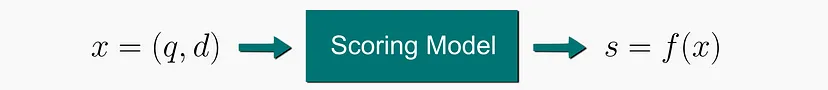
1. **Kĩ thuật xếp hạng**
2. **Định nghĩa**

Xếp hạng (ranking) là quá trình sắp xếp các tài liệu theo mức độ liên quan để tìm kiếm thông tin quan trọng nhất với một truy vấn cụ thể. Đây là một vấn đề cốt lõi trong truy xuất thông tin và 1 số ứng dụng khác như hệ thống gợi ý dựa trên hồ sơ người dùng và lịch sử mua hàng để xếp hạng các sản phẩm nhằm gợi ý sản phẩm phù hợp nhất, xếp phòng trống cho khách với bộ lọc là ngày check-in và check-out…

1. **Cách hoạt động**

Các mô hình xếp hạng dựa trên một hàm chấm điểm. Sau khi tính toán điểm liên quan, các tài liệu sẽ được sắp xếp dựa trên điểm số đó.



Hàm chấm điểm của kĩ thuật xếp hạng (Nguồn: [link](https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4))

* Mô hình không gian vector: Tính toán biểu diễn vector cho 𝑞 và 𝑑 (ví dụ: sử dụng Tf-Idf hoặc BERT) và đánh giá mức độ liên quan dựa trên cosine similarity giữa hai vector này.
* Học để xếp hạng (Learning to Rank): Sử dụng mô hình học máy để dự đoán

𝑠 và tối ưu hóa một hàm mất mát dựa trên thứ tự sắp xếp.

1. **Các tiêu chí đánh giá mô hình xếp hạng**

* Mean Average Precision (MAP)
* Discounted Cumulative Gain (DCG)

1. **Các mô hình học để xếp hạng**

Các mô hình này được chia thành ba phương pháp chính dựa trên cách tính hàm mất mát:

**4.1 Phương pháp PointWise**

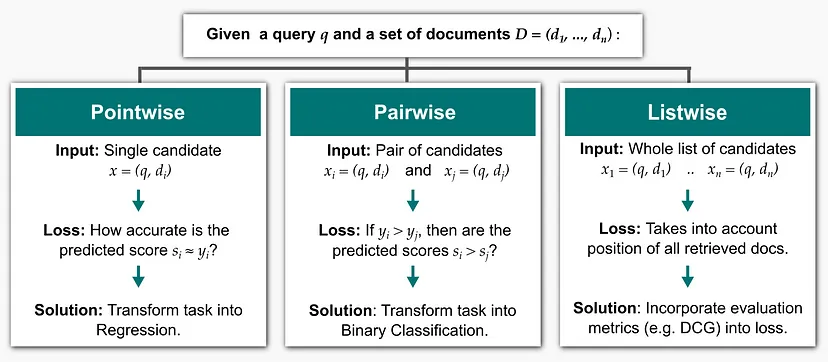
Phương pháp Pointwise tập trung vào việc dự đoán điểm liên quan si cho từng tài liệu di​ một cách độc lập. Điều này chuyển bài toán xếp hạng thành một bài toán hồi quy, trong đó mô hình học cách dự đoán mức độ liên quan của mỗi tài liệu dựa trên các đặc trưng đầu vào. Một ví dụ phổ biến là sử dụng hàm mất mát Mean Square Error (MSE) để tối ưu hóa mô hình. Mặc dù phương pháp này đơn giản và dễ triển khai, nhưng nó không tận dụng được mối quan hệ giữa các tài liệu trong danh sách, dẫn đến hiệu quả hạn chế trong các bài toán yêu cầu xếp hạng phức tạp.

**4.2 Phương pháp Pairwise**

Pairwise tập trung vào việc so sánh độ liên quan giữa hai tài liệu di và dj​ để xác định tài liệu nào quan trọng hơn. Nhiệm vụ này được biểu diễn như một bài toán phân loại nhị phân, với mục tiêu dự đoán mối quan hệ ưu tiên giữa hai tài liệu. Ví dụ, RankNet sử dụng sự khác biệt điểm số sij=si−sj sau đó ánh xạ qua hàm logistic để ước lượng xác suất. LambdaRank cải tiến bằng cách điều chỉnh gradient dựa trên độ quan trọng của thứ hạng, giúp tập trung hơn vào các tài liệu có thứ hạng cao. Pairwise vượt trội hơn Pointwise trong việc nắm bắt mối quan hệ giữa các tài liệu nhưng vẫn chưa tối ưu hóa toàn bộ danh sách.

**4.3 Phương pháp Listwise**

Listwise xử lý toàn bộ danh sách tài liệu cùng một lúc, tối ưu hóa trực tiếp trên thứ hạng dự đoán. Phương pháp này sử dụng các hàm mất mát được thiết kế để tính toán trên toàn bộ danh sách, chẳng hạn như SoftRank, áp dụng phân phối xác suất để làm mịn thứ hạng, hoặc ListNet, sử dụng Binary Cross-Entropy (BCE) giữa phân phối xác suất thực và dự đoán trong không gian các hoán vị. LambdaLoss, một khung tổng quát, kết hợp các ý tưởng từ các phương pháp trước đó và tối ưu hóa trực tiếp các tiêu chí đánh giá như NDCG. Listwise đạt hiệu quả cao nhất trong các bài toán xếp hạng nhờ khả năng xử lý toàn diện danh sách và tập trung vào các mục tiêu đánh giá.



Các phương pháp học để xếp hạng: pointwise, pairwise, listwise. (Nguồn: [link](https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4))

1. **Kết luận**

Bài toán xếp hạng xuất hiện ở khắp mọi nơi, từ truy xuất thông tin, hệ thống gợi ý đến đặt chỗ du lịch. Các tiêu chí đánh giá như MAP và NDCG đều xem xét cả thứ hạng và mức độ liên quan của các tài liệu được truy xuất, do đó rất khó để tối ưu hóa trực tiếp.

Các phương pháp học để xếp hạng (Learning to Rank) sử dụng các mô hình học máy để dự đoán điểm liên quan của tài liệu và được chia thành 3 loại chính: pointwise, pairwise, và listwise. Trong hầu hết các bài toán xếp hạng, các phương pháp listwise như LambdaRank và khung tổng quát LambdaLoss đạt hiệu quả tốt nhất.

**II. Machine translation (Dịch máy)**

Dịch máy là quá trình sử dụng trí tuệ nhân tạo để tự động dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác mà không cần sự tham gia của con người. Phương pháp dịch máy hiện đại không còn dừng lại ở việc dịch bám sát bản gốc để truyền đạt ý nghĩa đầy đủ của văn bản ngôn ngữ gốc bằng ngôn ngữ đích. Mà giờ đây phương pháp này còn phân tích tất cả các thành phần của văn bản, đồng thời nhận biết cả cách thức các từ ảnh hưởng đến nhau.

Trong dịch máy, văn bản hoặc ngôn ngữ gốc được gọi là ngôn ngữ nguồn và ngôn ngữ bạn muốn dịch sang được gọi là ngôn ngữ đích. Dịch vụ dịch máy hoạt động theo quy trình hai bước cơ bản:

* Giải mã ý nghĩa ngôn ngữ nguồn của văn bản gốc
* Mã hóa ý nghĩa sang ngôn ngữ đích.

1. **Dịch máy dựa theo quy tắc**

Các chuyên gia ngôn ngữ sẽ phát triển từ điển song ngữ và các quy tắc về ngữ nghĩa được tích hợp sẵn cho các ngành hoặc chủ đề cụ thể. Dịch vụ dịch máy dựa trên quy tắc sử dụng các từ điển này để dịch chính xác nội dung cụ thể. Sau đây là các bước trong quy trình:

1. Phần mềm dịch máy phân tích cú pháp văn bản đầu vào và tạo ra một dạng biểu diễn chuyển tiếp.
2. Sau đó chuyển đổi dạng biểu diễn này thành ngôn ngữ đích bằng cách sử dụng các quy tắc ngữ pháp và từ điển làm tài liệu tham khảo.

**Ưu và nhược điểm**

Chức năng dịch máy dựa trên quy tắc có thể được tùy chỉnh cho một ngành hoặc chủ đề cụ thể. Chức năng này cung cấp bản dịch có thể dự đoán được cũng như bản dịch chất lượng. Tuy nhiên, chức năng này lại tạo bản dịch kém chất lượng nếu văn bản nguồn có lỗi hoặc sử dụng các từ không có trong từ điển tích hợp sẵn. Cách duy nhất để cải thiện chức năng này đó là cập nhật từ điển theo cách thủ công một cách thường xuyên.

**2. Dịch máy theo cách thống kê**

Thay vì dựa vào các quy tắc ngôn ngữ, hình thức dịch máy theo cách thống kê sử dụng công nghệ máy học để dịch văn bản. Các thuật toán máy học sẽ phân tích một lượng lớn các bản dịch có sẵn của con người và tìm kiếm các mẫu thống kê. Sau đó, phần mềm sẽ đưa ra một phỏng đoán thông minh khi được yêu cầu dịch một văn bản nguồn mới. Hình thức dịch này đưa ra dự đoán dựa trên khả năng thống kê rằng một từ hoặc cụm từ cụ thể sẽ đi kèm với một từ hoặc cụm từ khác trong ngôn ngữ đích.

1. **Dịch máy dựa trên cú pháp**

Hình thức dịch máy dựa trên cú pháp là một danh mục phụ của hình thức dịch máy theo cách thống kê. Hình thức này sử dụng các quy tắc ngữ pháp để dịch các đơn vị cú pháp. Ngoài ra, hình thức này còn phân tích các câu để kết hợp các quy tắc cú pháp vào các mô hình dịch theo cách thống kê.

**Ưu và nhược điểm**

Phương pháp thống kê đòi hỏi phải có quá trình đào tạo hàng triệu từ cho mỗi cặp ngôn ngữ. Tuy nhiên, khi có đầy đủ dữ liệu, bản dịch máy sẽ rất chính xác.

1. **Dịch máy sử dụng mạng nơ-ron**

Dịch máy sử dụng mạng nơ-ron dùng trí tuệ nhân tạo để học ngôn ngữ và liên tục cải thiện kiến thức đó bằng cách sử dụng một phương pháp máy học cụ thể được gọi là mạng nơ-ron. Hình thưc này thường hoạt động kết hợp với các phương pháp dịch theo cách thống kê.

**Mạng nơ-ron**

Mạng nơ-ron là một tập hợp các nút được kết nối với nhau lấy cảm hứng từ bộ não của con người. Mạng này là một hệ thống thông tin nơi dữ liệu đầu vào đi qua một số nút được kết nối với nhau để tạo ra đầu ra. Phần mềm dịch máy sử dụng mạng nơ-ron dùng mạng nơ-ron để thao tác với các tập dữ liệu khổng lồ. Mỗi nút sẽ thực hiện một lần thay đổi được quy cho văn bản nguồn thành văn bản đích cho đến khi nút đầu ra đưa ra kết quả cuối cùng.

**Dịch máy sử dụng mạng nơ-ron và các phương pháp dịch khác**

Mạng nơ-ron xem xét toàn bộ câu đầu vào ở mỗi bước khi tạo câu đầu ra. Trong khi các mô hình dịch máy khác chia câu đầu vào thành các tập hợp từ và cụm từ, ánh xạ chúng thành một từ hoặc câu trong ngôn ngữ đích. Hệ thống dịch máy bằng mạng nơ-ron có thể giải quyết nhiều hạn chế của các phương pháp khác và thường tạo ra các bản dịch có chất lượng tốt hơn.

1. **Dịch máy kết hợp**

Công cụ dịch máy kết hợp sử dụng hai hoặc nhiều mô hình dịch máy trên một phần mềm. Bạn có thể sử dụng phương pháp kết hợp để cải thiện hiệu quả của một mô hình dịch đơn lẻ. Quá trình dịch máy này thường sử dụng các hệ thống con phụ trách việc dịch máy dựa trên quy tắc và theo cách thống kê. Bản dịch đầu ra cuối cùng là sự kết hợp của đầu ra từ tất cả các hệ thống con.

**Ưu và nhược điểm**

Các mô hình dịch máy kết hợp cải thiện thành công chất lượng bản dịch bằng cách khắc phục các vấn đề liên quan đến các phương pháp dịch đơn lẻ. (Nguồn: Amazon Web Service)

1. **Amazon machine translation**

Theo trang thông tin của Amazon Web Service, Amazon Translate là dịch vụ máy dịch ứng dụng mạng nơron để cung cấp bản dịch ngôn ngữ nhanh chóng, chất lượng cao, giá cả phải chăng và có thể tùy chỉnh. Bạn có thể bản địa hóa các nội dung như trang web và ứng dụng cho bộ phận người dùng đa dạng, dễ dàng dịch khối lượng lớn văn bản để phân tích và cho phép giao tiếp đa ngôn ngữ một cách hiệu quả giữa nhiều người dùng. Với, Amazon Translate, ta có thể:

* Dễ dàng tích hợp vào các ứng dụng của bạn bằng một lệnh gọi API đơn giản
* Tùy chỉnh bản dịch máy đầu ra bằng Custom Terminology (Thuật ngữ tùy chỉnh) để xác định cách dịch tên thương hiệu, tên mẫu mã và các thuật ngữ đặc biệt khác
* Mở rộng quy mô để đáp ứng nhu cầu dịch thuật với các bản dịch nhanh chóng và đáng tin cậy

1. Google machine translatation

Google Translate là một dịch vụ dịch thuật máy học đa ngôn ngữ do Google phát triển, cho phép dịch văn bản, tài liệu và trang web từ một ngôn ngữ sang ngôn ngữ khác. Dịch vụ này cung cấp giao diện trang web, ứng dụng di động dành cho Android và iOS, cùng với API hỗ trợ các nhà phát triển xây dựng tiện ích mở rộng trình duyệt và ứng dụng phần mềm. Tính đến tháng 1 năm 2025, Google Translate hỗ trợ 249 ngôn ngữ và phương ngữ ở nhiều cấp độ khác nhau. Vào tháng 5 năm 2013, dịch vụ này phục vụ hơn 200 triệu người dùng mỗi ngày, và đến tháng 4 năm 2016 đã đạt hơn 500 triệu người dùng với hơn 100 tỷ từ được dịch mỗi ngày. Google Translate có khả năng dịch nhiều hình thức văn bản và phương tiện, bao gồm văn bản, giọng nói, và văn bản trong hình ảnh tĩnh hoặc động. Cụ thể, các chức năng của nó bao gồm:

* Dịch từ viết: Một chức năng dịch từ hoặc văn bản viết sang ngôn ngữ khác.
* Dịch trang web: Chức năng dịch toàn bộ trang web sang các ngôn ngữ được chọn.
* Dịch tài liệu: Chức năng dịch tài liệu do người dùng tải lên sang ngôn ngữ được chọn. Các tài liệu phải ở định dạng: .doc, .docx, .odf, .pdf, .ppt, .pptx, .ps, .rtf, .txt, .xls, .xlsx.
* Dịch giọng nói: Chức năng dịch tức thì ngôn ngữ nói sang ngôn ngữ khác được chọn.
* Dịch trong ứng dụng di động: Năm 2018, Google ra mắt tính năng mới "Tap to Translate", cho phép dịch ngay lập tức trong bất kỳ ứng dụng nào mà không cần thoát hoặc chuyển đổi.
* Dịch hình ảnh: Chức năng nhận diện văn bản trong hình ảnh do người dùng chụp và dịch văn bản trên màn hình ngay lập tức.
* Dịch chữ viết tay: Chức năng dịch ngôn ngữ được viết tay trên màn hình điện thoại hoặc vẽ trên bàn phím ảo mà không cần bàn phím vật lý hỗ trợ.
* Dịch hội thoại song ngữ: Chức năng dịch các cuộc hội thoại bằng nhiều ngôn ngữ.
* Chuyển ngữ: Chức năng chuyển ngữ giọng nói thành văn bản trong các ngôn ngữ khác nhau.

Ngoài ra, API của Google Translate cung cấp dịch vụ trả phí cho các nhà phát triển. Tích hợp với Google Chrome, Google Assistant và các thiết bị như Google Nest và Pixel Buds, dịch vụ liên tục cải thiện nhờ công nghệ mạng thần kinh và phản hồi từ người dùng. (Nguồn: wikipedia)

**Nguồn tham khảo:**

1. [**https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4**](https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4)
2. [**https://en.wikipedia.org/wiki/Google\_Translate**](https://en.wikipedia.org/wiki/Google_Translate)
3. [**https://aws.amazon.com/translate/**](https://aws.amazon.com/translate/)