Online Traveling System: Technical Challenges at Tripi.vn

Trần Bình Giang
Co-founder
gtran@tripi.vn

www.tripi.vn

Outline

- Giới thiệu Tripi
- Các thách thức kỹ thuật tại Tripi
- Large Scale Search Engine
  - Tối ưu hoá trong bài toán crawling dữ liệu


Tripi TÌM·CHỌN·ĐI
TOUR DU LỊCH
VÉ MÁY BAY
KHÁCH SẠN
Giang Binh Tran
Hỗ trợ khách hàng (8h - 18h
(84-4) 3636 8333
VUI HÈ RỰC RỠ
Trải nghiệm như mơ
TOUR DU LỊCH
VÉ MÁY BAY
KHÁCH SẠN
Điểm đến hoặc tên công ty
Tháng khởi hành (bất kỳ)
Tìm kiếm
Tìm kiếm nâng cao
Tripi tìm kiếm các đối tác cung cấp khách sạn tốt nhất để giúp bạn tìm ra các ưu đãi tốt nhất.
FIDITOUR
etc
TRANSVIET
Travel
Vietravel
https://www.tripi.vn

Sàn du lịch trực tuyến lớn nhất Việt Nam

- Đầu tiên
- Hội tụ các yếu tố
  - *Du lịch trọn gói:* Tour, Landing Tours, Activities
  - *Lưu trú*: Khách sạn, Du Thuyền, Homestay
  - *Đưa đón:* Vé máy bay, car (thử nghiệm)
  - *Tripi Holidays:* Combo các sản phẩm Lưu trú + Vận tải + Activities

Sàn du lịch trực tuyến lớn nhất Việt Nam

800 gói du lịch

6000 khách sạn trong nước

100.000 khách sạn quốc tế
(chuẩn bị triển khai)

Gần **100** Công Ty Du Lịch

Outline

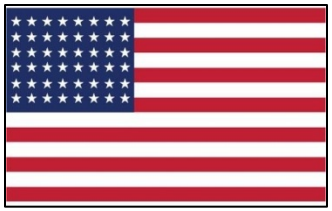
- Giới thiệu Tripi
  - Why Tripi?
- Các thách thức kỹ thuật tại Tripi
- Large Scale Search Engine

Thương mại điện tử B2C trên thế giới

Thị trường thế giới 2013 (tỷ $ USD)

$1,221

5 nước lớn nhất*



197 triệu người dùng

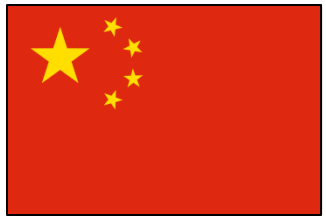
$ 395

(32%)

180 triệu người dùng

$ 230

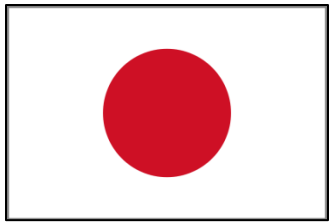
(19%)



309 triệu người dùng

$ 181

(15%)



86 triệu người dùng

$ 119

(10%)

Nguồn: AT Kiarney

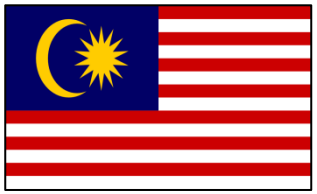
Thương mại điện tử B2C Đông Nam Á

Thị trường 6 nước lớn nhất trong Đông Nam Á 2013 (tỷ $ USD)

$ 7

3.2 triệu người dùng

$ 1.7



16 triệu người dùng

$ 1.3

5 triệu người dùng

$ 1.3

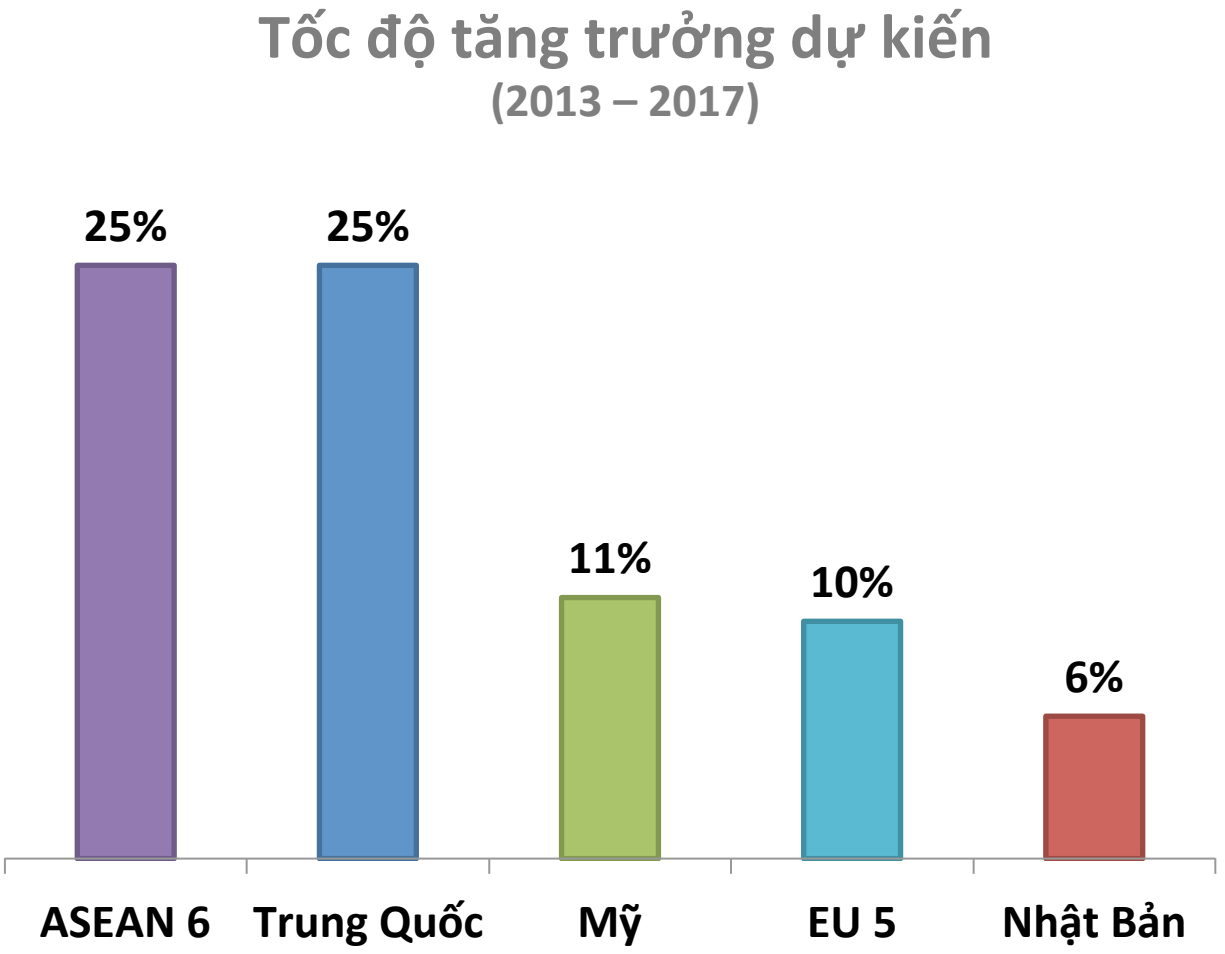
14 triệu người dùng

$ 0.9

~ 0.006% thị trường thế giới

Nguồn: AT Kiarney 2013

Thương mại điện tử B2C Đông Nam Á



Nguồn: AT Kiarney

Thương mại điện tử B2C Việt Nam

92 triệu dân số Việt Nam

44% sử dụng Internet

29% sử dụng 3G

65% truy cập bằng Smartphone

75% truy cập bằng laptop

58% tham gia giao dịch trực tuyến

Nguồn: Moore Corporation 2015 & VECITA 2014

Thương mại điện tử B2C Việt Nam

Các sản phẩm chính được mua trực tuyến

Quần áo

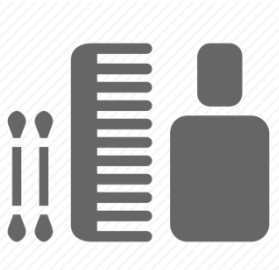
85%

Phụ kiện

27%

Đồ gia dụng

9%



Đồ dùng cá nhân

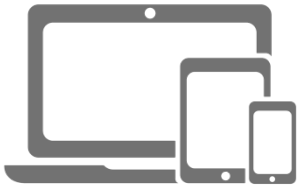
9%

Đồ ăn

6%

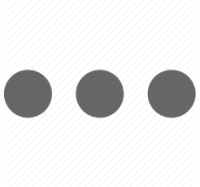
Phiếu giảm giá

4%



Đồ điện tử cá nhân

2%



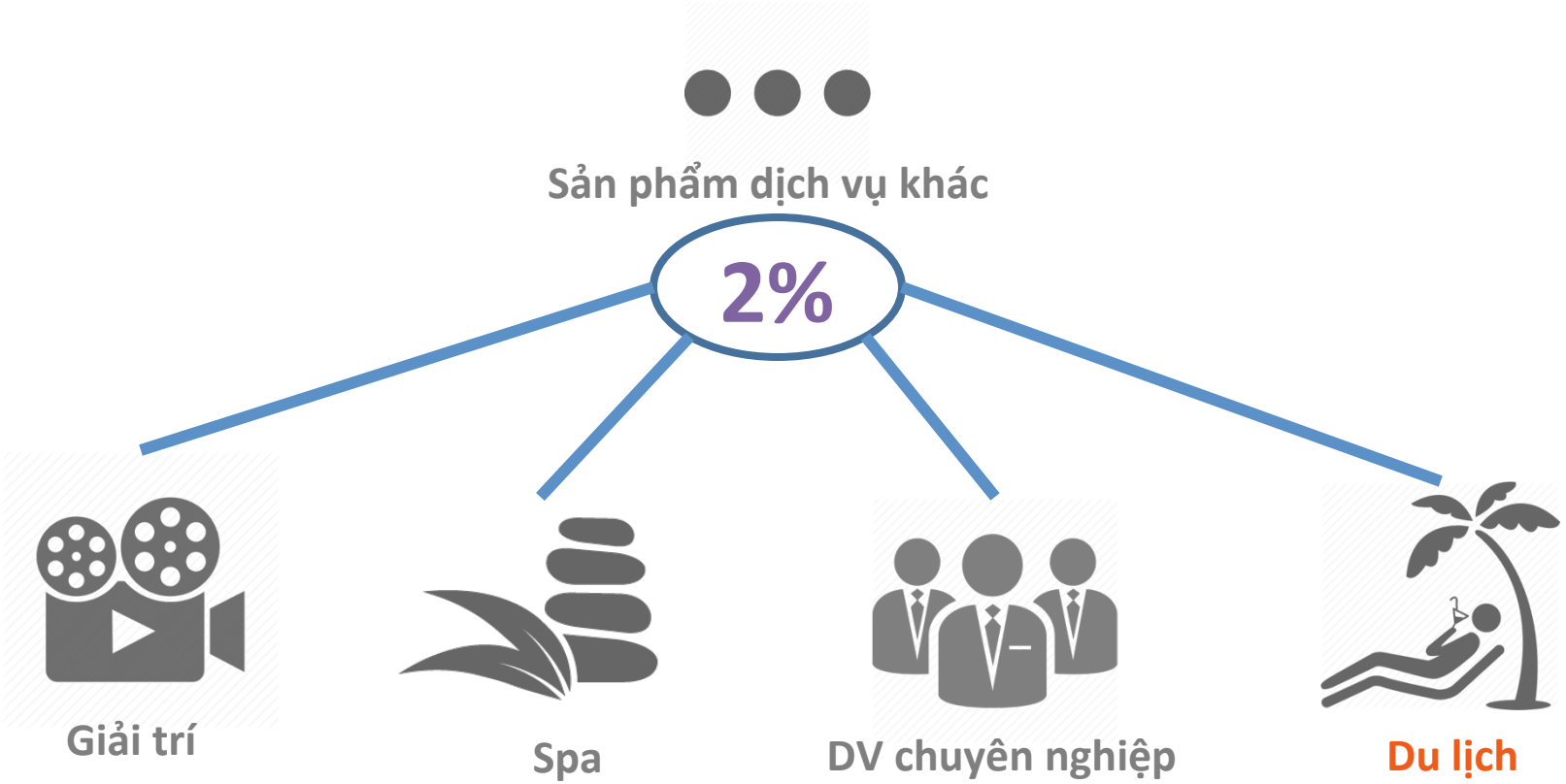
SPDV khác

2%

Nguồn: Moore Corporation

Thương mại điện tử B2C Việt Nam

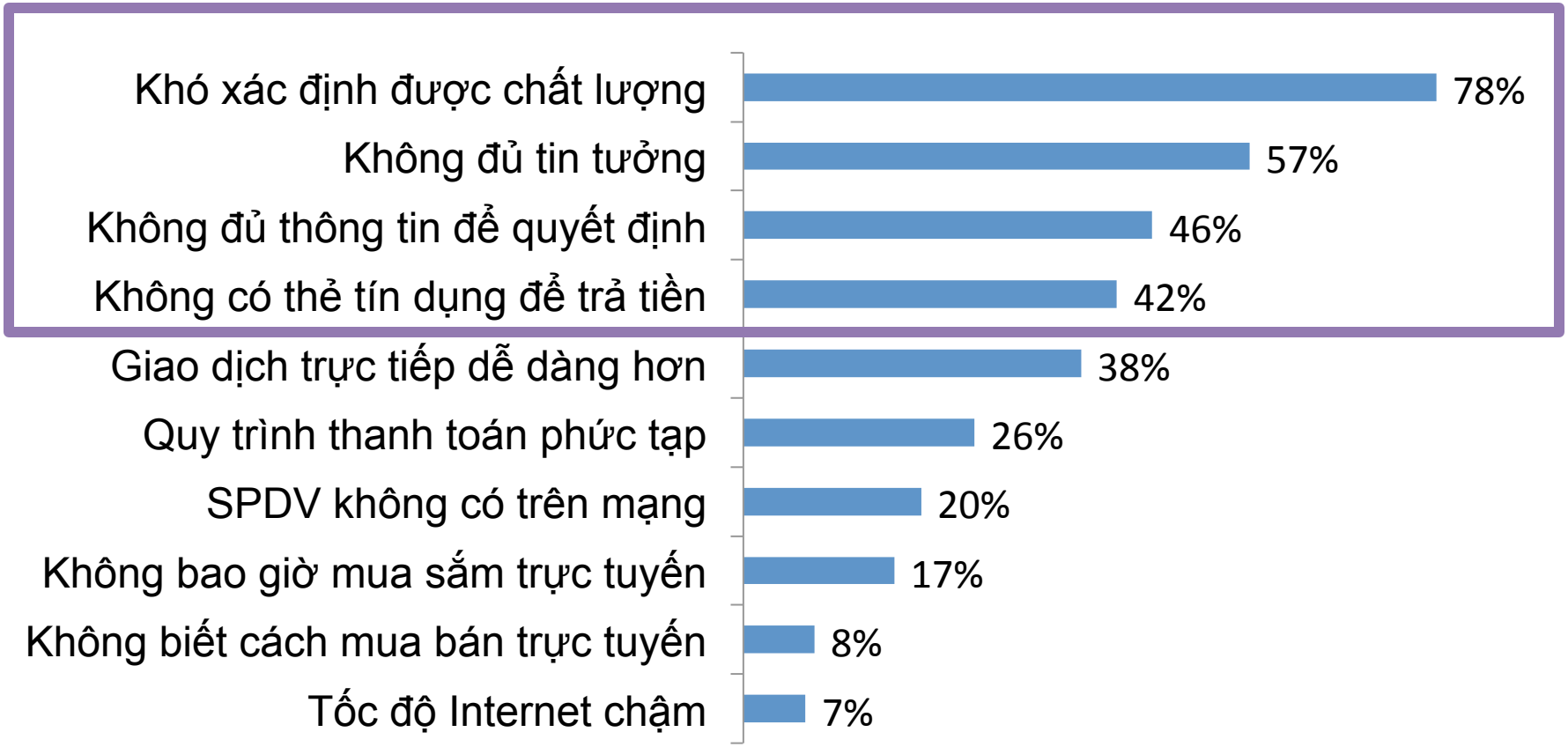
Các sản phẩm chính được mua trực tuyến



Nguồn: VECITA 2014

Thương mại điện tử B2C Việt Nam

Vấn đề chung trong thương mại điện tử Việt Nam



Nguồn: VECITA 2014

Giải pháp cho ngành du lịch

Một hệ thống thương mại điện tử du lịch, giúp kết nối người tiêu dùng với các nhà cung cấp để hỗ trợ và thúc đẩy:

Quyết định mua bán tốt hơn

Niềm tin người tiêu dùng

Uy tín sản phẩm dịch vụ du lịch

TRIPI.VN

Giúp người dùng tìm kiếm và lựa chọn:

Du lịch trọn gói

Vé máy bay

Khách sạn

Thuê xe

…từ những nhà cung cấp lớn nhất và uy tín nhất trong thị trường du lịch Việt Nam

Sứ mệnh của Tripi.vn

Xây dựng một hệ thống thương mại điện tử du lịch **thông minh** và **hữu ích** để giúp cho mọi hành trình của mọi người trở nên **đơn giản** hơn và **uy tín** hơn

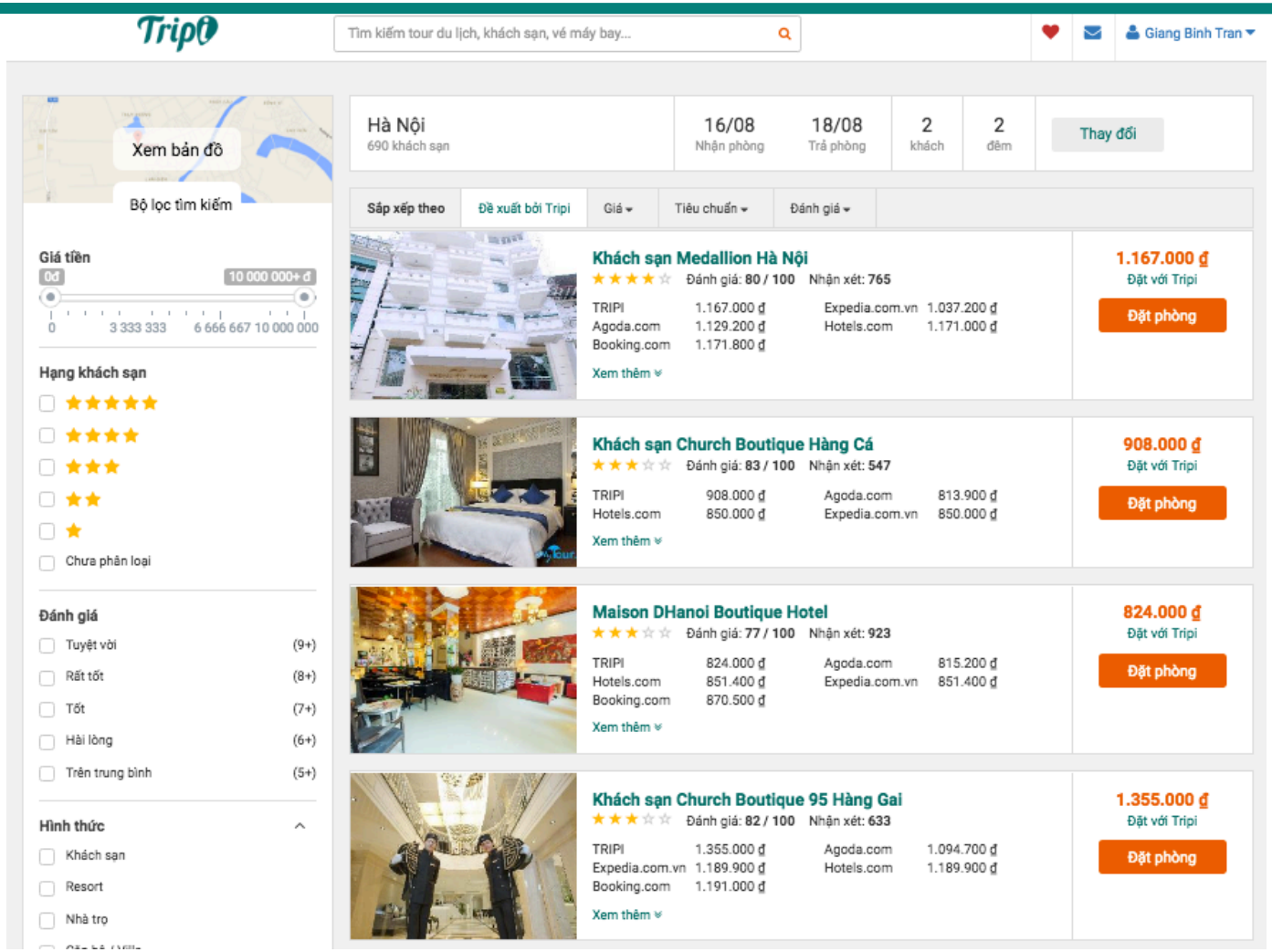
Outline

- Giới thiệu Tripi
- Các thách thức kỹ thuật tại Tripi
  - Các bài toán liên quan đến Data Science tại Tripi
- Large Scale Search Engine
- Summarization in News Media

Crawling @ Tripi

- Tech: Tripi là meta search engine
- Crawling Giá phòng khách sạn / Vé máy bay từ các nhà cung cấp
  - Airlines
  - Online Travel Agents (OTA)

Crawling @ Tripi



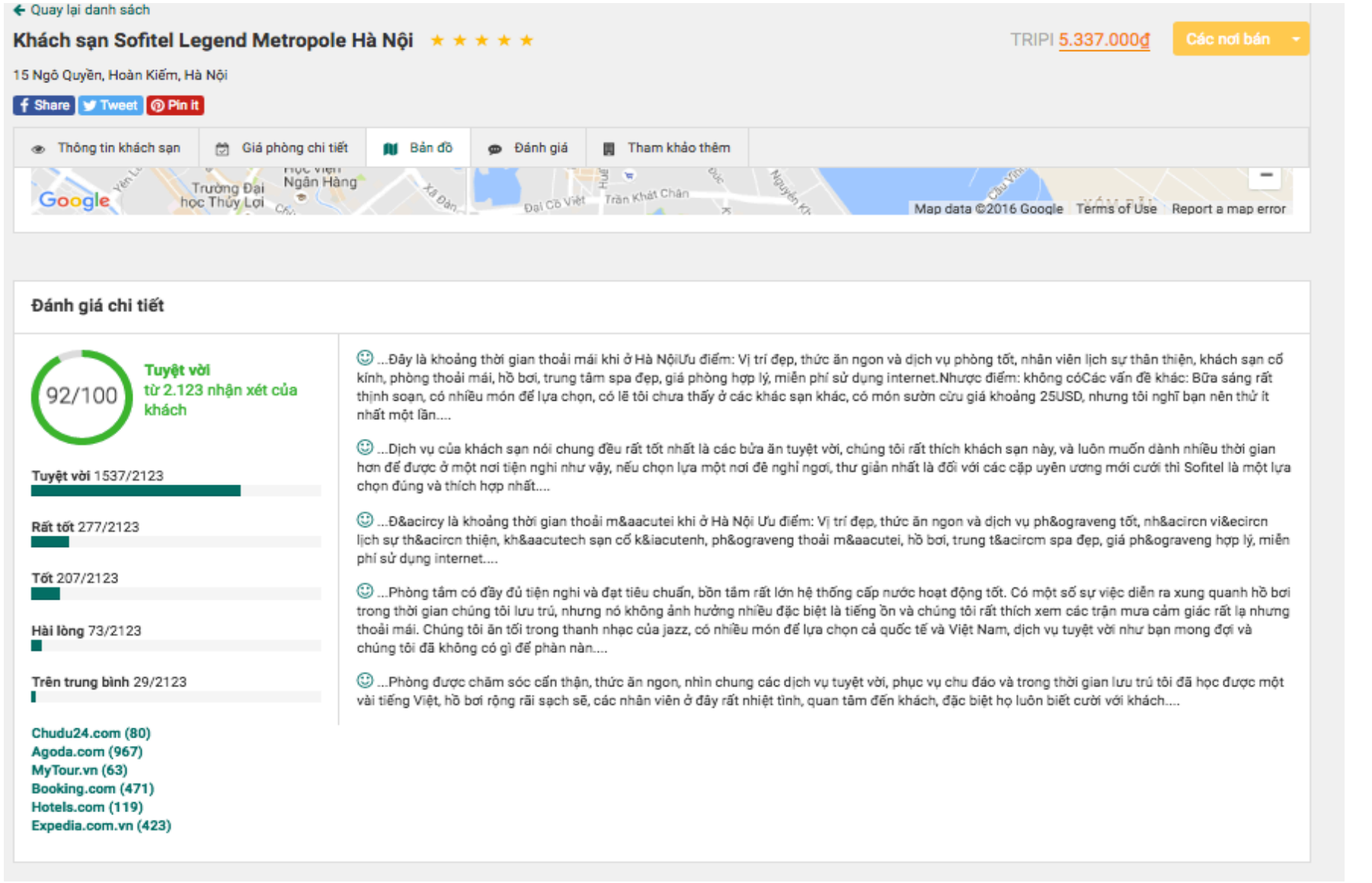
Crawl dữ liệu

- **Mục tiêu:**
  Crawl giá các khách sạn trước nhiều tháng
- **Thách thức:**
  Quá nhiều queries cần thực thi
  - H = 100K, Số lượng Hotels
  - R = 6 tháng, phạm vi thời gian để cache giá sẵn trong hệ thống
  - A = 10, số lượng các agents
  - **16.2B** queries
- Cần tối ưu hoá quá trình crawling

Crawl dữ liệu

- **Freshness**
  - Giá fresh nhất có thể
  - Xác suất 1 khách sạn thay đổi giá phòng, hết phòng trong một khoảng thời gian nào đó
  - Các yếu tố ảnh hưởng: thời gian, vị trí khách sạn, chất lượng khách sạn, comments, …
- **Politeness Policy**
  - Không vi phạm số lượng queries tối đa gửi tới các nhà cung cấp

Bóc tách thông tin



← Quay lại danh sách

Khách sạn Sofitel Legend Metropole Hà Nội ★ ★ ★ ★ ★

TRIPI 5.337.000₫ Các nơi bán

15 Ngô Quyền, Hoàn Kiếm, Hà Nội

Share Tweet Pin it

Thông tin khách sạn | Giá phòng chi tiết | Bản đồ | Đánh giá | Tham khảo thêm

Google Trường Đại học Thủy Lợi Học Viện Ngân Hàng Đại Cồ Việt Trần Khát Chân Map data ©2016 Google Terms of Use Report a map error

Đánh giá chi tiết

92/100 **Tuyệt vời** từ 2.123 nhận xét của khách

Tuyệt vời 1537/2123

Rất tốt 277/2123

Tốt 207/2123

Hài lòng 73/2123

Trên trung bình 29/2123

Chudu24.com (80)
Agoda.com (967)
MyTour.vn (63)
Booking.com (471)
Hotels.com (119)
Expedia.com.vn (423)

☺ ...Đây là khoảng thời gian thoải mái khi ở Hà NộiƯu điểm: Vị trí đẹp, thức ăn ngon và dịch vụ phòng tốt, nhân viên lịch sự thân thiện, khách sạn cổ kính, phòng thoải mái, hồ bơi, trung tâm spa đẹp, giá phòng hợp lý, miễn phí sử dụng internet.Nhược điểm: không cóCác vấn đề khác: Bữa sáng rất thịnh soạn, có nhiều món để lựa chọn, có lẽ tôi chưa thấy ở các khác sạn khác, có món sườn cừu giá khoảng 25USD, nhưng tôi nghĩ bạn nên thử ít nhất một lần....

☺ ...Dịch vụ của khách sạn nói chung đều rất tốt nhất là các bửa ăn tuyệt vời, chúng tôi rất thích khách sạn này, và luôn muốn dành nhiều thời gian hơn để được ở một nơi tiện nghi như vậy, nếu chọn lựa một nơi để nghỉ ngơi, thư giản nhất là đối với các cặp uyên ương mới cưới thì Sofitel là một lựa chọn đúng và thích hợp nhất....

☺ ...Đ&acircy là khoảng thời gian thoải m&aacutei khi ở Hà Nội Ưu điểm: Vị trí đẹp, thức ăn ngon và dịch vụ ph&ograveng tốt, nh&acircn vi&ecircn lịch sự th&acircn thiện, kh&aacutech sạn cổ k&iacutenh, ph&ograveng thoải m&aacutei, hồ bơi, trung t&acircm spa đẹp, giá ph&ograveng hợp lý, miễn phí sử dụng internet....

☺ ...Phòng tắm có đầy đủ tiện nghi và đạt tiêu chuẩn, bồn tắm rất lớn hệ thống cấp nước hoạt động tốt. Có một số sự việc diễn ra xung quanh hồ bơi trong thời gian chúng tôi lưu trú, nhưng nó không ảnh hưởng nhiều đặc biệt là tiếng ồn và chúng tôi rất thích xem các trận mưa cảm giác rất lạ nhưng thoải mái. Chúng tôi ăn tối trong thanh nhạc của jazz, có nhiều món để lựa chọn cả quốc tế và Việt Nam, dịch vụ tuyệt vời như bạn mong đợi và chúng tôi đã không có gì để phàn nàn....

☺ ...Phòng được chăm sóc cẩn thận, thức ăn ngon, nhìn chung các dịch vụ tuyệt vời, phục vụ chu đáo và trong thời gian lưu trú tôi đã học được một vài tiếng Việt, hồ bơi rộng rãi sạch sẽ, các nhân viên ở đây rất nhiệt tình, quan tâm đến khách, đặc biệt họ luôn biết cười với khách....

Bóc tách thông tin

- **Mục tiêu:**
  - Bóc tách những thông tin đặc trưng nhất của khách sạn từ reviews
    - Địa điểm
    - Đồ ăn
    - Dịch vụ
    - Giá cả …
- **Thách thức:**
  - Dữ liệu nhiễu, đa ngôn ngữ và không được gán nhãn trước

Dự đoán kinh doanh

- **Mục tiêu:**
  - Dự đoán giá cả, mức độ tiêu thụ phòng của các khách sạn, vé máy bay, car phục vụ cho kinh doanh
- **Thách thức:**
  - Giá cả và mức tiêu thụ thay đổi liên tục và phụ thuộc rất nhiều yếu tố khách quan khó đoán trước của thị trường
  - Mô hình dự đoán dựa vào dữ liệu trong quá khứ lẫn thông tin hiện tại để thích nghi nhanh
  - Cần độ chính xác cao

Learning to Rank

- **Mục tiêu:**
  - Sắp xếp các sản phẩm để tối ưu hoá tương tác (CTR) cũng như sales (CTA)
- **Thách thức:**
  - Mô phỏng và tham số hoá hành vi người dùng không hề dễ dàng

Outline

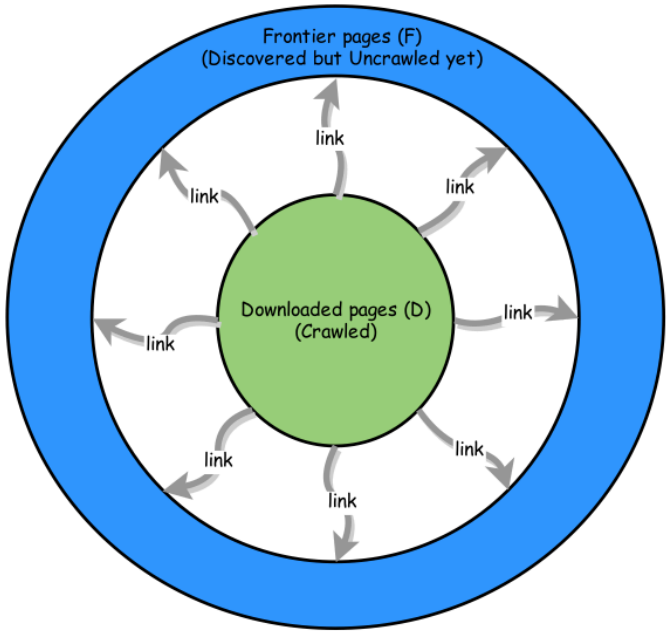
- Giới thiệu Tripi
- Các thách thức kỹ thuật tại Tripi
- Large Scale Search Engine
  - Tối ưu hoá trong bài toán crawling dữ liệu

A Random Walk Model for Optimization of Search Impact in Web Frontier Ranking

(SIGIR2015 – Best Paper Nominated)

Web Frontier Ranking Problem

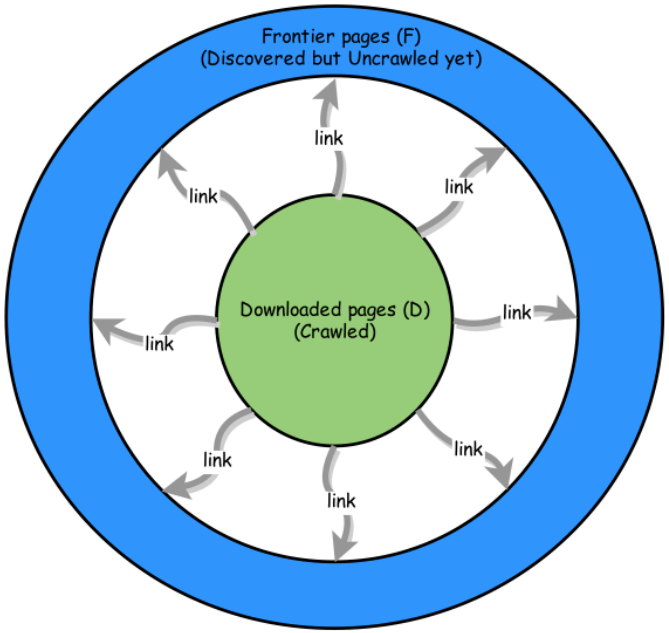
- Large-scale search engines cần crawl dữ liệu liên tục
  - Refreshing Process: fetch lại nội dung của những trang web đã được download từ trước đó
  - Discovery Process: fetch trang web mới



Ảnh hưởng trực tiếp tới kết quả tìm kiếm

Web Frontier Ranking Problem

- Large-scale search engines cần crawl dữ liệu liên tục
  - Refreshing Process: fetch lại nội dung của những trang web đã được download từ trước đó
  - Discovery Process: fetch trang web mới



"Chọn tập con các URL để tối ưu hoá lượng views / clicks trong tương lai"

DEFINITION 1 (FRONTIER PRIORITIZATION PROBLEM). *Given a set D of already crawled URLs, a set F of URLs in the frontier, a time period t, and a limit ℓ indicating the number of new URLs to be added to the collection,*[2] *find a subset $S \subset F$, such that $|S| \leq \ell$, and depending on the impact definition, either $\mathbb{E}_C(S) = \sum_{p \in S} I_C(p,t)$ or $\mathbb{E}_V(S) = \sum_{p \in S} I_V(p,t)$ is maximized, where $\mathbb{E}_C$ and $\mathbb{E}_V$ are two potential objective functions in the expected case.*

Các cách tiếp cận trước đó

- **Connectivity-based approach (ví dụ, PageRank)**
  - Sắp xếp các URL dựa vào mức liên kết
- **Search-centric approach**
  - Dự đoán xem một URL có khả năng được views/ clicks không trước khi crawl URL đó
  - Bài toán khó?
    - Nội dung của webpage chưa được download
    - Webpage chưa được hiển thị cho người dùng
    - …

Our approach: Combining two lines of work

- Kết hợp hai hướng giải quyết trên
- **Random Walk Model**
  - Kết hợp search impact vào trong hướng tiếp cận dựa trên connectivity
- **Machine learning-based enhancement**
  - Sử dụng machine learning trên các tính toán dựa vào Random Walk Model

Dữ liệu thực nghiệm

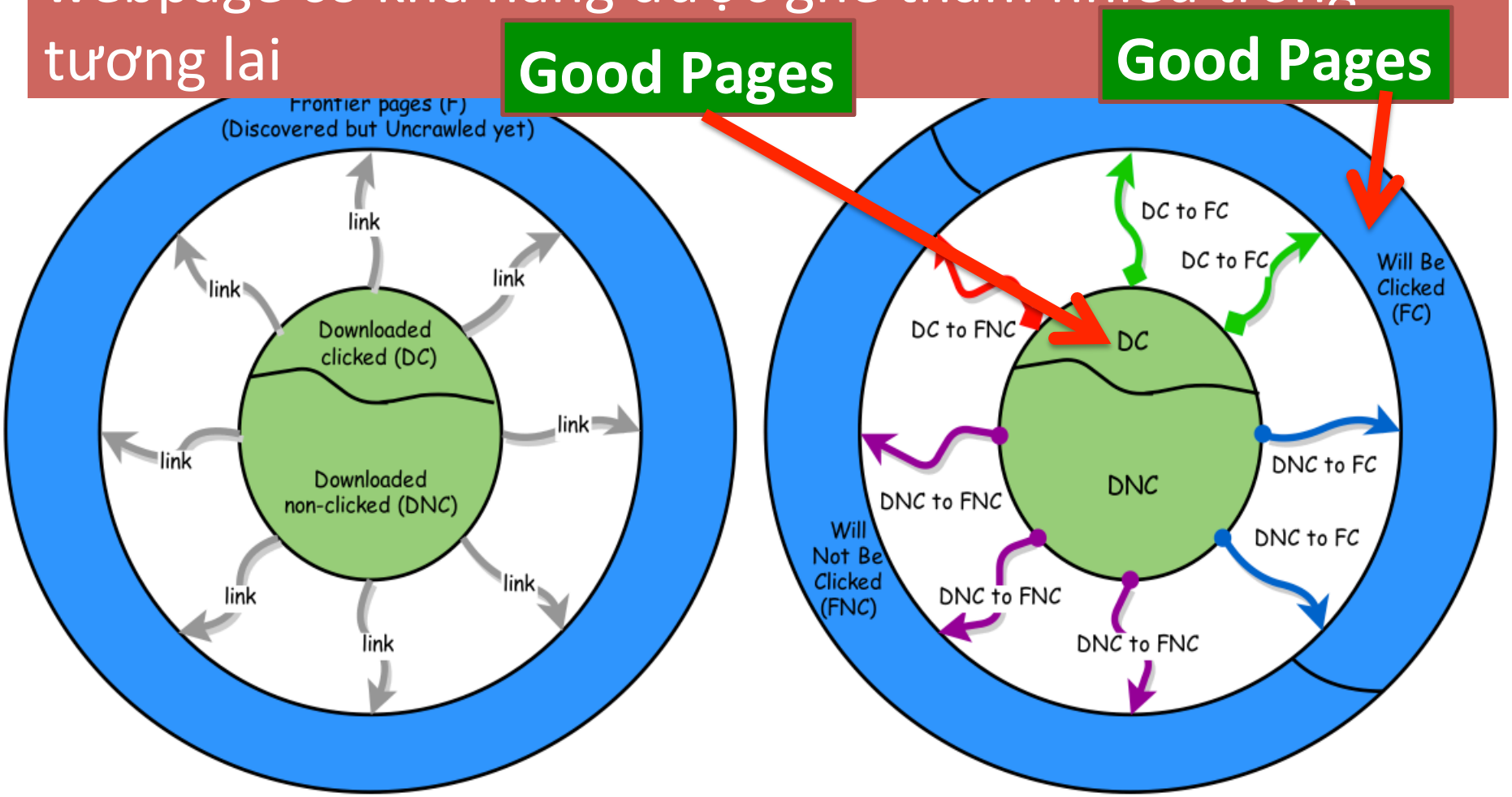
Dataset	total # of pages		# of clicked pages		# of viewed pages		# of links		frontier impact ratio	
	D	F	D_C	F_C	D_V	F_V	$D \to D$	$D \to F$	$\lvert F_C\rvert/\lvert F\rvert$	$\lvert F_V\rvert/\lvert F\rvert$
Wikipedia	476,805	127,411	98,709	34,926	148,812	45,213	1,532,121	347,071	27.4%	35.5%
WebCrawl	74,270,857	23,613,846	105,495	94,902	245,092	194,512	876,871,434	304,709,842	0.4%	0.8%

Phần rất nhỏ các URLs được coi là hữu ích từ mặt người dùng (click/view)

- **Wikipedia Dataset (~600K pages):**
- **Web crawl Dataset (~100M pages) - 6 tháng**

Impact-aware Frontier Prioritization

Webpage được ghé thăm nhiều thì thường link đến webpage có khả năng được ghé thăm nhiều trong tương lai

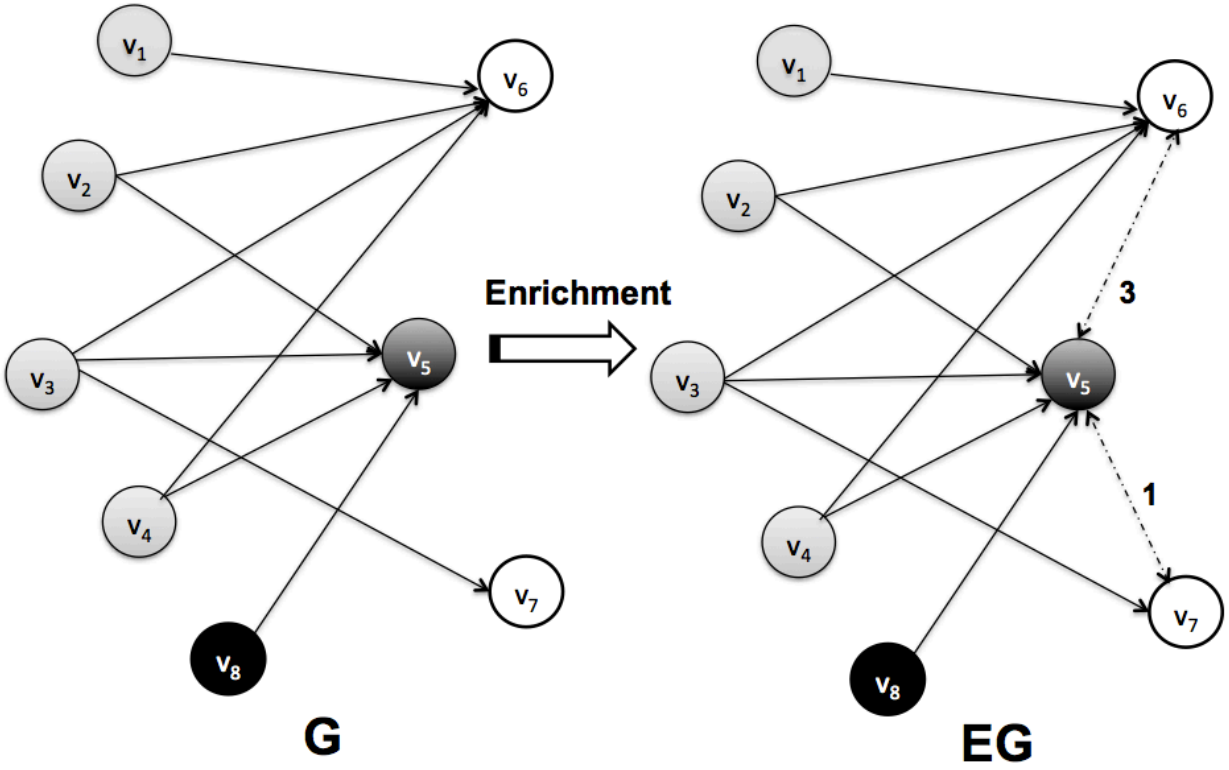


RW with Graph Enrichment

Jump Search Assumption:

"Người dùng sau khi thăm 1 webpage thì có khả năng tìm kiếm và đọc các webpage liên quan mặc dù không có link lẫn nhau"

Tạo ra các liên kết ảo giữa các webpages có độ tương đồng



RW-EG

Công thức

Đồ thị ban đầu

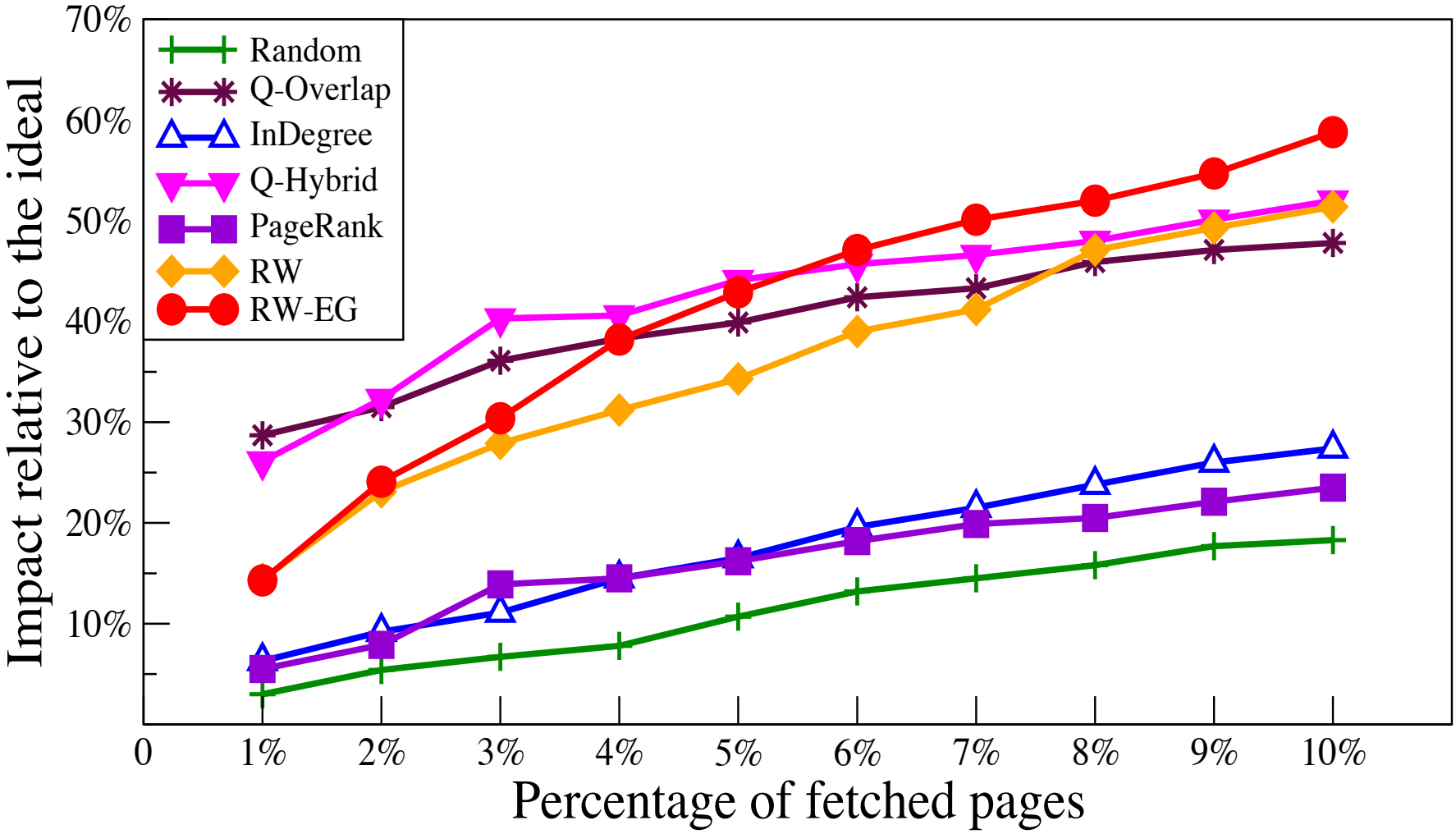
$$x_t(j) = \alpha\omega \sum_{i \in L_j^-} \mathcal{F}_i M_{ij} x_{t-1}(i)$$

$$+ \alpha(1-\omega) \sum_{i \in L_j^{-\prime}} \mathcal{F}'_i M'_{ij} x_{t-1}(i) \quad + (1-\alpha) v_j .$$

Đồ thị ảo

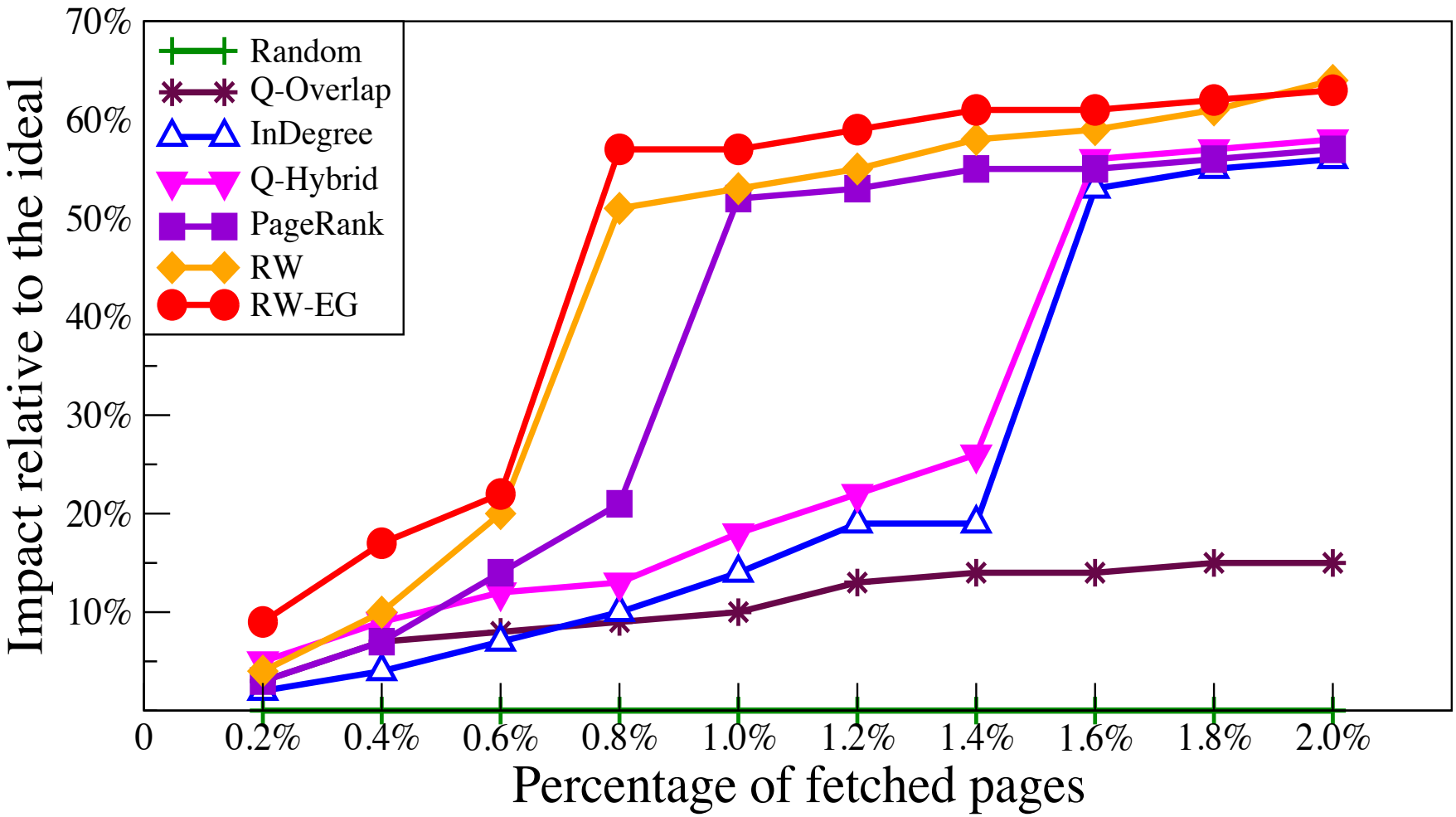
Chứng minh hội tụ

Click-Impact on Wikipedia Dataset



Similar results with *View Impact*

Click-Impact on WebCrawl Dataset



Similar results with *View Impact*

Takeaways

- Tripi là công ty công nghệ
- Rất nhiều bài toán thách thức cần lời giải tốt
- Rất nhiều data hữu ích
- Góc nhìn công nghiệp

  Large Scale Search Engine

  – Phương pháp Tối ưu hoá trong bài toán crawling dữ liệu

Thank you, folks !

– Thực tập
– R&D Engineers
– UI /UX Frontends

GTRAN@TRIPI.VN