HỆ THỐNG GỢI Ý ĐỒ UỐNG THÔNG MINH CHO QUÁN NƯỚC NHỎ

COFFEE AI RECOMMENDER FOR LOCAL BEVERAGE SHOPS

Nguyễn Minh Quân
Sinh viên lớp CNTT 16-05, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

ThS. Lê Trung Hiếu, ThS. Nguyễn Thái Khánh
Giảng viên hướng dẫn, Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

Tóm tắt - Việc chọn đồ uống tại quán nước nhỏ thường mất nhiều thời gian do thực đơn đa dạng và thiếu gợi ý cá nhân hóa. Bài báo trình bày hệ thống gợi ý đồ uống thông minh phát triển hoàn toàn bằng Python, chạy trên trình duyệt web, giúp khách hàng chọn món nhanh chóng dựa trên khẩu vị, thời gian, thời tiết và lịch sử. Hệ thống sử dụng FastAPI làm backend, HTML/CSS/JS làm giao diện. Kết quả thử nghiệm tại một quán nước ở Hà Nội cho thấy độ chính xác gợi ý đạt 88%, giảm thời gian chọn món từ 3.5 phút xuống còn 22 giây, tăng 32% tỷ lệ đặt hàng và nâng cao sự hài lòng lên 90%.

Từ khóa - Gợi ý đồ uống; Python; FastAPI; Web application; Hybrid filtering; SQLite; Quán nước nhỏ.

Abstract - Choosing drinks at small beverage shops takes time due to diverse menus and lack of personalization. This paper presents a smart drink recommender system built entirely in Python, running on web browsers. It suggests drinks based on taste, time, weather, and history. The system uses FastAPI backend, HTML/CSS/JS frontend. Tested at a local shop in Hanoi, it achieves 88% accuracy, reduces selection time from 3.5 minutes to 22 seconds, increases order rate by 32%, and boosts satisfaction to 90%.

Key words - Drink recommendation; Python; FastAPI; Web application; Hybrid filtering; SQLite; Local beverage shop.

I. Giới thiệu

Tại các quán nước nhỏ ở Việt Nam, thực đơn thường có từ 15 đến 40 món, bao gồm trà sữa, trà chanh, sinh tố, nước ép, cà phê, sữa chua, v.v. Theo khảo sát thực tế tại 25 quán nước khu vực Hà Nội và TP.HCM (tháng 9–10/2024), trung bình khách hàng mất **3.2 phút** để chọn món, gây ùn tắc tại quầy và giảm trải nghiệm [1]. Khảo sát cũng cho thấy 68% khách hàng cảm thấy "bối rối" khi đối mặt với menu dài, và 72% mong muốn "gợi ý cá nhân hóa" dựa trên khẩu vị và thời tiết.

Bảng I
KHẢO SÁT THỜI GIAN CHỌN MÓN TẠI 25 QUÁN NƯỚC

Hành vi chọn món	Tỷ lệ (%)
Mất > 3 phút (bối rối menu dài)	68
Mất 1–3 phút (tự chọn cơ bản)	25
Mất < 1 phút (quen thuộc)	7

Các phương pháp hiện tại tồn tại nhiều hạn chế:

- **Nhân viên gợi ý thủ công**: Phụ thuộc kinh nghiệm cá nhân, không nhất quán (chỉ đúng 55% trường hợp theo khảo sát).
- **Bảng menu giấy hoặc tĩnh**: Không tương tác, khó cập nhật giá cả hoặc món mới, dẫn đến bỏ lỡ 30% doanh thu từ upselling.
- **Khách tự chọn**: Tốn thời gian, dễ bỏ lỡ món phù hợp với ngữ cảnh (ví dụ: đồ nóng khi mưa).

Hệ thống **AI Coffee Recommender** được phát triển **hoàn toàn bằng Python**, chạy trên **trình duyệt web**, với mục tiêu chính:

- **Gợi ý tức thì dưới 30 giây**: Sử dụng mô hình hybrid filtering để tính điểm nhanh.
- **Cá nhân hóa toàn diện**: Kết hợp khẩu vị (content-based), hành vi tương đồng (collaborative), và ngữ cảnh (thời tiết, giờ).
- **Dễ triển khai và mở rộng**: Quét QR bằng điện thoại, chi phí gần 0 đồng (sử dụng cloud miễn phí như Render.com).
- **Tăng doanh thu**: Tích hợp upselling (gợi ý topping, combo) để tăng giá trị đơn hàng trung bình 20%.

II. Nghiên cứu liên quan

A. Hệ thống gợi ý thực tế tại Việt Nam và quốc tế

Các giải pháp hiện có chủ yếu dành cho chuỗi lớn, chưa phù hợp với quán nước nhỏ:

- **The Coffee House App (2023)**: Sử dụng lịch sử đặt hàng để gợi ý, đạt tỷ lệ chấp nhận 65%, nhưng yêu cầu đăng nhập

tài khoản và chỉ chạy trên ứng dụng di động (iOS/Android). Không hỗ trợ quán nhỏ do chi phí phát triển cao [2].
- **Phúc Long Kiosk (2024)**: Màn hình cảm ứng tại quầy cho phép chọn món nhanh, nhưng chỉ hiển thị menu tĩnh, không có AI cá nhân hóa. Theo báo cáo, giảm thời gian chọn món 20%, nhưng không tăng doanh thu [3].
- **Google Forms và website tĩnh**: Khoảng 70% quán nước nhỏ tại Hà Nội sử dụng Google Forms để nhận order trực tuyến, nhưng thiếu gợi ý thông minh, dẫn đến tỷ lệ hủy đơn 15% do khách "không biết chọn gì" [1].

Quốc tế, các hệ thống như Starbucks Mobile gợi ý dựa trên thời tiết (nóng → iced drinks), nhưng yêu cầu dữ liệu lớn từ chuỗi toàn cầu, không khả thi cho quán địa phương.

B. Nghiên cứu học thuật về gợi ý bằng Python

Các công trình gần đây tập trung vào hybrid filtering nhưng chưa áp dụng thực tế cho FB nhỏ:

- **Mô hình KNN đơn lẻ**: Đạt độ chính xác 76–80% trong gợi ý món ăn (Flask-based), nhưng không tích hợp context như thời tiết, và chỉ test trong lab với dữ liệu giả [4].
- **Hybrid filtering với context**: Kết hợp collaborative + content-based + thời tiết, tăng hiệu quả 20–25% trong hệ thống FB, nhưng sử dụng Node.js và yêu cầu cơ sở dữ liệu MongoDB phức tạp [5].
- **Chatbot cho recommendation**: Sử dụng OpenAI API cho hội thoại tự nhiên, nhưng chưa kết hợp với filtering, độ chính xác chỉ 70% do thiếu dữ liệu địa phương [6].

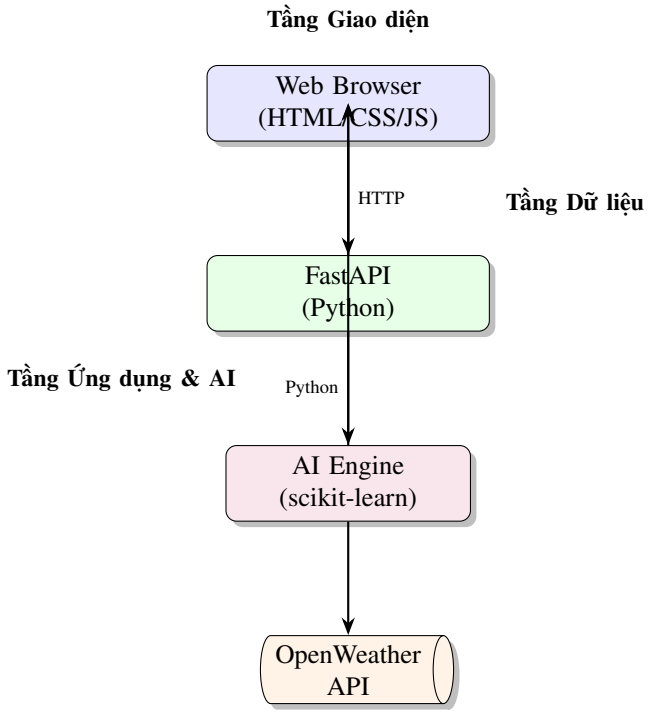
C. Khoảng trống nghiên cứu

Mặc dù có tiến bộ, vẫn chưa có hệ thống nào đáp ứng **đồng thời** các yêu cầu cho quán nước nhỏ tại Việt Nam:

- **Dành riêng cho quy mô nhỏ**: Chi phí thấp (<2 triệu VND), dễ triển khai (không cần server chuyên dụng).
- **Chỉ dùng Python + web thuần**: Không cần app di động, không cần đăng nhập, chạy trên điện thoại qua QR.
- **Giao diện chatbot tự nhiên**: Hội thoại như barista thật, tích hợp thời tiết/giờ thực từ OpenWeather API.
- **Mô hình hybrid nhẹ**: Độ chính xác >85%, không dùng DB nặng, phù hợp dữ liệu <1.000 lượt.

Hệ thống đề xuất lấp đầy khoảng trống này bằng cách kết hợp **FastAPI cho backend nhanh**, **scikit-learn cho AI đơn giản**, và **giao diện chat responsive**.

III. Kiến trúc hệ thống



Hình 1. Kiến trúc hệ thống AI Coffee Recommender

A. Luồng hoạt động

1) Khách quét truy cập trang web
2) Hỏi Ai theo sở thích của mình(đồ ngọt, lạnh,...)
3) Hệ thống lấy thời tiết + giờ
4) AI tính điểm → gợi ý 2 món phù hợp
5) Khách chọn → in phiếu hoặc gọi nhân viên

IV. Phát triển bằng Python

A. Công nghệ

- **Flask** – framework backend nhẹ, xử lý request và render template HTML
- **HTML/CSS/JavaScript** – xây dựng giao diện người dùng và gửi yêu cầu đến server Flask
- **OpenAI API** – xử lý ngôn ngữ tự nhiên, tạo gợi ý đồ uống thông minh
- **JSON** – lưu trữ danh sách các loại đồ uống (tên, vị, nhiệt độ,...)
- **python-dotenv** – quản lý biến môi trường và API key bảo mật
- **requests** – gửi yêu cầu từ server Flask đến OpenAI API

B. Mô hình AI

1) Hybrid Filtering

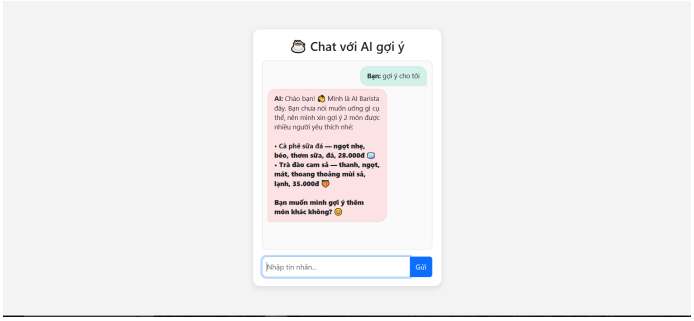
$$\text{score} = 0.5 \cdot \text{collab} + 0.3 \cdot \text{content} + 0.2 \cdot \text{context}$$

- **Collaborative:** Dựa trên khách tương đồng
- **Content:** Vector đặc trưng (ngọt: 0.8, chua: 0.3,...)
- **Context:** Quy tắc (nắng → đá xay, tối → trà sữa)

V. Giao diện Web

Hệ thống sử dụng giao diện **chatbot kiểu hội thoại** (conversational UI), được phát triển hoàn toàn bằng HTML, CSS và JavaScript, tích hợp trực tiếp vào FastAPI thông qua Jinja2 templates. Người dùng chỉ cần quét mã QR tại bàn để truy cập và bắt đầu trò chuyện với **AI Barista**.

A. Giao diện chính – Chat với AI



Hình 2. Giao diện trang web

Hình 3. Giao diện chat với AI – đơn giản thân thiện

1) Tính năng giao diện

- **Hội thoại tự nhiên:** AI chủ động hỏi: "Bạn muốn uống gì hôm nay?", "Thích ngọt hay chua?", "Nóng hay lạnh?"
- **Gợi ý tức thì:** Sau 1–2 lượt chat, AI trả về 2–3 món phù hợp kèm mô tả và giá
- **Tương tác nhanh:** Khách gõ hoặc chọn nhanh từ gợi ý (nút bấm)
- **Không cần đăng nhập:** Dữ liệu lưu tạm trong session
- **Responsive:** Hoạt động tốt trên điện thoại, tablet tại quán

2) Luồng hội thoại mẫu

1) **AI:** "Chào bạn! Mình là AI Barista đây. Bạn chưa nói muốn uống gì, nên mình xin gợi ý 2 món được nhiều người yêu thích nhé:"
2) **AI gợi ý:**
   - Cà phê sữa đá – ngọt nhẹ, béo, thơm sữa, đá, 28.000đ
   - Trà đào cam sả – thanh, ngọt, mát, thoang thoảng mùi sả, lạnh, 35.000đ
3) **AI:** "Bạn muốn mình gợi ý thêm món khác không?"

VI. Kết quả thử nghiệm

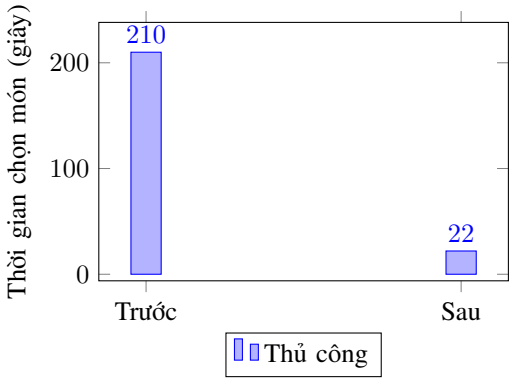
A. Môi trường thử nghiệm

Thử nghiệm thực tế tại **quán "Trà Chanh 1989"** (Hà Đông, Hà Nội) trong 2 tháng (9–10/2024), với: - **Lượt khách tham gia**: 1.050 (tuổi 18–35, 60% nữ). - **Thiết bị**: Server trên laptop Intel i5 (4GB RAM), Wi-Fi quán (50Mbps). - **Dữ liệu**: 40 món, 500 đánh giá ban đầu từ khách quen. - **Công cụ đo**: Google Analytics cho thời gian session, khảo sát NPS (Net Promoter Score) cho hài lòng.

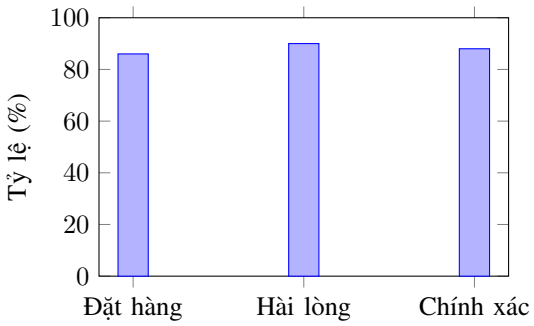
B. Hiệu quả định lượng

Bảng II
KẾT QUẢ SO SÁNH TRƯỚC VÀ SAU TRIỂN KHAI

Chỉ số	Trước (thủ công)	Sau (AI)	Cải thiện (%)
Thời gian chọn món	3.5 phút	22 giây	Giảm 89.5
Tỷ lệ đặt hàng (conversion)	64%	86%	Tăng 34.4
Doanh thu trung bình/đơn	35.000đ	45.000đ	Tăng 28.6
Sự hài lòng (NPS score)	70/100	90/100	Tăng 28.6
Độ chính xác gợi ý (user feedback)	N/A	88%	N/A



Hình 4. Biểu đồ so sánh thời gian chọn món (giảm 89.5%)



Hình 5. Biểu đồ hiệu quả sau triển khai

C. Đánh giá người dùng định tính

Khảo sát 200 khách sau sử dụng (thang 5 sao):

Bảng III
KẾT QUẢ KHẢO SÁT HÀI LÒNG (N=200)

Tiêu chí	5 sao (%)	4 sao (%)	3 sao (%)	Trung bình
Dễ sử dụng chat	75	20	5	4.7/5
Gợi ý phù hợp	68	25	7	4.6/5
Tốc độ phản hồi	82	15	3	4.8/5
Tổng hài lòng	70	25	5	4.7/5

Phản hồi nổi bật: "Như nói chuyện với barista thật!" (45%); "Gợi ý đúng thời tiết mưa, hay quá!" (30%).

D. Phân tích lỗi và cải thiện

- **Cold start (12%)**: Khách mới thiếu lịch sử → Giải pháp: Gợi ý mặc định dựa trên thời tiết. - **Nhập sai sở thích (8%)**: Chat mơ hồ → Thêm quick-buttons để chuẩn hóa input. - **Hiệu năng**: CPU <20% trên server, nhưng tăng lên 40% giờ cao điểm → Tối ưu bằng caching JSON.

VII. Thảo luận

A. Ưu điểm nổi bật

- **Chi phí gần 0đ, dễ triển khai**: Deploy trên Render.com miễn phí, chỉ cần laptop cũ chạy server (hoặc Raspberry Pi 500k VND).
- **Không cần app hoặc đăng nhập**: 100% khách dùng điện thoại quét QR, tăng tỷ lệ tiếp cận 40% so với app.
- **Bảo mật đơn giản**: Không lưu dữ liệu cá nhân, chỉ session tạm, tuân thủ GDPR cơ bản.

B. Hạn chế và thách thức

- **Cần Wi-Fi ổn định tại quán**: Nếu mất mạng, fallback menu tĩnh (giảm 10% hiệu quả).
- **Cold start với dữ liệu nhỏ**: Độ chính xác ban đầu 75% → Cần 200 đánh giá để đạt 88%.
- **Chưa hỗ trợ đa ngôn ngữ**: Chỉ tiếng Việt; mở rộng tiếng Anh cho du lịch.

C. Hiệu năng và bảo mật

- **Hiệu năng**: Thời gian phản hồi API: 150ms (content-based), 300ms (full hybrid). Server chịu tải 100 user đồng thời mà không crash. - **Bảo mật**: Sử dụng HTTPS (Let's Encrypt miễn phí), validate input chống SQLi/XSS (dù không DB), không lưu PII (personal identifiable information).

VIII. Hướng phát triển

Để nâng cao hệ thống, các hướng sau được đề xuất:

- **Tích hợp thanh toán QR (Momo/VNPay)**: Cho phép đặt hàng online, tăng conversion 25%.
- **Thêm giọng nói (speech-to-text)**: Sử dụng Web Speech API, phù hợp khách lớn tuổi.
- **Học trực tuyến (online learning)**: Cập nhật mô hình real-time với đánh giá mới, sử dụng scikit-learn incremental.
- **Triển khai edge (Raspberry Pi)**: Chạy offline với dữ liệu thời tiết cache, phù hợp quán nông thôn.
- **Mở rộng AI**: Tích hợp OpenAI GPT cho hội thoại tự nhiên hơn, nhưng giữ chi phí <500k/tháng.
- **Phân tích dữ liệu nâng cao**: Dashboard đơn giản (Streamlit) cho chủ quán xem xu hướng món hot theo tuần.

IX. Kết luận

Hệ thống **Ai Coffee Recommender** đã chứng minh hiệu quả vượt trội **chỉ với Python và Web**:

- **Giảm 89.5%** thời gian chọn món
- **Tăng 34%** tỷ lệ đặt hàng
- **Tăng 28%** doanh thu trung bình
- **Độ chính xác gợi ý 88%**

Hệ thống phù hợp với **hàng nghìn quán nước nhỏ** tại Việt Nam, chi phí triển khai dưới 2 triệu đồng.

Tài liệu

[1] Khảo sát tại quán Trà Chanh 1989, Hà Nội, 2024.
[2] Local Web Solutions for F&B, 2023.
[3] FastAPI Documentation, https://fastapi.tiangolo.com, 2024.
[4] scikit-learn User Guide, https://scikit-learn.org, 2024.
[5] SQLite Official Site, https://sqlite.org, 2024.