1. **Giới thiệu**

*Query Log Mining* là một phân nhánh của một ngành khoa học tổng quát hơn là *Web Analytics.* Thật ra có thể coi nó là một trường hợp của khai phá sử dụng web. Theo Web Analytics Association, *“Web Analytics là một cách đo lường, tập hợp, phân tích và báo cáo dữ liệu Internet cho mục đích hiểu và tối ưu hóa sử dụng Web”.*

Cụ thể hơn, khai phá query log có liên quan đến tất cả các kỹ thuật dùng để trích xuất các thông tin thú vị từ query log của bộ máy tìm kiếm với mục đích cải thiện cả hiệu quả lẫn hiệu năng của dịch vụ tìm kiếm trên mạng.

Khai phá query log không chỉ liên quan đến dịch vụ tìm kiếm (nơi xuất phát của hầu hết các câu truy vấn) mà còn với những dịch vụ tổng quát hơn như quảng cáo dựa vào tìm kiếm, hay web marketing.

* 1. **Bộ máy tìm kiếm web (Web search engines)**

Các hệ thống tương đương với các bộ máy tìm kiếm ngày nay bắt đầu xuất hiện từ năm 1994. Có thể kể đến như: *WWWW (World Wide Web Worm), AliWeb, AltaVista, Excite, Lycos, Yahoo!, Google, ASK, MSN*. Ngày nay tìm kiếm được coi là một trong những ứng dụng hữu ích nhất trên mạng internet.

Trong bài báo tổng kết lại những thách thức trong việc thiết kế các bộ máy tìm kiếm hiện đại, *Baeza-Yates et al* đã nói:

*“Thách thức chính là thiết kế ra một hệ thống phân tán ở quy mô lớn có khả năng* ***thỏa mãn những yêu cầu của người dùng*** *và đồng thời đảm bảo truy vấn phải* ***sử dụng tài nguyên một cách hiệu quả nhất****, nghĩa là phải giảm chỉ phí cho mỗi câu truy vấn”.*

Do vậy, hai vấn đề được chỉ ra của dịch vụ tìm kiếm là: (i) chất lượng của kết quả trả về và (ii) tốc độ nhận được kết quả trả về.

Bộ máy tìm kiếm web là một phần của một hệ thống Information Retrieval (IR). Hệ thống IR ra đời từ đầu những năm 1960 để đáp ứng hai nhu cầu chính. Thứ nhất là cho phép người dùng tìm kiếm trong các thư viện số. Thứ hai là tìm kiếm dữ liệu trong thư viện số của cá nhân người dùng lưu trữ.

Hệ thống IR là một phần mềm với nhiệm vụ chính là trả lại một danh sách các tài liệu là kết quả của một câu truy vấn. Điều này khiến nó có vẻ tương tự với một hệ cơ sở dữ liệu (DB system). Tuy nhiên sự khác nhau cơ bản nhất giữa hệ DB và IR là hệ DB trả lại kết quả chính xác với câu truy vấn của người dùng, trong khi đó hệ IR sử dụng câu truy vấn là ngôn ngữ tự nhiên khiến cho việc trả lại một kết quả hoàn hảo là điều bất khả thi, mỗi câu truy vấn có thể có ý nghĩa khác nhau với mỗi người dùng khác nhau. Lấy ví dụ đơn giản, câu truy vấn có từ *“meta”,* vậy *meta* nghĩa là gì? Ký tự *meta* hay phím *meta* trên bàn phím máy tính, hoặc trong tiếng Tây Ban Nha thì *meta* nghĩa là bàn thắng (*goal*).

Mở rộng ra, bộ máy tìm kiếm web là một hệ thống IR ở quy mô rất lớn (*very large scale*). Điểm khác biệt với các hệ thống IR quy mô nhỏ là nó phụ thuộc vào một lượng lớn thông tin được lưu trữ trong *query log*.

Một trong những cách để nâng cao trải nghiệm của việc tìm kiếm thông tin đó là khai phá kiến thức từ những câu truy vấn cũ của người dùng. Một query log điển hình sẽ chứa thông tin về người dùng, các câu truy vấn, kết quả được người dùng chọn, ... Những kiến thức này có thể được trích chọn để nâng cao chất lượng (hiệu quả và hiệu năng) của hệ thống. Hình 1.1 cho ta thấy một đoạn query log của AOL. Mỗi câu truy vấn có năm trường thông tin: user id, câu truy vấn, thời điểm thực hiện truy vấn, thứ hạng của kết quả được người dùng chọn và địa chỉ URL của kết quả đó.

Trong bài báo này, chúng ta sẽ nghiên cứu một vài kỹ thuật làm việc với query log và cách dùng chúng để cải thiện hoạt động của bộ máy tìm kiếm. Chúng ta sẽ tổng kết lại các kết quả cơ bản liên quan đến query log: phân tích, kỹ thuật trích xuất thông tin, các kết quả thú vị nhất, các ứng dụng hữu ích nhất, các vấn đề mở và khả năng mở rộng.

* 1. **Phác họa kiến trúc của Web Search Engine**
  2. **Sự thật thú vị về các câu truy vấn**

Vì lý do về mặt thương mại, việc truy cập đến query log không phải là chuyện dễ dàng. Một số query log đã từng được công bố ra công chúng, đó là query log của AltaVista, TodoBR và AOL. Hình 1.5 cho ta thấy 250 câu truy vấn phổ biến nhất trong query log của AOL.

Trong thế giới ngày này, con người rất dễ bị stress. Stress có thể dẫn đến những câu truy vấn khá ngớ ngẩn, như người dùng #427326 vào 2006-04-21 21:16:51 với từ khóa *“where is my computer”*. Truy vấn này trả về xấp xỉ 200,000 kết quả!!!

Các bộ máy tìm kiếm thường công bố danh sách những câu truy vấn thú vị nhất. Với Google danh sách này được gọi là Zeitgeist:

*“Một bản ghi các câu truy vấn thú vị được người dùng hỏi nhiều lần, trong một tên miền tìm kiếm quốc gia hoặc tại Google.com, có thể tiết lộ một chút về điều kiện sống của con người tại đây.”*

Zeitgeist không tiết lộ những câu truy vấn xuất hiện nhiều nhất mà chỉ những câu truy vấn có sự tăng lên bất thường trong cộng đồng. Ví dụ, vào cuối năm 2007, Zeitgeist của Ý xếp hạng *“federico calzolari”* là câu truy vấn “lạm phát” nhất. Nhiều người tờ báo Ý bắt đầu tìm hiểu xem ai là người có liên hệ với câu truy vấn này. Kết quả là một nghiên cứu sinh ở Pisa, người đã định kỳ truy vấn Google tên của mình. Kết quả là một sự tăng trưởng bất thường từ khóa này và cuối cùng nó đã xuất hiện trong Zeitgeist. Nhiều người cho rằng Federico Calzolari đã hack thuật toán xếp hạng của Google.

Điều thú vị ở trên có thể ngầm định rằng người ta có thể xác định danh tính của người dùng thông qua phiên truy vấn của họ. Tuy nhiên ở các phần sau ta có thể thấy đây là một thách thức không hề nhỏ. Mục đích của phần này là giúp chúng ta nhận ra: (i) sự đa dạng thông tin của các câu truy vấn và (ii) theo nguyên lý, các thông tin này có thể xuất phát từ một người dùng duy nhất.

* 1. **Các vấn đề riêng tư trong việc khai phá query log**