ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

GIẢI PHÁP XẾP HẠNG VÀ TÍNH TOÁN SONG SONG TRÊN NỀN TẢNG APACHE SPARK

Nguyễn Đông Đức

LUẬN ÁN THẠC SĨ
Bản tóm tắt
Chuyên ngành: Hệ Thống Thông Tin

12/2016

MỤC LỤC

Tóm tắt

Trong những năm gần đây, với sự phát triển nhanh chóng của WWW(World Wide Web) và những khó khăn trong việc tìm kiếm thông tin mong muốn, hệ thống tìm kiếm thông tin hiệu quả đã trở nên quan trọng hơn bao giờ hết, và các công cụ tìm kiếm đã trở thành một công cụ thiết yếu đối với nhiều người. Xếp hạng thông tin một thành phần không thể thiếu trong mọi công cụ tìm kiếm, thành phần này chịu trách nhiệm cho sự kết hợp giữa các truy vấn xử lý và tài liệu được lập chỉ mục. Ngoài ra, Ranking cũng là thành phần then chốt cho nhiều ứng dụng tìm kiếm thông tin khác, ví dụ như lọc cộng tác, tóm tắt văn bản và các hệ thống quảng cáo trực tuyến. Sử dụng mô hình học máy trong quá trình xếp hạng dẫn đến tạo ra cách mô hình các mô hình xếp hạng sáng tạo và hiệu quả hơn, và cũng dẫn đến phát triển một lĩnh vực nghiên cứu mới có tên là học máy xếp hạng (Learning to rank).

Trong mô hình mới này có rất nhiều cách tiếp cận như Pointwise , Pairwise, Listwise Luận văn này sẽ nghiên cứu các cách tiếp cận cho bài toán xếp hạng sử dụng Apache Spark và các thành phần bên trong nó cho việc phân tích dữ liệu đồng thời trên quy mô lớn có thể mở rộng dễ dàng cũng như khả năng chịu lỗi.

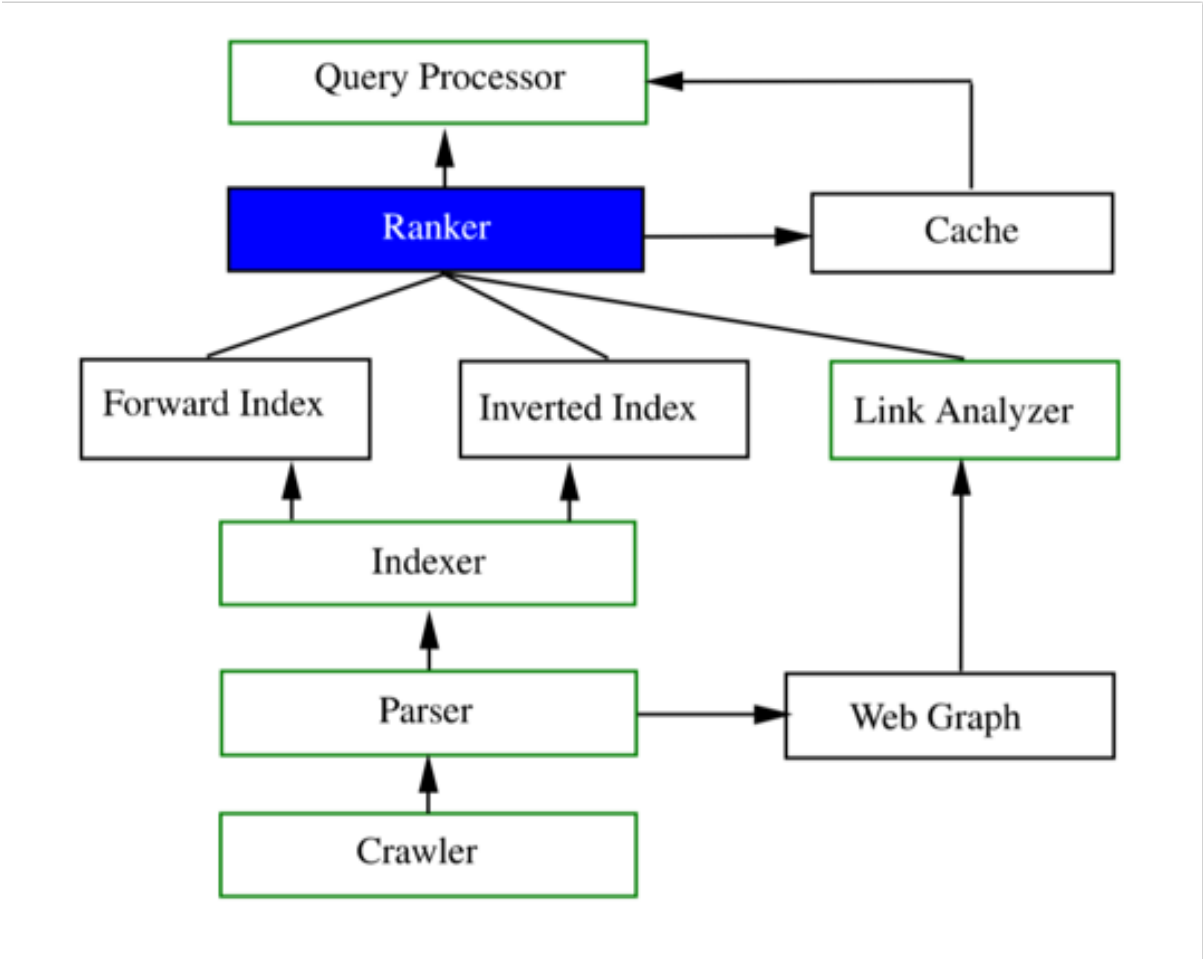
Chương 1
TỔNG QUAN VỀ HỌC MÁY XẾP HẠNG

1.1. Tổng quan về học máy xếp hạng

Với sự phát triển nhanh trong của thế giới Web dẫn đến tràn ngập thông tin trên mạng internet. Một nghiên cứu đã được tiến hành năm 2005 chỉ ra rằng thế giới Web chứa khoảng 11.5 tỉ tài liệu tại thời điểm tháng 1 năm 2005. Trong cùng năm đó, Yahoo đã thông báo rằng cỗ máy tìm kiếm của họ chứa khoảng hơn 19.2 tài liệu web. Ngày nay con số này đã lên đến hơn 50 triệu tỉ tài liệu đã được đánh chỉ mục trong các cỗ máy tìm kiếm. Từ những số liệu này chúng ta có thể thầy rằng số lượng tài liệu web đang tăng lên nhất nhanh.

Với kích thước cực kỳ lớn của thế giới Web rõ ràng rằng người dùng thông thường khó có thể tìm kiếm thông tin mà họ mong muốn bằng cách duyệt và tìm kiếm thông tin trên những trang web. Việc tìm kiếm và trích xuất thông tin đã trở nên quan trọng hơn bao giờ hết, và các công cụ tìm kiếm đã dần dần trở thành một công cụ thiết yếu mà mọi người dùng internet đều sử dụng.

Một kiến trúc điển hình của công cụ tìm kiếm được hiển thị trong Hình 1-1



Hình 1-1 - Hệ thống tìm kiếm tổng quát

1.2. Mô hình xếp hạng truyền thống

Trong các tài liệu của hệ thống truy hồi thông tin, rất nhiều mô hình xếp hạng đã được đề xuất **Error! Reference source not found.** có thể tạm phân loại 2 mô hình chính đó là mô hình xếp hạng dựa trên độ liên quan (Relevance Ranking Modal) và mô hình xếp hạng dựa trên độ quan trọng (Importance Ranking Models)

1.2.1. Mô hình xếp hạng dựa trên độ liên quan

Mục tiêu của mô hình xếp hạng liên quan là tạo ra một danh sách các tài liệu được xếp hạng theo mức độ liên quan giữa tài liệu và truy vấn. Sau đó sắp xếp tất các các tài liệu theo thứ tự giảm dần theo các chỉ số liên quan của chúng.

Mô hình xếp hạng liên quan trong hệ thống truy hồi thông tin đầu tiên được dựa trên sự xuất hiện các term của truy vấn trong tài liệu. Ví dụ điển hình cho mô hình này là mô hình Boolean **Error! Reference source not found.**. Về cơ bản mô hình có thể đoán một tài liệu là liên quan hay là không liên quan với truy vấn nhưng không đo được mức độ liên quan.

Một mô hình về đo độ liên quan mới là mô hình không gian Vector (Vector Space modal – SVM) được đưa ra **Error! Reference source not found.**. Cả tài liệu và truy vấn được định nghĩa như những vector trong một không gian Euclid, trong đó tích trong của 2 vector được sử dụng để đo mức độ tương tự của truy vấn và tài liệu. Để tạo ra vector hiệu quả đại diện truy vấn và tài liệu thì mỗi từ trong không gian vector sẽ có một trọng số, có nhiều phương pháp xếp hạng khác nhau, nhưng tf-idf (term frequency–inverse document frequency) **Error! Reference source not found.** là một phương pháp phổ biến để đánh giá và xếp hạng một từ trong một tài liệu. Về cơ bản thì tf-idf là một kỹ thuật (cụ thể là ranking function) giúp chuyển đổi thông tin dưới dạng văn bản thành một Vector space model thông qua các trọng số. Vector space model và tf-idf được phát triển bởi Gerard Salton vào đầu thập niên 1960s.

TF của một term t trong một vector được định là số lần xuất hiện của nó trong tài liệu.

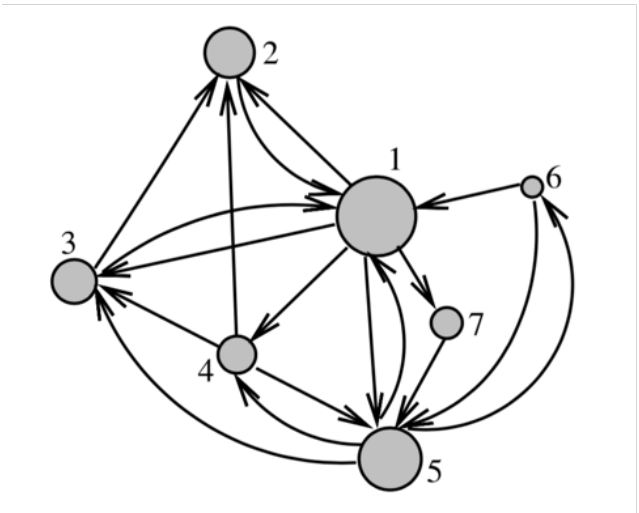
IDF được định nghĩa như sau

$$IDF(t) = log\frac{N}{n(t)} \quad (2.1)$$

trong đó N là số lượng tài liệu liệu trong tập hợp truy vấn, và n(t) là só lượng tài liệu mà chứa term *t*

1.2.2. Mô hình xếp hạng dựa trên độ quan trọng

Trong tài liệu truy hồi thông tin, cũng có rất nhiều mô hình mà xếp hạng các tài liệu dựa trên độ quan trọng của chúng. Một mô hình rất nổi tiếng đó là PageRank, mô hình này được áp dụng đặc biệt hệ thống tìm kiếm Web bởi vì nó sử dụng cấu trúc siêu liên kết Web để xếp hạng.



Hình 1-2 – Minh họa thuật toán PageRank

Mô hình này được Page và các đồng tác giả đã đưa ra ý tưởng là độ quan trọng của một trang chịu ảnh hưởng của độ quan trọng từ các trang liên kết đến nó.

1.3. Phương pháp đánh giá mô hình xếp hạng

1.3.1. Phương pháp MRR (Mean Reciprocal Rank)

Cho truy vấn q, vị trí xếp hạng của phần tử liên quan đầu tiên trong danh sách xếp hạng mà mô hình xếp hạng trả về được biểu diễn là r_1. Sau đó $\frac{1}{r_1}$ được định nghĩa như như một MRR cho truy vấn q. Có thể hiểu rằng những tài liệu được xếp hạng dưới r_1 xe không được quan tâm trong MRR.

Query = learning to rank

1. http://research.microsoft.com/~letor/	Relevant
2. http://www.learn-in-china.com/rank.htm	Irrelevant
3. http://web.mit.edu/shivani/www/Ranking-NIPS-05/	Relevant
... ...	

Hình 1-3 – Kết quả truy hồi cho truy vấn "learning to rank"

Nhìn vào ví dụ được minh họa trong Hình 1-3. Vì vị trí đầu tiền trong kết quả truy hồi liên quan là r_1 =1. Do đó MRR cho truy vấn này là 1.

1.3.2. Phương pháp đánh giá MAP (Mean Average Precision)

Để định nghĩa MAP **Error! Reference source not found.**, thì cần phải định nghĩa độ chính xác tại vị trí *k (P@k)* đầu tiên. Giả sử rằng chúng có điểm số đánh giá nhị phân cho tài liệu, v.v, Nhãn là 1 cho tài liệu liên quan và nhãn là 0 cho tài liệu không liên quan. Sau đó P@k được định nghĩa như sau

$$P@k(\pi, l) = \frac{\sum_{t \le k} I_{\{l_{\pi^{-1}(t)}=1\}}}{k}, \quad (2.10)$$

Trong đó $I_{\{\cdot\}}$ là hàm đặc trưng và $\pi^{-1}(j)$ biểu thị vị trí tài liệu xếp hạng tại vị trí j of the list π.

Sau đó trung bình độ chính xác (AP) được định nghĩa như sau

$$\mathrm{AP}(\pi, l) = \frac{\sum_{k=1}^{m} P@k \cdot I_{\{l_{\pi^{-1}(k)}=1\}}}{m_1}, \tag{2.11}$$

Trong đó m là tổng số tài liệu tương ứng với truy vấn q, và m_1 là số lượng tài liệu mà có nhán là 1.

Giá trị trung bình của AP cho toàn bộ test truy vấn được gọi là độ chính xác trung bình (MAP)

Với ví dụ được chỉ ra trong Hình 1-3. Vì tài liệu đầu tiên trong kết quả truy hồi là liên quan, rõ ràng P@1 = 1. Bởi vì tài liệu thứ 2 là không liên quan nên P@2 = $\frac{1}{2}$. Cuối cùng tài liệu thứ 3 là liên quan từ đó suy ra P@3 = $\frac{2}{3}$.

Từ đó ta có $AP = \frac{1}{2}\left(1 + \frac{2}{3}\right) = \frac{5}{6}$

1.3.3. Phương pháp đánh giá DCG (Discounted Cumulative Gain)

DCG **Error! Reference source not found.** là một phương pháp đánh giá có thể sử dụng điểm số đánh số đánh giá cho nhiều loại đánh giá có thứ tự, vì nó đã có hệ số chiết khấu vị trí rõ ràng trong định nghĩa của nó. Định ngĩa một cách chính thức như sau, giả sử danh sách sắp xếp cho truy vấn q là π, sau đó DCG tại vị trí k được định nghĩa như sau:

$$\mathrm{DCG@}k(\pi, l) = \sum_{j=1}^{k} G(l_{\pi^{-1}(j)})\eta(j), \tag{2.12}$$

Trong đó $G(\cdot)$ là đánh giá của một tài liệu hàm này thường được gán bằng $G(z) = (2^z - 1)$ và η(j) là hệ số chiết khấu được được gán bằng η(j) = 1/ log(j + 1)

Bằng cách bình thường hóa DCG@k với giá trị cực đại có thể Zk, chúng ta sẽ có được một phương pháp đo khác là Normalized DCG (NDCG).

$$\mathrm{NDCG@}k(\pi, l) = \frac{1}{Z_k} \sum_{j=1}^{k} G(l_{\pi^{-1}(j)})\eta(j). \quad (2.13)$$

Dó đó NDCG sẽ luôn luôn có giá trị từ 0 và 1

Với ví dụ được chỉ ra ở Hình 1-3. Có thể dễ dàng tính được DCG@3 = 1.5, và Z_3=1.63. Cuối cùng $\mathrm{NDCG} = \mathrm{NDCG@3} = \frac{1.5}{1.63} = 0.92$

1.4. Học máy xếp hạng

1.4.1. Nền tảng cơ sở của học máy

Có rất nhiều thuật toán học máy xếp hạng sử trong đó có ba cách tiếp cận cho mô hình học máy đó là các tiếp cận pointwise, pairwise và listwise.

1. 4. 1. 1. Hướng tiếp cận Pointwise

Theo hướng này, các đối tượng x_i trong dữ liệu học có một điểm số hay thứ tự y_i. Tiếp đó, học xếp hạng có thể được xấp xỉ bởi hồi quy (hồi quy có thứ tự). Với D = {(x_i, y_i)}, hàm tính hạng h(x) thỏa mãn, r(x_i) = y_i. Một số thuật toán học xếp hạng như: OPRF [4], SLR [7], ...

1.4.1.2. Hướng tiếp cận Pairwise

Có *D = {(x_i, x_j)}* là tập các cặp đối tượng được sắp thứ tự, với mỗi cặp (x_i, x_j) có thứ hạng của x_i cao hơn thứ hạng của x_j, hay x_i phù hợp hơn x_j: x_i> x_j). Tìm r(x):

$$\forall(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{S}\ \textbf{có}\ \boldsymbol{x_i} > \boldsymbol{x_j}\ \boldsymbol{th}\textbf{ì}\ \boldsymbol{r(x_i)} > \boldsymbol{r(x_j)} \quad (2.14)$$

Một số thuật toán học xếp hạng như SVM-rank, RankRLS ...

1. 4. 1. 3. Hướng tiếp cận Listwise

Các thuật toán theo hướng này cố gắng trực tiếp sắp xếp tất cả các đối tượng trong dữ liệu học. Điều này thực sự khó khăn. Khi thứ hạng của K đối tượng đầu tiên được xác định thì tất cả các đối tượng khác đều có hạng thấp hơn.

Với $D=\{x_1,x_2...,x_m\}$ có sắp thứ tự: $x_1 > x_2 > ... > x_m$, tìm hàm tính hạng r(x) sao cho $r(x_1) > r(x_2) > ... > r(x_m)$.

Một số thuật toán học xếp hạng như ListMLE, Listnet, PermuRank

Luận văn này sẽ sử dụng cách tiếp này để nghiên cứu và thực nghiệm xếp hạng.

Chương 2

GIẢI PHÁP XẾP HẠNG KẾT QUẢ TÌM KIẾM

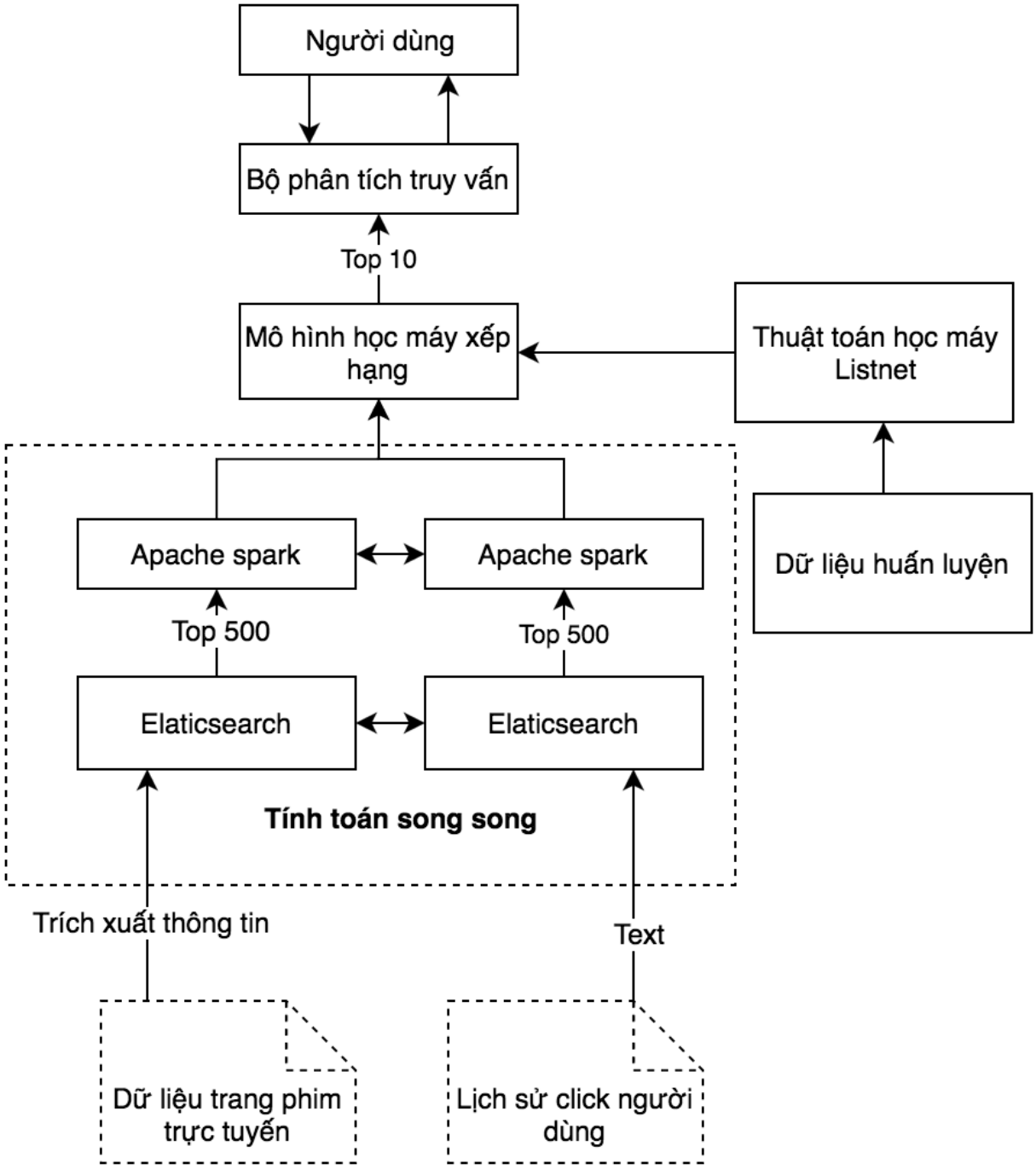
2.1. Công nghệ

Như đã trình bày ở trên chúng tôi thực hiện xây dựng hệ thống xếp hạng có thể tính toán song song trên nhiều máy tính làm rút ngắn thời gian truy vấn, huấn luyện dữ liệu. Bên cạch hệ thống cần phải chạy theo thời gian thực, khả năng mở rộng và khả năng chịu lỗi. Sau đây là các công nghệ đã được sử dụng trong hệ thống này.

Hệ thống sử dụng Elasticsearch và ApacheSpark để tính toán và xử lý song song

2.2. Mô hình hệ thống

Phần này sẽ giới thiệu toàn bộ mô hình từ thu thập dữ liệu, huấn luyến mô hình, và phục vụ tìm kiếm các bộ phim cho hệ thống tìm kiếm tại Cốc Cốc.



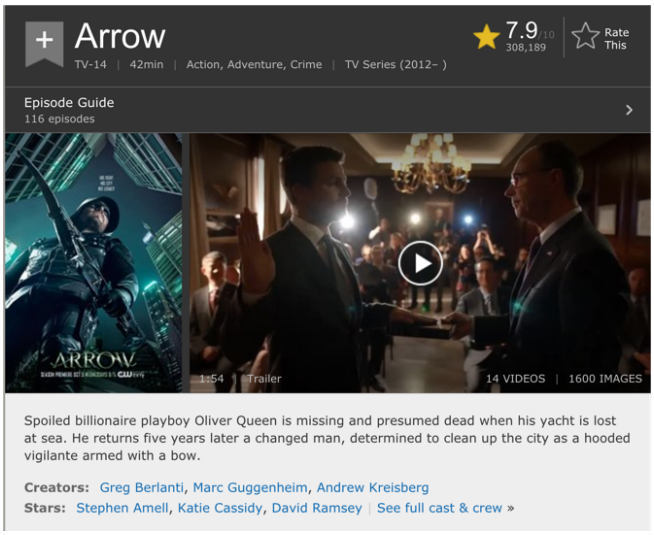
Hình 2-1 – Mô hình giải pháp xếp hạng và tính toán song song

2.3. Thu thập và xử lý dữ liệu

Dữ liệu sẽ được thu thập từ nhiều trang web và thông tin của người dùng từ hệ thống crawler và search của cốc cốc

2.3.1. Thu thập dữ liệu phim

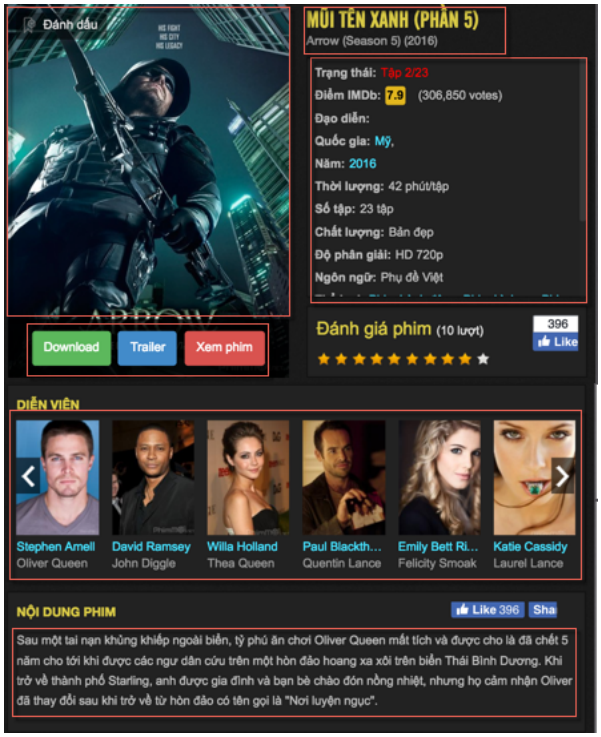
Đầu tiên là hệ thống sẽ trích xuất các thông tin từ trang web đánh giá phim IMDb (Internet Movie Database).



Hình 2-2 – Thông tin phim trên trang IMDb

IMDb là một website trực tuyến nó đóng vai trò như một thư viện, nơi lưu trữ những thông tin chi tiết về các tác phẩm điện ảnh nổi tiếng, ngoài ra IMDb còn là website uy tín đóng vai trò như một nhà phê bình. IMDb cũng là nơi tổng hợp những ý kiến đánh giá, xếp hạng của một tác phẩm điện ảnh dựa trên các yếu tố như kịch bản, công tác đạo diễn, bối cảnh, hiệu quả hình ảnh, kỹ thuật quay phim…IMDb rất có uy tín với giới độc giả Internet, cũng như các tín đồ của môn nghệ thuật thứ 7. Ngoài nội dung phê bình đánh giá về các tác phẩm thuộc lĩnh vực điện ảnh, IMDb còn đánh giá những tác phẩm truyền hình hay những ngôi sao điện ảnh, nhà sản xuất phim…

Thứ Hai là hệ thông sẽ bóc tách các thông tin như được miêu tả ở Hình 2-3 (Các vùng được bôi đỏ sẽ được trích xuất thông tin) các thống tin được trích xuất bao gồm những thành phần được bôi đỏ như sau.

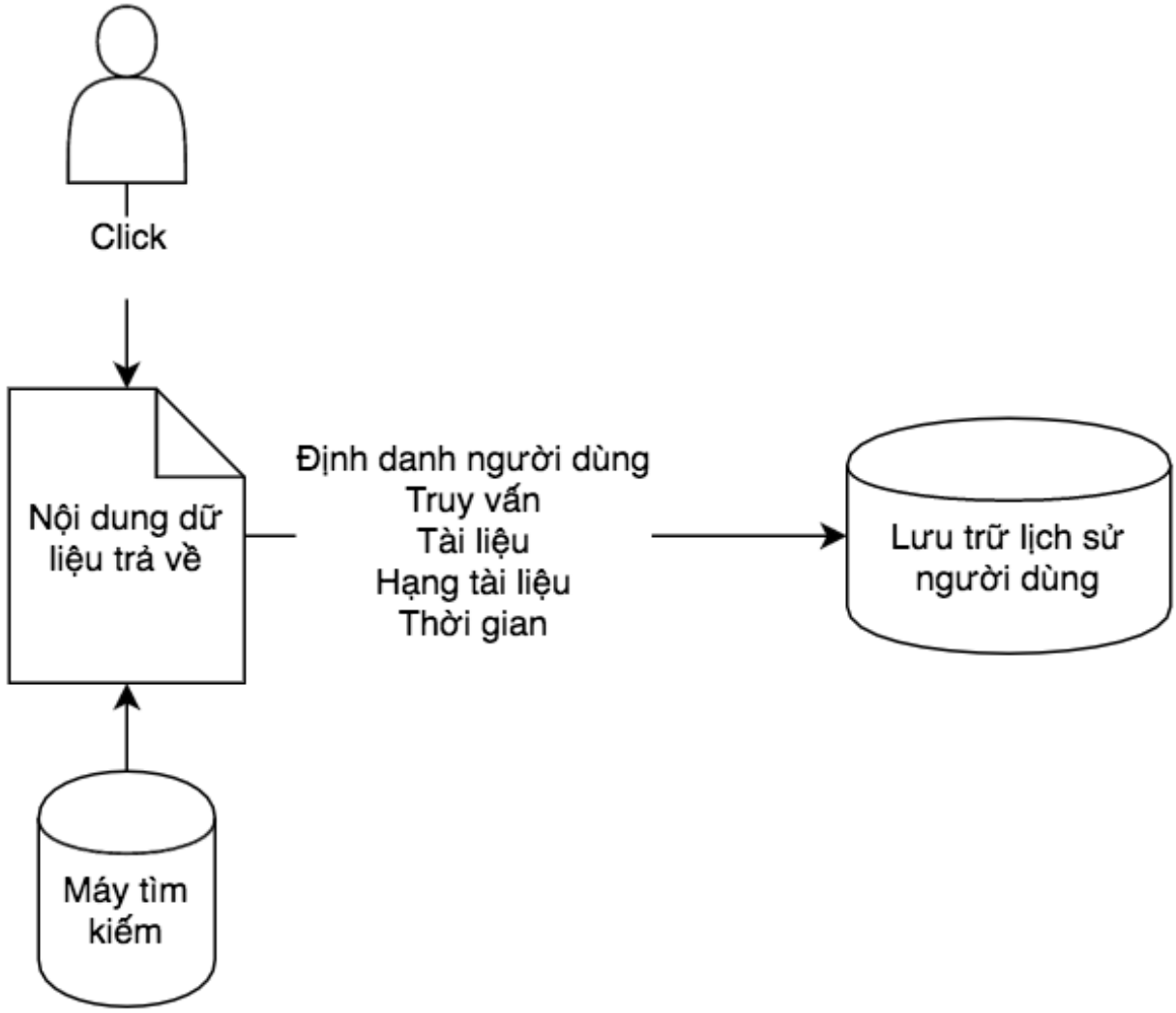


Hình 2-3 – Thông tin được trích xuất trong trang phim trực tuyến.

2.3.2. Thu thập lịch sử click của người dùng

Đây là dữ liệu có được có được khi hệ thống đã được đưa ra để sử dụng, dữ liệu này cũng được coi như là một tham số trong vector đặc điểm dùng để huấn luyện mô hình. Dữ liệu thông tin lịch sử được thu thập bao gồm: truy vấn, định danh người dùng, liên kết phim được click, hạng được click.

Khi hệ thống chưa được đưa ra sử dụng thì thông này sẽ được thu thập từ hệ thống tìm kiếm của cốc cốc và trích xuất thông tin click của người dùng từ những trang phim được định trước.



Hình 2-4 – Mô hình lưu trữ lịch sử của người dùng

2.3.3. Đánh chỉ mục cho dữ liệu

Tất cả các thông tin được thu được như thông tin phim, dữ liệu IMDb, lịch sử click của người dùng được đánh chỉ mục vào các document trong hệ thống Elasticsearch như sau:

2.4. Xác định vector đặc trưng cho mô hình

Vector đặc trưng được sử dụng trong mô hình huấn luyện bao gồm các giá trị điểm số được tính toán dữa trên truy vấn và tài liệu, các thuộc tính thuộc tính của vector đặc trược được biểu diễn trong bảng dưới đây

Bảng 2.1 – Bảng mô tả vector đặc trinh cho mô hình học máy xếp hạng

Số thứ tự	Mô tả

1	Tổng điểm số TF của tiêu đề phim
2	Độ dài của tiêu đề phim
3	Điểm số BM25 của truy vấn và tiêu đề phim
4	Tổng điểm số TF của nội dụng phim
5	Độ dài của nội dung phim
6	Điểm số BM25 của truy vấn và nội dung phim.
7	Hạng trang web của tài liệu
8	Hạng của domain gốc của tài liệu
9	Điểm số IMDB của tài liệu
10	Tổng số lượt click của tài liệu
11	Thơi gian sản xuất phim (Năm hiện tại – Năm sản xuất)

Chương 3

THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Dữ liệu

Mô hình sử dụng 130.380 thông tin phim được trích xuất từ IMDb và 93.270 liết kết phim được trích xuất

Mô hình sử dụng query log của hệ thống tìm kiếm tại Cốc Cốc được phân loại theo chủ đề phim. Query log là thành phần quan trọng của một bộ máy tìm kiếm, đây là dữ liệu thu thập lại hành vi của người sử dụng qua từng truy vấn mà người dùng đó thao tác trên bộ máy tìm kiếm. Dữ liệu log này không chứa tài liệu quảng cáo mà được hiển thị ra cho người sử dụng. Đây cũng là dữ liệu cho bộ huấn luyện cũng như đánh giá. Dữ liệu về query log cũng được tổng hợp theo hàng tuần và được lưu trữ như sơ đồ hình Hình 2-4

Dữ liệu huấn luyện sử dụng lịch sử ba tháng query log của người dùng được lọc theo nội dụng truy vấn và liên kết của tài liệu để xác định có phải là truy vấn để truy hồi thông tin phim trực tuyến hay không. Sau khi đã trích chọn thu được 20 triệu truy vấn dữ liệu click. Dữ liệu bao gồm các thông tin sau đây:

- **User**: định danh của người dùng
- **Query**: nội dung query, đây là nội dung query được người dùng nhập vào.
- **Time**: thời điểm người dùng click vào URL.
- **URL**: URL được người dùng click.
- **Position**: vị trí của url được click trong danh sách kết quả trả về.

3.2. Môi trường thực nghiệm

3.2.1. Cấu hình phần cứng

Quá trình thực nghiệm được tiến hành trên máy tính có cấu hình phần cứng như sau:

Bảng 3.1 - Thông số máy chủ sử dụng trong thực nghiệm.

STT	Thông số	Số lượng
1	OS: Debian 8.0 HDD: 2TB RAM: 32GB CPU: 2.7 GHz x 24 Core	3
2	OS: Debian 8.0 HDD: 1TB RAM: 64GB CPU: 2.7 GHz x 24 Core	1

3.2.2. Các công cụ được sử dụng

Dưới đây là các công cụ mã nguồn mở được sử dụng

Bảng 3.2 - Danh sách phần mềm mã nguồn mở được sử dụng

STT	Tên phần mềm	Nguồn	Phiên bản
1	elasticsearch-hadoop	https://www.elastic.co/downloads/hadoop	2.4.0
2	Apache Spark	http://spark.apache.org/downloads.html	2.0.1
3	Ranklib	https://sourceforge.net/p/lemur/wiki/RankLib/	2.7

3.3. Quá trình thực nghiệm.

Quá trình thực nghiệm gồm các bước chính sau đây:

- Xử lý dữ liệu: tiền xử lý dữ liệu, xây dựng tập tài liệu học cho mô hình, véc tơ hóa dữ liệu. Đánh chỉ mục cho dữ liệu.
- Xây dựng hàm xếp hạng: tiến hành training trên tập dữ liệu đã có bằng thuật toán ListNet trong tự viện RankLib 2.7
- Đánh giá kết quả mô hình: Đánh giá thời gian thực thi của mô hình bằng một máy tính và ba máy tính.

3.3.1. Tiền xử lý dữ liệu

Sau khi khi trích xuất được các thông tin phim như điểm Imdb và các tài liệu phim online dữ liệu sẽ được đánh chỉ mục vào trong Elasticsearch theo các trường được miêu tả tại **Error! Reference source not found.** và **Error! Reference source not found.**

Tiếp đến là phần xác định điểm Imdb cho tất cả tài liệu phim online đã được đánh chỉ mục. Mỗi tên bộ phim trong liệu phim online sẽ được tìm kiếm tương ứng bảng dữ liệu Imdb, các dữ liệu trong imdb như đạo diễn, diễn viên, năm sản xuất sẽ được làm yếu tố để đánh giá xem một dữ liệu phim imdb sẽ phù hợp nhất với dữ liệu phim online nào.

Bước tiếp theo là lọc các truy vấn và liên kết phim online phù hợp dựa vào các liên kết phim online có phải là tên miền hợp lệ không với quy tắc là mỗi truy vấn phải có 10 liên kết được người dùng click là domain phim. Tại bước này ta thu được gần 430.000 truy vấn liên quan tới phim.

Từ các dữ liệu được trích xuất trên ta tiến hành trích xuất vector đặc trưng với các phần tử được miêu tả trong Bảng 2.1

3.3.2. Tiến hành thực nghiệm

Để có thể đánh giá thời gian thực thi và làm rõ mục tiêu của luận văn là xây dựng mô hình xếp hạng bằng tính toán song song. Cách thức thực nghiệm sẽ được chia thành hai phần một phần là so sánh hiệu quả thời gian một phần là so sánh về chất lượng của phương pháp xếp hạng.

3.3.2.1. So sánh hiệu quả thời gian

Để so sánh hiệu quả thời gian tôi tiến hành chạy các bước thực nghiệm trên một máy đơn và ba máy tính có thống số trong **Error! Reference source not found.**. Kết quả của quá trình thực nghiệm này được biểu diễu dưới đây

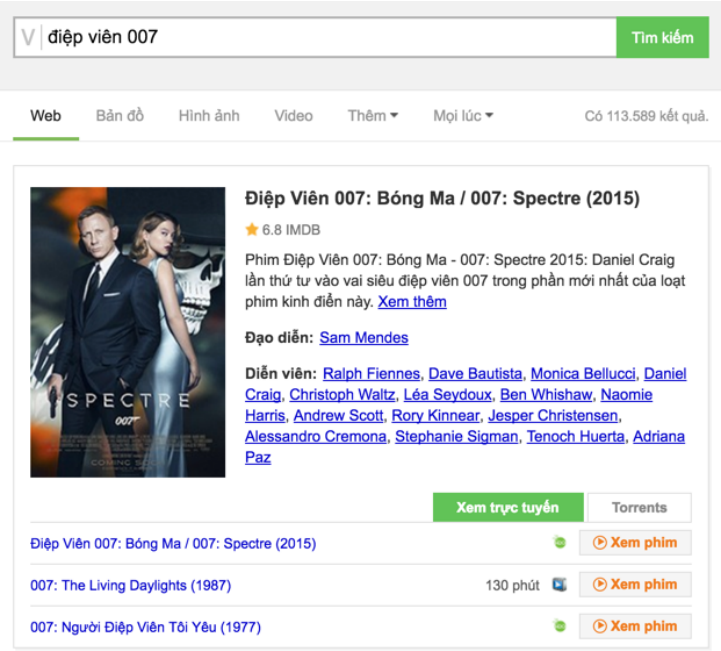
Bảng 3.3 – Bảng đánh giá hiệu quả về mặt thời gian

	Một máy tính	Ba máy tính
Đánh chỉ mục dữ liệu cho hơn 3 triệu bản ghi	32 phút 15s	13 phút 27s
Huấn luyện mô hình cho gần 500.000 truy vấn và tài liệu	2h 30phút	44 phút
Chạy 1.000.000 truy vấn của người dùng	45 phút 23s	18phút 09s

Từ bảng kết quả trên cho thấy với ba máy tính tốc độ xử lý đã tăng lên rất nhiều do đã tận dụng sức mạnh của nhiều máy tính trong cùng một khoảng thời gian. Mô hình cũng cho phép có thể kế nối với nhiều máy hơn nữa để giảm thời gian chạy hoặc tăng khối lượng tính toán.

3.3.2.2. So sánh chất lượng xếp hạng

Mô hình đã được chạy trên hệ thống Cốc Cốc như một thành phần của hệ thống tìm kiếm.



Hình 3-1 – Hệ thống tìm kiếm phim online trên cốc cốc

Hình 3-1 biểu diễn chức năng tìm kiếm phim với truy vấn "diep vien 007". Sau khi áp dụng mô hình xếp hạng mới và giải pháp tính toán song song, tốc độ và chất lượng của hệ thống tìm kiếm phim online cụ thể là điểm số CTR(Click through Rate) đã được cải thiện đáng kể. Dưới đây là bảng thống kê về chỉ số CTR trước và sau 10 ngày sau khi triển khai mô hình mới.

Bảng 3.4 – Tỉ lệ CTR trước vào sau khi áp dụng mô hình

Kết quả trước và sau 10 ngày	Số lần hiển thị	Số lần nhấp chuột	CTR
Trước khi áp dụng mô hình (03/09/2016 – 13/09/2016)	923.070	79,107	8,57%
Sau khi áp dụng mô hình (14/09/2016 – 24/09/2016)	1.110.402	136.579	12,3%

Chương 4

KẾT LUẬN

Tính toán song song đang là xu thế của công nghệ cũng là lĩnh vực đang được rất quan tâm. Để có thể đáp ứng phục vụ ngày càng nhiều người dùng và ngày các nhiều dữ liệu trên WWW. Tính toán song song đã giúp việc xử lý dữ liệu lớn trên nhiều máy tính khác nhau để mở rộng khả năng tính toán, mở rộng khả năng chịu lỗi.

Luận văn này đã tiếp cập vấn đề học máy xếp hạng và nghiên cứu, đưa ra mô hình, áp dụng vào máy tìm kiếm Cốc Cốc để nâng cao chất lượng của bộ máy tìm kiếm.

Luận văn đã được những kết quả:

- Đưa ra cái nhìn tổng quát về bộ máy tìm kiếm và các thành phần bên trong một bộ máy tìm kiếm
- Trình bày các mô xếp hạng truyến thống và học máy xếp và các phương pháp đánh giá chất lượng của mô hình xếp hạng.
- Tìm hiểu nghiên cứu Apache Spark và Elasticsearch hai phần mềm mã nguồn mở cho lưu trữ và tính toán song song.
- Đưa ra mô hình xếp hạng phim trực tuyến cho máy tìm kiếm tại Cốc Cốc có khả năng mở rộng và khả năng tính toán song song và nâng cao chất lượng cũng như tỉ lệ CTR.