**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

---


# BÁO CÁO

# GRADUATE RESEARCH I

## ĐỀ TÀI

## HỆ THỐNG GỢI Ý WEBSITE CHO MÁY TÌM KIẾM BẰNG PHƯƠNG PHÁP KHAI PHÁ QUERY LOG


**SINH VIÊN : CHU QUANG VIÊN**

**LỚP : AS1-K54 HEDSPI**

**SHSV : 20093236**

**GVHD : PGS.TS LÊ THANH HƯƠNG**


***Hà Nội, 5-2013***

# Mục lục

# I. Tổng quan về hệ gợi ý (Recommender System)

Mục đích tìm hiểu: Trong phần này em tìm hiểu khái quát về hệ gợi ý.

1.Mục đích xây dựng hệ gợi ý

2.Tìm hiểu hệ gợi ý thông qua bài toán gợi ý

3.Phân loại hệ gợi ý

Em đã đọc một số tài liệu trên mạng và tài liệu cô gửi để có cái nhìn khái quát về hệ gợi ý.Trong phần báo cáo này em sử dụng hầu hết kiến thức được dịch từ cuốn sách ”**Towards the Next Generation of Recommender Systems**”. Sau đây em xin báo cáo những gì em đã đọc và tìm hiểu được.

Em sẽ chia báo cáo của phần I.Tổng quan về hệ gợi ý ra làm 3 phần:

1.Sơ lược về hệ gợi ý

2.Phân loại hệ gợi ý

## 1. Khái niệm sơ lược về hệ gợi ý

### 1.1. Sơ lược về hệ gợi ý

Trong cuộc sống hàng ngày, trong rất nhiều trường hợp, người ta đưa ra các lựa chọn dựa trên những ý kiến hay lời khuyên của mọi người xung quanh, có thể qua lời nói, các bản đánh giá sản phẩm, khảo sát thị trường, thư giới thiệu ... Nhưng trong kỉ nguyên thông tin, hàng triệu thông tin được đưa lên internet mỗi ngày, điều này dẫn tới yêu cầu phải có các phương pháp tự động thu thập thông tin và đưa ra lời khuyên để hỗ trợ cho các phương pháp truyến thống trên . Hệ gợi ý (recommender system) là một giải pháp như vậy.Hệ thống này đưa ra gợi ý dựa trên những gì người dùng đã làm trong quá khứ, hoặc dựa trên tổng hợp ý kiến của những người dùng khác. Hệ gợi ý đã trở thành một ứng dụng quan trọng và thu hút được sự quan tâm lớn của các nhà nghiên cứu cũng như các doanh nghiệp.

Như vậy ta có thể hiểu nôm na như sau:" **Hệ gợi ý(Recommender system) là một thành phần trong hệ thống thông tin. Mục đích của nó là hỗ trợ người dùng tìm kiếm được đúng thông tin cần thiết phù hợp với mục đích của họ**."

Một vài hệ gợi ý nổi tiếng :

- Phim / TV/ âm nhạc: MovieLens, EachMovie, Morse, Firefly, Flycasting, Ringo…
- Tin tức / báo chí: Tapestry, GroupLens, Lotus Notes, Anatagonomy…
- Sách / Tài liệu: Amazon.com, Foxtrot, InfoFinder…
- Web: Phoaks, Gab, Fab, IfWeb, Let's Browse …
- Nhà hàng: Adaptive Place Advisor, Polylens, Pocket restaurent finder…
- Du lịch: Dietorecs, LifestyleFinder\

## 1.2. Bài toán gợi ý

Theo Adomavicius và Tuzhilin trong cuốn sách họ viết: "**Towards the Next Generation of Recommender Systems**" thì trong hầu hết các trường hợp, bài toán tư vấn được coi là bài toán ước lượng trước hạng (rating) của các sản phẩm (phim, cd, nhà hàng …) chưa được người dùng xem xét. Việc ước lượng này thường dựa trên những đánh giá đã có của chính người dùng đó hoặc những người dùng khác.Những sản phẩm có hạng cao nhất sẽ được dùng để tư vấn.

Một cách hình thức, bài toán tư vấn được mô tả như sau:

Gọi C là tập tất cả người dùng; S là tập tất cả các sản phẩm có thể tư vấn. Tập S có thể rất lớn, từ hàng trăm ngàn (sách, cd…) đến hàng triệu (như website). Tập C trong một số trường hợp cũng có thể lên tới hàng triệu.

Hàm u(c,s) đo độ phù hợp (hay hạng) của sản phẩm s với user c khi đó

u: $C$ × S → R với R là tập được sắp thứ tự. Với mỗi người dùng c∈$C$, cần tìm sản phẩm s'∈S sao cho hàm u(s', c) đạt giá trị lớn nhất:

∀c∈$C$, s'= arg $max_{s' \in S}$ (c,s)

Trong hệ tư vấn, độ phù hợp của một sản phẩm thường được cho bằng điểm, ví dụ người dùng A đánh giá bộ phim "Harry Poster và Căn phòng bí mật" được điểm 7/10. Tuy nhiên, nhìn chung độ phù hợp có thể là một hàm bất kì tùy thuộc vào ứng dụng cụ thể. Giá trị của hàm u có thể được xác định bởi người dùng hoặc được tính toán bởi công thức nào đó.

Mỗi người dùng trong không gian C được xác định bởi một hồ sơ (profile). Hồsơ này có thể gồm rất nhiều loại thông tin: tuổi, giới tính, thu nhập, … hoặc có thể chỉgồm một trường mã số người dùng (user id) duy nhất. Tương tự, mỗi sản phẩm

trong không gian S cũng được xác định bởi một tập các đặc trưng. Ví dụ, trong hệ thống tưvấn phim, đặc trưng của mỗi bộ phim có thể là : tên phim, thể loại, đạo diễn, năm sản xuất, diễn viên chính ….

Vấn đề chính của hệ tư vấn là hàm u không được xác định trên toàn không gian $C \times S$ mà chỉ trên một miền nhỏ của không gian đó. Điều này dẫn tới việc hàm u phải được dự đoán(ngoại suy) trong không gian $C \times S$ . Thông thường, độ phù hợp được thể hiện bằng điểm và chỉ xác định trên tập các sản phẩm đã từng được người dùng đánh giá từ trước (thường khá nhỏ).

Ví dụ:

| Phim / Người dùng | Harry Poster | IronMan3 | The Ring | Musical HighSchool |
|---|---|---|---|---|
| An | 6 | 9 | 8 | 4 |
| Mai | 7 | ∅ | 7 | 8 |
| Hạnh | 8 | 8 | ∅ | 6 |
| Quỳnh | 9 | ∅ | 7 | ∅ |

Bảng 1: Đánh giá của người dùng về một số bộ phim mà họ đã xem

Bảng 1 là đánh giá của một số người dùng với các phim mà họ đã xem (thang điểm từ 0-10, kí hiệu ∅ nghĩa là bộ phim chưa được người dùng cho điểm). Từnhững thông tin đó, hệ thống tư vấn phải dự đoán (ngoại suy) điểm cho các bộ phim chưa được người dùng đánh giá, từ đó đưa ra những gợi ý phù hợp nhất.

Giải thích thuật ngữ "Ngoại suy":Ngoại suy là sự mở rộng các kết luận về một bộ phận nào đó của hiện tượng sang bộ phận khác hay sang hiện tượng khác, sang tương lai… Cụ thể hơn, đó là sự suy diễn trên giả thuyết một định luật, một hàm số

hay một đại lượng ngoài phạm vi thời gian mà nó đã được điều tra hay quan sát một cách khách quan.

## 2. Phân loại hệ gợi ý

Có rất nhiều cách để dự đoán, ước lượng hạng/điểm cho các sản phẩm như sử dụng học máy, lí thuyết xấp sỉ, các thuật toán dựa trên kinh nghiệm… Theo Adomavicius và Tuzhilin thì các hệthống gợi ý thường được phân thành ba loại dựa trên cách nó dùng để ước lượng hạng của sản phẩm:

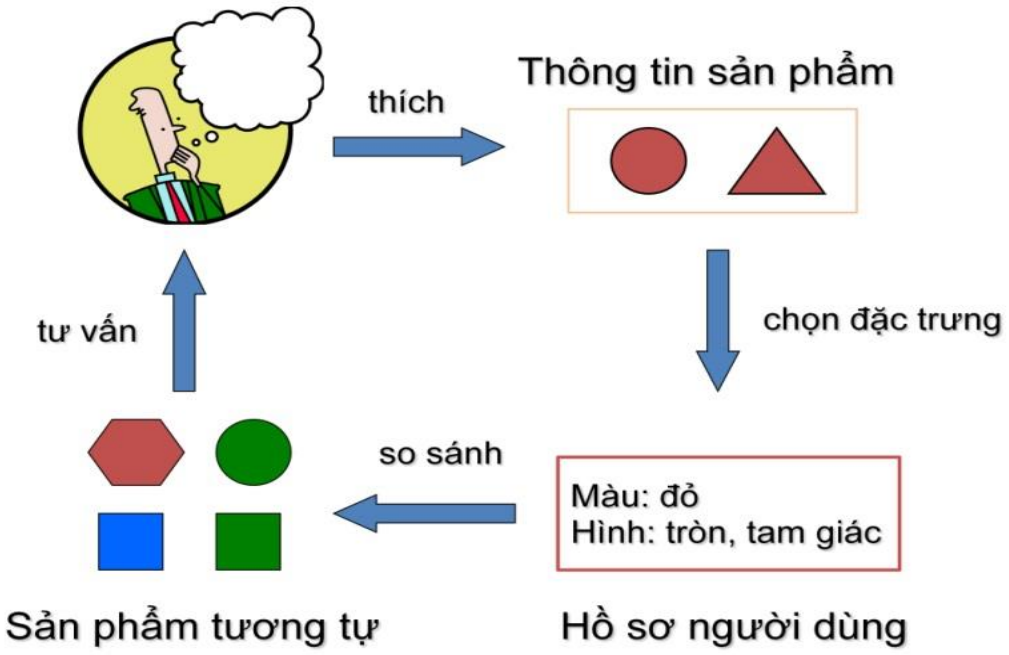
- Dựa trên nội dung (content-based): người dùng được gợi ý những sản phẩm tương tự như các sản phẩm từng được họ đánh giá cao.
- Cộng tác (collaborative): người dùng được gợi ý những sản phẩm mà những người cùng sở thích với họ đánh giá cao.
- Lai ghép (hybrid): kết hợp cả phương pháp dựa trên.

Ta sẽ đi phân tích từng loại như sau:

### 2.1. Phương pháp gợi ý dựa trên nội dung

Với phương pháp gợi ý dựa trên nội dung, độ phù hợp u(c, s) của sản phẩm s với người dùng c được đánh giá dựa trên độ phù hợp u(c, $s_i$), trong đó , $s_i \epsilon$ S và "tương tự" như s. Ví dụ, để gợi ý một bộ phim cho người dùng c, hệ thống gợi ý sẽtìm các đặc điểm của những bộ phim từng được c đánh giá cao (như diễn viên, đạo diễn…); sau đó chỉ những bộ phim tương đồng với sở thích của c mới được giới thiệu. Hướng tiếp cận dựa trên nội dung bắt nguồn từ những nghiên cứu về thu thập thông tin (IR - information retrieval) và lọc thông tin (IF - information filtering). Do đó, rất nhiều hệ thống gợi ý dựa trên nội dung hiện nay tập trung vào gợi ý các đối tượng chứa dữ liệu text như văn bản, tin tức, website(URL)… Những tiến bộ so với hướng tiếp cận cũ của IR là do việc sử dụng hồ sơ về người dùng

(chứa thông tin về sở thích, nhu cầu…) . Các thông tin hồ sơ có thể được biết trực tiếp từ người dùng một cách rõ ràng thông qua bảng trả lời câu hỏi hoặc gián tiếp (do khai phá thông tin từ các giao dịch của người dùng) theo thời gian.




Hình 1: Phương pháp gợi ý dựa trên nội dung

Để cụ thể hơn, đặt Content(s) là tập thông tin (hay tập các đặc trưng) về sản phẩm s. Do hệ thống dựa trên nội dung được thiết kế chủ yếu để dành cho các sản phẩm là text, nên nội dung sản phẩm thường được biểu diễn bởi các từ khóa (keyword): Content(s) = ($w_{1s}$, …$w_{ks}$), với $w_{1s}$,..,$w_{ks}$ là trọng số của các từ khóa từ 1 tới k (có thể được tính bằng "Term frequency/ Inverse document frequency" được viết tắt là TF-IDF). Ví dụ, hệ gợi ý website Fab biểu diễn nội dung các trang web bằng 100 từ quan trọng nhất.Tương tự, hệ thống Syskill & Webert biểu diễn văn bản bằng 128 từ có trọng số cao nhất.

Nói rõ hơn về TF-IDF ta có:

- TF: là tần số xuất hiện của một từ trong một văn bản. Giả sử $f_{i,j}$ là số lần từ khóa $k_i$ xuất hiện trong tài liệu $d_j$và $\max\{f_{z,j}: z \in d_j\}$ là số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản $d_j$.

  Khi đó, $TF_{i,j}$ – tần số xuất hiện của từ khóa $k_i$ trong tài liệu $d_j$ được định nghĩa như sau:

$$TF_{i,j} = \frac{f_{i,j}}{\max\{f_{z,j}: z \in d_j\}}$$

  Như vậy nói đơn giản thì TF là thương của số lần xuất hiện 1 từ trong văn bản và số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản đó. (giá trị sẽ thuộc khoảng [0, 1]).Tuy nhiên việc từ khóa xuất hiện nhiều trong văn bản không phải là hữu ích trong việc phân biệt giữa một tài liệu có liên quan và một tài liệu không liên quan tới mục đích gợi ý với người dùng.Vì thế cần sử dụng them IDF.

- IDF: là tần số nghịch của một từ trong một văn bản.Tính IDF để giảm giá trị của những từ phổ biến và mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản.Công thức tính $IDF_i$ của từ khóa $k_i$được định nghĩa như sau:

$$IDF_i = \log\frac{N}{n_i}$$

  Trong đó:

  - N là tổng số văn bản có thể gợi ý cho người dùng
  - $n_i$ là số văn bản có chứa từkhóa nhất định

    $n_i = |\{d_j \in N, k_i \in d_j\}|$

Nếu $k_i$ không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho 0 không hợp lệ, vì thế người ta thường thay bằng mẫu thức 1+ $n_i$

Cơ số logarit trong công thức này không thay đổi giá trị của 1 từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi 1 số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Tuy nhiên việc thay đổi khoảng giá trị sẽ giúp tỷ lệ giữa IDF và TF tương đồng

- TF-IDF: xác định trong số của 1 từ khóa $k_i$ trong văn bản $d_j$ ,được xây dựng theo công thức sau:

  $w_{i,j} = TF_{i,j}$ x $IDF_i$

  Sự kết hợp hai tiêu chuẩn này cho biết: tầm quan trọng của một từ khóa (do TF mang lại) và sự phân biệt giữa các từ khóa (do IDF mang lại). Một từ khóa có tầm quantrọng lớn hơn thì giá trị$w_{i,j}$của nó phải lớn hơn.Việc này giúp lọc ra những từ khóa phổ biến và giữ lại những từ khóa có giá trị cao (từ khoá quan trọng của văn bản đó).

  Giá trị trọng số này:

  - Cao nhất khi từ đó xuất hiện nhiều lần trong một tập nhỏ các tài liệu.
  - Nhỏ hơn khi xuất hiện vài lần trong một tài liệu và xuất hiện trong nhiều tài liệukhác.

- Nhỏ nhất khi từ này xuất hiện gần như trong tất cả các tài liệu.

Đặt Profile(c) là hồ sơ về người dùng c, bao gồm các thông tin về sở thích của c. Những thông tin này có được bằng cách phân tích nội dung của các sản phẩm từng được c đánh giá (cho điểm) trước đó. Phương pháp được sử dụng thường là các kĩ thuật phân tích từ khóa của IR, do đó, Profile(c) cũng có thể được định nghĩa như một vector trọng số:

Profile(c) = ($w_{1s}$, …,$w_{ks}$) với $x_{ic}$biểu thị độ quan trọng của từ khóa i với người dùng c.

Trong hệ thống tư vấn dựa trên nội dung, độ phù hợp u(c,s) được xác định bởi công thức:

u(c,s) = score(Profile(c), Content(s))

Cả Profile(c), Content(s) đều có thể được biểu diễn bằng vector trọng số từ TF-IDF (tương ứnglà$\overrightarrow{w_c}$ , $\overrightarrow{w_s}$ ) nên có thể đo độ tương đồng của chúng bằng độ đo cosin:

$$u(c,s) = \cos(\overrightarrow{w_c}, \overrightarrow{w_s}) = \frac{\overrightarrow{w_c} . \overrightarrow{w_s}}{||w_c| \times |w_s||} = \frac{\sum_{i=1}^{K} w_{i,c} . w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2} \sqrt{\sum_{i=1}^{K} w_{i,s}^2}}$$

Trong đó K – tổng số từ khóa có trong hệ thống

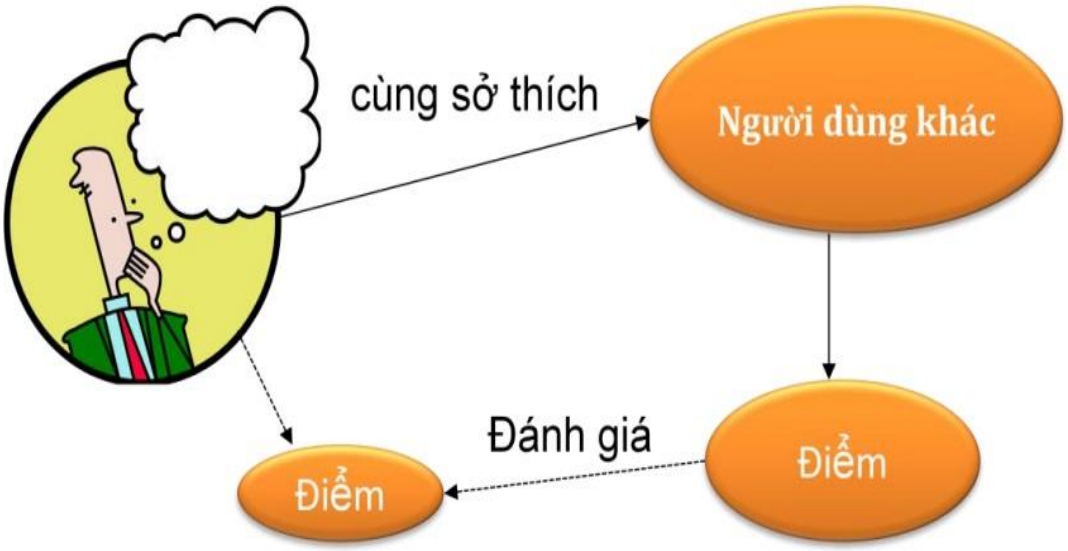
Ví dụ, nếu c đọc nhiều bài báo thuộc lĩnh vực sinh học thì các từ khóa liên quan tới sinh học (như gen, protein, tế bào, ADN…) trong Profile(c) sẽ có trọng số cao. Hệquả là với các bài báo s cũng thuộc lĩnh vực này sẽ có độ phù hợp u(c,s) cao hơn với người dùng c.Bên cạnh các phương pháp IR, hệ tư vấn dựa trên nội dung còn sử dụng nhiều phương pháp học máy khác như: phân lớp Bayes, cây

quyết định, mạng nơron nhân tạo… Các phương pháp này khác với các phương pháp của IR ở chỗ nó dựa trên các mô hình học được từ dữ liệu nền. Ví dụ, dựa trên tập các trang web đã được người dùng đánh giá là có nội dung "tốt" hoặc "xấu" có thể sử dụng phân lớp Bayes để phân loại các trang web chưa được đánh giá.

### 2.2. Phương pháp gợi ý cộng tác

Theo như Adomavicius và Tuzhilin,không giống như phương pháp tư vấn dựa trên nội dung, hệ thống cộng tác godự đoán độ phù hợp u(c,s) của một sản phẩm s với người dùng c dựa trên độ phù hợp u($c_i$, s) giữa người dùng$c_i$ và s, trong đó $c_i$ là người có cùng sở thích với c. Ví dụ, đểgợi ý một bộ phim cho người dùng c, đầu tiên hệ thống cộng tác tìm những người dùng khác có cùng sở thích phim ảnh với c. Sau đó, những bộ phim được họ đánh giá cao sẽđược dùng để tư vấn cho c.

Có rất nhiều hệ thống cộng tác đã được phát triển như: Grundy, GroupLens (tin tức), Ringo (âm nhạc), Amazon.com (sách), Phoaks (web)… Các hệ thống này có thểchia thành hai loại: dựa trên kinh nghiệm (heuristic-based hay memory-based) và dựa trên mô hình (model-based).




Hình 2: Phương pháp gợi ý dựa vào cộng tác

**a. Hệ thống cộng tác dựa trên kinh nghiệm**

Các thuật toán dựa trên kinh nghiệm dự đoán hạng của một sản phẩm dựa trên toàn bộ các sản phẩm đã được đánh giá trước đó bởi người dùng. Nghĩa là, hạng của sản phẩm s với người dùng c .Khi đó $r_{c,s}$ được tổng hợp từ đánh giá của những người dùng khác về s (thường là N người có sở thích tương đồng nhất với c)

$r_{c,s} = agrr\, r_{c',s}$ với$c' \in \widehat{C}$ (tập N người dùng có cùng sở thích với c)

Một số ví dụ về hàm tổng hợp (aggregate):

(a) $r_{c,s} = \frac{1}{N} \sum_{c' \in \hat{C}} r_{c',s}$

(b) $r_{c,s} = k \times \sum_{c' \in \hat{C}} sim(c, c') \times r_{c',s}$

(c) $r_{c,s} = \overline{r_c} + k \times \sum_{c' \in \hat{C}} sim(c, c') \times r_{c',s}$

Với: k= hệ số chuẩn hóa

sim$(c, c')$=độ tương đồng (về sở thích) giữa người dùng $c$ và $c'$

$\overline{r_c}, \overline{r_{c'}}$= trung bình của các đánh giá được cho bởi người dùng $c$ và $c'$

Có nhiều cách để tính độ tương đồng (về sở thích) giữa hai người dùng, nhưng trong hầu hết các phương pháp, độ tương đồng chỉ được tính dựa trên các sản phẩm được cả hai người cùng đánh giá.Hai phương pháp phổ biến nhất là dựa trên độ tương quan (correlation-based) và dựa trên cosin (consine-based).

Đặt $S_{xy} = \{s \in S | r_{x,s} \neq \emptyset \& r_{y,s} \neq \emptyset\}$ là tập hợp các sản phẩm được đánh giá bởi cả hai người dùng $x, y$.

Công thức dựa trên độ tương quan của Pearson [27]:

$$\sin x, y = \frac{\sum_{s \in S_{xy}} (r_{x,s} - \overline{r_x}) \times (r_{y,s} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{x,s} - \overline{r_x})^2 \times \sum_{s \in S_{xy}} (r_{y,s} - \overline{r_y})^2}}$$

Với phương pháp dựa trên cosin, hai người dùng được biểu diễn bởi 2 vector m chiều, với m=$|S_{xy}|$. Độ tương đồng giữa 2 vector được tính bởi công thức:

$$\sin(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x}.\vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{s \in S_{xy}} r_{x,s} \times r_{y,s}}{\sqrt{\sum_{s \in S_{xy}} {r_{x,s}}^2} \times \sqrt{\sum_{s \in S_{xy}} {r_{y,s}}^2}}$$

### a. Hệ thống gợi ý cộng tác dựa mô hình

Khác với phương pháp dựa trên kinh nghiệm, phương pháp dựa trên mô hình (model-based) sự dụng kĩ thuật thống kê và học máy trên dữ liệu nền (các đánh giá đã biết) để xây dựng nên các mô hình.Mô hình này sau đó sẽ được dùng để dự đoán hạng của sản phẩm chưa được đánh giá.

Breese trong cuốn sách của ông "**Empirical analysis of predictive algorithms for collaborative filtering**" đề xuất hướng tiếp cận xác suất cho lọc cộng tác (collaborative filtering), trong đó công thức sau ước lượng đánh giá của người dùng $c$ về sản phẩm $s$ (thang điểm đánh giá từ 0 đến n):

$$r_{c,s} = E(r_{c,s}) = \sum_{i=0}^{n} i \times \Pr(r_{c,s} = i | r_{c,s'}, s' \in S_c)$$

Billsus và Pazzani trong "**Learning collaborative information filters**" đề xuất phương pháp lọc công tác trên nền học máy, trong đó rất nhiều các kĩ thuật học máy (như mạng nơron nhân tạo) và các kĩ thuật trích chọn đặc trưng (như SVD – một kĩ thuật đại số nhằm làm giảm số chiều của ma trận) có thể được sử dụng

Ngoài ra còn nhiều hướng tiếp cận khác như mô hình thống kê, mô hình bayes mô hình hồi quy tuyến tính, mô hình entropy cực đại…

Hệ thống gợi ý cộng tác khắc phục được nhiều nhược điểm của hệ thống dựa trên nội dung.Một điểm quan trọng là nó có thể xử lý mọi loại dữ liệu và gợi ý mọi loại sản phẩm, kể cả những sản phẩm mới, khác hoàn toàn so với những gì người dùng từng xem.

## 2.3. Phương pháp lai ghép

Một vài hệ gợi ý kết hợp cả phương pháp cộng tác và dựa trên nội dung nhằm tránh những hạn chế của cả hai. Có thể phân thành bốn cách kết hợp sau:

- Cài đặt hai phương pháp riêng rẽ rồi kết hợp dự đoán của chúng.
- Tích hợp các đặc trưng của phương pháp, dựa trên nội dung vào hệ thống cộng tác.
- Tích hợp các đặc trưng của phương pháp cộng tác vào hệ thống dựa trên đặc trưng.
- Xây dựng mô hình hợp nhất, bao gồm các đặc trưng của cả hai phương pháp.

### a. Kết hợp hai phương pháp riêng rẽ

Có hai kịch bản cho trường hợp này :

- Cách 1: kết hợp kết quản của cả hai phương pháp thành một kết quả chung duy nhất, sử dụng cách kết hợp tuyến tính (linear combination) hoặc voting scheme
- Cách 2: Tại mỗi thời điểm, chỉ chọn phương pháp cho kết quả tốt hơn (dựa trên một số độ đo chất lượng tư vấn nào đó). Ví dụ, hệ thống DailyLearner system chọn phương pháp nào đưa ra gợi ý với độ chính xác (confidence) cao hơn.

### b. Thêm đặc trưng của mô hình dựa trên nội dung vào mô hình cộng tác

Một số hệ thống lai (như Fab) dựa chủ yếu trên các kĩ thuật cộng tác nhưng vẫn duy trì hồ sơ về người dùng (theo dạng của mô hình dựa trên nội dung). Hồ sơ này được dùng để tính độ tương đồng giữa hai người dùng, nhờ đó giải quyết được

trường hợp có quá ít sản phẩm chung được đánh giá bởi cả hai người. Một lợi ích khác là các gợi ý sẽ không chỉ giới hạn trong các sản phẩm được đánh giá cao bởi những người cùng sở thích (gián tiếp), mà còn cả với những sản phẩm có độ tương đồng cao với sở thích của chính người dung đó (trực tiếp).

**c. Thêm đặc trưng của mô hình cộng tác vào mô hình dựa trên nội dung**

Hướng tiếp cận phổ biến nhất là dùng các kĩ thuật giảm số chiều trên tập hồ sơ của phương pháp dựa trên nội dung. Ví dụ sử dụng phân tích ngữ nghĩa ẩn (latent semantic analysis) để tạo ra cách nhìn cộng tác (collaborative view) với tập hồ sơ người dùng (mỗi hồ sơ được biểu diễn bởi một vector từ khóa).Cách này đã được viết khá rõ ràng trong cuốn "Combining content and collaboration in text filtering" của Soboroff, I. and C. Nicholas

**d. Mô hình hợp nhất hai phương pháp**

Trong những năm gần đây đã có khá nhiều nghiên cứu về mô hình hợp nhất. Basu trong cuốn "Recommendation as classification: Using social and content-based information in recommendation" đề xuất kết hợp đặc trưng của cả hai phương pháp vào một bộ phân lớp dựa trên luật (rule-based classifier). Popescul và cộng sự đưa ra phương pháp xác suất hợp nhất dựa trên phân tích xác suất ngữ nghĩa ẩn (probabilistic latent semantic analysis). Ansari giới thiệu mô hình hồi quy Bayes sử dụng dây Markov Monte Carlo để ước lượng tham số. Độ chính xác của hệ thống gợi ý lai ghép có thể được cải tiến bằng cách sử dụng các kĩ thuật dựa trên tri thức (knowledge-based) như case-based reasoning. Ví dụ, hệ thống Entrée dùng những tri thức về nhà hàng, thực phẩm (như: đồ biển không phải là thức ăn chay).. để gợi ý nhà hàng thích hợp cho người dùng. Hạn chế chính của hệ thống dạng này là nó cần phải thu thập đủ tri thức, đây cũng là nút thắt cổ chai (bottle-neck) của rất nhiều hệ thống trí tuệ nhân tạo khác. Tuy nhiên, các hệ

thống tư vấn dựa trên tri thức hiện đang được phát triển trên các lĩnh vực mà miền tri thức của nó có thể biểu diễn ở dạng mà máy tính đọc được (như ontology). Ví dụ, hệ thống Quickstep và Foxtrot sử dụng ontology về chủ đề của các bài báo khoa học để gợi ý những bài báo phù hợp cho người dùng. Một vài bài báo như "Fab: Content-based, collaborative recommendation" của Balabanovic đã thực hiện so sánh hiệu năng của hệ thống lai ghép với các hệ thống dựa trên nội dung hoặc cộng tác thuần túy và cho thấy hệ thống lai ghép có độ chính xác cao hơn.

| Phương pháp | Các kĩ thuật sử dụng | |
|---|---|---|
| | Dựa trên kinh nghiệm | Dựa trên mô hình |
| Dựa trên nội dung | +TF-IDF<br>+Phân cụm | +Phân lớp bayes<br>+Phân cụm<br>+Cây quyết định<br>+Mạng nơron nhân tạo |
| Cộng tác | +k-Láng giềng gần nhất<br>+Phân cụm<br>+Lí thuyết đồ thị | +Mạng bayes<br>+Phân cụm<br>+Mạng nơron nhân tạo<br>+Hồi quy tuyến tính<br>+Mô hình xác suất |
| Lai ghép | +Kết hợp tuyến tính kết quả | +Tích hợp đặc trưng của một phương pháp vào mô hình của phương pháp còn lại.<br>+Xây dựng mô hình hợp nhất hai phương pháp. |

Hình 3: Ba phương pháp gợi ý

# II. Tìm hiểu sơ bộ về hệ thống tìm kiếm(Search Engine)

Trong phần này thì em mới tìm hiểu sơ qua,chưa đi sâu vào nội dung của hệ thống tìm kiếm.Hầu như nội dung trong phần này em đã tham khảo trong khóa luận "Ứng dụng Query Log trong hệ thống tìm kiếm thông minh " của anh Nguyễn Sinh Thành- K50.

Mục đích tìm hiểu của em trong chương này là có cái nhìn khái quát về hệ thống tìm kiếm để có thể có kiến thức cơ bản trong việc xây dựng một máy tìm kiếm thông ming có tích hợp hệ gợi ý trong truy vấn bằng cách sử dụng Query Log.

Sau đây em xin trình bày những nội dung đã nghiên cứu

1.Khái niệm về hệ thống tìm kiếm
2.Phân loại hệ thống tìm kiếm
3.Kiến trúc hệ thống tìm kiếm
4.Modul thu thập tài liệu
5.Modul đánh chỉ mục tài liệu
6.Modul tìm kiếm

## 1. Khái niệm hệ thống tìm kiếm

Bắt đầu từ thập kỉ 90 của thế kỉ 18 hàng loạt hệ thống tìm kiếm(search engine) lớn xuất hiện như Google(1998) ,Yahoo, Bing,…Các công cụ tìm kiếm này đã phần nào đáp ứng nhu cầu của người dùng internet.

Như vậy những người dùng internet chắc chắn đều có một hình dung sơ bộ về hệ thống tìm kiếm.Em xin đưa ra khái niệm cơ bản về hệ thống tìm kiếm: "**Hệ thống tìm kiếm là một hệ thống hỗ trợ người dùng sử dụng tìm kiếm thông tin một cách nhanh chóng và dễ dàng. Thông tin có thể bao gồm tài liệu,hình ảnh, âm thanh, video và những đối tượng đa phương tiện khác.**"

## 2. Phân loại hệ thống tìm kiếm

### 2.1. Theo mô hình hoạt động

Phân loại theo mô hình hoạt động thì tồn tại 2 dạng hệ thống tìm kiếm sau:

- **Hệ thống tìm kiếm thông thường**: Đây là hệ thống được xây dựng hoàn chỉnh, có cơ sở dữ liệu chỉ mục, hệ thống thu thập tài liệu. Hệ thống này thường rất lớn về quy mô và dữ liệu, đặc biệt là nó không sử dụng lại dữ liệu của các hệ thống khác. Ví dụ như hệ thống tìm kiếm của Google với hơn 1000 tỷ trang web được đánh chỉ mục(7/2008) và hơn 450000 máy chủ(số liệu không chính thức năm 2000).Một số máy tìm kiếm thông thường như: Google, Bing, Yahoo Search,….
- **Hệ thống Meta-Search(máy tìm kiếm liên hợp):** Xuất phát từ ý tưởng tận dụng tối đa nguồn tài nguyên từ các hệ thống tìm kiếm lớn và giảm chi phí

xây dựng, các hệ thống tìm kiếm liên hợp được đưa ra như là một giải pháp khả thi.Một máy tìm kiếm liên hợp không tự xây dựng bất cứ thành phần nào như thu thập tài liệu, cơ sở dữ liệu chỉ mục như các hệ thống tìm kiếm thông thường. Thay vào đó, với mỗi câu truy vấn của người dùng, máy tìm kiếm liên hợp sẽ chuyển đến các máy tìm kiếm khác như Google, Bing,.. sau đó xử lý kết quả trả về từ các máy tìm kiếm này trước khi đưa kết quả ra cho người dùng.Một số máy tìm kiếm liên hợp như: Clussty, KartOO, Mamma, Search, Excite,..

### 2.2. Theo miền dữ liệu

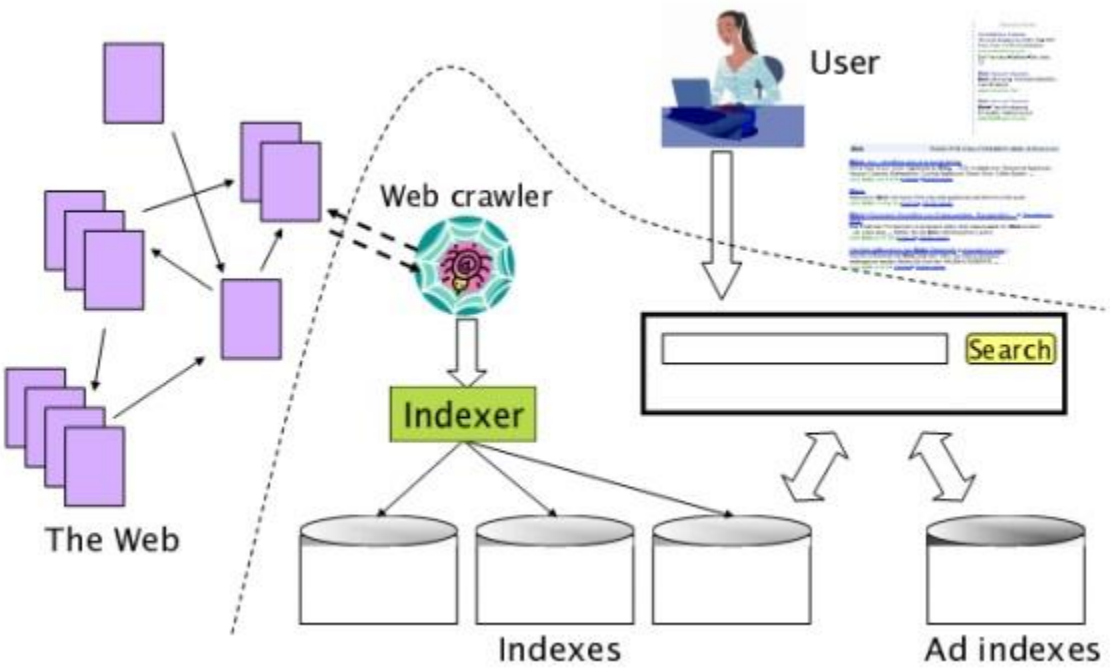
Phân loại theo miền dữ liệu thì có 2 dạng:

- **Hệ thống tìm kiếm tổng quát**: Đây là hệ thống tìm kiếm trên tất cả các lĩnh vực, trên tất cả các dữ liệu. Miền dữ liệu tìm kiếm không bị giới hạn trong bất cứ một lĩnh vực hẹp nào. Các hệ thống tìm kiếm lớn như Google đều cung cấp tính năng tìm kiếm tổng quát này.
- **Hệ thống tìm kiếm chuyên sâu**: Điểm đặc biệt của hệ thống này là chỉ tìm kiếm trong một miền dữ liệu giới hạn. Ví dụ: hệ thống tìm kiếm sách, hệ thống tìm kiếm nhà hàng, hệ thống tìm kiếm thông tin thể thao,… Bên cạnh đó, còn có hệ thống tìm kiếm theo các định dạng dữ liệu nhất định như tìm kiếm ảnh, tìm kiếm video, tìm kiếm theo định dạng file,… Hệ thống tìm kiếm của Google cũng cung cấp một số chức năng tìm kiếm chuyên sâu theo kiểu dữ liệu.

## 3. Kiến trúc hệ thống tìm kiếm

Về cơ bản, hệ thong máy tìm kiếm gồm 3 thành phần chính:

- **Bộ thu thập thông tin (Web crawler):** Bộ phận này có nhiệm vụ tự động thu thập, tải về các trang web trên Internet.
- **Bộ đánh chỉ mục (Indexer):** Bộ phận này tiến hành phân tích các trang web được tải về, sau đó thực hiện đánh chỉ mục cho các trang web đó.
- **Bộ tìm kiếm thông tin (Search):** Chức năng tìm kiếm các tài liệu phù hợp với yêucầu của người dùng, việc tìm kiếm được dựa trên các từ khóa.




Hình 4: Các thành phần cơ bản của hệ thống tìm kiếm

Ba thành phần này phụ thuộc nhau về dữ liệu nhưng độc lập về hoạt động. Chi tiết các thành phần được trình bày trong các phần tiếp theo. Ngoài 3 thành phần trên, hệ thống tìm kiếm còn nhiều thành phần chức năng bổ sung như xếp hạng tài liệu,….Kiến trúc hệ thống trên được đề xuất chung cho các hệ thống tìm kiếm trên Internet hiện nay. Tuy nhiên, với từng ngôn ngữ khác nhau thì bổ sung

thêm một số chức năng đặc biệt. Với tiếng Việt, hệ thống cần bổ sung một thành phần đặc biệt là tách từ tiếng Việt để phục vụ cho chức năng đánh chỉ mục sau đó.

## 4. Module thu thập tài liệu

Việc thu thập tài liệu được bắt đầu với một vài URL. Hệ thống tự động duyệt qua các URL này để thu thập tài liệu, sau đó lại trích rút các siêu liên kết và đoạn văn bản. Các siêu liên kết này được đưa qua các bộ kiểm tra và lọc nhằm tránh trùng lặp, sau đó được đưa vào hàng đợi URL để chờ xử lý.

Các tính năng mà một bộ thu thập cần phải có:

- **Robustness (Mạnh mẽ):** Một số trang web đặt bẫy các bộ thu thập, bộ thu thập cóthể chạy vô hạn trong đó. Cần có cơ chế phát hiện và tránh bẫy kiểu này.
- **Politeness (Lịch sự):** Các máy chủ có các quy định cho tốc độ truy nhập củacrawler(bộ thu thập). Bộ thu thập cần tôn trọng tất cả các quy định này.

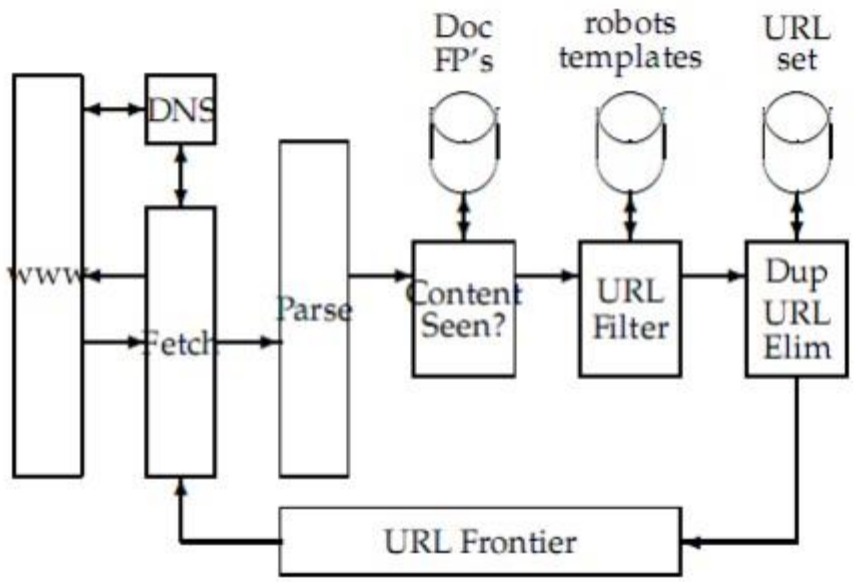
Các tính năng mà một bộ thu thập nên có:

- **Phân tán**:Bộ thu thập nên có khả năng chạy trong môi trường phân tán trên nhiều máy khác nhau.
- **Khả năng tăng cường**: Kiến trúc bộ thu thập nên cho phép mở rộng tốc độ thu thập bằng cách thêm máy mở rộng và tốc độ.
- **Hiệu năng và hiệu quả**:Hệ thống cần sử dụng hiệu quả tài nguyên hệ thống như bộ xử lý, lưu trữ, tốc độ mạng.
- **Chất lượng:** Bộ thu thập nên duyệt những trang "có ích" đầu tiên.

- **Làm mới:** Hệ thống nên thu thập một trang với tốc độ xấp xỉ tốc độ thay đổi của trang.
- **Khả năng mở rộng**: Kiến trúc được thiết kế để mở rộng với nhiều kiểu dữ liệu, nhiều giao thức lấy file.

Có 3 chiến lược thu thập dữ liệu: (thứ tự duyệt URL)

- Tìm kiếm theo chiều rộng
- Tìm kiếm theo chiều sâu
- Tìm kiếm ngẫu nhiên




Hình 5: Kiến trúc tổng quát của một hệ thống thu thập

Trong đó:

- URL Frontier: chứa các URL chưa được tải về trong lần thu thập hiện tại.
- DNS Resolution(phân giải DNS): xác định các web server để tải về một trang cụ thể thông qua URL.
- Fetch(tải về):Sử dụng giao thức Http để tải về các trang web tại địa chỉ URL.

- Parse(Phân tích): Trích xuất các đoạn văn bản và tập các liên kết từ trang web được tải về.
- Duplicate URL Elimination: xác định những liên kết được trích rút ra đã có trong URL Frontier hoặc đã được tải về hay chưa.

**5. Module đánh chỉ mục tài liệu**

Sau khi đã tách được các từ trong văn bản, ta tiến hành lập chỉ mục. Tuy nhiên, không phải từ nào cũng được sử dụng để đánh chỉ mục. Đầu tiên cần loại bỏ những từ không mang nhiều ý nghĩa (từ dừng) như : vậy thì, nếu, mà, …Những từ này đã có một danh sách các từ dừng cho tiếng Việt. Hoặc những từ xuất hiện với tần suất nhiều cũng không có nhiều ý nghĩa (xuất hiện trong hơn một nửa tổng số tài liệu). Sau khi đã loại bỏ những từ dừng ta tiến hành đánh trọng số cho các từ, trọng số này phản ánh độ quan trọng của từ trong mỗi tài liệu.

Trong việc lưu trữ thông tin, tập tin nghịch đảo được sử dụng để tăng tốc độ tìm kiếm. Tập tin nghịch đảo: mỗi từ có một danh sách các tài liệu kèm theo mà chứa từ đó. Vì vậy, chẳng hạn khi người dùng nhập vào một từ, hệ thống sẽ nhanh chóng lấy được danh sách các tài liệu chứa từ đó. Trong khi đó tập tin trực tiếp: mỗi tài liệu có một danh sách các từ mà nó chứa. Sử dụng cách lưu trữ này sẽ không hiệu quả.

Có rất nhiều phương pháp trong việc đánh trọng số cho từ. Tiêu biểu là phương pháp TF-IDF đã được nói đến ở trên.

**6. Module tìm kiếm**

Module này có chức năng cung cấp giao diện tìm kiếm cho người dùng. Với mỗi chuỗi đầu vào của người dùng, chức năng này sẽ tiến hành tách từ dựa trên các từ đã được đánh chỉ mục trong cơ sở dữ liệu. Đối với câu truy vấn tiếng Việt, việc tách từ có khác so với tiếng Anh (giải thuật tách từ sẽ được trình bày ngay bên

dưới). Sau khi có được danh sách các từ, tiến hành lấy thông tin về các tài liệu trả lại kết quả cho người dùng. Với số lượng kết quả lớn, chức năng tự động phân trang để người dùng dễ dàng theo dõi.

Bên cạnh cung cấp kết quả tìm kiếm, module này còn có chức năng gợi ý truy vấn cho người dùng. Mỗi khi người dùng bắt đầu nhập hoặc nhập xong câu truy vấn thì đều nhận được những câu gợi ý từ hệ thống.

## III. Query Log và Ứng dụng khai phá Query gợi ý truy vấn cho người dùng

### 1. Query Log là gì ?

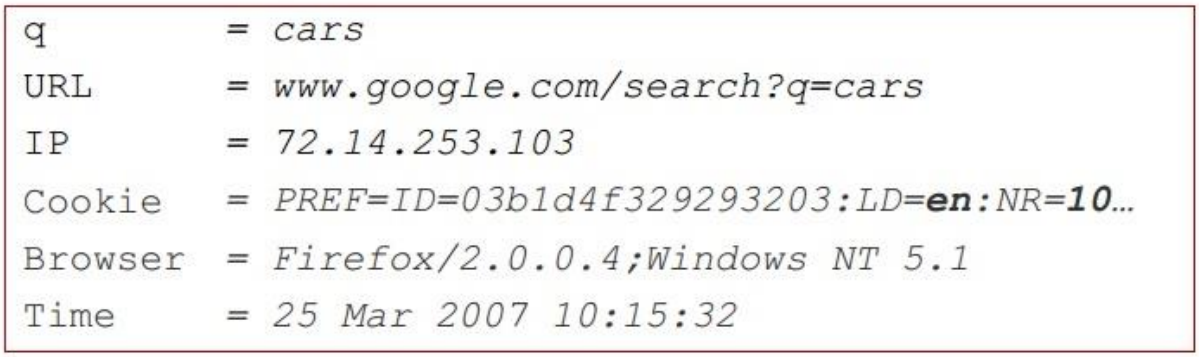
QueryLog là một tập các bản ghi, mỗi bản ghi bao gồm thông tin về những lượt tìm kiếm của người dùng được máy tìm kiếm lưu lại. Khác với server log thông thường,thì query log bao gồm câu truy vấn của người dùng, tập các kết quả cho câu truy vấn đó và tập các liên kết URL trong tập kết quả được người dùng click. Mỗi máy tìm kiếm có một cách lưu log khác nhau và thường rất ít khi công bố ra ngoài (một lí do là vì vi phạm sự riêng tư của người dùng).

Sau đây là một phần query log được AOL công bố năm 2006

| AnonID | Query | QueryTime | Rank | ClickURL |
|---|---|---|---|---|
| 479 | family guy movie references | 2006-03-03 22:37:46 | 1 | http://www.familyguyfiles.com |
| 479 | top grossing movies of all time | 2006-03-03 22:42:42 | 1 | http://movieweb.com |
| 479 | top grossing movies of all time | 2006-03-03 22:42:42 | 2 | http://www.imdb.com |
| 479 | car decals | 2006-03-03 23:20:12 | 4 | http://www.decaljunky.com |
| 479 | car decals | 2006-03-03 23:20:12 | 1 | http://www.modernimage.net |
| 479 | car decals | 2006-03-03 23:20:12 | 5 | http://www.webdecal.com |
| 479 | car window decals | 2006-03-03 23:24:05 | 9 | http://www.customautotrim.com |
| 479 | car window sponsor decals | 2006-03-03 23:27:17 | 3 | http://www.streetglo.net |
| 479 | bose | 2006-03-03 23:30:11 | 1 | http://www.bose.com |
| 479 | bose car decal | 2006-03-03 23:31:48 | 1 | http://stickers.signprint.co.uk |
| 479 | bose car decal | 2006-03-03 23:31:48 | 1 | http://stickers.signprint.co.uk |
| 479 | bose car decal | 2006-03-03 23:31:48 | 7 | http://www.motorcitydecals.com |
| 479 | chicago the mix | 2006-03-04 22:11:31 | 1 | http://www.wtmx.com |
| 479 | chicago the drive | 2006-03-04 22:14:51 | 2 | http://www.wdrv.com |

Hình 6.Một phần query log do AOL cung cấp

## 2. Cấu trúc Query Log




Hình 7. Cấu trúc query log của Google

Tuy cách lưu log ở mỗi máy tìm kiếm là khác nhau nhưng query log thường có các trường sau:

Xét truy vấn mà người dùng gửi tới máy tìm kiếm.

Ví dụ: "mon ngon ha noi", "dai học bach khoa ha noi", "com ga 123" …

Một số máy tìm kiếm giới hạn số từ trong query ví dụ như Google cho phép query dài tối đa 32 từ.

- **Url được click và vị trí của url**

Địa chỉ url người dùng click và vị trí của nó (trường ItemRank của AOL query log) trong danh sách kết quả máy tìm kiếm trả về cho query vừa được gửi.Ví dụ, với query "champion league", các url được click là: www.uefa.com (ở vị trí 1) và soccernet.espn.go.com (ở vị trí 4, theo kết quả của Google).

- **Địa chỉ IP:**

Địa chỉ IP của người dùng (ví dụ:141.243.1.172) hoặc tên DNS (ví dụ: wpbfl2-45.gate.net). Từ IP có thể biết được địa chỉ (quốc gia, vùng) của người dùng và

nhà cung cấp dịch vụ internet cho họ (Internet Service Provider). Khi công bố query log ra công chúng, các máy tìm kiếm buộc phải ―nặc danh hóa‖ (anonymizing) trường này để không làm lộ danh tính và các thông tin cá nhân của người dùng. Như ở trên, trong query log được AOL công bố, trường IP được thay thế bằng AnonID (định danh ẩn).

- **Phần mềm sử dụng ở máy của người dùng (user agents):**

Trường này lưu thông tin về tên, phiên bản của trình duyệt cũng như tên, phiên bản của hệ điều hành được người dùng sử dụng.Ví dụ:―Firefox/2.0.0.4;Windows NT 5.1".

- **Thời gian:**

Thời gian người dùng gửi query tới máy tìm kiếm. Thông thường, như trong Google hay AOL, thời gian được ghi theo định dạng

[DD/Mon/YYYY/: HH:MM:SS offset] với:

- DD/Mon/YYYY: chỉ ngày tháng năm.
- HH:MM:SS : thể hiện 24h trong ngày.
- Offset: chỉ độ lệch múi giờ so với giờ GMT (Greenwich Mean Time).

Ví dụ:" 22/May/2009:16:03:00 +0700" chỉ thời điểm 16:03:00 ngày 22 tháng 5 năm 2009, tại múi giờ GMT+7 (Bangkok-Hanoi-Jakarta). Ở một số máy tìm kiếm khác, như AltaVista, trường thời gian được lưu ở dạng timestamp, là số milli giây từ một mốc thời gian trong quá khứ (baseline) đến thời điểm query được gửi. Ví dụ, nếu chọn mốc thời gian là 00:00:00 ngày 1/1/1995 thì thời điểm 12:00:02 28/10/2004 có timestamp = 20822005

- **Cookie:**

Được máy tìm kiếm lưu ở máy người dùng để nhận biết một số thông tin về họ.

Ví dụ, trường cookie của Google lưu sở thích của người dùng về ngôn ngữ tìm kiếm và số kết quả mong muốn trong mỗi trang.

"Cookie = PREF=ID=03b1d4f329293203:LD=en:NR=10…"

Để đảm bảo tính bí riêng tư, sau 18 tháng, Google sẽ xóa thông tin về cookie và IP của người dùng. Ví dụ, các thông tin đó sẽ được đưa về dạng

IP=72.14.253.XX và Cookie=PREF=XXXXXXXX.

**3. Đặc điểm của Query Log**

- Kích thước của QueryLog tăng lên một cách chóng mặt theo thời gian.
- Chỉ những câu truy vấn mà người dùng có nhấn vào một tài liệu trong kết quả mới được ghi nhận.