình nghiên cứu bao gồm các bước sau:

***2.2.1. Tiền xử lý dữ liệu:***

* Loại bỏ các bản ghi không hợp lệ và chuẩn hóa dữ liệu để đảm bảo tính đồng nhất.
* Biểu diễn dữ liệu dưới dạng đồ thị hai phía (bipartite graph), trong đó các nút đại diện cho người dùng và dịch vụ công, còn các cạnh biểu diễn mối quan hệ tương tác giữa chúng.

***2.2.2. Mô hình LightGCN:***

Mô hình LightGCN được thiết kế nhằm tối ưu hóa việc gợi ý bằng cách tập trung vào lan truyền và tổng hợp thông tin từ đồ thị. Cụ thể, mô hình bao gồm các thành phần và công thức như sau:

**Lan truyền thông tin (Graph Convolution):**

LightGCN thực hiện lan truyền thông tin giữa các nút (người dùng và dịch vụ) trên đồ thị để cập nhật biểu diễn của các nút. Công thức lan truyền được định nghĩa như sau:

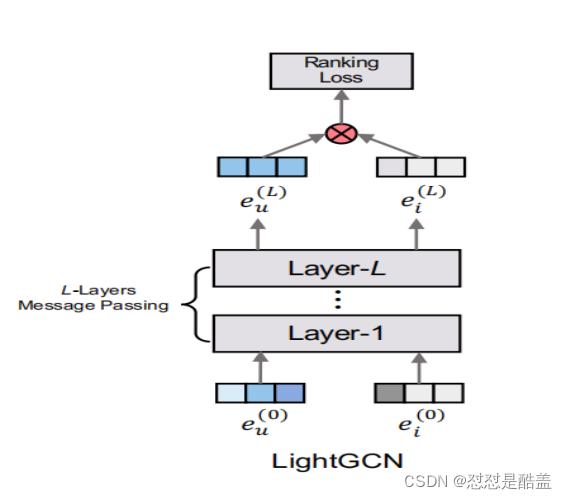
A group of mathematical equations

Description automatically generated

Trong đó:

*  : vector nhúng của người dùng u và dịch vụ i tại lớp k.
*  : tập các nút lân cận của người dùng u và dịch vụ i.
*  : hệ số chuẩn hóa (normalization coefficient), giúp cân bằng ảnh hưởng của các nút có số lượng lân cận khác nhau.

Quá trình lan truyền này được thực hiện qua K lớp, giúp tổng hợp thông tin từ các nút lân cận xa hơn.



**Tổng hợp biểu diễn (Embedding Aggregation):**

Sau khi thực hiện lan truyền qua K lớp, biểu diễn cuối cùng của một nút được tính bằng cách lấy trung bình cộng các biểu diễn qua tất cả các lớp. Cụ thể, công thức tổng hợp được định nghĩa như sau:

A close up of a white background

Description automatically generated

Trong đó:

*  : biểu diễn cuối cùng của người dùng u và dịch vụ i.
* ​: vector nhúng của người dùng u và dịch vụ i tại lớp k.
* K+1: số lượng lớp lan truyền, bao gồm cả biểu diễn ban đầu (k=0).

A diagram of a light graph convolution

Description automatically generated

Quá trình tổng hợp này giúp tích lũy thông tin từ các tầng khác nhau của mạng, đảm bảo rằng biểu diễn cuối cùng của một nút phản ánh đầy đủ thông tin từ các nút lân cận trực tiếp và gián tiếp.

**Hàm mất mát (Loss Function):**

LightGCN sử dụng hàm mất mát Bayesian Personalized Ranking (BPR), một hàm phổ biến trong hệ thống gợi ý, để tối ưu hóa thứ tự gợi ý giữa các cặp dịch vụ đã được tương tác và chưa được tương tác. Công thức được định nghĩa như sau:



Trong đó:

* D: tập dữ liệu huấn luyện, bao gồm các bộ ba (u, i, j) trong đó:
  + u: người dùng.
  + i: dịch vụ mà người dùng uuu đã tương tác.
  + j: dịch vụ mà người dùng uuu chưa tương tác.
* ​ : hàm sigmoid.
*  : vector nhúng cuối cùng của người dùng u, dịch vụ i, và dịch vụ j sau quá trình tổng hợp.
* λ : hệ số điều chuẩn (regularization coefficient) nhằm giảm overfitting.
*  : chuẩn L2​ của các vector nhúng.

Hàm mất mát này được thiết kế để học thứ tự ưu tiên giữa các dịch vụ, thay vì dự đoán giá trị tuyệt đối, giúp tăng độ chính xác của hệ thống gợi ý.

***2.2.3. Thực nghiệm và đánh giá:***

* Dữ liệu được chia thành tập huấn luyện (75%) và kiểm tra (25%).
* Các chỉ số Precision@K, Recall@K và NDCG@K được sử dụng để đánh giá hiệu quả của mô hình.
* So sánh kết quả của LightGCN với các mô hình khác như SAR (Smart Adaptive Recommendations) và NCF (Neural Collaborative Filtering) để chứng minh sự vượt trội của LightGCN.

**2.3. Phương pháp phân tích**

**Phân tích chất lượng gợi ý:**

* Sử dụng các chỉ số đánh giá phổ biến như Precision@K, Recall@K và NDCG@K để đo lường mức độ phù hợp của các đề xuất so với nhu cầu thực tế của người dùng.
* Tập trung vào việc phân tích kết quả dựa trên các nhóm người dùng có tần suất tương tác cao, trung bình và thấp để kiểm tra khả năng khái quát của mô hình.

**Phân tích hiệu suất mô hình:**

* Đánh giá thời gian huấn luyện và dự đoán của LightGCN so với các mô hình khác, bao gồm SAR (Smart Adaptive Recommendations) và NCF (Neural Collaborative Filtering).
* Kiểm tra sự tiêu tốn tài nguyên tính toán khi số lượng người dùng và dịch vụ công tăng lên, để đảm bảo mô hình có thể mở rộng cho hệ thống thực tế.

**Phân tích ảnh hưởng của tham số:**

* Tiến hành thử nghiệm với các giá trị khác nhau của siêu tham số như số lớp lan truyền, kích thước embedding, và hệ số học (learning rate) để xác định cấu hình tối ưu.
* Phân tích độ nhạy của mô hình đối với các thay đổi này nhằm đảm bảo tính ổn định trong các kịch bản sử dụng khác nhau.

**2.4. Công cụ và môi trường thực nghiệm**

**Ngôn ngữ lập trình và thư viện:**

* **Ngôn ngữ lập trình:** Python là ngôn ngữ chính được sử dụng để triển khai mô hình.
* **Thư viện:** các thư viện như PyTorch, NumPy Pandas và thư viện Recommenders.

**Môi trường thực nghiệm:**

* **Google Colab**: Nền tảng đám mây miễn phí hỗ trợ thực thi code Python.
* **GPU NVIDIA Tesla T4 hoặc tương đương**: Được kích hoạt trong Google Colab để tăng tốc huấn luyện mô hình.

**3. Kết quả**

**3.1. So sánh hiệu suất giữa các mô hình**

Các mô hình được so sánh gồm **LightGCN**, **SAR (Smart Adaptive Recommendations)** và **NCF (Neural Collaborative Filtering)** trên tập dữ liệu giả lập với 1 triệu bản ghi tương tác giữa 50.000 người dùng và 4.417 dịch vụ công. Bảng dưới đây tóm tắt các kết quả thực nghiệm:

| **Mô hình** | **Thời gian huấn luyện (s)** | **Thời gian gợi ý (s)** | **MAP** | **NDCG** | **Precision@K** | **Recall@K** |
| --- | --- | --- | --- | --- | --- | --- |
| LightGCN | 1015.6 | 2.3 | 0.210 | 0.387 | 0.370 | 0.231 |
| SAR | 452.3 | 3.4 | 0.308 | 0.347 | 0.332 | 0.206 |
| NCF | 3628.4 | 12.3 | 0.254 | 0.352 | 0.342 | 0.207 |

**3.2. Đánh giá kết quả**

* **Hiệu quả gợi ý (Precision@K, Recall@K):** LightGCN đạt **Precision@K** cao nhất (0.370), tiếp theo là NCF (0.342) và SAR (0.332). Kết quả này cho thấy LightGCN có khả năng đề xuất chính xác nhất so với nhu cầu thực tế của người dùng. Tương tự, **Recall@K** của LightGCN (0.231) cũng vượt trội, nhấn mạnh khả năng bao phủ tốt hơn trong các gợi ý.
* **Độ phù hợp tổng thể (MAP và NDCG):** Mặc dù SAR đạt **MAP** cao nhất (0.308), LightGCN vượt trội hơn về **NDCG** (0.387) – một chỉ số quan trọng để đánh giá thứ tự gợi ý. Điều này cho thấy LightGCN ưu tiên các mục quan trọng với người dùng ở các vị trí đầu danh sách đề xuất.
* **Hiệu suất thời gian:**
* Thời gian huấn luyện của LightGCN (**1015.6s**) thấp hơn NCF (**3628.4s**), chứng minh khả năng tối ưu tính toán nhờ thiết kế gọn nhẹ.
* Thời gian gợi ý của LightGCN (**2.3ms**) nhanh nhất, phù hợp cho các hệ thống yêu cầu phản hồi thời gian thực.

**3.3. Tổng kết**

LightGCN vượt trội về độ chính xác gợi ý và khả năng đáp ứng nhanh, dù thời gian huấn luyện dài hơn SAR. Kết quả này khẳng định LightGCN là lựa chọn phù hợp để xây dựng hệ thống gợi ý dịch vụ công trực tuyến, đặc biệt trong các ứng dụng có yêu cầu cao về hiệu suất và tính cá nhân hóa.

**4. Thảo luận**

**4.1. Ý nghĩa kết quả**

Kết quả thực nghiệm cho thấy **LightGCN** là mô hình tối ưu cho hệ thống gợi ý dịch vụ công trực tuyến, nhờ vào khả năng khai thác cấu trúc đồ thị một cách hiệu quả. So với các mô hình khác như **SAR** và **NCF**, LightGCN đạt độ chính xác cao hơn (Precision@K = 0.370) và thời gian gợi ý nhanh nhất (2.3ms). Điều này rất quan trọng trong bối cảnh hệ thống dịch vụ công cần xử lý khối lượng lớn yêu cầu trong thời gian ngắn, đồng thời đưa ra gợi ý phù hợp với nhu cầu cá nhân hóa của người dùng.

**4.2. So sánh với các nghiên cứu trước**

So với **SAR**, mô hình **LightGCN** cải thiện đáng kể về **Recall@K** và **NDCG**, mặc dù thời gian huấn luyện dài hơn. Điều này cho thấy LightGCN tận dụng tốt hơn mối quan hệ giữa người dùng và dịch vụ công thông qua đồ thị, vượt qua hạn chế của SAR – vốn chỉ dựa trên các quy luật tương quan đơn giản.

Trong khi đó, **NCF** đạt thời gian huấn luyện cao nhất (3628.4s) và thời gian gợi ý chậm nhất (12.3ms), nhấn mạnh chi phí tính toán lớn của các mô hình deep learning với cấu trúc phức tạp. LightGCN, nhờ việc loại bỏ các bước không cần thiết như kích hoạt phi tuyến và ma trận biến đổi, duy trì sự cân bằng giữa độ chính xác và hiệu suất tính toán.

**4.3. Hạn chế**

Mặc dù đạt kết quả tốt, mô hình LightGCN vẫn có một số hạn chế:

* **Thời gian huấn luyện:** Dù nhanh hơn NCF, LightGCN vẫn tiêu tốn thời gian huấn luyện gấp đôi SAR (1015.6s so với 452.3s). Điều này có thể là một thách thức khi mở rộng hệ thống cho dữ liệu lớn hơn.
* **Khả năng tổng quát:** Kết quả gợi ý chủ yếu tập trung vào nhóm người dùng có tần suất tương tác cao. Người dùng có ít tương tác có thể nhận được gợi ý kém chính xác hơn do thiếu dữ liệu đầu vào.

**4.4. Định hướng nghiên cứu tương lai**

Để cải thiện và mở rộng nghiên cứu này, các hướng đi sau có thể được xem xét:

* **Tăng cường dữ liệu:** Kết hợp các nguồn dữ liệu khác nhau, như thông tin người dùng hoặc đặc điểm dịch vụ, để tăng khả năng tổng quát của mô hình.
* **Tối ưu hóa hiệu suất:** Sử dụng các kỹ thuật phân tán hoặc triển khai mô hình trên các cụm GPU để giảm thời gian huấn luyện.
* **Tích hợp yếu tố thời gian:** Áp dụng các kỹ thuật Temporal Graph để xử lý sự thay đổi theo thời gian trong tương tác giữa người dùng và dịch vụ.

**4.5. Ứng dụng thực tế**

Hệ thống gợi ý dựa trên LightGCN có tiềm năng ứng dụng cao trong môi trường dịch vụ công trực tuyến, đặc biệt trong việc gợi ý các dịch vụ liên quan đến nhu cầu cá nhân của người dân. Điều này không chỉ nâng cao trải nghiệm người dùng mà còn giúp các cơ quan quản lý phân bổ nguồn lực hiệu quả hơn.

**5. Kết luận**

Báo cáo đã trình bày quá trình xây dựng hệ thống gợi ý dịch vụ công trực tuyến dựa trên mô hình **LightGCN**. Kết quả nghiên cứu cho thấy LightGCN là một lựa chọn phù hợp và hiệu quả trong việc khai thác cấu trúc đồ thị để tối ưu hóa các gợi ý, đặc biệt trong bối cảnh dữ liệu lớn và yêu cầu thời gian xử lý nhanh.

**LightGCN** vượt trội so với các mô hình truyền thống như **SAR** và **NCF** về độ chính xác gợi ý (Precision@K, Recall@K, NDCG) cũng như thời gian gợi ý, chứng minh tính hiệu quả của việc đơn giản hóa các thành phần không cần thiết trong Graph Convolution Network. Kết quả thực nghiệm cũng khẳng định rằng việc áp dụng LightGCN vào hệ thống dịch vụ công trực tuyến có thể cải thiện đáng kể trải nghiệm người dùng, hỗ trợ cá nhân hóa và tăng hiệu quả cung cấp dịch vụ.

Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế, bao gồm thời gian huấn luyện còn cao và khả năng gợi ý cho nhóm người dùng ít tương tác còn hạn chế. Những thách thức này có thể được giải quyết trong các nghiên cứu tương lai bằng cách tích hợp thêm dữ liệu bổ sung, cải thiện hiệu năng tính toán và sử dụng các kỹ thuật đồ thị động để xử lý yếu tố thời gian.

Với tiềm năng ứng dụng cao, hệ thống gợi ý dựa trên LightGCN không chỉ hữu ích trong môi trường dịch vụ công trực tuyến mà còn có thể mở rộng sang các lĩnh vực khác như giáo dục, y tế, và thương mại điện tử, góp phần thúc đẩy chuyển đổi số toàn diện và hiệu quả.

**6. Tài liệu tham khảo**

[1] Sun, N., Chen, T., Guo, W., & Ran, L. (2021). Enhanced Collaborative Filtering for Personalized E-Government Recommendation. *Sustainability, 13*(10), 1–12. doi:10.3390/su13105400

[2] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M. (2020). LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, 639–648. doi:10.1145/3397271.3401063

[3] Microsoft Recommenders Team. (2025). **Recommenders: Best Practices for Building Recommendation Systems**. GitHub Repository. Available at: <https://github.com/recommenders-team/recommenders>