Tìm kiếm thông tin

Chương 12. Vượt qua giới hạn tập từ
(Các xu hướng)

Mối quan hệ giữa các từ

- Trong thực tế sự xuất hiện của 1 từ có thể liên quan đến sự xuất hiện của 1 từ khác (các từ không độc lập).
- Các mối liên hệ có thể phức tạp
- van Rijsbergen (1979) đưa ra mô hình phụ thuộc hình cây đơn giản
  - Tree Augmented Naive Bayes, Friedman và Goldszmidt, 1996.

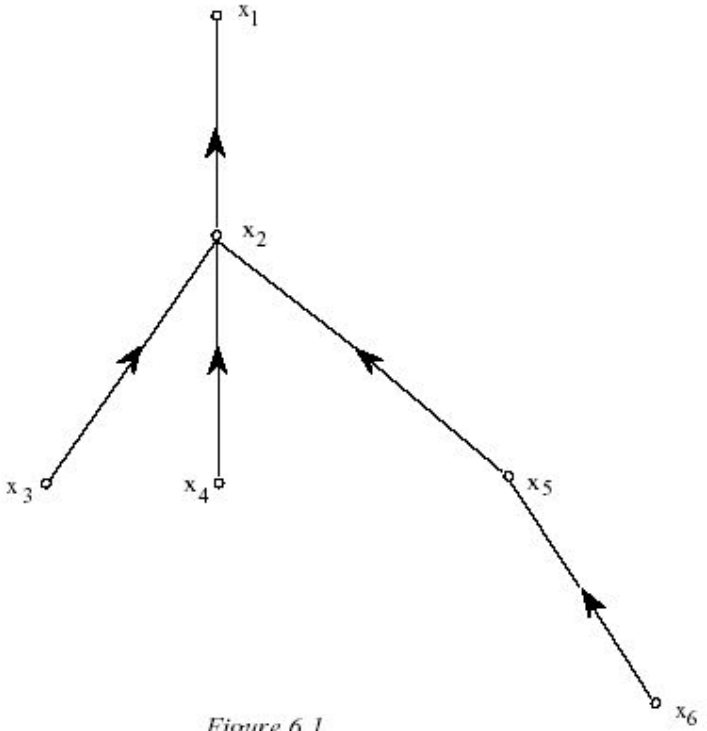


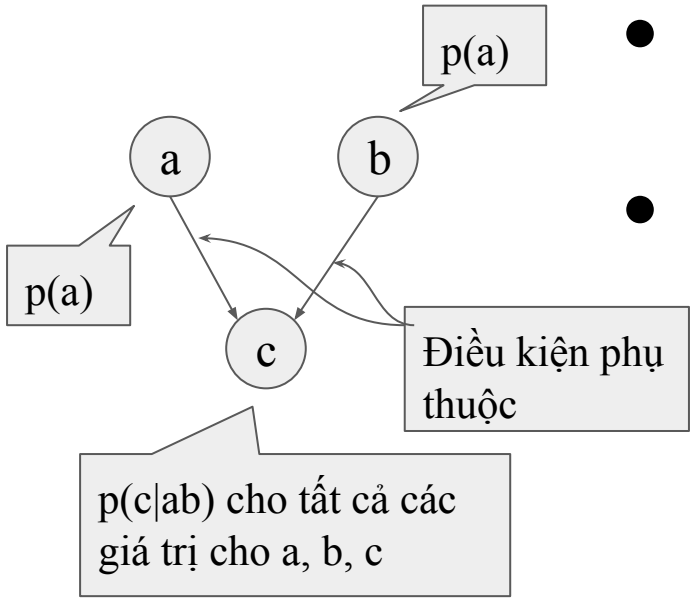
Figure 6.1.

- Mỗi từ phụ thuộc vào một từ khác
- Trong những năm 1970, mô hình này còn tồn tại vấn đề ước lượng các giá trị.

Mạng Bayes cho tìm kiếm văn bản

- Turtle và Croft 1990
- Mạng Bayes là gì?
  - Đồ thị có hướng không chứa chu trình
  - Các đỉnh
    - Sự kiện hoặc biến
  - Liên kết
    - Mô hình hóa sự phụ thuộc trực tiếp giữa các nút

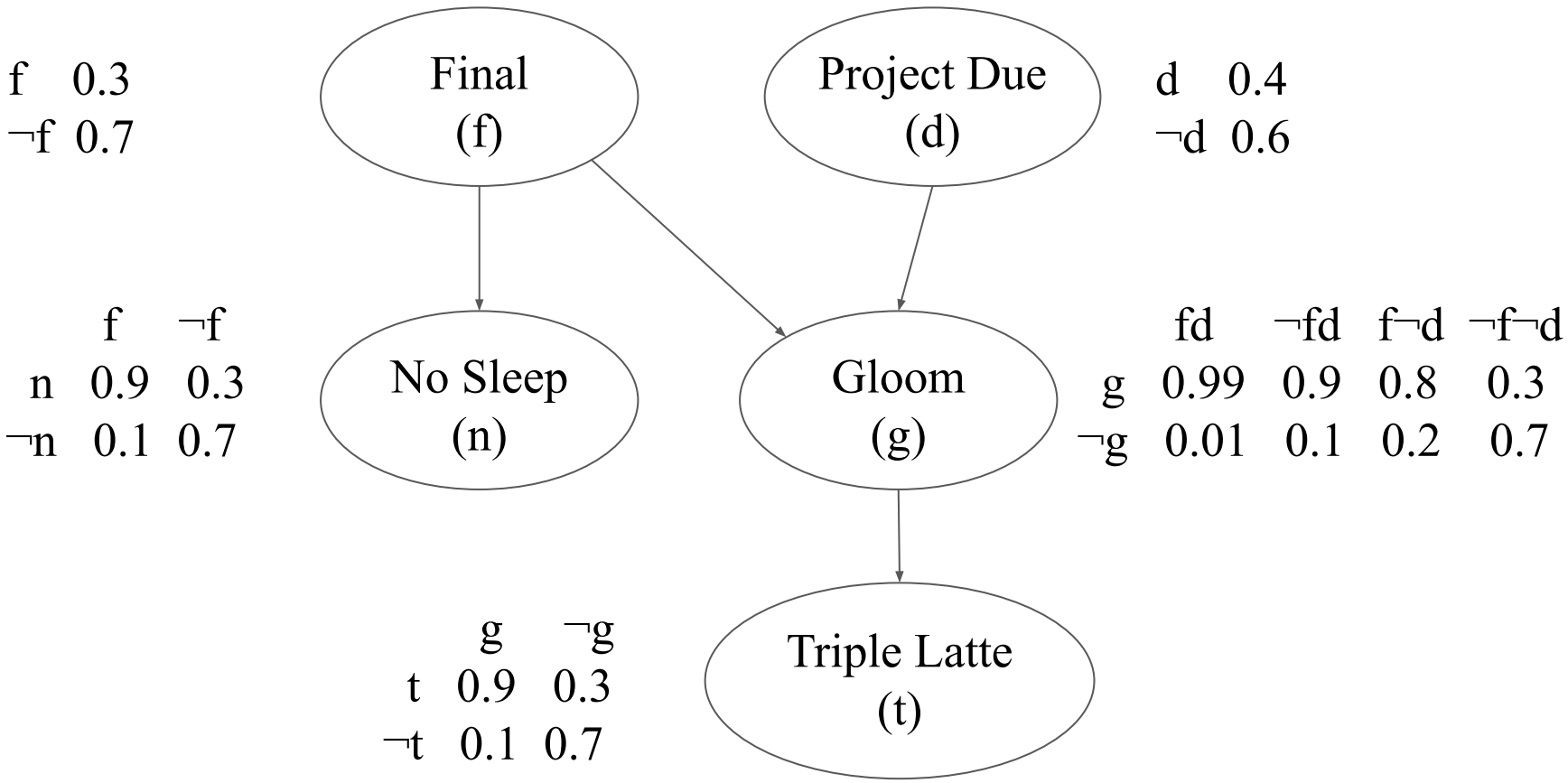
Mạng Bayes



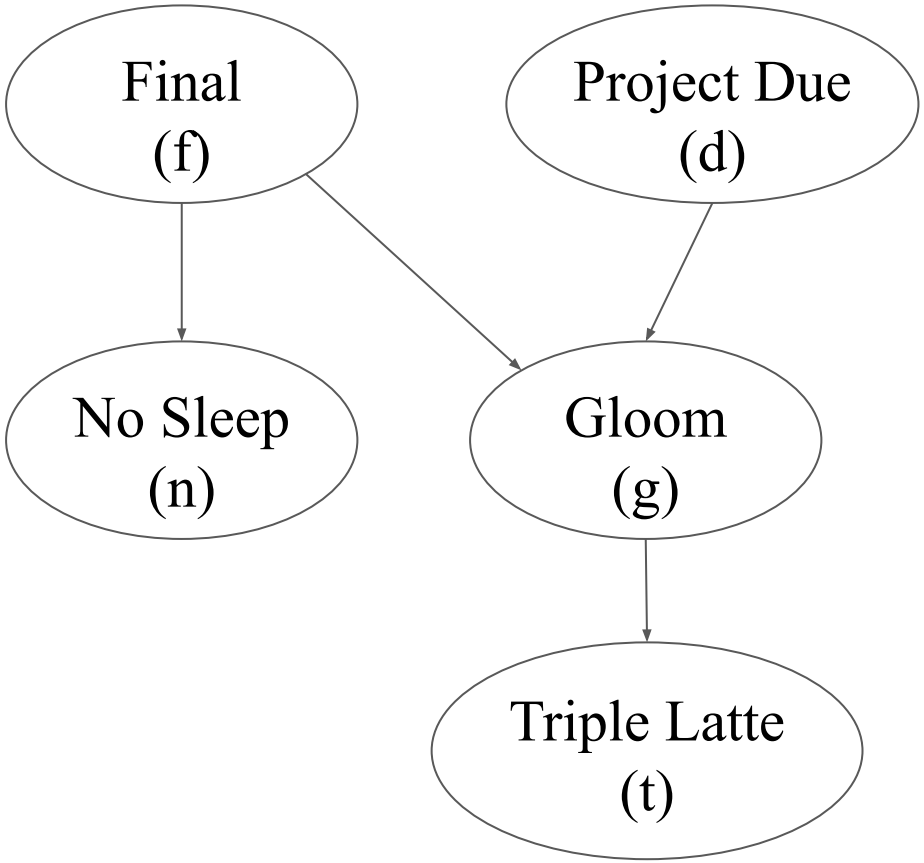
- Mô hình mạng Bayes mô phỏng mối quan hệ giữa các sự kiện
- Suy diễn trong mạng Bayes:
  - Cho phân bố xác suất của gốc và các xác suất có điều kiện có thể tính xác suất hậu nghiệm của bất kỳ sự kiện nào
  - Giả thuyết điều chỉnh, ví dụ sau quan sát được b sẽ dẫn tới tính toán lại các giá trị xác suất

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. 1999. Probabilistic Networks and Expert Systems. Springer Verlag.
J. Pearl. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan-Kaufman.

Ví dụ 12.1. mạng Bayes



Giả thuyết độc lập

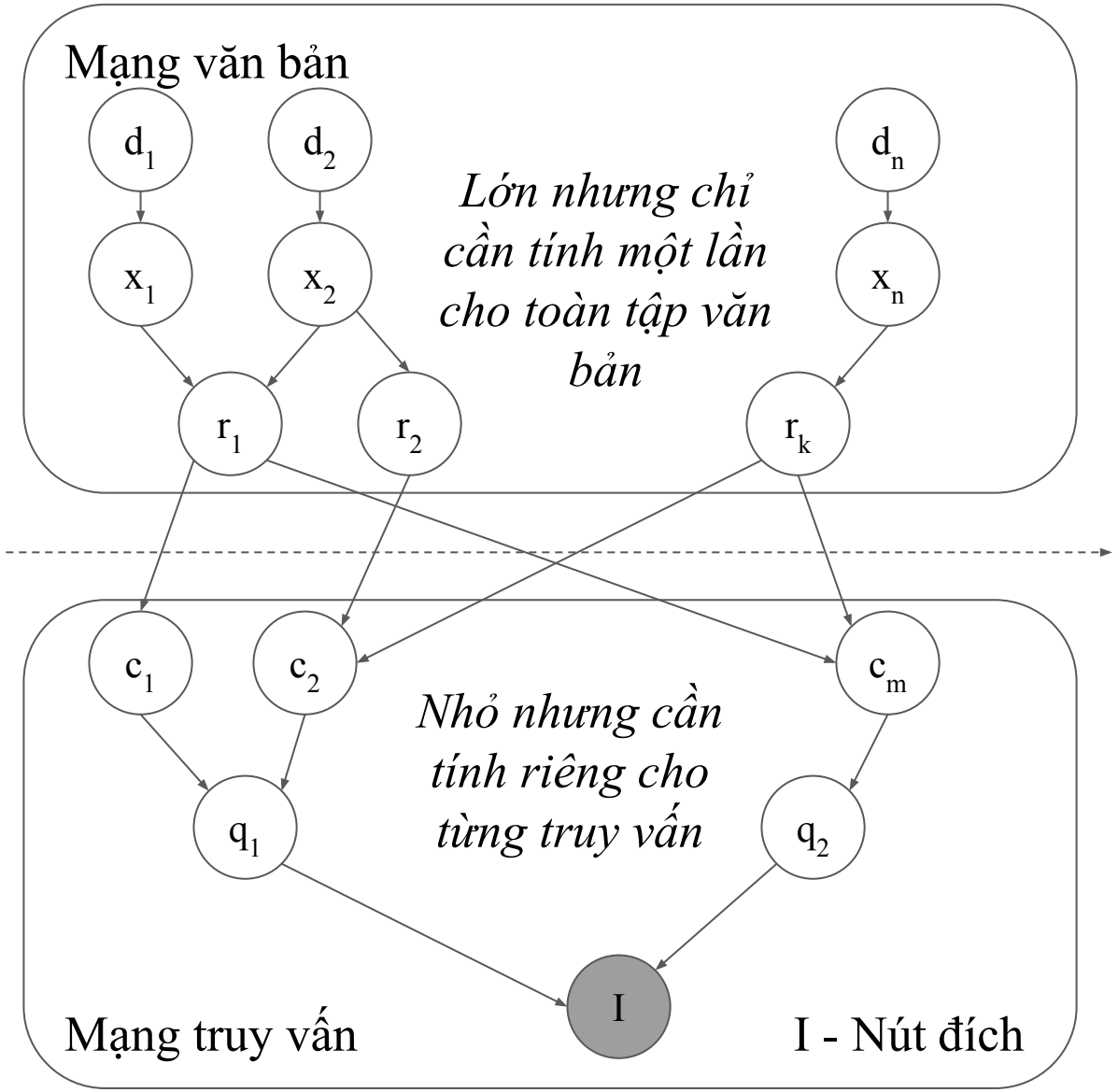


- Giả thuyết độc lập:
  $P(t|g, f) = P(t|g)$
- Xác suất kết hợp
  $P(f, d, n, g, t) = P(f)P(d)$
  $P(n|f)P(g|f, d)P(t|g)$

Các mô hình

- Mô hình văn bản trong mạng văn bản
- Mô hình nhu cầu thông tin trong mạng truy vấn

Mạng Bayes cho IR: Ý tưởng



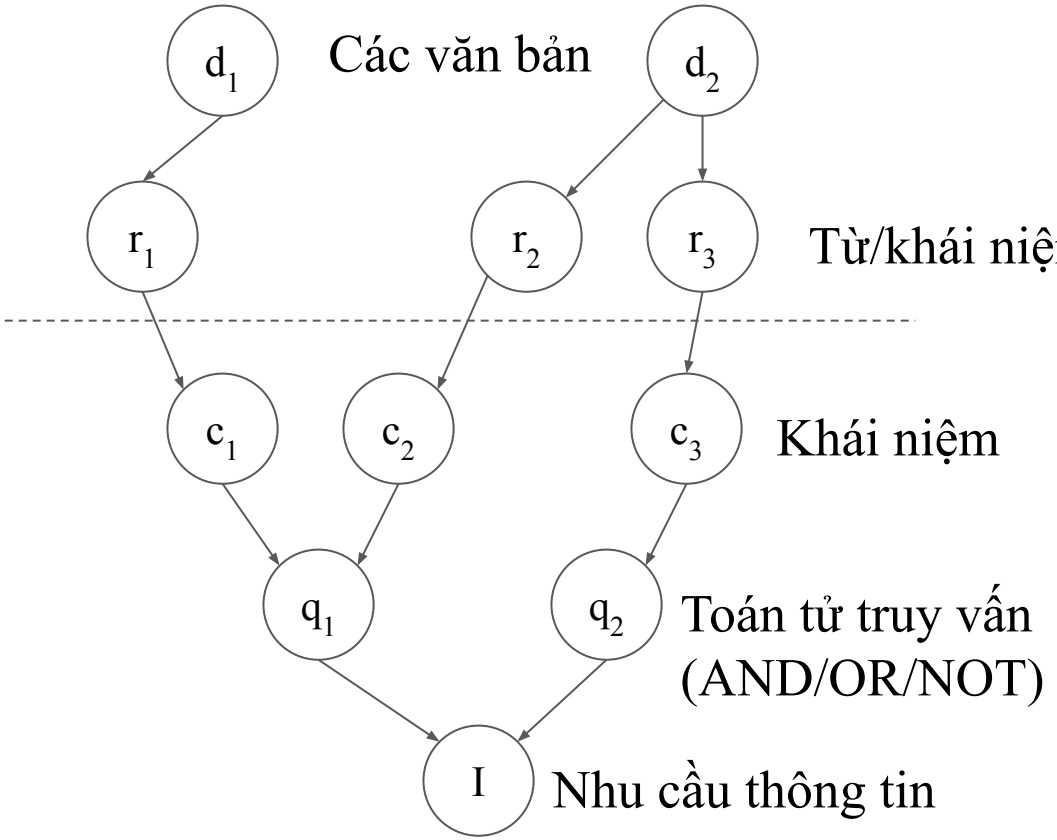
d_i - văn bản
x_i - biểu diễn văn bản
r_i - khái niệm

c_i - khái niệm trong truy vấn
q_i - khái niệm bậc cao trong truy vấn

Mạng Bayes cho IR

- Xây dựng mạng văn bản (1 lần)
- Với mỗi truy vấn
  - Xây dựng mạng truy vấn tốt nhất
  - Gắn kết nó với mạng văn bản
  - Tìm tập con của d_i sao cho giá trị xác suất của nút I đạt cực tại (tập con tốt nhất)
  - Trả về các d_i tìm được

Mạng Bayes cho tìm kiếm văn bản



Mạng văn bản

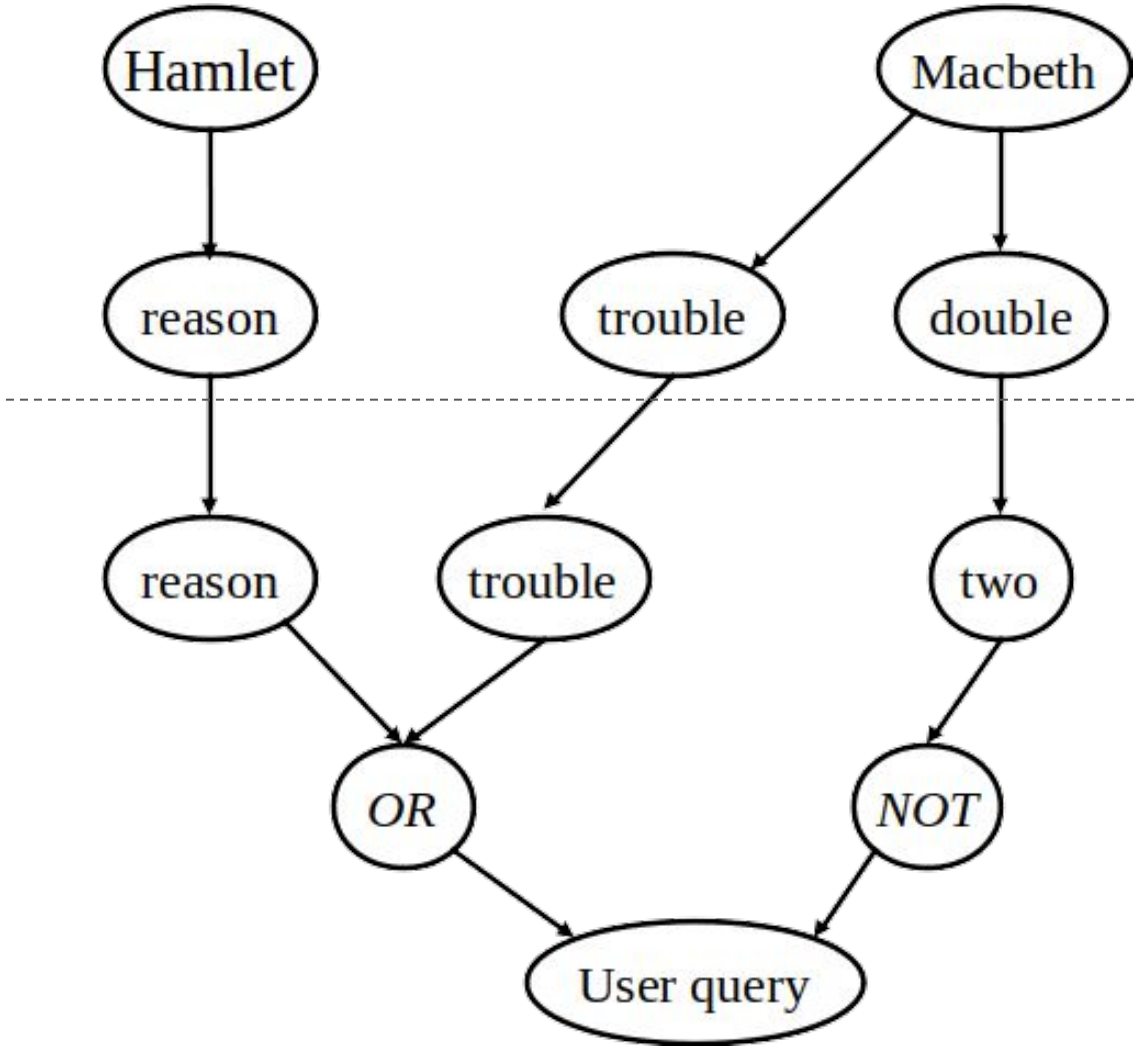
Mạng truy vấn

Các ma trận liên kết và xác suất

- Xác suất tiền nghiệm văn bản
  P(d) = 1/n
- P(r|d)
  - Tần suất từ trong văn bản
  - tf x idf - cơ bản
- P(c|r) - ánh xạ 1-1, từ điển đồng nghĩa
- P(q|c)- dạng chuẩn của toán tử truy vấn
  - Luôn sử dụng các toán tử AND và NOT - không bao giờ lưu bảng xác suất có điều kiện đầy đủ.

Ví dụ 12.2. Mạng Bayes

"Reason trouble - two"



Mạng văn bản

Mạng truy vấn

Mở rộng khái niệm ẩn

- Các khái niệm ẩn là các từ hoặc các câu có trong ý niệm của người dùng nhưng không được sử dụng tường minh trong câu truy vấn của họ.
- Các đại lượng xác suất được ước lượng cho các từ mở rộng

Ví dụ 12.3. Mở rộng khái niệm ẩn (LCE)

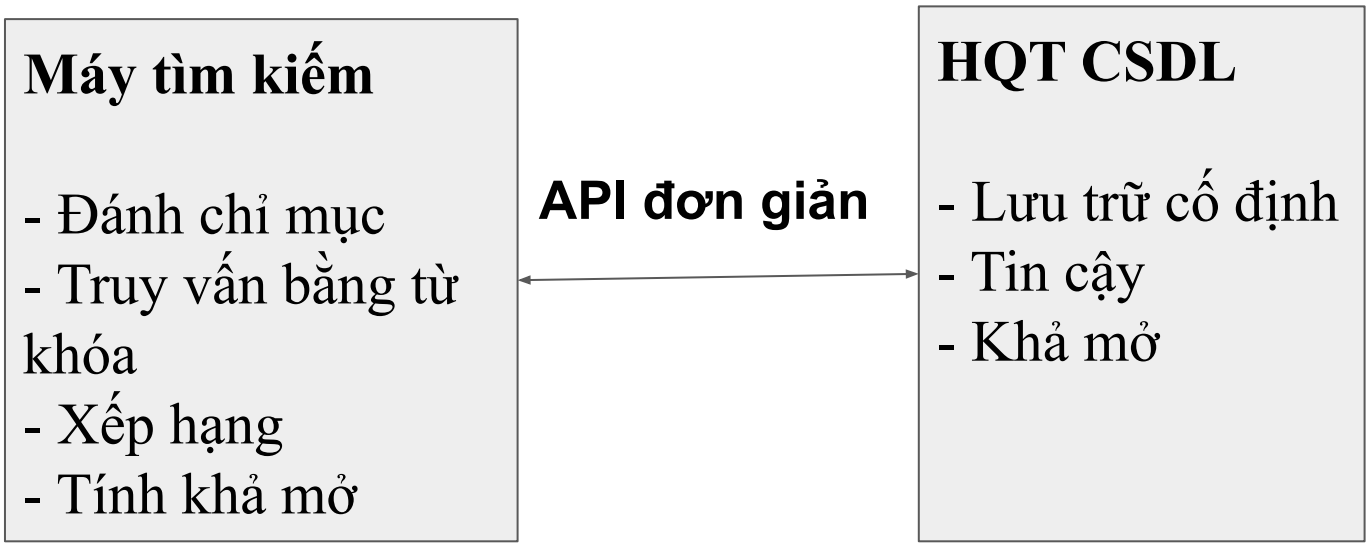
1-word concepts	*2-word concepts*
telescope	hubble telescope
hubble	space telescope
space	hubble space
mirror	telescope mirror
NASA	telescope hubble
launch	mirror telescope
astronomy	telescope NASA
shuttle	telescope space
test	hubble mirror
new	NASA hubble
discovery	telescope astronomy
time	telescope optical
universe	hubble optical
optical	telescope discovery
light	telescope shuttle

Tích hợp CSDL và TKTT

- Một số cách tiếp cận:
  - Mở rộng mô hình CSDL để có thể hỗ trợ các cơ chế xếp hạng và truy xuất dựa trên từ khóa
    - Ví dụ, fulltext search trong MySQL
  - Mở rộng các mô hình TKTT để có thể xử lý các cấu trúc phức tạp và dữ liệu quan hệ
    - Ví dụ, giao diện SQL cho Solr
  - Phát triển 1 mô hình và hệ thống hợp nhất
- Các ứng dụng như thương mại điện tử, khai phá dữ liệu
  - Kết hợp hệ thống TKTT với các hệ quản trị CSDL khác.

Máy tìm kiếm và HQT CSDL

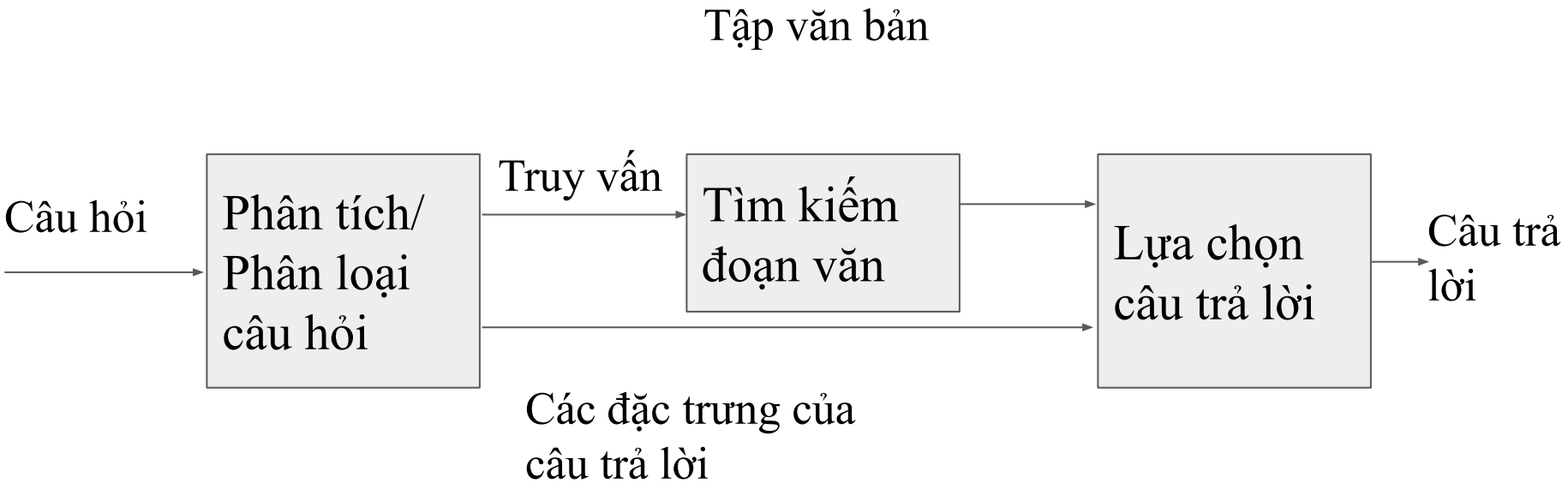
Ví dụ tìm kiếm sản phẩm cho các hệ thống thương mại điện tử



Hỏi đáp

- Đưa ra các câu trả lời trực tiếp thay vì một danh sách văn bản được xếp hạng
  - Sinh câu trả lời
  - Trích rút câu trả lời từ một nguồn lớn
- Hỏi đáp dựa trên các dữ kiện
  - Giới hạn trong phạm vi các câu hỏi ngắn và đơn giản
  - Ví dụ, ai?ở đâu? cái gì?
  - ... Dựa trên các biểu diễn ngữ nghĩa hình thức và các cơ chế suy diễn tự động

Kiến trúc hỏi đáp



Hỏi đáp dựa trên các dữ kiện

- Câu hỏi được phân loại theo loại câu trả lời được mong đợi
  - Hầu hết các danh mục tương ứng với các thực thể có tên
- Danh mục được sử dụng để xác định đoạn văn có khả năng là câu trả lời
- Các xử lý ngôn ngữ tự nhiên và suy diễn ngữ nghĩa được sử dụng để xếp hạng các đoạn văn và xác định câu trả lời

Ví dụ 12.4. Các danh mục

Example Question	*Question Category*
What do you call a group of geese?	Animal
Who was Monet?	Biography
How many types of lemurs are there?	Cardinal
What is the effect of acid rain?	Cause/Effect
What is the street address of the White House?	Contact Info
Boxing Day is celebrated on what day?	Date
What is sake?	Definition
What is another name for nearsightedness?	Disease
What was the famous battle in 1836 between Texas and Mexico?	Event
What is the tallest building in Japan?	Facility
What type of bridge is the Golden Gate Bridge?	Facility Description
What is the most popular sport in Japan?	Game
What is the capital of Sri Lanka?	Geo-Political Entity
Name a Gaelic language.	Language
What is the world's highest peak?	Location
How much money does the Sultan of Brunei have?	Money
Jackson Pollock is of what nationality?	Nationality
Who manufactures Magic Chef appliances?	Organization
What kind of sports team is the Buffalo Sabres?	Org. Description
What color is yak milk?	Other
How much of an apple is water?	Percent
Who was the first Russian astronaut to walk in space?	Person
What is Australia's national flower?	Plant
What is the most heavily caffeinated soft drink?	Product
What does the Peugeot company manufacture?	Product Description
How far away is the moon?	Quantity
Why can't ostriches fly?	Reason
What metal has the highest melting point?	Substance
What time of day did Emperor Hirohito die?	Time
What does your spleen do?	Use
What is the best-selling book of all time?	Work of Art

Các phương tiện khác

- Có nhiều phương tiện truyền thông tin khác quan trọng đối với các ứng dụng tìm kiếm
  - Văn bản được biểu diễn như ảnh, ví dụ bản scan của sách
  - Bản ghi âm,
  - Âm nhạc,
  - Hình ảnh, video
  - v.v.
- Thường không có văn bản đính kèm
  - Ghi chú được tạo bởi con người có vai trò quan trọng trong 1 số trường hợp ứng dụng
- Các giải thuật tìm kiếm dựa trên các đặc trưng được tách ra từ tài nguyên thông tin

Văn bản nhiễu

- Nhận diện ký tự quang học (OCR) và nhận dạng giọng nói (Speech to Text) tạo ra nhiều văn bản chứa nhiễu
  - Văn bản cùng với các ký tự vô nghĩa do lỗi nhận dạng
- Với những mô hình tìm kiếm đủ tốt, thông tin nhiễu thường gây ra những ảnh hưởng không đáng kể tới hiệu quả tìm kiếm
  - Dựa trên lượng lớn nội dung văn bản
  - Xử lý các vấn đề với các văn bản ngắn

Ví dụ 12.5. OCR

Original:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums of all sizes.

OCR:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums ot aH sizes~

Original:

*** This work was carried out under the sponsorship of National Science Foundation Grants NSF-GN-380 (Studies in Indexing Depth and Retrieval Effectiveness) and NSF-GN-432 (Requirements Study for Future Catalogs).**

OCR:

This work was carried out under the sp011J!0rship 01 NatiolUl1 Setenee Foundation 0rant. NSF·0N·SB0 (Studl .. In Indexing Depth and Retrieval Eflccth"ene&&) and NSF·0N·482 (Requirements Study lor Future 'Catalogs)•

Ví dụ 12.6. Chuyển giọng nói thành văn bản

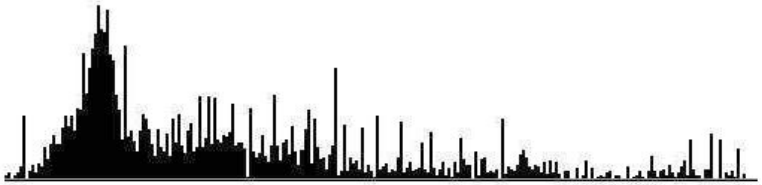
Transcript:
French prosecutors are investigating former Chilean strongman Augusto Pinochet. The French justice minister may seek his extradition from Britain. Three French families whose relatives disappeared in Chile have filed a Complaint charging Pinochet with crimes against humanity. The national court in Spain has ruled crimes committed by the Pinochet regime fall under Spanish jurisdiction.

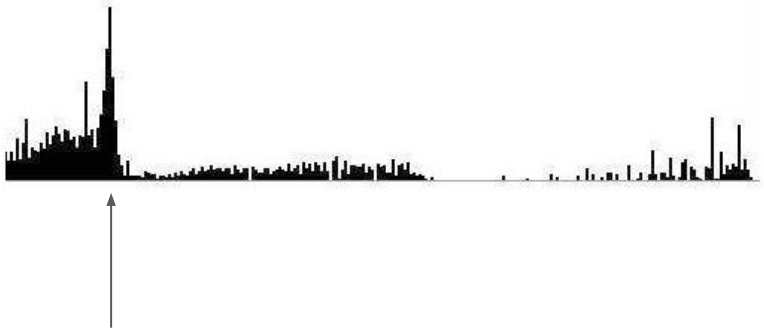
Speech recognizer output:
french prosecutors are investigating former chilean strongman of coastal fish today the french justice minister may seek his extradition from britain three french families whose relatives disappeared until i have filed a complaint charging tenants say with crimes against humanity the national court in spain has ruled crimes committed by the tennessee with james all under spanish jurisdiction

Hình ảnh và video

- Trích rút các đặc trưng từ hình ảnh và video thường khó hơn so với văn bản
- Các đặc trưng ở tầng thấp ít gắn kết trực tiếp với nội dung của hình ảnh như các mô tả văn bản
- Các đặc trưng thông dụng gắn với: Mầu sắc, hoa văn, hình khối
  - Ví dụ biểu đồ mầu

Ví dụ 12.7. Biểu đồ mầu





Có đỉnh là mầu vàng

Hoa văn và hình khối

- Hoa văn là sự xếp đặt các thang xám của hình ảnh trong khung ảnh
- Các đặc trưng hình khối mô tả hình dạng của đường viền của đối tượng và các cạnh
- Ví dụ:

← Hình khối

Video

- Được phân đoạn thành các hình hoặc phân cảnh
  - Chuỗi liên tục các khung hình có gắn kết trực quan
  - Biên được xác định bởi sự gián đoạn trực quan
- Video được biểu diễn bởi các khung hình chính
  - Ví dụ những hình ảnh đầu tiên.

Ghi chú ảnh

- Cho các dữ liệu huấn luyện, có thể học mô hình xác suất liên kết cho các từ mô tả với các đặc trưng hình ảnh
- Cho phép tự động tạo ghi chú văn bản cho hình ảnh
  - Cho phép tìm kiếm hình ảnh bằng từ
  - Các kỹ thuật hiện có hiệu quả ở mức trung bình

people, pool, swimmers, water

cars, formula, tracks, wall

clouds, jet, plane, sky

fox, forest, river, water

Lỗi

Âm nhạc

- Âm nhạc còn ít liên kết với từ hơn so với hình ảnh
- Có nhiều biểu diễn khác nhau
  - Ví dụ, audio, MIDI
- Tìm kiếm dựa trên các đặc trưng như
  - Giai điệu tương đối,
  - Nhịp điệu
  - Thang âm, hợp âm