**RECOMMENDER SYSTEM TUTORIAL**

**PART 1**

**1. Recommender System là gì ?**

Một hệ thống gợi ý (Recommender System) là một ứng dụng máy học/khai thác dữ liệu nhằm đưa ra những đề xuất cá nhân hóa cho người dùng dựa trên hồ sơ, sở thích hoặc hành vi trước đó của họ. Thay vì chờ người dùng nhập truy vấn như trong công cụ tìm kiếm, hệ thống gợi ý “đẩy” nội dung phù hợp đến họ, giúp khám phá sản phẩm, dịch vụ hoặc thông tin mà họ có thể chưa biết đến.

So sánh Recommender System và Search Enginee:

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Search Enginee** | **Recommender System** |
| **Cơ chế hoạt động** | Pull: người dùng chủ động nhập truy vấn | Push: hệ thống tự động đề xuất, không cần truy vấn |
| **Dữ liệu đầu vào** | Từ khóa (keywords), cú pháp truy vấn | Hồ sơ người dùng, lịch sử tương tác, đặc trưng nội dung |
| **Mục tiêu** | Tìm chính xác nội dung theo yêu cầu của người dùng | Khám phá và cá nhân hóa, tăng tính gắn kết và hài lòng |
| **Khả năng khám phá** | Giới hạn ở những gì người dùng biết để tìm | Mở rộng tầm nhìn, gợi ý nội dung người dùng chưa nghĩ tới |
| **Thời điểm tương tác** | Khi người dùng có nhu cầu rõ ràng | Liên tục, trong mọi ngữ cảnh (trang chủ, email, thông báo) |

**2. Dữ liệu hành vi người dùng**

|  |  |  |
| --- | --- | --- |
| **Dữ liệu hành vi người dùng** | **Thu thập** | **Tiền xử lý** |
| **Clickstream** | - Gắn SDK/JavaScript tracker (ví dụ Google Analytics, Snowplow) trên front‑end để log sự kiện click.  - Gửi event vào hệ thống message queue (Kafka) ngay khi người dùng nhấp.  - Mã hóa/ẩn danh user‑ID trước khi gửi. | - Sessionization: gom các click liên tiếp trong một phiên (ví dụ timeout 30 phút).  - Hash hoặc tokenize user‑ID, item‑ID để loại bỏ PII.  - Tính tần suất click, tỷ lệ click‑through (CTR) theo user/item.  - Lưu vào feature store dạng bảng tổng hợp (e.g. hourly, daily) để dễ scale. |
| **Purchase events** | - Log transaction server‑side (đảm bảo độ tin cậy cao hơn client).  - Mã hóa dữ liệu thanh toán nhạy cảm (số thẻ, địa chỉ) ngay tại nguồn.  - Chỉ giữ lại các trường cần thiết: user‑ID (đã ẩn danh), item‑ID, timestamp, giá trị đơn hàng. | - Aggregate theo user và theo item (tổng số lần mua, tổng doanh thu).  - Áp dụng k‑anonymity cho nhóm user có số lượng giao dịch rất ít (< k) để tránh định danh.  - Chuẩn hóa (normalize) giá trị đơn hàng nếu dùng làm feature đầu vào. |
| **Ratings (đánh giá – explicit feedback)** | - Cho phép user submit rating (1–5 sao) qua API với cơ chế xác thực token.  - Lưu log kèm timestamp, user‑ID (ẩn danh), item‑ID, rating. | - Loại bỏ outlier (rating bất thường, spam) bằng cách kiểm tra phân phối rating.  - Chuẩn hóa rating (z‑score) nếu kết hợp với các feature khác.  - Dùng differential privacy khi publish thống kê rating tổng hợp (ví dụ thêm noise). |
| **Dwell time / View time (thời gian xem nội dung)** | - Tính thời gian giữa event “page load” và “page unload” (hoặc chuyển tab).  - Gửi heartbeat event (ví dụ mỗi 5 giây) để biết user vẫn còn tương tác.  - Ẩn danh session bằng session‑ID không liên kết trực tiếp đến PII. | - Cắt ngưỡng: loại bỏ dwell time quá ngắn (< 2 giây) hoặc quá dài (ngủ quên tab).  - Tính trung bình, phân vị (percentile) dwell time theo user/item.  - Chuẩn hóa và mã hóa trước khi lưu vào feature store. |
| **Search queries (từ khóa tìm kiếm)** | - Log query server‑side kèm session‑ID và timestamp.  - Áp dụng filtering loại bỏ stop‑words, từ ngữ nhạy cảm.  - Mã hóa query hash nếu cần bảo mật thêm. | - Tokenization + stemming/lemmatization để chuẩn hóa từ khóa.  - Xây bộ từ vựng (vocabulary) và chuyển thành vector (TF‑IDF hoặc embedding).  - Giữ lại chỉ những query xuất hiện đủ tần suất để giảm noise và scale tốt. |
| **Add‑to‑cart / Wishlist** | - Log sự kiện khi user thêm vào giỏ hoặc wishlist.  - Gửi event qua API với user‑ID ẩn danh, item‑ID, loại sự kiện, timestamp. | - Tính tần suất add‑to‑cart, tỷ lệ chuyển đổi từ add‑to‑cart → purchase.  - Gom nhóm theo user/item để xây feature “tình trạng giỏ hàng” (empty, partial, full).  - Áp dụng aggregation window (hourly/daily) để giảm kích thước dữ liệu. |

- Các nguyên tắc chung để đảm bảo riêng tư và mở rộng:

+ Pseudonymization / Anonymization: luôn thay thế user‑ID gốc bằng hash hoặc token.

+ Minimize Data: chỉ thu thập và lưu trữ các trường thật sự cần thiết.

+ Encrypt in transit & at rest: HTTPS/TLS cho truyền, AES‑256 cho lưu trữ.

+ Differential Privacy: thêm noise khi chia sẻ analytics tổng hợp.

+ Streaming & Batch Hybrid: dùng Kafka + Spark/Flink để xử lý real‑time và batch, đảm bảo scale với khối lượng lớn.

+ Feature Store: lưu trữ feature đã xử lý, tái sử dụng cho training và serving, tránh tính toán lại nhiều lần.

**3. 5 chỉ số then chốt để giám sát một hệ thống gợi ý đang chạy**

3.1. Click‑Through Rate (CTR)

- Công thức:

- Instrumentation:

+ Gửi event recommendation\_impression mỗi khi render list gợi ý.

+ Gửi event recommendation\_click khi user click vào một item trong list.

+ Dồn về hệ thống Kafka → Spark/Flink Streaming để tính CTR theo window (1h, 24h).

- Cảnh báo:

+ Nếu CTR giảm >20 % so với trung bình 7 ngày (rolling), tự động push alert lên Slack/PagerDuty.

+ Dùng anomaly detection (ví dụ Holt‑Winters) để phát hiện biến động bất thường.

3. 2. Conversion Rate (CVR)

- Công thức:

- Instrumentation:

+ Event recommendation\_click gắn recommendation\_id.

+ Event purchase gắn source=recommendation nếu đơn hàng đến từ click đó.

+ Tính CVR theo session/user trên batch (hàng ngày) và real‑time (giờ).

- Cảnh báo:

+ CVR xuống dưới ngưỡng 5 % (hoặc giảm >15 % so với tuần trước) → gửi email cho team sản phẩm.

+ Kiểm tra trend CVR hàng giờ, nếu có spike giảm đột ngột, kích hoạt webhook.

3.3. Novelty (Tính mới)

- Định nghĩa: Tỷ lệ item được đề xuất mà user chưa từng tương tác (click/mua/đánh giá) trước đó.

- Công thức:

- Instrumentation:

+ Với mỗi batch gợi ý, so sánh item\_id với lịch sử tương tác lưu trong feature store.

+ Tính tỷ lệ này theo user trung bình, tổng hợp daily.

- Cảnh báo:

+ Nếu Novelty < 20 % (quá “an toàn”) hoặc > 80 % (quá “xa lạ”) so với target, gửi alert để điều chỉnh tham số mô hình.

+ Dùng threshold-based alerting trên Grafana.

4. Diversity (Đa dạng)

- Định nghĩa: Độ khác biệt giữa các item trong cùng một list gợi ý.

- Công thức (ví dụ):

- Instrumentation:

+ Lấy embedding hoặc taxonomy của item.

+ Tính pairwise dissimilarity cho mỗi recommendation list (K=10–20).

+ Trung bình hóa theo session/user, lưu metric hourly.

- Cảnh báo:

+ Diversity giảm dưới 0.3 (theo scale [0,1]) hoặc lệch >15 % so với baseline → alert.

+ Kết hợp anomaly detection để phát hiện sudden drop.

5. Catalog Coverage (Phủ sóng danh mục)

- Định nghĩa: Tỷ lệ item trong toàn bộ catalog từng được đề xuất ít nhất một lần trong kỳ.

- Công thức:

- Instrumentation:

+ Log recommendation\_impression ở cấp item.

+ Hàng ngày aggregate số item duy nhất xuất hiện.

- Cảnh báo:

+ Coverage < 30 % hoặc giảm > 10 % so với 30‑day rolling average → thông báo team dữ liệu.

+ Thiết lập alert batch job failure (nếu pipeline không chạy).

- Công cụ gợi ý triển khai cảnh báo

+ Metrics pipeline: Prometheus + Grafana hoặc Datadog

+ Alerting:

* Threshold-based (static) cho các ngưỡng cố định.
* Anomaly detection (dynamic) dùng EWMA/Holt‑Winters hoặc machine‑learning.
* Notification qua Slack/PagerDuty/Email.