ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐÔNG ĐỨC

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

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN ĐÔNG ĐỨC

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

Ngành: Công Nghệ Thông Tin

Chuyên ngành: Hệ thống Thông Tin

Mã số chuyên ngành: 60480104

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Nguyễn Ngọc Hóa

Hà Nội – 2016

LỜI CAM ĐOAN

" Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Các số liệu, kết quả trình bày trong luận văn này là trung thực và chưa từng được ai công bố trong bất kỳ công trình luận văn nào trước đây."

Chữ ký:……

PHÊ DUYỆT CỦA GIÁO VIÊN HƯỚNG DẪN

"Tôi xin cam đoan rằng luận án đã đảm bảo đúng yêu cầu của chương trình đào Thạc sĩ Công nghệ Thông Tin của trường Đại học Công Nghệ."

Chữ ký:……

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, xếp hạng 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.

Lời cảm ơn

Trước tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Phó Giáo sư Tiến sĩ Nguyễn Ngọc Hóa, người đã tận tình chỉ bảo và hướng dẫn tôi trong suốt quá trình thực hiện khoá luận tốt nghiệp.

Tôi chân thành cảm ơn các thầy, cô đã tạo cho tôi những điều kiện thuận lợi để học tập và nghiên cứu tại trường Đại Học Công Nghệ.

Tôi cũng xin gửi lời cảm ơn tới các anh chị và các đồng nghiệp tại **Cốc Cốc** đã giúp đỡ và hỗ trợ tôi rất nhiều về kiến thức chuyên môn trong quá trình làm việc.

Cuối cùng, tôi muốn gửi lời cảm vô hạn tới gia đình và bạn bè, những người thân yêu luôn bên cạnh và động viên tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp.

Tôi xin chân thành cảm ơn!.

Danh sách các hình

Danh sách các bảng

Danh sách các từ viết tắt

BM25	Best Match 25
CTR	Click Through Rate
IDF	Inverse Document Frequency
LETOR	LEarning TO Rank
LMIR	Language Model for Information Retrieval
LSI	Laten Semantic Indexing
MRR	Mean Reciprocal Rank
SIGIR	Special Interest Group on Information Retrieval
SVD	Singular Value Decomposition
TF	Term srequency
WWW	World Wide Web

Chương 1.Giới thiệu chung

1.1 Động lực nghiên cứu

Với sự phát triển bùng nổ của công nghệ thông tin khi một người sử dụng internet có thể rất bối rối khi tìm kiếm thông tin do khối lượng đồ sộ của nó. Với nhiều nhu cầu tìm kiếm thông tin của người dùng các kết quả được trả về từ các hệ thống tìm kiếm cần được chính xác và chuyên biệt hóa thông tin. Nhận thấy nhu cầu giải trí đặc biệt là nhu cầu tìm kiếm phim online là một nhu cầu lớn trên bộ máy tìm kiếm trên Cốc Cốc với hàng triệu lượt truy vấn mỗi tuần. Cốc đã đã đưa ra ý tưởng là xây dựng một thành phần tìm kiếm phim trực tuyến. Để có thể cập nhật thông tin phim các bộ phim mới nhất cũng như hiển thị nhiều thông tin tới người dùng, Cốc Cốc đã xây dựng một hệ thống tìm kiếm chuyên biệt bên trong hệ thống tìm kiếm của Cốc Cốc để có thể hiển thị trực quan hóa hơn và hiển thị các thông tin như trailer, nội dung phim, đạo diễn, diễn viên, điểm imdb của bộ phim, kèm theo đó là những liên kết tới các trang web xem phim trực tuyến.

Với thiết kế hệ thống ban đầu hệ thống tìm kiếm phim trực tuyến được thiết kế và tính toán trên một máy chủ, với mô hình thiết kế này hệ thống có thể đáp ứng tốt trong thời gian đầu. Hệ thống có thể trả về kết quả các liên kết phim và xếp hạng chúng hiệu quả. Nhưng do dữ liệu ngày càng lớn để đáp ứng khả năng mở rộng khi cơ sở dữ liệu phim ngày càng lớn cần một mô hình tính toán song song trên nhiều máy tính và tính ổn định chịu lỗi khi nâng cấp hoặc có sự cố trên một máy tính xảy ra.

Cũng trong thời gian đầu các hệ số nhân của các yếu tố trong hệ thống xếp hạng phim được cố định trước và được điều chỉnh bằng cảm quan ban đầu. điều này dẫn đến tình trạng quá khớp với một số trường hợp tìm kiếm, nên cần một mô hình xếp hạng tổng quan có thể tìm ra tham số thích hợp nhất với từng truy vấn và có thể áp dụng cho nhiều loại truy vấn khác nhau không chỉ riêng tìm kiếm phim ảnh.

1.2 Mục tiêu và nội dung của luận văn

Luận văn này sẽ nghiên cứu các cách tiếp cận mô hình học máy xếp hạng áp dụng cho bài toán xếp hạng trang web xem phim trên **Cốc Cốc** sử dụng Apache Spark và Elasticsearch cho lưu trữ, phân tích dữ liệu đồng thời trên quy mô lớn có thể mở rộng dễ dàng cũng như khả năng chịu lỗi.

- Nghiên cứu, khảo sát bài toán xếp hạng tổng quát và nền tảng Apache Spark
- Phân tích, đánh giá một số kỹ thuật Listwise trong học xếp hạng
- Xây dựng giải pháp triển khai kỹ thuật học xếp hạng kiểu Listwise trên nền Apache Spark

- Thực nghiệm và đánh giá khả năng xử lý xếp hạng trên Apache Spark thông qua bài toán xếp hạng phim tích hợp trong dịch vụ tìm kiếm của Cốc Cốc.

1.3 Tổ chức của luận văn

Khóa luận bao gồm năm chương sau đây là mô tả vắn tắt các chương:

Chương 1. Giới thiệu chung. Chương này giới thiệu về mục tiêu và động lực nghiên cứu của luận văn.

Chương 2. Tổng quan về xếp hạng. Chương này trình bày tổng quan về các mô hình xếp hạng truyền thống được sử dụng và phân loại các mô hình xếp hạng.

Chương 3. Tổng quan về học máy xếp hạng. Chương này trình bày nền các mô hình học máy xếp hạng được sử dụng trong các hệ thống truy hồi thông tin

Chương 4. Giải pháp xếp hạng kết quả tìm kiếm. Chương này trình bày các công nghệ tính toán song song và đưa ra giải pháp cho bài toán xếp hạng và tính toán song song sử dụng Apache Spark và Elasticsearch.

Chương 5. Thực nghiệm và đánh giá. Chương này trình bày về dữ liệu được sử dụng, các giai đoạn xử lý dữ liệu và thực nghiệm, đưa ra kết quả của mô hình, nhận xét và phân tích kết quả thu được.

Chương 6. Kết luận. Chương này tổng kết và tóm lược nội dung chính của khóa luận.

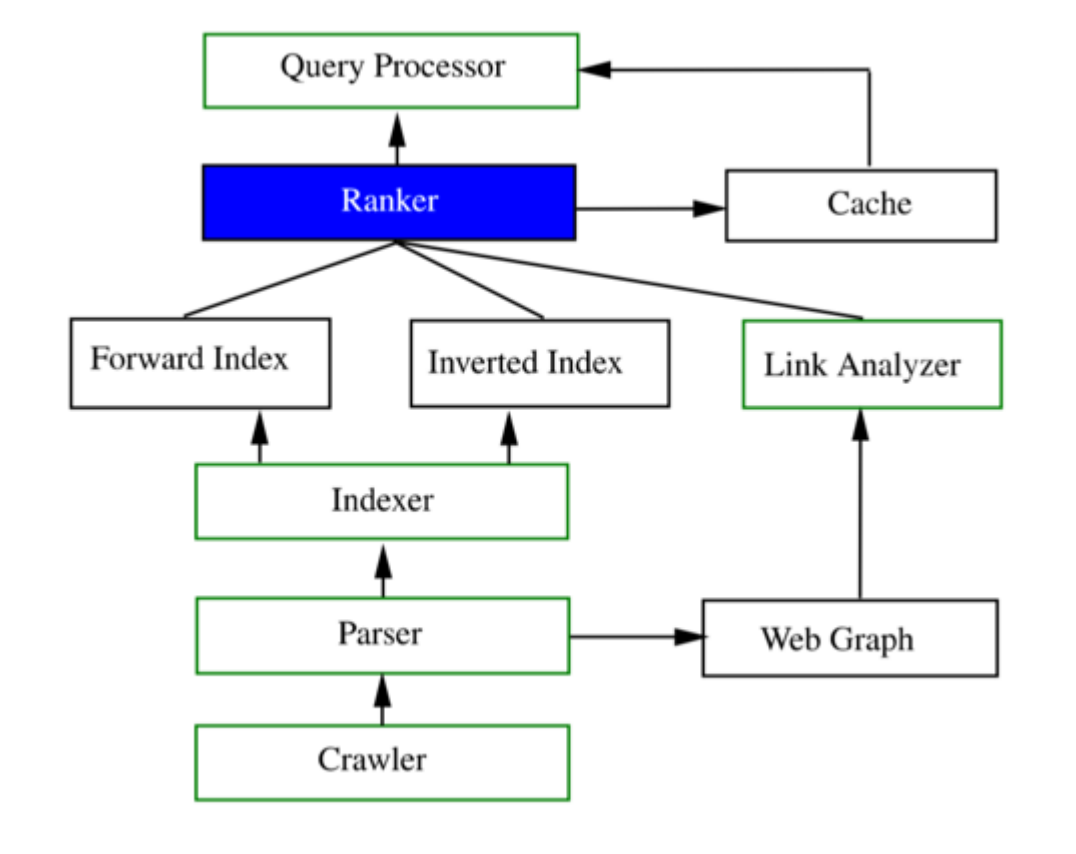
Chương 2.Tổng quan về xếp hạng

2.1 Tổng quan về xếp hạng

Sự phát triển bùng nổ thông tin của thế giới Web dẫn đến tràn ngập thông tin trên mạng internet. Một nghiên cứu đã được tiến hành năm 2005[23] chỉ ra rằng thế giới Web chứa khoảng 11.5 tỉ tài liệu tại thời điểm con số được thống kê là 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 (số liệu được lấy từ trang http://www.worldwidewebsize.com/) . Từ những số liệu này chúng ta có thể thấy rằng số lượng tài liệu web đang tăng lên ngày một 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 theo hình dưới đây



Hình 2-1 Hệ thống tìm kiếm tổng quát [24]

Có 6 thành phần chính trong một hệ thống tìm kiếm (Search Engine) là:

- **Crawler** (Bộ thu thập dữ liệu): Thu thập dữ liệu từ trang web và các tài liệu khác từ mạng internet theo sự ưu tiên.
- **Parser** (Bộ bóc tách dữ liệu): Lấy tài liệu từ crawler đánh chỉ mục và tạo đồ thị liên kết chứa các đường dẫn tới trang web (Hyperlink graph).
- **Indexer** (Bộ đánh chỉ mục): Có vai trò lấy dữ liệu từ Parser và tạo các chỉ mục từ (term) và các cấu trúc dữ liệu cho phép có thể tìm kiếm nhanh các tài liệu web.
- **Link Analyzer** (Bộ phân tích liên kết): Lấy dữ liệu từ đồ thị siêu liên kết và xác định độ quan trọng cho mỗi trang web. Kết quả này có thể để tạo độ ưu tiên được sử dụng cho việc cập nhật lại trang web thông qua Crawler hoặc để xác định như một tham số đặc trưng để xếp hạng.
- **Query processor** (Bộ xử lý truy vấn): Thành phần này nhận các truy vấn từ người dùng sau đó truy vấn được xử lý như loại bỏ các từ phổ biến, sửa lỗi cho truy vấn… sau đó chuyển chúng thành các từ (term) mà hệ thống tìm kiếm có thể hiểu được.
- **Ranker** (Bộ xếp hạng): Đây là thành phần trung tâm của hệ thống tìm kiếm nó chịu trách nhiệm tìm ra tài liệu thích hợp nhất từ truy vấn của người dùng và các tài liệu được đánh mục lục. Bộ xếp hạng có thể lấy trực tiếp các truy vấn và các tài liệu để tính toán một điểm số (score) sử dụng các công thức heuristic, hoặc cũng có thể trích xuất những đặc điểm giữa các cặp tài liệu và truy vấn để tạo ra điểm số được kết hợp từ những đặc điểm đó.

Hệ thống xếp hạng là một thành phần có vai trò trung tâm trong máy tìm kiếm do đó các công ty công nghệ lớn như **Yahoo**, **Google**, **Microsoft** trên thế giới và **Cốc Cốc** tại Việt Nam thì các thuật toán xếp hạng để cải thiện chất lượng của các cỗ máy tìm kiếm luôn là nhưng lĩnh vực được nghiên cứu nhiều nhất

Ngoài ra bộ xếp hạng cũng là thành phần trung tâm của rất nhiều ứng dụng truy hồi thông tin khác như lọc cộng tác, hệ thống hỏi đáp, truy hồi đa phương tiện, tóm tắm văn bản, và các hệ thống quảng cáo trực tuyến. Để giải quyết vấn đề của hệ thống truy hồi thông tin, rất nhiều mô hình xếp hạng heuristic đã được đề xuất và sử dụng trong hệ thống truy hồi thông tin.

Trong những năm gần đây, Học máy xếp hạng đã trở thành định hướng nghiên cứu nổi bật trong truy hồi thông tin và một số lượng lớn các bài báo khoa học về vấn đề học máy xếp hạng được xuất bản trong các hội nghị đứng đầu về học máy và truy hồi thông tin. Hàng năm có rất nhiều các chuyên đề trong hội nghị SIGIR được dành riêng cho chủ đề học máy xếp hạng, Các dataset như LETOR được sử dụng cho chủ đề này cũng được công bố để thuận tiện cho nghiên cứu học máy xếp hạng. Rất nhiều bài báo đã sử dụng dataset này cho việc thực nghiệm và nghiên cứu. Qua đó cũng thấy được tầm quan trọng cũng như mức độ phổ biến của học máy xếp hạng trong các hệ thống truy hồi thông tin.

Trong các tài liệu của hệ thống truy hồi thông tin, rất nhiều mô hình xếp hạng đã được đề xuất [5] 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)

2.2 Mô hình xếp hạng dựa trên độ liên quan

Mục tiêu của mô hình xếp hạng dựa trên độ 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 mức độ liên quan của chúng.

Mô hình xếp hạng dựa trên độ 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 [5]. 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) được đưa ra [5]. Trong mô hình này tài liệu và truy vấn được nghĩa như là các vector trong một không gian Euclid, trong đó tích trong của 2 vector được sử dụng để đo mức độ liên quan giữa truy vấn và tài liệu. Để tạo ra vector có kết quả tốt nhất thì mỗi term 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) [6] 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 hàm xếp hạng giúp chuyển đổi văn bản thành mô hình không gian vector thông qua các trọng số. Mô hình không gian vector và tf-idf được phát triển bởi Gerard Salton vào đầu thập niên 1960s.

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

IDF được định nghĩa như sau

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

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

Trong khi mô hình không gian vector bao hàm giả định về việc phụ thuộc giữa các term, Thì mô hình LSI (Laten Semantic Indexing) cố tránh giả định này. Cụ thể, SVD (Singular Value Decomposition) được sử dụng để chuyển đổi không gian tuyến tính các đặc trưng ban đầu thành không gian ngữ nghĩa ẩn (Latent semantic space). Không gian mới này cũng tương tự như mô hình không gian vector nó được sử dụng để định nghĩa độ liên quan giữa truy vấn và tài liệu.

Khi so sánh với mô hình dựa trên xác suất đã tạo được nhiều sự chú ý hơn và đạt được nhiều thành công trong thập kỷ qua. Mô hình nổi tiếng như MB25 và mô hình LMIR (Language model for information retrieval) cả hai có thể phân loại như là mô hình xếp hạng xác suất.

Ý tưởng cơ bản của BM25 là xếp hạng tài liệu bằng log và chỉ số odds của mức độ liên quan. Thực sự thì BM25 không giống như mô hình riêng rẽ, nhưng lại định nghĩa ra hàng loạt mô hình xếp hạng với sự khác nhau giữa các thành phần và các tham số trong công thức. Một trong những cách triển khai phổ biến chỉ số BM25 của một tài liệu d được tính như sau.

$$BM25(d,q) = \sum_{i=1}^{m} \frac{IDF(t_i).TF(t_i,d).(k_1+1)}{TF(t_i,d) + k_1.(1-b+b.\frac{LEN(d)}{avdl})} \quad (2.2)$$

trong đó **q** là một truy vấn chứa các term t1,…,tm, **TF**(*t,d*) là tần suất xuất hiện của term *t* trong tài liệu *d*, ***LEN***(*d*) là tổng độ dài (số các từ) của tài liệu *d*. và avdl là độ dài trung bình của tài liệu trong tập hợp được lấy ra. k_1 và b là tham số tự chọn, ***IDF***(*t*) là trọng số **IDF** của term t được tính bằng công thức trên.

LMIR là một ứng dụng của mô hình ngôn ngữ thống kê trong truy hồi thông tin. Một mô hình ngôn ngữ thống kê gán một xác suất đến một chuỗi các term. Khi sử dụng trong hệ thống truy hồi thông tin, một mô hình ngôn ngữ được liên kết với một tài liệu. Với đầu vào là truy vấn q các tài liệu được xếp hạng dựa trên sự hợp lý (likelihood) của truy vấn đó hoặc xác suất mà mô hình ngôn ngữ của tài liệu sẽ tạo ra term đó trong truy vấn (i.e., P(q|d)). Bằng cách tiếp tục giả định sự độc lập giữa các term do đó

$$P\left(\frac{q}{d}\right) = \prod_{i=1}^{M} P(ti|d) \quad (2.3)$$

nếu như query q chứa term $t1,\ldots,t_M$

Để học mô hình ngôn ngữ của tài liệu, một mô hình hợp lý cực đại (maximum likelihood) được sử dụng, như nhiều phương pháp hợp lý cực đại khác, vấn đề của mình làm mịn ước tính là rất quan trọng. Thông thường một mô hình ngôn ngữ nền tảng ước tính sử dụng toàn bộ tập hợp dữ liệu cho mục đích này. Sau đó, mô hình ngôn ngữ của tài liệu có thể được tạo ra như sau

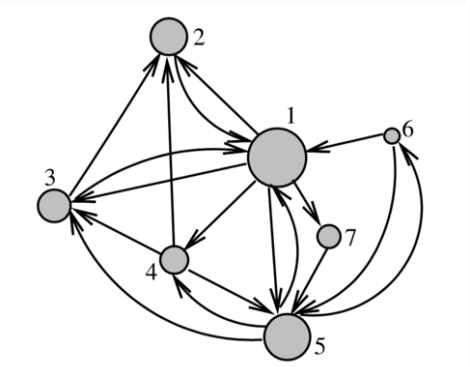
$$P(t_i,d) = (1-\lambda)\frac{TF(t_{i,d})}{LEN(d)} + \lambda p(t_i|C) \quad (2.4)$$

Trong đó $p(t_i|C)$ là là mô hình ngôn ngữ nền tảng của term t_i *và* $\lambda \in [0,1]$ nhân tố làm mịn.

Ngoài các mô hình trên cũng có nhiều các mô hình đã được đưa ra để tính toán liên quan giữa các truy vấn và tài liệu, mô hình lấy lân cận giữa các truy vấn và term làm mối quan tâm, một vài mô hình khác lại quan tâm tới sự tương đồng giữa tài liệu và term, cấu trúc của các siêu liên kết, cấu trúc website, và sự đa dạng của chủ đề.

2.3 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 thông tin trên Web bởi vì nó sử dụng cấu trúc siêu liên kết Web để xếp hạng.



Hình 2-2 Minh họa thuật toán PageRank [24]

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. Và công thức tính PageRank cho một trang u, gọi là π_u được tính như sau:

$$\pi_u = \sum_{i \in B_I(i)} \frac{\pi_i}{N_i} \tag{2.5}$$

Với $B_I(i)$là tập hợp các trang có liên kết đến trang I và Ni là số trang liên kết ra từ trang i. Biểu diễn đồ thị Web bởi ma trận chuyển P, khi đó phương trình 2.5 được viết lại dưới dạng ma trận:

$$\pi = \pi P \tag{2.6}$$

Trong đó: π = (π1, π2, . . . πn) là véc-tơ hạng các trang web, với thành phần πi là hạng của trang i.

Từ 2.6 cho thấy véc-tơ hạng trang π chính là véc-tơ riêng của ma trận chuyển P tương ứng với giá trị riêng λ = 1.

Do tính chất của chuỗi Markov, để tính véc-tơ riêng của P thuật toán giả thiết rằng đồ thị trang web là liên thông, tức với cặp hai trang web i, j bất kì luôn có đường đi từ i tới j và ngược lại. Tuy nhiên thực tế trên World Wide Web (WWW) vẫn tồn tại không ít các trang web không có liên kết đến hoặc liên kết ra nên việc giả thiết đồ thị Web liên thông là không hợp lý. Và trong ma trận P vẫn tồn tại hàng chỉ toàn số 0, nên không tồn tại một phân phối xác suất dừng ổn định của P hay chính là véc-tơ hạng trang. Vì vậy cần phải biến đổi ma trận P thành P′ sao cho phù hợp.

Định nghĩa véc-tơ v, được chuẩn hóa $\| v \| = 1$, xác định xác suất phân phối vớI vi là xác suất trang web i được gọi đến ở lần duyệt web đầu tiên. véc-tơ v có vai trò trong việc hướng kết quả PageRank theo chủ đề, lĩnh vực mong muốn. Khi không xét đến ngữ cảnh đó có thể chọn $vi = \frac{1}{n}$ $với \; \forall i = 1,2..n$.

Gọi d là véc-tơ n × 1 xác định các trang không có liên kết ra (dangling nút trên đồ thị Web):

$$d_i = \begin{cases} 1 \; nếu \; N(i) = 0 \\ 0 \; ngược \; lại \end{cases} \tag{2.7}$$

Ma trận P′ được xác định:

$$\mathbf{P'} = \mathbf{P} + \mathbf{dv} \tag{2.8}$$

Khi thay đổi ma trận P như vậy tức thêm các liên kết ảo từ các dangling nút tới tất cả các nút khác trong đồ thị Web theo phân phối xác suất v. Điều đó giúp tránh việc khi duyệt các trang không có liên kết ra sẽ không duyệt tiếp được.

Để đảm bảo phân phối dừng ổn định (duy nhất), chuỗi Markov tương ứng với quá trình duyệt Web của người dùng cần có tính chất ergodic, tức từ một trang web người dùng có thể chuyển tới một trang bất kì khác. Do vậy ma trận Markov $\check{P}$ được xác định như sau:

$$\check{P} = \alpha P' + \frac{(1 - \alpha)}{J} \tag{2.9}$$

α thường được chọn giá trị 0.85, với ý nghĩa tại mỗi bước duyệt Web người dùng có thể chuyển tới một trang trong các liên kết ra từ trang hiện tại với xác suất α và chuyển tới các trang khác trong đồ thị Web với xác suất (1 − α) theo phân phối v. Với : $J = [1]_{n \times 1} v$ và α: là hệ số hãm. Khi đó, thay vì tính vector riêng của ma trận P ta tính vector riêng π của ma trận

$\check{P}$: $\pi = \pi \check{P}$. Theo tính chất của chuỗi Markov, tổng các thành phần của véc-tơ π bằng 1. Vậy véc-tơ hạng trang chính là véc-tơ riêng của ma trận $\check{P}$.

Đã có rất nhiều các thuật toán được phát triển để mở rộng hơn nữa độ chính xác và độ hiệu quả của PageRank. Một số tập trung vào tăng tốc độ tính toán [11][12][13] trong khi một số khác lại tập trung vào chất lượng xếp hạng cho các mô hình. Ví dụ: Pagerank trong chủ đề nhạy cảm (topic-sensitive PageRank) [14] và PageRank trong truy vấn phụ thuộc [15] giới thiệu các chủ đề và cho rằng sự ủng hộ từ một trang thuộc cùng một chủ đề lớn hơn là từ một trang thuộc về một chủ đề khác nhau. Các biến thể khác của PageRank bao gồm những thay đổi của các 'vector cá nhân hóa'.

Thuật toán mà có thể tạo ra độ quan trọng xếp hạng để chống lại việc spam liên kết cũng được đưa ra. Ví dụ: TrustRank [16] là thuật toán quan trọng xem xét độ tin cậy của trang web khi tính đến tầm quan trọng của trang. Trong TrustRank, một tập hợp các trang đáng tin cậy đầu tiên được xác định là các trang có độ tin cậy cáo. Sau đó, sự tin tưởng của một trang hạt giống là tuyên truyền để các trang khác trên trang web liên kết đồ thị. Kể từ khi việc nhân giống trong TrustRank bắt đầu từ các trang tin cậy, TrustRank thể có thể phát hiện được nhiều spam hơn PageRank.

Chương 3.Học máy xếp hạng

Đã có rất nhiều mô hình học máy xếp hạng đã được giới thiệu ở các phần trước, đa phần trong số chúng chứa các tham số. Ví dụ tham số k1 và b trong BM25, và tham số λ trong LMIR tham số α trong PageRank. Để có thể có được hiệu suất xếp hạng tốt (đánh giá bằng các phương pháp xếp hạng ở trên), chúng ta cần tinh chỉnh thông số này sử dụng các luật bằng đánh giá cảm tính. Tuy nhiên việc điều chỉnh các tham số này là không hề đơn giản. Ngoài ra một mô hình cho kết quả tốt trên bộ test đôi khi cho kết quả kém trên tập truy vấn mới hơn, vấn đề này thường được gọi là over-fitting. Một vấn đề khác đề cập đến việc kết hợp những mô hình xếp hạng là với rất nhiều mô hình được đưa ra làm thế nào để kết hợp những mô hình này tạo ra một mô hình mới hiệu quả hơn và tốt hơn.

Trong khi các nghiên cứu về truy hồi thông tin chưa tìm được giải phát tốt nhất để giải quyết các vấn đề trên thì học máy đã chứng đỏ rằng là một mô hình mới hiệu quả hơn trong việc tự điều chỉnh các tham số, kết hợp với nhiều đặc trưng, và tránh tình trạng quá khớp "over-fitting". Do đó rất có triển vọng để áp dụng các công nghệ học máy cho các vấn đề xếp hạng nói trên.

3.1 Nền tảng cơ sở của học máy

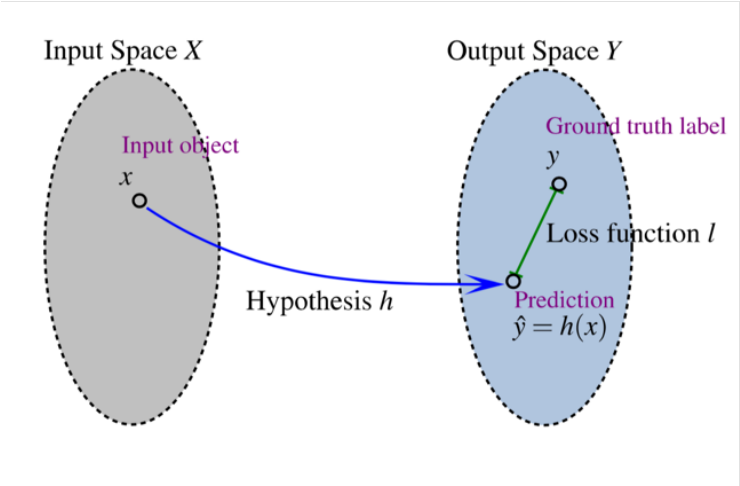
Có nhất nhiều các nghiên cứu đã chú ý đến các thành phần quan trọng sau.

- Không gian đầu vào (input space) chứa các đối tượng cần nghiên cứu, Thông thường các đối tượng này được đại diện bằng các vecto đặc trưng được trích xuất từ dữ liệu ban đầu.

- Không gian đầu ra (output space), chứa các mô hình dữ liệu được tính toán từ dữ liệu đầu vào. Có hai thứ liên quan nhưng lại khác nhau về định nghĩa trong không gian đầu ra trong học máy. Đầu tiên là không gian đầu ra của một chức năng thì lại phụ thuộc vào ứng dụng. Ví dụ trong hồi quy thì không gian đầu ra là không gian các số thực R; trong phân lớp là tập hợp lớp riêng biệt $\{1, 2, \ldots, K\}$. Thứ 2 là không gian đầu ra được thiết kế để thuật tiện cho quá trình học máy. Điều này khiến các chức năng học không giống nhau về không gian đầu ra. Ví dụ như khi sử dụng phương pháp hồi quy để giải quyết vấn đề phân lớp, không gian đầu ra làm cho thuận tiện cho việc học là tập hợp các số thực chứ không phải là tập hợp các lớp riêng biệt.
- Không gian theo giả thuyết (Hypothesis) định nghĩa các hàm chuyển đổi giữa đầu vào và đầu ra. Điều đó có ý nghĩa là những chức năng này hoạt động trên những vector đặc trưng của không gian đầu vào và dự đoán theo định dạng của không gian đầu ra.
- Để học máy tối ưu trên không gian giả thuyết, một bộ huấn luyện sẽ được sử dụng, nó chứa các đối tượng và các nhãn thực sự của nó, bộ huấn luyện sẽ cho dữ liệu đầu

vào và và kết quả của quá trình học máy từ dữ liệu đầu vào. Một hàm chí phí (loss function) đo độ dự đoán được tạo ra bởi không gian giả thuyết với nhãn thực sự. Ví dụ hàm chi phí cho các mô hình phân lớp như lũy thừa, logistic. Hàm chí phí cũng có thể được coi như vai trò trung tâm trong học máy, vì nó cho ta biết mô hình dự đoán là chính xác hay không và có hiệu quả hay không.

Có thể nhìn thấy sự liên quan giữa 4 thành phần trong hình dưới đây

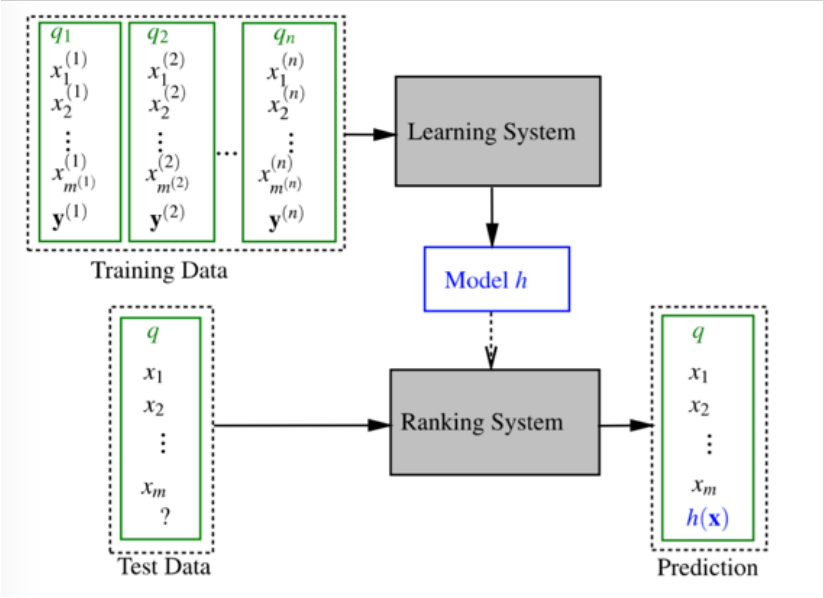


Hình 3-1 Nền tảng cơ sở của học máy [24]

3.2 Nền tảng cơ sở của học máy xếp hạng.

Hình 3-2 biểu diễn luồng dữ liệu của một hệ thống học máy xếp hạng điển hình. Từ hình chúng ta có thể nhìn thấy rằng học máy xếp hạng là một loại học máy có giám sát với một tập dữ liệu huấn luyện. Tạo ra một bộ huấn luyện cũng giống như một bộ đánh giá. Ví dụ, một bộ huấn luyện điển hình bao gồm n truy vấn huấn luyện q_i (i = 1, . . . , n), chúng liên kết với các tài liệu được đại diện bởi vector đặc trưng $s^{(i)} = \{x_j^{(i)}\}_{j=1}^{m^{(i)}}$ trong đó $m^{(i)}$ là số lượng các tài liệu liên quan đến quy vấn q_i và điểm số đánh giá liên quan. Sau đó một mô hình học máy xếp hạng cụ thể được cài đặt từ bộ dữ liệu huấn luyện ban đầu để cho ra danh sách các tài liệu được sắp xếp theo độ ưu tiên càng chính xác càng tốt, để có để biết kết quả của mô hình ta sử dụng hàm chi phí để so sánh độ chính xác. Trong giai đoạn kiểm tra các truy vấn mới được đưa vào mô mình đã được huấn luyện để sắp xếp các tài liệu và trả về các danh sách xếp hạng tương ứng với truy vấn.

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



Hình 3-2 Nền tảng cơ sở của học máy xếp hạng[24]

3. 2. 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], ...

3.2.2 Hướng tiếp cận Pairwise

Có *D = {(xi, xj)}* 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):

$$\forall(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{S}\ \boldsymbol{có}\ \boldsymbol{x_i} > \boldsymbol{x_j}\ \boldsymbol{thì}\ \boldsymbol{r(x_i)} > \boldsymbol{r(x_j)} \qquad (3.1)$$

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

3. 2. 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

3.3 Tổng kết chương

Chương này đã giới thiệu chung nền tảng cơ sở về học máy xếp hạng. Đồng thời cũng nêu ra hướng tiếp cận học máy xếp hạng là ba cách tiếp cận Pointwise, Pairwise, ListWise. Luận văn sẽ sử dụng cách tiếp cận ListWise, chương sau giới thiệu về cách triển khai hướng tiếp cận này và đưa ra mô hình xếp hạng và tính toán song song cho máy tìm kiếm phim tại Cốc Cốc.

Chương 4.Giải pháp xếp hạng và tính toán song song trên nền apache spark

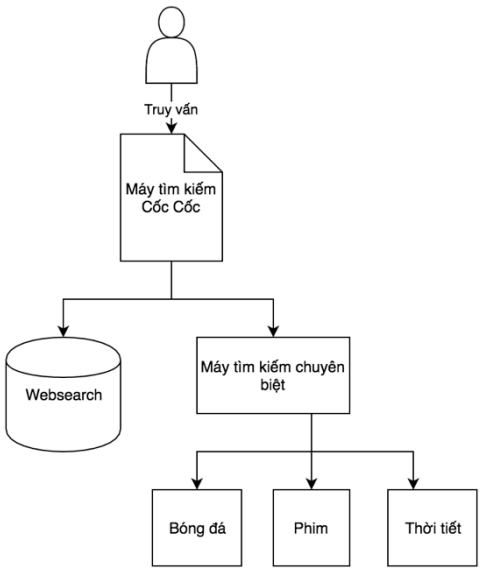
Trong chương này, khóa luận trình bày chi tiết về mô hình hệ thống tìm kiếm và xếp hạng phim ảnh sử dụng tính toán song song trên nền tảng Apache Spark. Đây là hệ thống sẽ trích xuất thông tin phim từ những trang web được crawler của Cốc Cốc tải về. Dữ liệu phim ảnh được lấy bóc tách bao gồm thông tin phim, đạo diễn, năm sản xuất, v.v. Những thông tin này sẽ được sử dụng chuyên biệt cho hệ thống tìm kiếm phim trong máy tìm kiếm **Cốc Cốc**

4.1 Bài toán đặt ra

Ban đầu do thiết kế hệ thống ban đầu hệ thống tìm kiếm phim trực tuyến được thiết kế và tính toán cho một máy chủ tính toán, với mô hình thiết kế này hệ thống có thể đáp ứng tốt trong thời gian đầu. Nhưng do dữ liệu ngày càng lớn để đáp ứng khả năng mở rộng khi cơ sở dữ liệu phim ngày càng lớn cần một mô hình tính toán song song trên nhiều máy tính và tính ổn định chịu lỗi khi nâng cấp hoặc có sự cố trên một máy tính xảy ra. Hơn thế nữa cũng cần một mô hình xếp hạng thống nhất để có thể áp dụng vào các hệ thống tìm kiếm chuyên biệt sau này.

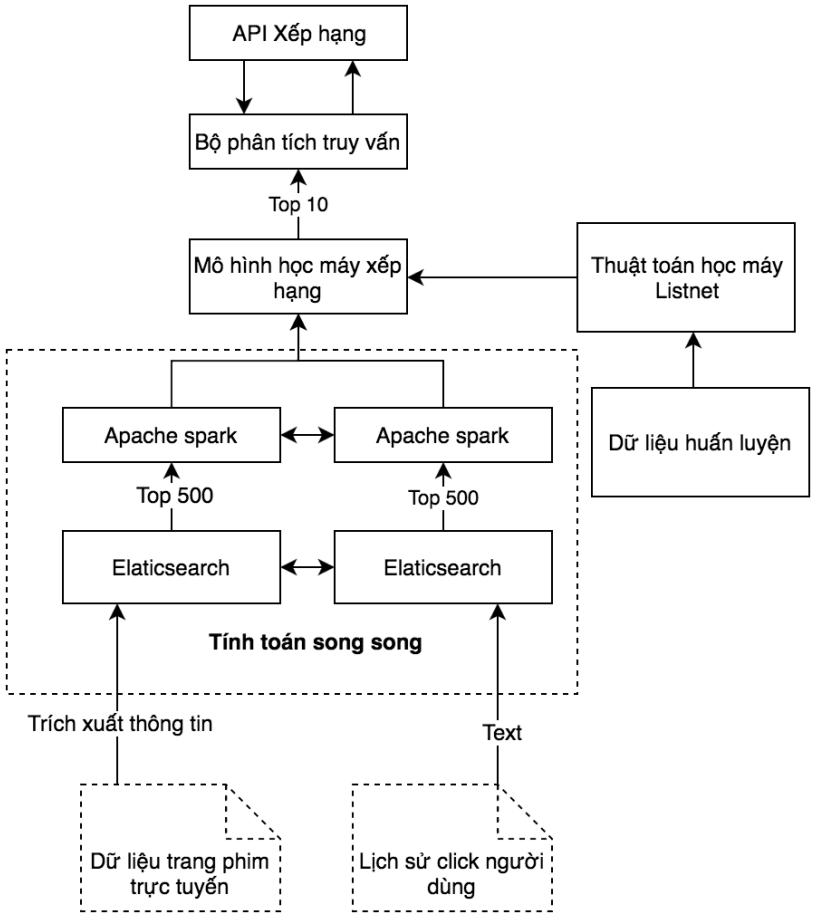
4.2 Mô hình đặt ra

Với bài toán đặt ra ở trên phần này sẽ giới thiệu toàn bộ mô hình từ thu thập dữ liệu, huấn luyện mô hình, và phục vụ tìm kiếm các bộ phim cho hệ thống tìm kiếm tại **Cốc Cốc**. Với mô hình hiện tại của máy tìm kiếm Cốc khi có truy vấn của người dùng thì truy vấn sẽ được gửi đi song song, một tới thành phần tìm kiếm trang web, hai là tới máy tìm kiếm chuyên biệt. Với mỗi máy tìm kiếm chuyên biệt bao gồm nhiều loại truy vấn khác nhau.



Hình 4-1 Cấu trúc thành phần máy tìm kiếm tại Cốc Cốc

Sau đây là mô hình mới được đưa ra cho thành phần tìm kiếm chuyên biệt cho phim ảnh.



Hình 4-2 Mô hình giải pháp xếp hạng và tính toán song song

Mồ hình này sẽ gồm ba thành phần lớn:

Thành phần thu thập dữ liệu: Dữ liệu sẽ được thu thập từ hệ thống crawl của Cốc Cốc từ các domain phim sẽ được trích xuất nội dung và được đánh chỉ mục vào hệ thống tìm kiếm full-tex. Tại hệ thống này chúng ta trích xuất các thông tin nội dung phim từ trang imdb, điểm imdb sẽ là một phần trong vector đặc trưng được sử dụng trong quá trình học máy xếp hạng.

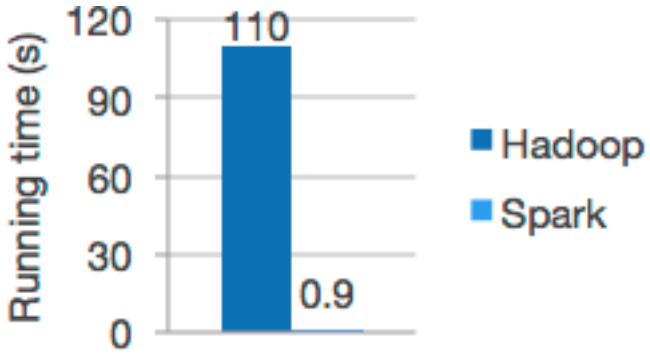
Thành phần lưu trữ và tìm kiếm full-text: Thành phần này sử dụng Elaticsearch và Apache Spark để tìm kiếm full-text search cho toàn bộ dữ liệu thu thập như tiêu đề phim thông tin tác giả diễn viên, nội dung sau đó trên mỗi máy chủ tính toán sẽ thu thập 500 bản ghi liên quan tới truy vấn, Mô hình sử dụng Elasticsearch và Apache Spark đã được sử dụng rộng rãi trong xử lý tính toán song song trong nhiều ngiên cứu tìm kiếm và phân tích dữ liệu lớn như các bài báo [2][3][4]

Thành phần học máy xếp hạng: Đây là thành phần đóng vai trò trung tâm của hệ thống chịu trách nhiệm giữa người dùng và hệ thống xếp hạng tính toán song song.

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 cạnh đó hệ thống cần phải chạy theo thời gian thực, có khả năng mở rộng và khả năng chịu lỗi. Sau đây là các công nghệ đã được sử dụng trong hệ thống này.

4.3 Apache Spark

Ngày nay có rất nhiều hệ thống xử lý dữ liệu thông tin đang sử dụng Hadoop rộng rãi để phân tích dữ liệu lớn. Ưu điểm lớn nhất của **Hadoop** là được dựa trên một mô hình lập trình song song với xử lý dữ liệu lớn là **MapReduce**, mô hình này cho phép khả năng tính toán có thể mở rộng, linh hoạt, khả năng chịu lỗi, chi phí rẻ. Điều này cho phép tăng tốc thời gian xử lý các dữ liệu lớn nhằm duy trì tốc độ, giảm thời gian chờ đợi khi dữ liệu ngày càng lớn. Hadoop đã được nền tảng tính toán cho rất nhiều cho một bài toàn xử lý dữ liệu lớn [9] và các vấn đề về mở rộng tính toán song song trong các bài toàn xếp hạng [7]. **Apache Haddop** cũng được sử dụng tại rất nhiều công ty lớn như **Yahoo**, **Google** và tại **Cốc Cốc** cũng đang sử dụng **Apache Hadop** để lưu trữ cho hệ thống crawler. Dù có rất nhiều điểm mạnh về khả năng tính toán song song và khả năng chịu lỗi cao nhưng **Apache Haddop** có một nhược điểm là tất cả các thao tác đều phải thực hiện trên ổ đĩa cứng điều này đã làm giảm tốc độ tính toán đi gấp nhiều lần. Để khắc phục được nhược điểm này thì **Apache Spark** được ra đời. **Apache Spark** có thể chạy nhanh hơn 10 lần so với **Haddop** ở trên đĩa cứng và 100 lần khi chạy trên bộ nhớ **RAM** [8], hình dưới biểu thị thời gian chạy của tính toán hồi quy **Logistic** trên **Haddop** và **Spark**. (Nguồn https://spark.apache.org/)



Hình 4-3 Thời gian chạy của tính toán hồi quy Logistic trên Hadoop và Spark

Apache Spark là một open source cluster computing framework được phát triển sơ khởi vào năm 2009 bởi **AMPLab** tại đại học **California**, **Berkeley**. Sau này, **Spark** đã được trao cho **Apache Software Foundation** vào năm 2013 và được phát triển cho đến nay. **Apache Spark** được phát triển nhằm tăng tốc khả năng tính toán xử lý của **Haddop**.

Spark cho phép xây dựng và phân tích nhanh các mô hình dự đoán. Hơn nữa, nó còn cung cấp khả năng truy xuất toàn bộ dữ liệu cùng lúc, nhờ vậy ta không cần phải lấy mẫu dữ liệu đòi

hỏi bởi các ngôn ngữ lập trình như R. Thêm vào đó, **Spark** còn cung cấp tính năng streaming, được dùng để xây dựng các mô hình real-time bằng cách nạp toàn bộ dữ liệu vào bộ nhớ [10].

Khi ta có một tác vụ nào đó quá lớn mà không thể xử lý trên một laptop hay một server, Spark cho phép ta phân chia tác vụ này thành những phần dễ quản lý hơn. Sau đó, Spark sẽ chạy các tác vụ này trong bộ nhớ, trên các cluster của nhiều server khác nhau để khai thác tốc độ truy xuất nhanh từ RAM. Spark sử dụng API Resilient Distributed Dataset (RDD) để xử lý dữ liệu.

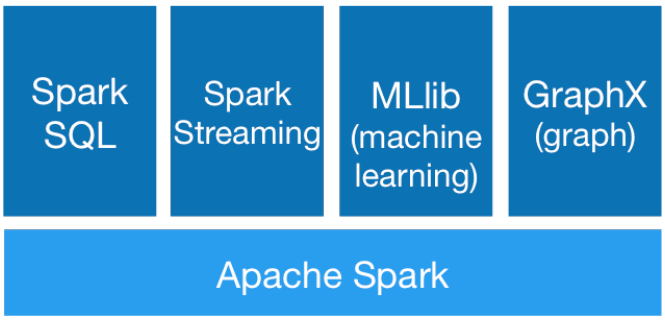
Spark nhận được nhiều sự hưởng ứng từ cộng đồng Big data trên thế giới do cung cấp khả năng tính toán nhanh và nhiều thư viện hữu ích đi kèm như Spark SQL (với kiểu dữ liệu DataFrames), Spark Streaming, MLlib (machine learning: classification, regression, clustering, collaborative filtering, và dimensionality reduction) và GraphX (tính toán song song trên dữ liệu đồ thị).

4.3.1 Tính năng của Apache Spark

Apache Spark có các tính năng đặc trưng sau đây.

- **Tốc độ**: **Spark** có thể chạy trên cụm Hadoop và có thể chạy nhanh hơn 100 lần khi chạy trên bộ nhớ **RAM**, và nhanh hơn 10 lần khi chạy trên ổ cứng. Bằng việc giảm số thao tác đọc nghi lên đĩa cứng. Nó lưu trữ trực tiếp dữ liệu xử lý lên bộ nhớ
- **Hỗ trợ đa ngôn ngữ**: **Spark** cung cấp các **API** có sẵn cho các ngôn ngữ **Java**, **Scala**, hoặc **Python**. Do đó, bạn có thể viết các ứng dụng bằng nhiều các ngôn ngữ khác nhau. **Spark** đi kèm 80 truy vấn tương tác mức cao.
- **Phân tích nâng cao**: Spark không chỉ hỗ trợ '**Map'** và '**Reduce'**. Nó còn hỗ trợ truy vấn **SQL**, xử lý theo **Stream**, học máy, và các thuật toán đồ thị (**Graph**)

4.3.2 Các thành phần của Apache Spark



Hình 4-4 Các thành phần Apache Spark [25]

- **Apache Spark Core**: Spark Core là thành phần cốt lõi thực thi cho tác vụ cơ bản làm nền tảng cho các chức năng khác. Nó cung cấp khả năng tính toán trên bộ nhớ và datase trong bộ nhớ hệ thống lưu trữ ngoài.
- **Spark SQL**: Là một thành phần nằm trên **Spark Core** nó cung cấp một sự ảo hóa mới cho dữ liệu là **SchemaRDD**, hỗ trợ các dữ liệu có cấu trúc và bán cấu trúc.
- **Spark Streaming**: Cho phép thực hiện phân tích xử lý trực tuyến xử lý theo lô.
- **MLlib (Machine Learning Library)**: **MLlib** là một nền tảng học máy phân tán bên trên Spark do kiến trúc phân tán dựa trên bộ nhớ. Theo các so sánh benchmark **Spark MLlib** nhanh hơn chín lần so với phiên bản chạy trên **Hadoop** (**Apache Mahout**)
- **GrapX:** Grapx là nền tảng xử lý đồ thị dựa trên **Spark**. Nó cung cấp các Api để diễn tả các tính toán trong đồ thị bằng cách sử dụng **Pregel Api**.

4.3.3 Resilient Distributed Datasets

Resilient Distributed Datasets (RDD) là một cấu trúc dữ liệu cơ bản của **Spark**. Nó là một tập hợp bất biến phân tán của một đối tượng. Mỗi dataset trong **RDD** được chia ra thành nhiều phần vùng logical. Có thể được tính toán trên các node khác nhau của một cụm máy chủ (cluster). **RDDs** có thể chứa bất kỳ kiểu dữ liệu nào của **Python**, **Java**, hoặc đối tượng **Scala**, bao gồm các kiểu dữ liệu do người dùng định nghĩa.

Thông thường, **RDD** chỉ cho phép đọc, phân mục tập hợp của các bản ghi. **RDDs** có thể được tạo ra qua điều khiển xác định trên dữ liệu trong bộ nhớ hoặc **RDDs**, **RDD** là một tập hợp có khả năng chịu lỗi mỗi thành phần có thể được tính toán song song.

Có hai cách để tạo **RDDs**

- Tạo từ một tập hợp dữ liệu có sẵn trong ngôn ngữ sử dụng như **Java**, **Python**, **Scala**
- Lấy từ dataset hệ thống lưu trữ bên ngoài như **HDFS**, **Hbase** hoặc các cơ sở dữ liệu quan hệ.

4.4 Elasticsearch

Elasticsearch được phát triển bởi Shay Banon vào năm 2010 và dựa trên Apache Lucene, Elasticsearch được phát hành theo Giấy phép Apache 2.0.

Hình 4-5 Logo của Elasticsearch

Elasticsearch là hệ thống phân tán theo thời gian thực là một hệ thống tìm kiếm full-text, phân tích mã nguồn mở. Có thể sử dụng thông qua **RESTfull** sử dụng **JSON** (**JavaScript Object Notation)** để chứa dữ liệu. Được viết bằng ngôn ngữ Java điều này cho phép Elasticsearch có thể chạy trên nhiều nền tảng khác nhau. Cho phép người sử dụng truy vấn dữ liệu lớn với tốc độ cao.

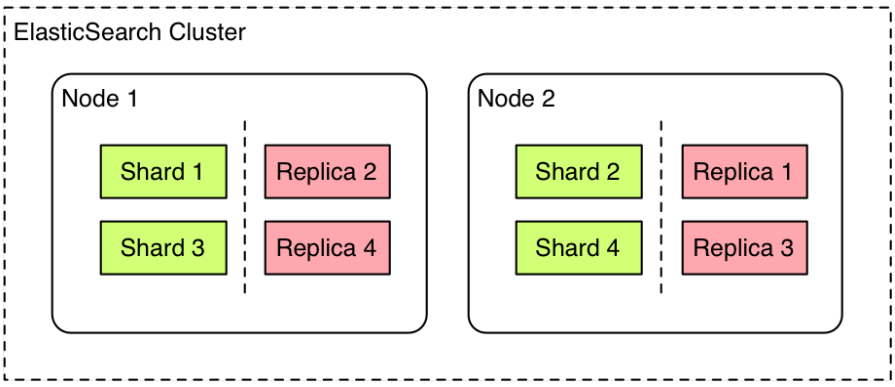
4.4.1 Tính năng tổng quát

- **Elasticsearch** có thể được mở rộng lên đến Petabyte dữ liệu có cấu trúc và không có cấu trúc
- **Elasticsearch** có thể được sử dụng như một thay thế cho các lưu trữ tài liệu như **MongoDb** hay **RavenDb**.
- **Elasticsearch** được sử dụng để cải thiện hiệu năng tìm kiếm, đặc biệt là tìm kiếm full-text.
- **Elasticsearch** là một máy tìm kiếm phổ biến nhất được sử dụng bởi nhiều tổ chức lớn như **Wikipedia**, **The Guardian**, **StakOverflow**, **GitHub**, v.v

4.4.2 Khái niệm cơ bản

- **Nút (Node)**: Nó là thể hiện của một chương trình chạy độc lập của **Elasticsearch**. Một máy chủ vật lý và máy chủ ảo có thể chứa nhiều node phụ thuộc vào tài nguyên của chúng như RAM, CPU và bộ nhớ ngoài
- **Cụm (Cluster)**: Là tập hợp của một hoặc nhiều node. Cluster cung cấp tập hợp khả năng đánh chỉ mục và tìm kiếm xuyên qua toàn bộ node cho toàn bộ dữ liệu.
- **Chỉ mục (Index)**: Là tập hợp các kiểu dữ liệu khác nhau của tài liệu và thuộc tính tài liệu. Index cũng sử dụng khái niệm shard để cải thiện hiệu năng tính toán. Ví dụ như một bộ các tài liệu chứa dữ liệu cho mạng xã hội.
- **Type/Mapping**: Nó là tập hợp cái tài liệu mô tả các miêu tả của trường dữ liệu trong cùng index. Ví dụ một Index chứa dữ liệu cho ứng dụng mạng xã hội, và có các kiểu dữ liệu cụ thể cho dữ liệu thông tin người dùng, dữ liệu tin nhắn, bình luận.
- **(Tài liệu) Document**: Là một tập hợp các trường dữ liệu được xác định cụ thể trong định dạng JSON. Mỗi tài liệu thuộc về một Type và lưu trữ bên trong mỗi index. Mỗi tài liệu được liên kết với một định dạng duy nhất là UID.
- **Shard**: Các Index được mở rộng theo chiều ngang bằng cách chia thành nhiều shards. Điều này nghĩa là mỗi Shard chứa tất cả các thuộc thuộc tính của một Document, nhưng chứa ít đối tượng JSON hơn Index. Việc phân chia theo chiều ngang làm shard là một node độc lập và có thể chứa trên bất kỳ node nào. Shard chính là phần gốc của mỗi phần phân chia và những phần này được nhân bản.

- **Replicas**: **Elasticsearch** cho phép người sử dụng tạo nhiều nhân bản với các index và shard. Sự nhân bản này không chỉ giúp tăng sẵn sàng cho dữ liệu trong trường hợp xảy ra lỗi mà còn nâng cao hiệu năng tìm kiếm bằng tìm kiếm song song trong những phần nhân bản này



Hình 4-6 Minh họa một Cluster trong Elasticsearch

4.4.3 Ưu điểm của Elasticsearch

- **Elasticsearch** được phát triển trên Java điều này cho phép nó có thể chạy trên hầu hết mọi nền tảng
- **Elasticsearch** có thể hoạt động một cách trực tuyến, nghĩa là việc thêm tài liệu được cập nhập và tìm kiếm ngay lập tức.
- Xử lý đa người sử dụng trong **Elasticsearch** là dễ dàng hơn so với **Apache Solr**.
- **Elasticsearch** sử dụng định dạng JSON cho truy vấn và kết quả trả về do đó dễ dàng gọi Elasticsearch từ nhiều ngôn ngữ lập trình khác nhau.
- **Elasticsearch** hỗ trợ nhiều loại kiểu dữ liệu khác nhau như văn bản, ngày tháng, số thực, số nguyên, địa chỉ IP… và nhiều truy vấn phức tạp.

4.4.4 Nhược điểm của Elasticsearch

- Một nhược điểm cố hữu của Elasticsearch là do Elasticsearch sử dụng cấu trúc của Apache Lucene cho mỗi shard do đó không thể thay đổi số lượng shard khi bạn đã tạo index do đó chúng ta cần tính toán kỹ số shard của mỗi index vì nếu shard nhiều sẽ ảnh hưởng đến hiệu năng ngược lại sẽ làm giảm khả năng mở rộng khi dữ liệu liệu tăng. Do đó cần tính toán kỹ số lượng shard.
- Elasticsearch không hỗ trợ nhiều định dạng trả về khác ngoài JSON không giống như trong Apache Solr nó hỗ trợ các định dạng như CSV, XML và JSON.

4.5 Tính toán song song trên ElasticSearch và Apache Spark

Elasticsearch là một hệ thống tìm kiếm và phân tích trực tuyến có khả năng mở rộng theo chiều ngang, dữ liệu trong Elasticsearch được chia thành các Shard, mỗi Shard là một thành phần độc lập với nhau, mỗi khi có dữ liệu mới được index, đầu tiên dữ liệu sẽ được gửi đến máy Master tại đây máy Master sẽ kiểm tra thông tin về các Shard như kích thước, nơi lưu trữ sau đó sẽ phân bổ tài nguyên giao task chia công việc index ra thành nhiều Shard khác nhau nhằm tăng tốc thời gian đánh chỉ mục cho dữ liệu. Khi có yêu cầu truy vấn trên Elasticsearch truy vấn sẽ được gửi về máy Master sau đó máy này sẽ đưa truy vấn đến tất cá chác Shard chứa tài liệu liên quan đến truy vấn và thực thi tìm kiếm xếp hạng kết quả trả về trên mỗi Shard, sau đó toàn bộ liệu sẽ được hợp nhất lại để đưa ra kết quả cuối cùng.

Apache Spark là một nền tảng tính toán song song nó có rất nhiều module để truy cập đến nhiều nguồn chứa dữ liệu khác nhau như HDFS thao tác dữ liệu trên hệ thống tập tin của Hadoop, Spark SQL là một module cho phép thao tác lấy dữ liệu song song trong các cơ sở dữ liệu. Với Elasticsearch thì có module Apache Spark connector giúp cho phép tạo và sử dụng đối tượng Spark Context có thể đánh chỉ mục và truy vấn dữ liệu song song bộ thự viện này được cung cấp thông qua phần mở rộng là elasticsearch-hadoop. Trong mô hình này Apache sẽ tìm thông tin phim dựa trên truy vấn của người dùng tại mỗi máy đơn lẻ, quá trình này hoàn toàn được thực hiện song song. Với mỗi máy đơn lẻ Apache Spark sẽ lấy 100 truy vấn liên quan nhất, sau đó 100 truy vấn này sẽ được đưa vào học máy lúc này sẽ có 100 truy vấn được sắp xếp trên mỗi máy, khi tất cả các máy chủ tính toán song lúc này Apache Spark sẽ tiến hành tổng hợp kết quả trên các máy chủ đơn và đưa ra danh sách cuối cùng được xếp hạng theo độ liên quan giữa truy vấn và tài liệu.

Như vậy mô hình sẽ được tính toán song song dựa trên 2 quá trình

- Tìm kiếm song song trên Elasticsearch
- Tính toán, áp dụng mô hình học máy xếp hạng song song với Apache Spark.

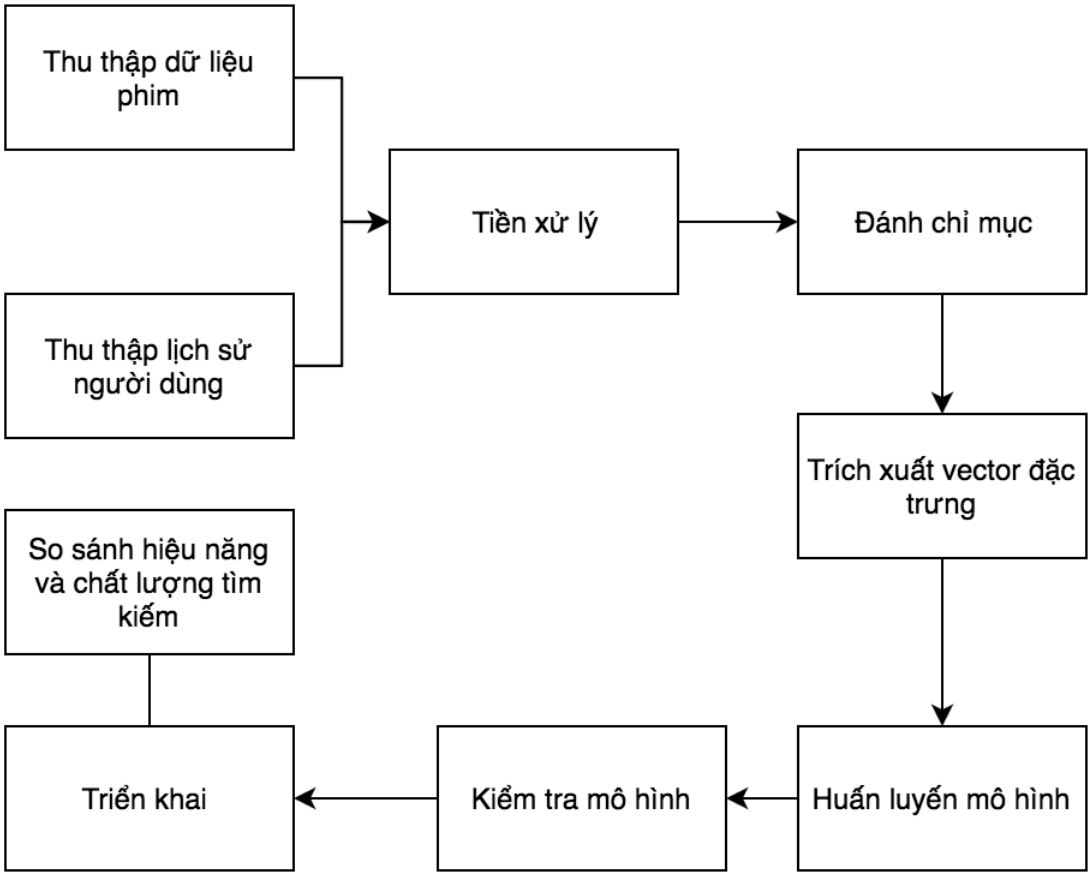
4.6 Tổng kết chương

Chương này đã đưa ra giải pháp xếp hạng và tính toán song song cho máy tìm kiếm thông tin phim ảnh dựa trên giải pháp tính toán song song thông qua nền tảng Apache Spark và công cụ quản trị dữ liệu Elasticsearch. Từ mô hình giải pháp này, chúng tôi sẽ tiến hành vận dụng, xây dựng hệ thống thử nghiệm và tiến hành đánh giá ở chương sau.

Chương 5.Thực nghiệm và đánh giá

5.1 Mô hình thực nghiệm

Đây là mô hình thực nghiệm đã được áp dụng vào thực tế trong hệ thống tìm kiếm phim trực tuyến riêng biệt tại **Cốc Cốc**. Tất cả các quá trình toàn bộ được thu thập từ dữ liệu và lịch sử người dùng sử dụng của các dịch vụ tại đây. Sau đây là chi tiết mô hình giải pháp xếp hạng và tính toán song song trên nền tảng Apache Spark được đưa ra ở trên, sau đây tôi xin được đưa ra mô hình thực nghiệm và đã được sử dụng trên hệ thống tìm kiếm phim trực tuyến tại **Cốc Cốc.**



Hình 5-1 Mô hình thực nghiệm

Toàn bộ quá trình thực nghiệm được chia thành các giai đoạn như sau:

- Bước đầu tiên là thu thập dữ liệu người dùng từ lịch sử click và các dữ liệu phim được bóc tách.
- Bước thứ hai là tiền xử lý dữ liệu bao gồm chuẩn hóa dữ liệu và đánh chỉ mục cho toàn bộ dữ liệu vào Elasticsearch, và trích xuất các vector đặc trưng làm dữ liệu đầu vào cho phần huấn luyện mô hình.

- Bước thứ 3 là huấn luyện mô hình bước này sẽ sử dụng thuật toán Listnet sử dụng các vector đặc trưng được trích xuất ở phần trên.
- Bước thứ 4 là triển khai mô hình vào máy tìm kiếm phim tại **Cốc Cốc**.

Chi tiết các bước và môi trường thực hiện sẽ được miêu tả chi tiết hơn ở phần sau.

5.2 Môi trường thực nghiệm

5.2.1 Hạ tầng tính toán

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

Bảng 5-1 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

5.2.2 Các công cụ được sử dụng

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

Bảng 5-2 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
	Elasticsearch-Jdbc	https://github.com/jprante/elasticsearch-jdbc	2.3.4.1
	Jsoup	https://jsoup.org/	1.10.1

5.3 Thực nghiệm

Quá trình thực nghiệm học máy xếp hạng gồm các bước chính sau đây:

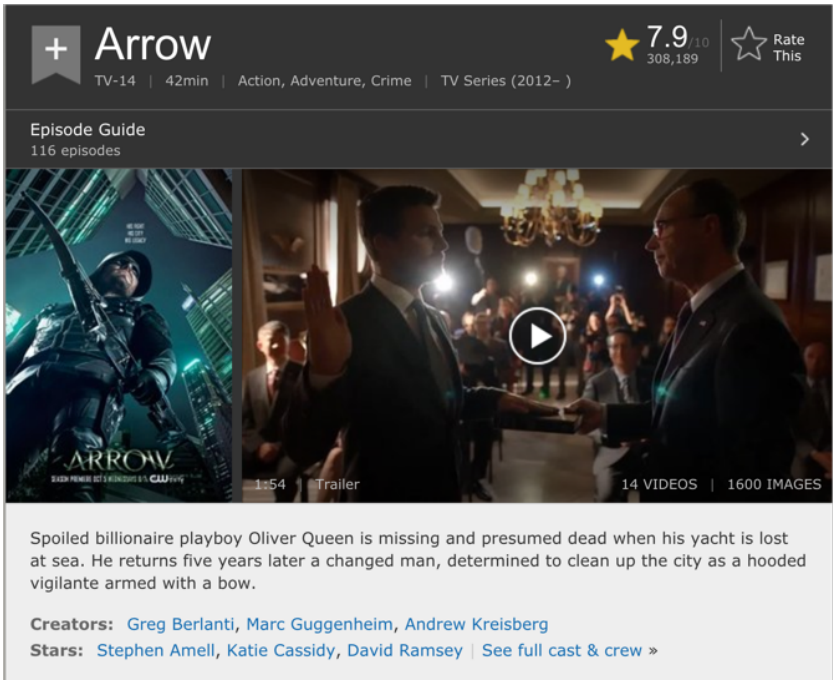
- Thu thập dữ liệu: thu thập toàn bộ dữ liệu về phim và dữ liệu lịch sử của người dùng trong hệ thống tìm kiếm **Cốc Cốc**
- Xử lý dữ liệu: tiền xử lý dữ liệu, đánh chỉ mục cho dữ liệu, xây dựng tập tài liệu học cho mô hình, véc tơ hóa dữ liệu.
- Xây dựng hàm xếp hạng: tiến hành training trên tập dữ liệu đã có bằng thuật toán ListNet trong tự viện RankLib 2.7

5.3.1 Thu thập dữ liệu phim

Tất các các 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** hệ thống được chạy hàng ngày ngay khi có tất cả các dữ liệu thêm mới, bộ phân tích sẽ tự động bóc tách (sử dụng Jsoup để bóc tách dữ liệu html đây là công cụ cho phép dùng cú pháp css để chọn các thẻ và thuộc tính html) và lưu trữ vào cơ sở dữ liệu.

a. Thu thập dữ liệu phim IMDb

Đầu tiên là hệ thống sẽ trích xuất các thông tin từ trang web đánh giá phim IMDb (Internet Movie Database). Dưới đây là thông tin một bộ phim được trích xuất từ imdb. (Nguồn http://www.imdb.com/title/tt2193021/)



Hình 5-2 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…

Các thông tin trên trang được trích xuất trên trang IMDb bao gồm

Tên phim, năm sản xuất

Đạo diễn, diễn viên

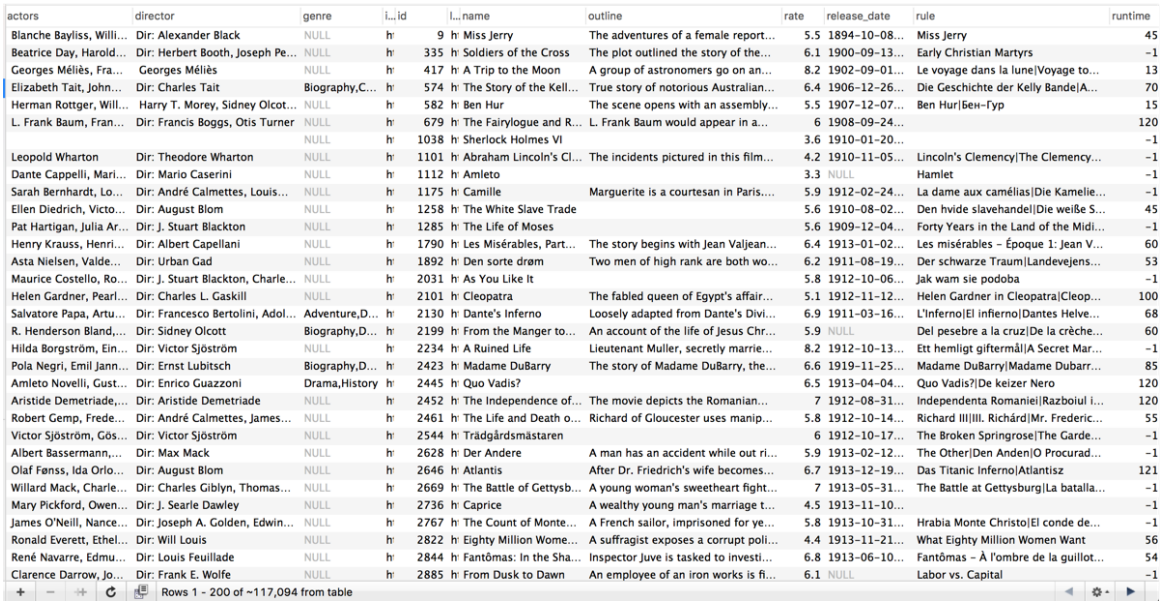
Nội dung phim, thể loại, điểm số rating.

Bước này thu thập được **117.094** thông tin phim IMDb dữ liệu ban đầu được chứa vào cơ sở dữ liệu MySQL, và được chứa theo định dạng sau.

Bảng 5-3 Định dạng trường dữ liệu thông tin phim IMDb trong cơ sở dữ liệu

Tên trường	Miêu tả
id	Định danh của IMDb
director	Đạo diễn
genre	Thể loại
image_link	Poster
link	Link trên IMDb
name	Tên phim
outline	Nội dung
year	Năm
release_date	Ngày phát hành
actor	Diễn viên
runtime	Thời lượng
ratingCount	Tổng số đánh giá
rate	Điểm đánh giá trung bình

Dưới đây là một vài thông tin phim đã thu thập được.



actors	director	genre	i...	id	l...	name	outline	rate	release_date	rule	runtime
Blanche Bayliss, Willi...	Dir: Alexander Black	NULL	h	9	h	Miss Jerry	The adventures of a female report...	5.5	1894-10-08...	Miss Jerry	45
Beatrice Day, Harold...	Dir: Herbert Booth, Joseph Pe...	NULL	h	335	h	Soldiers of the Cross	The plot outlined the story of the...	6.1	1900-09-13...	Early Christian Martyrs	-1
Georges Méliès, Fra...	Georges Méliès	NULL	h	417	h	A Trip to the Moon	A group of astronomers go on an...	8.2	1902-09-01...	Le voyage dans la lune|Voyage to...	13
Elizabeth Tait, John...	Dir: Charles Tait	Biography,C...	h	574	h	The Story of the Kell...	True story of notorious Australian...	6.4	1906-12-26...	Die Geschichte der Kelly Bande|A...	70
Herman Rottger, Will...	Harry T. Morey, Sidney Olcot...	NULL	h	582	h	Ben Hur	The scene opens with an assembly...	5.5	1907-12-07...	Ben Hur|Бен-Гур	15
L. Frank Baum, Fran...	Dir: Francis Boggs, Otis Turner	NULL	h	679	h	The Fairylogue and R...	L. Frank Baum would appear in a...	6	1908-09-24...		120
		NULL	h	1038	h	Sherlock Holmes VI		3.6	1910-01-20...		-1
Leopold Wharton	Dir: Theodore Wharton	NULL	h	1101	h	Abraham Lincoln's Cl...	The incidents pictured in this film...	4.2	1910-11-05...	Lincoln's Clemency|The Clemency...	-1
Dante Cappelli, Mari...	Dir: Mario Caserini	NULL	h	1112	h	Amleto		3.3	NULL	Hamlet	-1
Sarah Bernhardt, Lo...	Dir: André Calmettes, Louis...	NULL	h	1175	h	Camille	Marguerite is a courtesan in Paris....	5.9	1912-02-24...	La dame aux camélias|Die Kamelie...	-1
Ellen Diedrich, Victo...	Dir: August Blom	NULL	h	1258	h	The White Slave Trade		5.6	1910-08-02...	Den hvide slavehandel|Die weiße S...	45
Pat Hartigan, Julia Ar...	Dir: J. Stuart Blackton	NULL	h	1285	h	The Life of Moses		5.6	1909-12-04...	Forty Years in the Land of the Midi...	-1
Henry Krauss, Henri...	Dir: Albert Capellani	NULL	h	1790	h	Les Misérables, Part...	The story begins with Jean Valjean...	6.4	1913-01-02...	Les misérables - Époque 1: Jean V...	60
Asta Nielsen, Valde...	Dir: Urban Gad	NULL	h	1892	h	Den sorte drøm	Two men of high rank are both wo...	6.2	1911-08-19...	Der schwarze Traum|Landevejens...	53
Maurice Costello, Ro...	Dir: J. Stuart Blackton, Charle...	NULL	h	2031	h	As You Like It		5.8	1912-10-06...	Jak wam sie podoba	-1
Helen Gardner, Pearl...	Dir: Charles L. Gaskill	NULL	h	2101	h	Cleopatra	The fabled queen of Egypt's affair...	5.1	1912-11-12...	Helen Gardner in Cleopatra|Cleop...	100
Salvatore Papa, Artu...	Dir: Francesco Bertolini, Adol...	Adventure,D...	h	2130	h	Dante's Inferno	Loosely adapted from Dante's Divi...	6.9	1911-03-16...	L'Inferno|El infierno|Dantes Helve...	68
R. Henderson Bland,...	Dir: Sidney Olcott	Biography,D...	h	2199	h	From the Manger to...	An account of the life of Jesus Chr...	5.9	NULL	Del pesebre a la cruz|De la crèche...	60
Hilda Borgström, Ein...	Dir: Victor Sjöström	NULL	h	2234	h	A Ruined Life	Lieutenant Muller, secretly marrie...	8.2	1912-10-13...	Ett hemligt giftermål|A Secret Mar...	-1
Pola Negri, Emil Jann...	Dir: Ernst Lubitsch	Biography,D...	h	2423	h	Madame DuBarry	The story of Madame DuBarry, the...	6.6	1919-11-25...	Madame DuBarry|Madame Dubarr...	85
Amleto Novelli, Gust...	Dir: Enrico Guazzoni	Drama,History	h	2445	h	Quo Vadis?		6.5	1913-04-04...	Quo Vadis?|De keizer Nero	120
Aristide Demetriade,...	Dir: Aristide Demetriade	NULL	h	2452	h	The Independence of...	The movie depicts the Romanian...	7	1912-08-31...	Independenta Romaniei|Razboiul i...	120
Robert Gemp, Frede...	Dir: André Calmettes, James...	NULL	h	2461	h	The Life and Death o...	Richard of Gloucester uses manip...	5.8	1912-10-14...	Richard III|III. Richárd|Mr. Frederic...	55
Victor Sjöström, Gös...	Dir: Victor Sjöström	NULL	h	2544	h	Trädgårdsmästaren		6	1912-10-17...	The Broken Springrose|The Garde...	-1
Albert Bassermann,...	Dir: Max Mack	NULL	h	2628	h	Der Andere	A man has an accident while out ri...	5.9	1913-02-12...	The Other|Den Anden|O Procurad...	-1
Olaf Fønss, Ida Orlo...	Dir: August Blom	NULL	h	2646	h	Atlantis	After Dr. Friedrich's wife becomes...	6.7	1913-12-19...	Das Titanic Inferno|Atlantisz	121
Willard Mack, Charle...	Dir: Charles Giblyn, Thomas...	NULL	h	2669	h	The Battle of Gettysb...	A young woman's sweetheart fight...	7	1913-05-31...	The Battle at Gettysburg|La batalla...	-1
Mary Pickford, Owen...	Dir: J. Searle Dawley	NULL	h	2736	h	Caprice	A wealthy young man's marriage t...	4.5	1913-11-10...		-1
James O'Neill, Nance...	Dir: Joseph A. Golden, Edwin...	NULL	h	2767	h	The Count of Monte...	A French sailor, imprisoned for ye...	5.8	1913-10-31...	Hrabia Monte Christo|El conde de...	-1
Ronald Everett, Ethel...	Dir: Will Louis	NULL	h	2822	h	Eighty Million Wome...	A suffragist exposes a corrupt poli...	4.4	1913-11-21...	What Eighty Million Women Want	56
René Navarre, Edmu...	Dir: Louis Feuillade	NULL	h	2844	h	Fantômas: In the Sha...	Inspector Juve is tasked to investi...	6.8	1913-06-10...	Fantômas - À l'ombre de la guillot...	54
Clarence Darrow, Jo...	Dir: Frank E. Wolfe	NULL	h	2885	h	From Dusk to Dawn	An employee of an iron works is fi...	6.1	NULL	Labor vs. Capital	-1

Rows 1 - 200 of ~117,094 from table

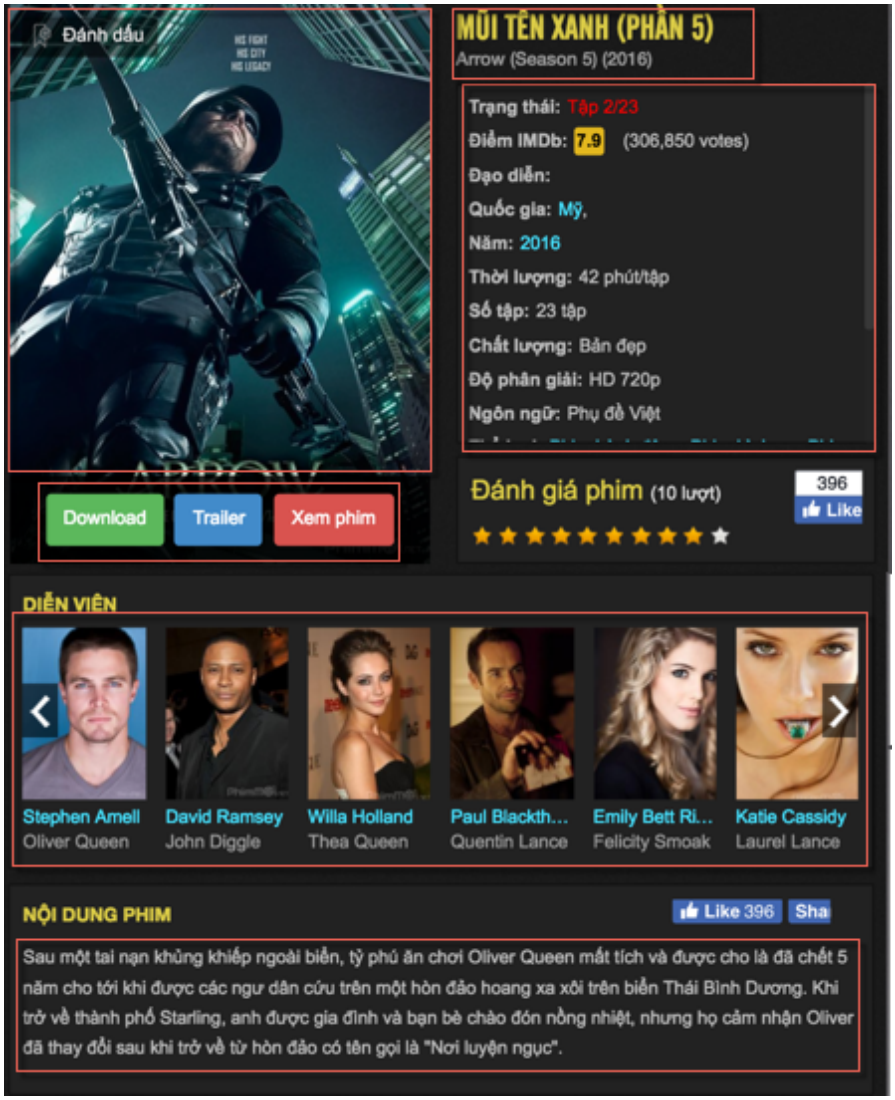
Hình 5-3 Dữ liệu IMDb trong cơ sở dữ liệu Mysql.

b. Thu thập dữ liệu trên trang chiếu phim trực tuyến

Các dữ liệu trên trang chiếu phim trực tuyến sẽ được trích xuất hàng ngày do hệ thống crawler của **Cốc Cốc** thu thập về từ các domain sau đây.

- http://phim3s.net/
- http://hayhaytv.vn/
- http://phim14.net/
- http://hdviet.com/
- http://www.phimmoi.net/
- http://hdonline.vn/
- http://bomtan.org/

Thông tin về bộ phim được bóc tách từ HTML của các trang bên trên, dưới đây phần khoanh đỏ là thông tin phim được bóc tách của trang "http://www.phimmoi.net/phim/mui-ten-xanh-phan-5-4268/"



Hình 5-4 Dữ liệu thông tin phim trên trang phimmoi.net

Dữ liệu thông tin thu thập về được lưu trữ vào cơ sở dữ liệu MySQL theo bảng dưới đây

Bảng 5-4 Định dạng trường dữ liệu dữ liệu phim trực tuyến trong cơ sở dữ liệu

Tên trường	Miêu tả
id	Định danh
director	Đạo diễn
genre	Thể loại
image_link	Poster
imdb_id	Đinh danh IMDb

outline	Nội dung
year	Năm
release_date	Ngày phát hành
actor	Diễn viên
runtime	Thời lượng
nameVn	Tên phim tiếng việt
nameEn	Tên phim tiếng anh

Bước này thu thập được **213.253** dữ liệu mẫu cho phim online và được mô tả dưới đây

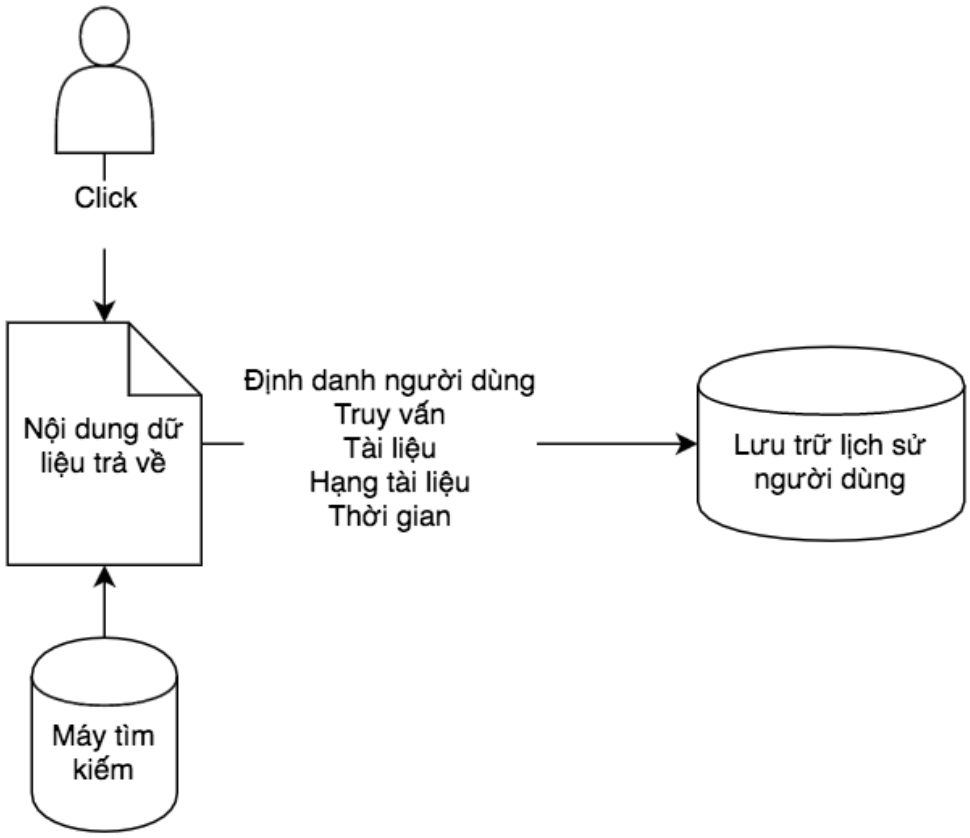
id	actors	c..	director	genre	imageLink	imdbId	l...	l...	nameVn
97655	Lục Nghị, Tào Cách, Dương Uy, N...	2		Phim Hài Hước, Phim Phiêu Lưu, P...	http://cdn.phim3s.net/images/fil...	0	N	h	[TQ] Bố Ơi Mình Đi Đâu Thế Mùa 2
109758	Lâm Chí Dĩnh, Điền Lượng, Quách...	2		Phim Hài Hước, Phim TV Show	http://cdn.phim3s.net/images/fil...	0	N	h	[TQ] Bố Ơi, Mình Đi Đâu Thế
147874	Tom Pelphrey,Natalie Hall,Method...	2	Brendan Gabriel Murphy	Phim Hài Hước	http://static.hdonline.vn/i/resour...	2538204	N	h	#Lucky Number
127055	Pierce Brosnan,Halle Berry,Rosamu...	2	Lee Tamahori	Phim Hành Động	http://static.hdonline.vn/i/resour...	246460	N	h	007 Chết Vào Một Ngày Khác
127039	Roger Moore,Yaphet Kotto,Jane Se...	2	Guy Hamilton	Phim Hành Động, Phim Phiêu Lưu	http://static.hdonline.vn/i/resour...	70328	N	h	007 Cuộc Chiến Trên Đảo Thuốc P..
126674	Daniel Craig,Gemma Arterton,Jeffr...	2	Marc Forster	Phim Hành Động, Phim Hình Sự	http://static.hdonline.vn/i/resour...	830515	N	h	007 Định Mức Khuấy Khoả
127048	Timothy Dalton,Robert Davi,Carey...	2	John Glen	Phim Hành Động	http://static.hdonline.vn/i/resour...	97742	N	h	007 Lệnh Hành Quyết
127054	Sean Connery,Gert Fröbe,Honor Bl...	2	Guy Hamilton	Phim Hành Động	http://static.hdonline.vn/i/resour...	58150	N	h	007 Ngón Tay Vàng
127057	Roger Moore,Lois Chiles,Michael L...	2	Lewis Gilbert	Phim Hành Động, Phim Hình Sự	http://static.hdonline.vn/i/resour...	79574	N	h	007 Người Đi Tìm Mặt Trăng
157229	Sean Connery,Claudine Auger,Ado...	2	Terence Young	Phim Hành Động	http://static.hdonline.vn/i/resour...	59800	N	h	007 Quả Cầu Sấm Sét
127046	Roger Moore,Carole Bouquet,Topol	2	John Glen	Phim Hành Động	http://static.hdonline.vn/i/resour...	82398	N	h	007 Riêng Cho Đôi Mắt Em
127041	Daniel Craig,Eva Green,Judi Dench...	2	Martin Campbell	Phim Hành Động, Phim Phiêu Lưu,...	http://static.hdonline.vn/i/resour...	381061	N	h	007 Sòng Bạc Hoàng Gia
127049	Pierce Brosnan,Sophie Marceau	2	Michael Apted	Phim Hành Động, Phim Hình Sự	http://static.hdonline.vn/i/resour...	143145	N	h	007 Thế Giới Không Đủ
127050	Sean Connery,Daniela Bianchi,Lois...	2	Terence Young	Phim Hành Động	http://static.hdonline.vn/i/resour...	57076	N	h	007 Tình Yêu Đến Từ Nước Nga
117720	Daniel Craig, Javier Bardem, Naom...	2	Sam Mendes	Phim Hành Động	http://cdn.phim3s.net/images/fil...	1074638	N	h	007 Tử Địa Skyfall
127044	Roger Moore,Maud Adams,Louis J...	2	John Glen	Phim Hành Động, Phim Phiêu Lưu	http://static.hdonline.vn/i/resour...	86034	N	h	007 Vòi Bạch Tuộc
127042	Timothy Dalton,Maryam Dabo,Jero...	2	John Glen	Phim Hành Động	http://static.hdonline.vn/i/resour...	93428	N	h	007: Ánh Sáng Chết Người
119600	Pierce Brosnan,Halle Berry,Rosamu...	2	Lee Tamahori	Phim Hành Động Phiêu Lưu,Phim...	NULL	246460	N	h	007: Hẹn Chết Ngày Khác
123402	Pierce Brosnan,Sean Bean,Izabella...	2	Martin Campbell	Phim Hành Động Phiêu Lưu,Phim...	NULL	113189	N	h	007: Mắt Vàng
127040	Pierce Brosnan,Sean Bean,Izabella...	2	Martin Campbell	Phim Hành Động	http://static.hdonline.vn/i/resour...	113189	N	h	007: Mắt Vàng
122163	Pierce Brosnan,Jonathan Pryce,Mic...	2	Roger Spottiswoode	Phim Hành Động Phiêu Lưu,Phim...	NULL	120347	N	h	007: Ngày Mai Không Tàn Lụi
127038	Pierce Brosnan,Jonathan Pryce,Du...	2	Roger Spottiswoode	Phim Hành Động, Phim Phiêu Lưu	http://static.hdonline.vn/i/resour...	120347	N	h	007: Ngày Mai Không Tàn Lụi
127052	Barbara Bach,Curd Jürgens,Roger...	2	Lewis Gilbert	Phim Hành Động, Phim Phiêu Lưu,...	http://static.hdonline.vn/i/resour...	76752	N	h	007: Người Điệp Viên Tôi Yêu
123422	Roger Moore,Carole Bouquet,Topol	2	John Glen	Phim Hành Động Phiêu Lưu,Phim...	NULL	82398	N	h	007: Riêng Cho Đôi Mắt Em
127053	Roger Moore,Christopher Lee,Britt...	2	Guy Hamilton	Phim Hành Động	http://static.hdonline.vn/i/resour...	71807	N	h	007: Sát Thủ Với Khẩu Súng Vàng
117269	Pierce Brosnan, Sophie Marceau, R...	2	Michael Apted	Phim Hành Động, Phim Phiêu Lưu	http://cdn.phim3s.net/images/fil...	143145	N	h	007: Thế Giới Không Đủ
127059	Sean Connery,Jack Lord,Ursula An...	2	Terence Young	Phim Hành Động, Phim Hình Sự	http://static.hdonline.vn/i/resour...	55928	N	h	007: Tiến Sĩ No
147378	Kim Jong Min,Cha Tae Hyun,Kim J...	2		Phim TV-Show	http://static.hdonline.vn/i/resour...	0	N	h	1 Đêm 2 Ngày Phần 3
142368	Sawajiri Erika, Nishikido Ryou, Nar...	2		Phim tình cảm		0	N	h	1 Lít Nước Mắt
114988	Yoshimi Ashikawa, Mitsuo Hamad...	2	Riki Okamura	Phim Tâm Lý	http://cdn.phim3s.net/images/fil...	0	N	h	1 Lít Nước Mắt
116619	Erika Sawajiri,Ryô Nishikido,Hirok...	2		Phim Tình Cảm,Phim Tâm lý	NULL	494724	N	h	1 Lít Nước Mắt
141185	Eugene Levy, Jo Hyun Jae, Shin Su...	2	Lee Jae Sang	Phim tình cảm, phim hài hước		0	N	h	1 Mẹ 3 Bố
107366	Anne Hathaway, Jim Sturgess	2	Scherfig	Phim Tâm Lý	http://cdn.phim3s.net/images/fil...	1563738	N	h	1 Ngày
113245		2		Phim Phiêu Lưu, Phim Kinh Dị	http://cdn.phim3s.net/images/fil...	444682	N	h	10 Đại Dịch Của Chúa
132232	Hilary Swank,David Morrissey,Idris...	2	Stephen Hopkins	Phim Kinh Dị	http://static.hdonline.vn/i/resour...	444682	N	h	10 Đại Dịch Của Chúa
123451	Heath Ledger,Julia Stiles,Joseph G...	2	Gil Junger	Phim Hài,Phim Tình Cảm,Phim Tâm lý	NULL	147800	N	h	10 Điều Em Ghét Anh
132419	Heath Ledger,Julia Stiles,Joseph G...	2	Gil Junger	Phim Tình Cảm	http://static.hdonline.vn/i/resour...	147800	N	h	10 Điều Em Ghét Anh
163993	Joseph Gordon-Levitt, Heath Ledg...	2	Gil Junger	Phim hài hước, phim tình cảm		0	N	h	10 Điều Em Ghét Anh
118958	Joseph Gordon Levitt, Heath Ledg...	2	Gil Junger	Phim Hài Hước	http://cdn.phim3s.net/images/fil...	147800	N	h	10 Điều Em Ghét Anh

Hình 5-5 Thông tin được trích xuất trong trang phim trực tuyến.

5.3.2 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 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 5-6 Mô hình lưu trữ lịch sử của người dùng

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 trên.

Dữ liệu huấn luyện sử dụng lịch sử ba tháng query log của người dùng được lọc theo nội dung 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 **583,129** truy vấn dữ liệu click. Dữ liệu bao được lưu trữ theo định dạng dưới đây

Bảng 5-5 Các trường dữ liệu được đánh chỉ mục của lịch sử click của người dùng

Tên trường	Miêu tả
query_id	Định danh truy vấn

user_id	Định danh của người dùng
link	Liên kết được click
order	Hạng của liên kết
time	Thơi gian được click

5.3.3 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 từ cơ sở dữ liệu **MySQL** sử dụng thư viện **Elasticsearch-Jdbc** sử dụng cấu hình từ **Mysql** đến một cụm máy chủ **Elasticsearch** tất cả các bước được đánh chỉ mục được thực hiện cùng một cấu hình theo **Error! Reference source not found.** trên một máy chủ đơn và 2 máy chủ cùng đồng thời đánh chỉ mục.

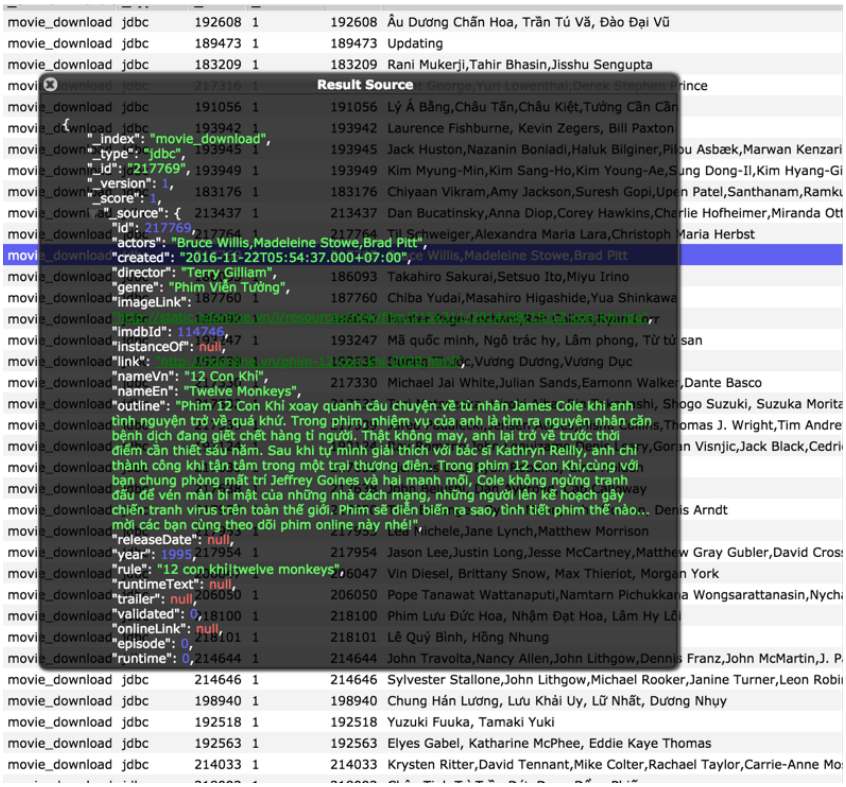
```
#!/bin/sh

DIR="$( cd "$( dirname "${BASH_SOURCE[0]}" )" && pwd )"
bin=${DIR}/../bin
lib=${DIR}/../lib

echo '
{
    "type" : "jdbc",
    "jdbc" : {
        "url" : "jdbc:mysql://fsearcher1v.itim.vn:3306/querypattern",
        "user" : "vsearch",
        "password" : "*****",
        "sql" :  "select *, id as _id from movie_download",
        "treat_binary_as_string" : true,
        "elasticsearch" : {
             "cluster" : "cluster",
             "host" : "rank2v.dev.itim.vn",
             "port" : 9300
        },
        "index" : "movie_download"
      }
}
' | java \
    -cp "${lib}/*" \
    -Dlog4j.configurationFile=${bin}/log4j2.xml \
    org.xbib.tools.Runner \
    org.xbib.tools.JDBCImporter
```

Hình 5-7 Cấu hình đánh chỉ mục từ Mysql sang cụm ElasticSearch

Sau bước này toàn bộ dữ liệu được đánh chỉ mục lên Elasticsearch và có thể tìm kiếm dùng các API tìm kiếm của Elasticsech.



Hình 5-8 Dữ liệu được đánh chỉ mục lên Elasticsearch

5.3.4 Trích xuất dữ liệu huấn luyện

Toàn bộ dữ liệu huấn luyện được thu thập từ lịch sử click biểu thị ra sự liên quan giữa truy vấn và click của người dùng. Các dữ liệu này sẽ được lọc và chỉ lấy dữ liệu các truy vấn và click liên quan tới chủ để phim trực tuyến, và đã sẽ sắp xếp theo số lượng click.

Ví dụ như truy vấn phim "quá nhanh quá nguy hiểm"

Bảng 5-6 Dữ liệu huấn luyện cho mô hình

Hạng	Tài liệu	Lượt Click
1	http://hdonline.vn/phim-qua-nhanh-qua-nguy-hiem-7-7454.html	1534
2	http://phim3s.net/phim-le/qua-nhanh-qua-nguy-hiem-7_8389/?utm_source=CocCoc	876

3	http://www.phimmoi.net/phim/qua-nhanh-qua-nguy-hiem-5-70/?utm_source=CocCoc	781

Sau khi trích chọn được thông tin của hạng tài liệu giữu các truy vấn ta sẽ tiến hành trích xuất vector đặc trưng để làm dữ liệu huấn luyện. tại bước này thu được 583,129 truy vấn dữ liệu giữa truy vấn của người dùng và liên kết trang web được click.

5.3.5 Trích xuất 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 dựa trên truy vấn và tài liệu, các thuộc tính thuộc tính của vector đặc trược được biểu diễn trong bảng dưới đây

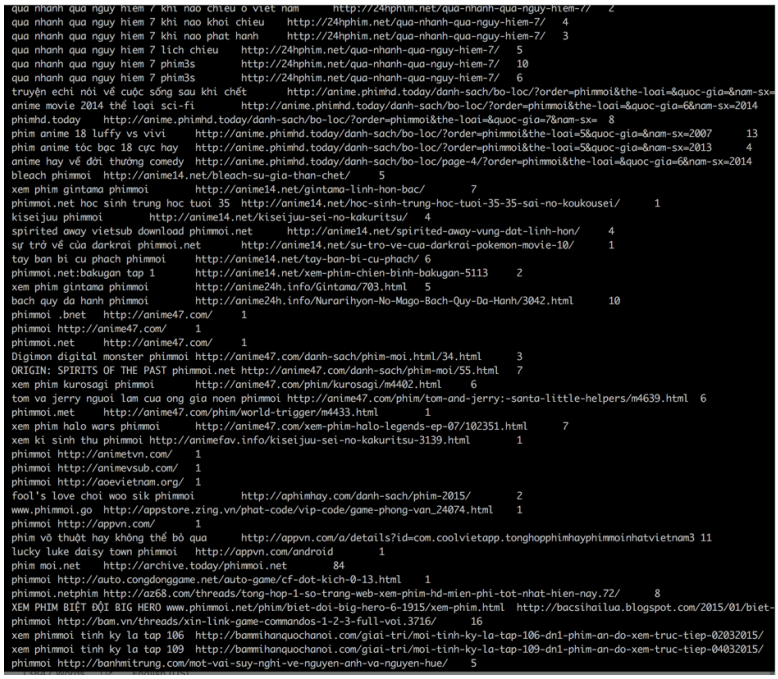
Bảng 5-7 Bảng mô tả vector đặc trưng cho mô hình học máy xếp hạng

Số thứ tự	Mô tả
1	IDF của tiêu đề phim
2	Độ dài của tiêu đề phim
3	Điểm số BM25 của truy vấn và tiêu đề phim
4	IDF của nội dung 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	Thời gian sản xuất phim (Năm hiện tại – Năm sản xuất)

Tại bước này sẽ tiến hành thu thập toàn bộ dữ liệu truy vấn của người dùng và thứ tự xếp hạng của các truy vấn xem phim mà người dùng nhập vào hệ thống tìm kiếm Cốc Cốc. Dữ liệu lịch sử thu được sẽ biểu diễn là tên truy vấn, liên kết được click và số lượng click. Để nhận biết truy vấn nào là truy vấn phim ta dựa vào hai tiêu chí sau đây.

Tiêu để truy vấn: Tiêu đề của truy vấn là những truy vấn mà xuất hiện trong cơ sở dữ liệu phim đã được đánh chỉ mục trong Elasticsearch.

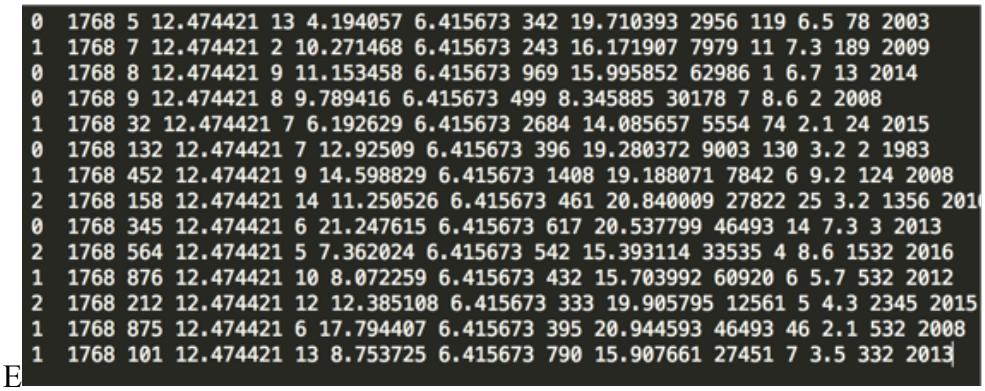
Liên kết được click: Các domain trong các liên kết được click phải nằm trong các trang web xem phim online như sau.



Hình 5-9 Lịch sử click của người dùng

Sau khi trích chọn được các truy vấn xem phim và sắp xếp theo thứ tự lượt click của người dùng ta coi đây là danh sách các liên kết phim có liên quan tới truy vấn. Tham số đầu vào của mô hình huấn luyện được biểu diễn như sau:

[**độ liên quan của truy vấn và liên kết phim**, **id của truy vấn**, **id của liên kết phim**, (11 thuộc tính được tính tóan dữ trên truy vấn và liên kết phim gốc)]. dưới đây mô tả bảng vector đặc trưng giữa truy vấn và liên kết phim theo thứ tự chỉ số được miêu tả bên trên.



E

Hình 5-10 Vector đặc trưng giữa truy vấn và liên kết phim

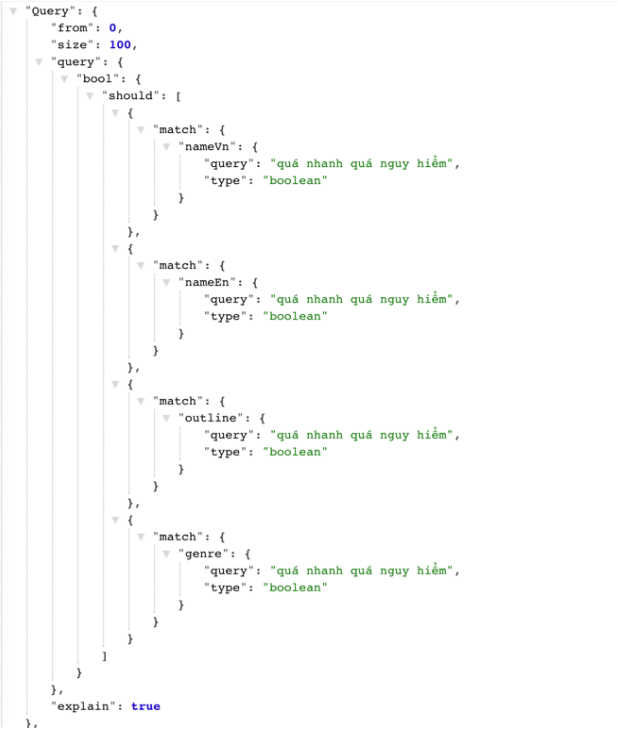
Sau khi có được bảng vector đặc trưng giữa truy vấn và liên kết phim ta tiến hành huấn luyện cho mô hình. Mô hình sẽ sử dụng thuật toán Listnet trong thư viện RankLib với các tham số huấn luyện dành cho thuật toán Listnet tham khảo tại

https://sourceforge.net/p/lemur/wiki/RankLib%20How%20to%20use/#eval

5.3.6 Xây dựng hệ thống xếp hạng và tính toán song song

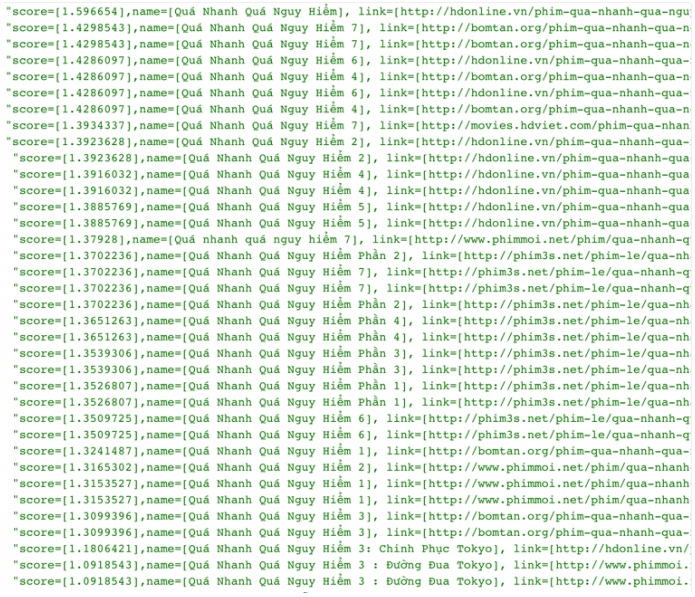
Sau khi huấn luyện mô hình học máy, tiếp đến là bước tính hợp mô hình vào hệ thống tìm kiếm phim trực tuyến. Với mỗi truy vấn của người dùng hệ thống sẽ được gửi cho bộ tìm kiếm thô, ở đây sử dụng **Apache Spark** để có thể truy vấn và tìm kiếm song song trong **Elasticsearch** để lấy về top 500 truy vấn trên mỗi máy

Dữ liệu sau khi được đánh chỉ mục trong Elasticsearch, chúng ta có thể tìm theo tên tiếng việt, tên tiếng anh, nội dung và thể loại phim dưới đây là cú pháp truy vấn cho truy vấn "quá nhanh quá nguy hiểm":

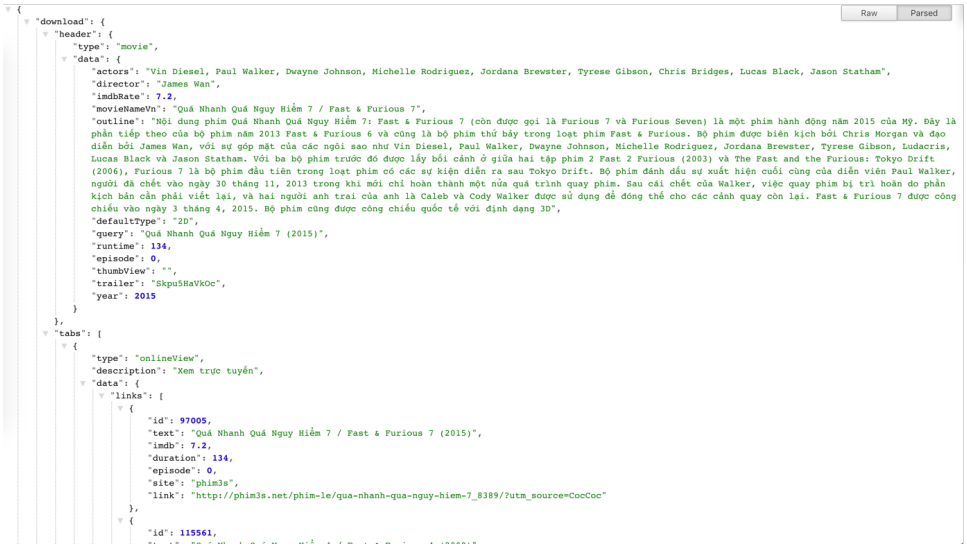


```
"Query": {
    "from": 0,
    "size": 100,
    "query": {
        "bool": {
            "should": [
                {
                    "match": {
                        "nameVn": {
                            "query": "quá nhanh quá nguy hiểm",
                            "type": "boolean"
                        }
                    }
                },
                {
                    "match": {
                        "nameEn": {
                            "query": "quá nhanh quá nguy hiểm",
                            "type": "boolean"
                        }
                    }
                },
                {
                    "match": {
                        "outline": {
                            "query": "quá nhanh quá nguy hiểm",
                            "type": "boolean"
                        }
                    }
                },
                {
                    "match": {
                        "genre": {
                            "query": "quá nhanh quá nguy hiểm",
                            "type": "boolean"
                        }
                    }
                }
            ]
        }
    },
    "explain": true
},
```


Sau khi gửi mẫu truy vấn này bộ phân tích truy vấn của Elasticserch ta có thể thu thập được kết quả là danh sách các bộ phim như dưới đây.



Sau khi thu được top 500 kết quả thu thập được ta sẽ tiến hành thực hiện trích xuất cho vector đặc trưng và đưa vào mô hình học máy xếp hạng Listnet đã được tính toán trước đó và đưa ra kết quả cuối cùng đến người dùng thông qua **Json Web Service**

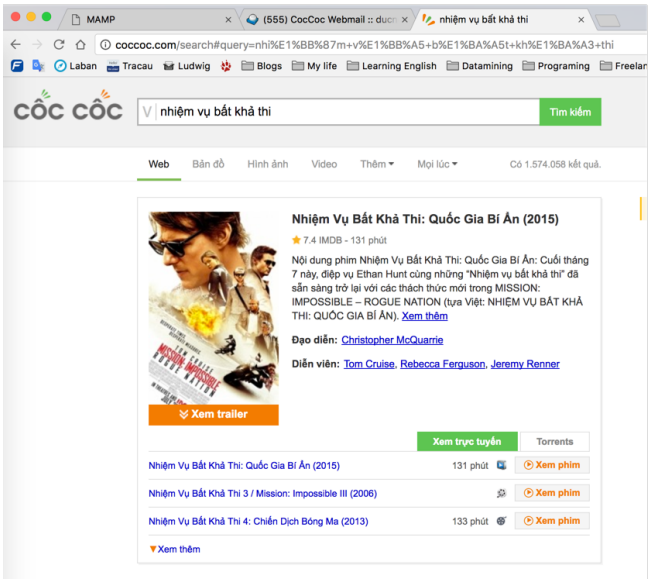


Hình 5-11 Dữ liệu trả về từ service tìm kiếm phim trực tuyến tại Cốc Cốc

5.3.7 Kết quả thực nghiệm

Kết quả của quá trình thực nghiệm được áp dụng trong quá trình xây dựng chức năng tìm kiếm riêng biệt hóa chức năng xem phim online trên trình duyệt Cốc Cốc. Đây là tính năng cho phép người dùng có thể nhanh trong xem được nội dung phim như tiêu đề tiếng anh, tiếng việt, năm sản xuất, liên kết xem phim trực tuyến một các trực quan hóa tất cả các bộ phim sẽ được sắp xếp theo

mô hình học máy xếp hạng được trình bày bên trên. Dưới đây là minh họa cho truy vấn phim "nhiệm vụ bất khả thi" http://coccoc.com/search#query=nhiệm vụ bất khả thi



Hình 5-12 Minh họa chức năng tìm kiếm phim trực tuyến

Đây là một tính năng hữu ích cho người dùng hay tìm kiếm thông tin phim nội dung phim, người dùng có thể chọn lựa dễ dàng giữa các nhà cung cấp phim trực tuyến đã được xếp hạng để có thể hiện thị nội dung phù hợp với người dùng hơn.

5.4 Đánh giá

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. Phần đánh giá thực nghiệm này 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.

5.4.1 Hiệu năng

Để so sánh hiệu quả thời gian tôi tiến hành chạy các bước thực nghiệm trên một máy đơn và ba máy tính có thông số như sau. Kết quả của quá trình thực nghiệm này được biểu diễu dưới đây

Bảng 5-8 Bảng đánh giá hiệu quả về mặt thời gian

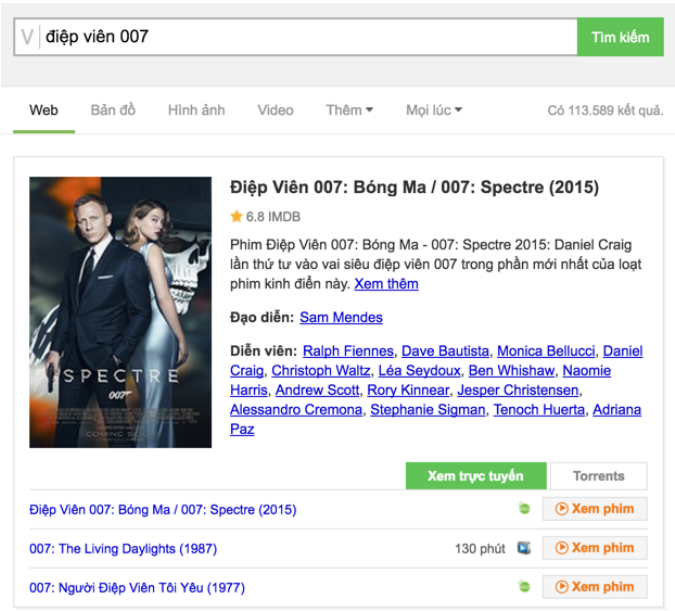
Công việc thực hiện	Một máy tính	Ba máy tính
Đánh chỉ mục dữ liệu cho 117.094 bản ghi IMDb, 213.253 phim online, 583,129 truy vấn dữ liệu click	32 phút 15s	13 phút 27s

Huấn luyện mô hình 230.000 truy vấn và tài liệu	2h 30phút	44 phút
Chạy 930.321 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.

5.4.2 Chất lượng xếp hạng

Mô hình đã được chạy trên hệ thống **Cốc Cốc** như một thành phần của hệ thống tìm kiếm. Hệ thống tìm kiếm mới đã có thể bổ sung thêm giao diện trực quan hóa do đó người dùng đã có thể dễ dàng tìm và chọn ra nhưng bộ phim phù hợp nhất thông qua nhưỡng dữ liệu đã được hiện thị thêm.



Hình 5-13 Hệ thống tìm kiếm phim online trên Cốc Cốc

Hình trên 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 5-9 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%

5.5 Tổng kết chương

Qua xây dựng và đánh giá mô hình thực nghiệm. Các kết quả thu được cho thấy hiệu quả rõ rệt về mặt thời gian khi sử dụng phương pháp tính toán song song, chất lượng tìm kiếm cũng được mở rộng khi tỉ lệ CTR đã tăng từ 8,57% lên tới 12,3% khi áp dụng mô hình mới tại máy tìm phim.

Kết luận chung

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.
- Hướng phát triển tiếp theo:
- Tiếp tục tham khảo nhiều thuật toán học máy xếp hạng khác để so sánh và nâng cao chất lượng tìm kiếm hơn nữa.
- Áp dụng mô hình cho nhiều máy tìm kiếm chuyên biệt tại **Cốc Cốc** như tìm kiếm tin tức, sản phẩm mua sắm

Tài liệu tham khảo

[1] ITU, “Internet protocol data communication service – IP packet transfer and availability performance parameters,” ITU-T Recommendation Y.1540, Feb. 1999.

[2] M. Winlaw, M. B. Hynes, A. Caterini and H. D. Sterck, "Algorithmic Acceleration of Parallel ALS for Collaborative Filtering: Speeding up Distributed Big Data Recommendation in Spark," Parallel and Distributed Systems (ICPADS), 2015 IEEE 21st International Conference on, Melbourne, VIC, 2015, pp. 682-691.

[3] X. M. Li and Y. Y. Wang, "Design and Implementation of an Indexing Method Based on Fields for Elasticsearch," 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Qinhuangdao, 2015, pp. 626-630.

[4] P. P. I. Langi, Widyawan, W. Najib and T. B. Aji, "An evaluation of Twitter river and Logstash performances as elasticsearch inputs for social media analysis of Twitter," Information & Communication Technology and Systems (ICTS), 2015 International Conference on, Surabaya, 2015, pp. 181-186.

[5] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)

[6] Singhal, A.: Modern information retrieval: a brief overview. IEEE Data Engineering Bulletin 24(4), 35–43 (2001)

[7] Tax, Niek (2014) Scaling Learning to Rank to Big Data: Using MapReduce to parallelise Learning to Rank.

[8] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, Learning Spark: Lightning-Fast Big Data Analysis. Sebastopol, CA, USA: O’Reilly Media, Inc., 2015.

[9] C. Avery, “Giraph: Large-scale graph processing infrastructure on hadoop,” Proceedings of the Hadoop Summit. Santa Clara, 2011.

[10] M. Gates, H. Anzt, J. Kurzak and J. Dongarra, "Accelerating collaborative filtering using concepts from high performance computing," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 667-676.

[11] Amento, B., Terveen, L., Hill, W.: Does authority mean quality? Predicting expert quality ratings of web documents. In: Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 296– 303 (2000)

[12] Haveliwala, T.: Efficient computation of pageRank. Tech. rep. 1999-31, Stanford University (1999)

[13] McSherry, F.: A uniform approach to accelerated pagerank computation. In: Proceedings of the 14th International Conference on World Wide Web (WWW 2005), pp. 575–582. ACM, New York (2005)

[14] S. Hatakenaka and T. Miura, "Query and Topic Sensitive PageRank for general documents," 2012 14th IEEE International Symposium on Web Systems Evolution (WSE), Trento, 2012, pp. 97-101.

[15] Richardson, M., Domingos, P.: The intelligent surfer: probabilistic combination of link and content information in pagerank. In: Advances in Neural Information Processing Systems 14 (NIPS 2001), pp. 1441–1448. MIT Press, Cambridge (2002)

[16] Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Pro- ceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004), pp. 576–587 (2004). VLDB Endowment

[17] Voorhees,E.M.:The philosophyof information retrieval evaluation. In: Lecture Notes in Computer Science (CLEF 2001), pp. 355–370 (2001)

[18] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans- actions on Information Systems 20(4), 422–446 (2002)

[19] IEEE Reference Format [Online] http://www.ieee.org/auinfo03.pdf

[20] B. Callaghan, Voices from the Margins: Postmodernism and Latin American Fiction, Master thesis, University College Cork, 1994.

[21] H. Schimanski and C. Thanner, “Raiders of the lost ark,” IEEE Trans. Electromagnetic Compatibility, vol. 51, no. 5, pp. 543–547, May 2003.

[22] J. Matula and R. Franck, “A case for two,” in Proc. 15th Int. Zurich Symposium and Technical Exhibition on Electromagnetic Compatibility, Zurich, Switzerland, Feb. 2003, vol. 1, pp. 347–350.

[23] Signorini. The Indexable Web is More than 11.5 Billion Pages, University of Iowa, Computer Science, 2005

[24] Tie-Yan Liu.Learning to Rank for Information Retrieval, 2011

[25] http://spark.apache.org/