ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

MỤC TIÊU MÔN HỌC

1) Hiểu các khái niệm, các vấn đề trong truy xuất thông tin
2) Áp dụng các kiến thức liên quan để đề xuất phương pháp phân tích văn bản.
3) Cài đặt thử nghiệm các mô hình truy xuất thông tin cơ bản
4) Đánh giá và so sánh các mô hình truy xuất thông tin
5) Xây dựng một Search Engine đơn giản.

NỘI DUNG MÔN HỌC

- CHƯƠNG I: DẪN NHẬP
- CHƯƠNG II: MÔ HÌNH KHÔNG GIAN VECTOR
- CHƯƠNG III: ĐÁNH GIÁ MÔ HÌNH TRUY XUẤT THÔNG TIN
- **CHƯƠNG IV: XÂY DỰNG SEARCH ENGINE**
- **CHƯƠNG V: MỘT SỐ MÔ HÌNH TRUY XUẤT THÔNG TIN KHÁC**
- **CHƯƠNG VI: PHÂN LỚP VĂN BẢN**

ĐÁNH GIÁ MÔN HỌC

- ĐỒ ÁN: 50% điểm tổng kết.
  - Thuyết trình theo nhóm: 50% điểm đồ án
  - Báo cáo + phần mềm: 50% điểm đồ án
- THI CUỐI KỲ: 50% điểm tổng kết

QUY ĐỊNH:

- Tham dự trên 80% số buổi học lý thuyết.
- Không thuyết trình thì không chấm điểm đồ án.
- Mỗi nhóm chỉ 2 đến 3 sinh viên. Nếu không tìm được nhóm có thể làm một mình nhưng không khuyến khích.

TÀI LIỆU

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

THẢO LUẬN

- Trên lớp
- Website môn học trên moodle của trường:
  - Địa chỉ: courses.uit.edu.vn
  - Đăng nhập bằng tài khoản của nhà trường
  - Thảo luận tất cả các vấn đề về môn học
- Email cá nhân (không khuyến khích)
  - Địa chỉ: chinhnt@uit.edu.vn
  - Chỉ sử dụng khi cần phải trao đổi khi không thể dùng website môn học.

ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG I - DẪN NHẬP

Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

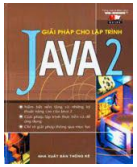
NỘI DUNG TRÌNH BÀY

- GIỚI THIỆU
- CÁC MÔ HÌNH TRUY XUẤT THÔNG TIN CĂN BẢN
- LẬP CHỈ MỤC
- TRUY VẤN CHỈ MỤC

GIỚI THIỆU

❖KHÁI NIỆM

Lập trình java



HỆ THỐNG TRUY XUẤT THÔNG TIN

GIỚI THIỆU

❖KHÁI NIỆM

Truy xuất thông tin là tìm kiếm

- Vật liệu chứa thông tin (tài liệu – Document).
- Phi cấu trúc hoặc bán cấu trúc: Free Text, XML
- Từ các tập lưu trữ lớn (Collections).
- Thỏa yêu cầu.

GIỚI THIỆU

❖KHÁI NIỆM

Các dạng tài liệu:

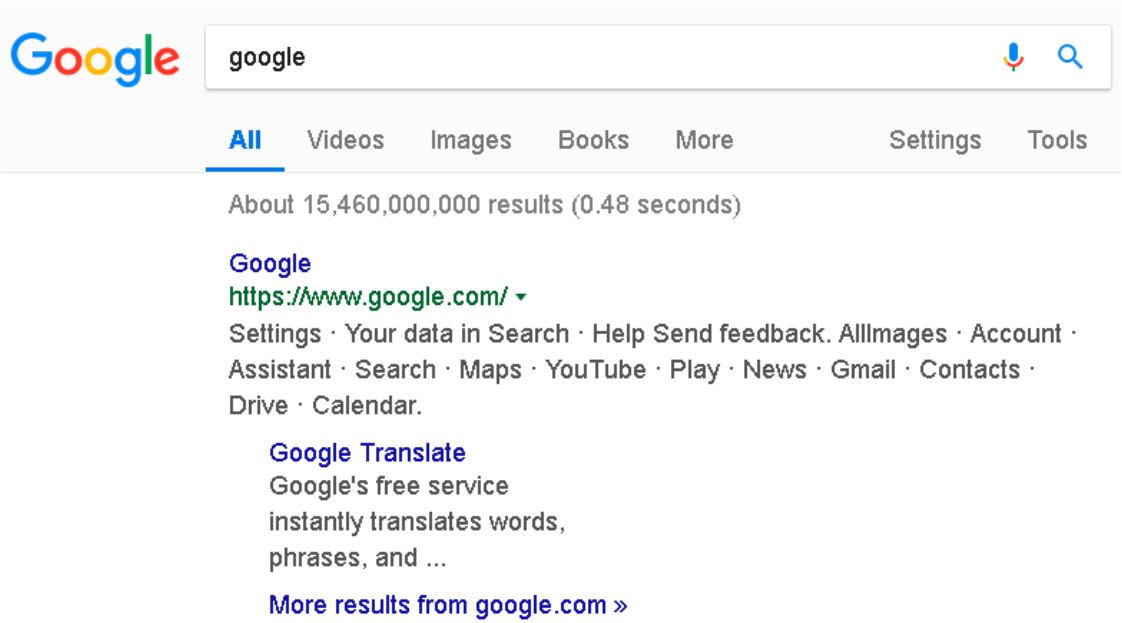
- Email
- Tập tin trên máy tính cá nhân
- Hệ thống văn bản pháp lý
- Các cơ sở tri thức
- Hình ảnh
- Âm thanh
- Video
- …..

GIỚI THIỆU

❖KHÁI NIỆM

Một số hệ thống truy xuất thông tin:

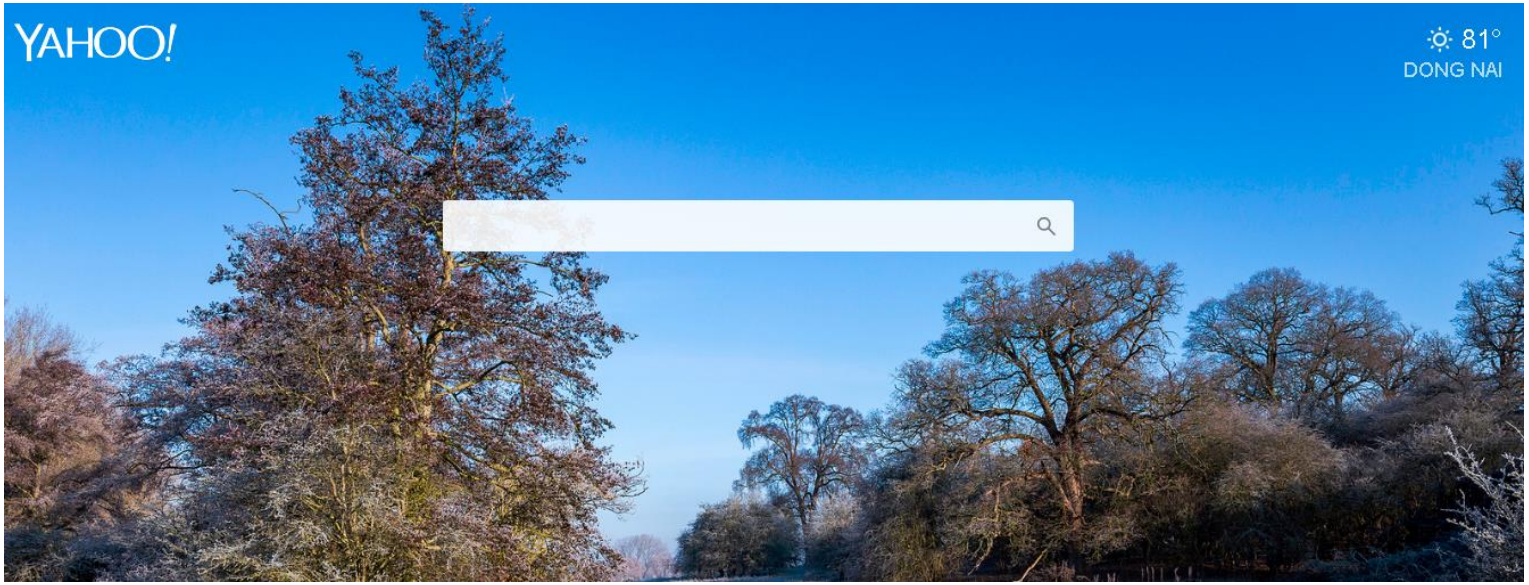
- Internet search engines:
  - Google



GIỚI THIỆU

❖KHÁI NIỆM

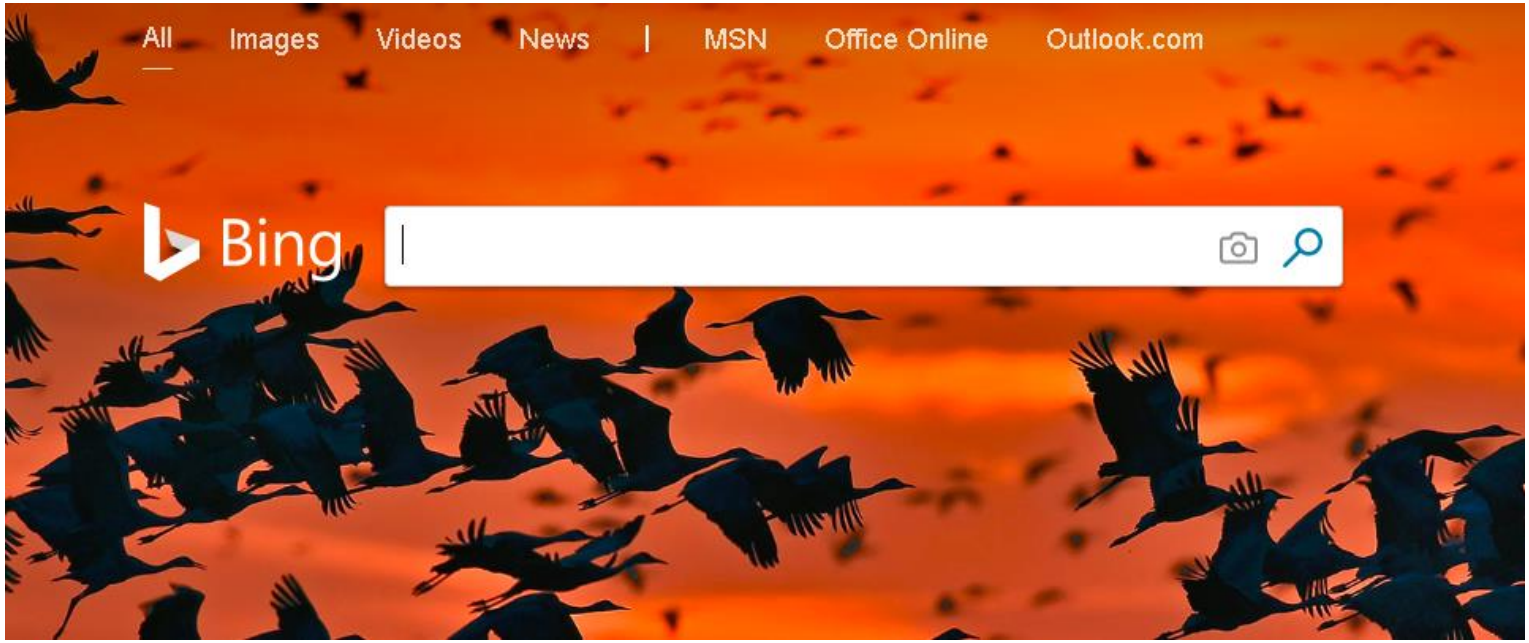
- Internet search engines:
  - Yahoo! Web search



GIỚI THIỆU

❖KHÁI NIỆM

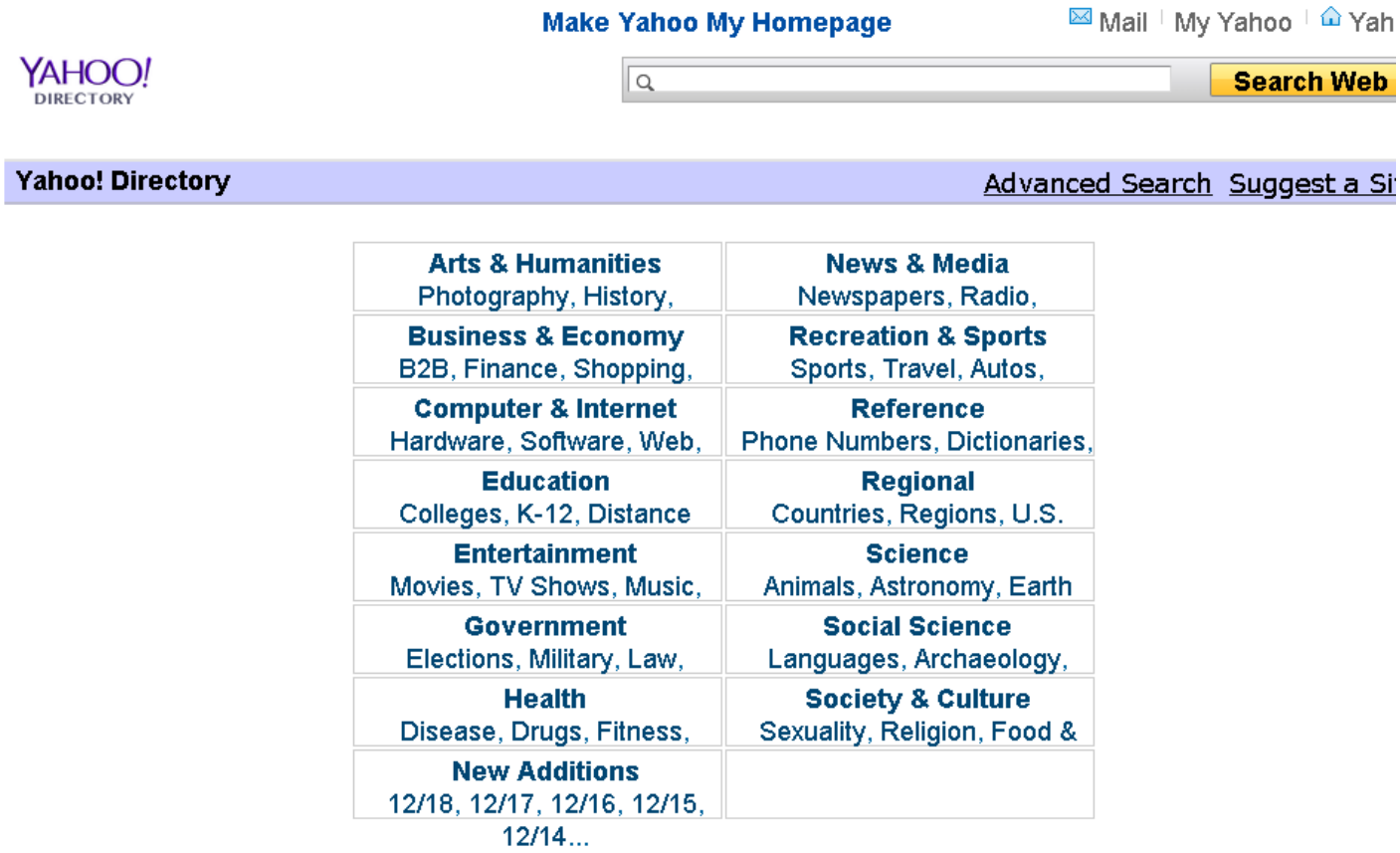
- Internet search engines:
  - Bing



GIỚI THIỆU

❖KHÁI NIỆM

- Danh bạ thư viện số: Yahoo!Directory



GIỚI THIỆU

❖KHÁI NIỆM

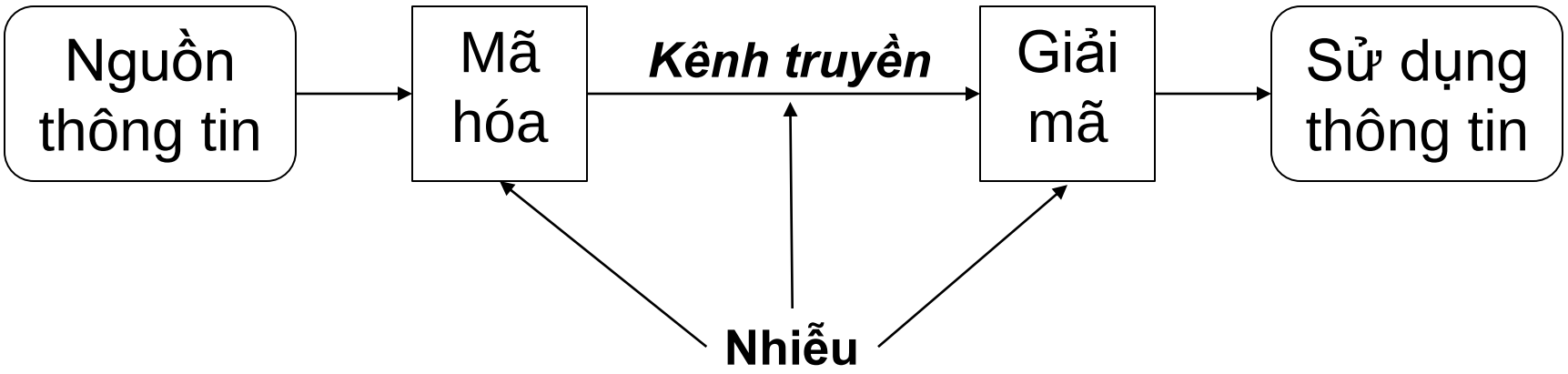
Một số ứng dụng của truy xuất thông tin:

- Truy xuất thông tin xuyên ngôn ngữ (Cross-lingual IR)
- Truy xuất giọng nói, bản tin của đài phát thanh.
- Phân loại văn bản
- Tóm tắt văn bản
- Truy xuất thông tin có cấu trúc (XML)
- Truy xuất thông tin địa lý (Geographic IR)

GIỚI THIỆU

❖KHÁI NIỆM

Mô hình truyền tin trong Lý thuyết thông tin:

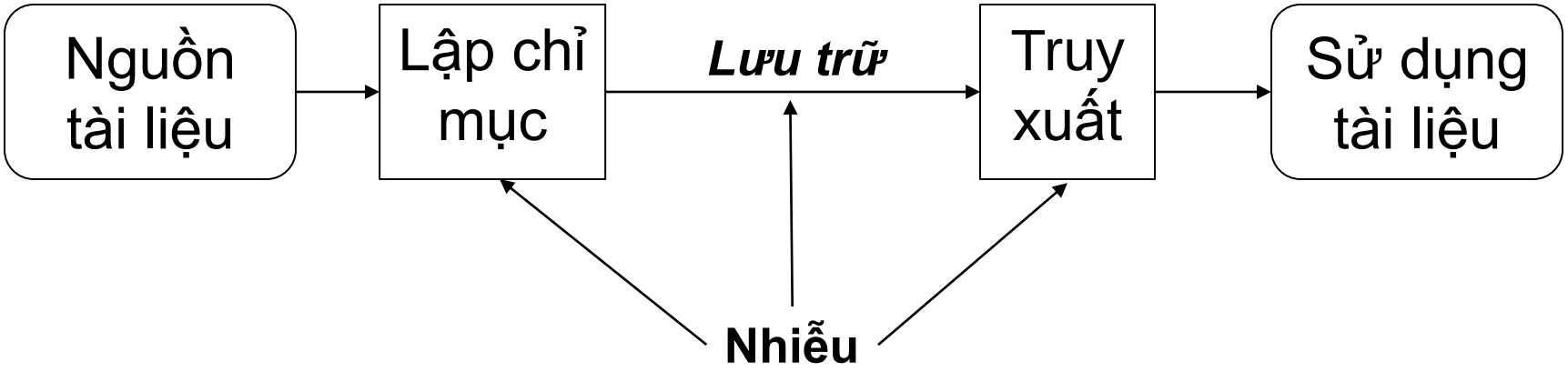


GIỚI THIỆU

❖KHÁI NIỆM

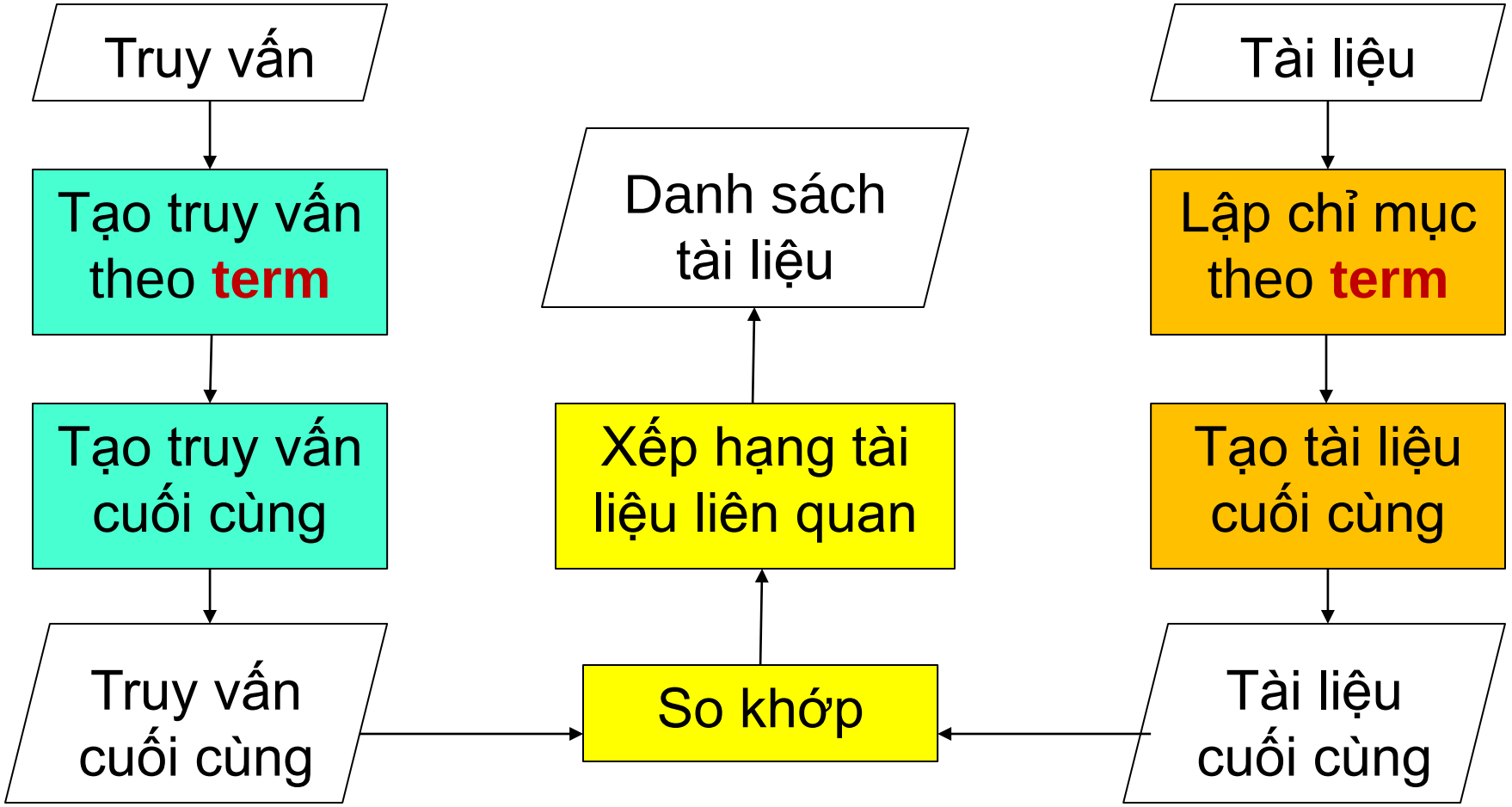
Truy xuất thông tin bắt nguồn từ lý thuyết thông tin:

Trong đó, chữ viết là dạng mã hóa của thông tin.



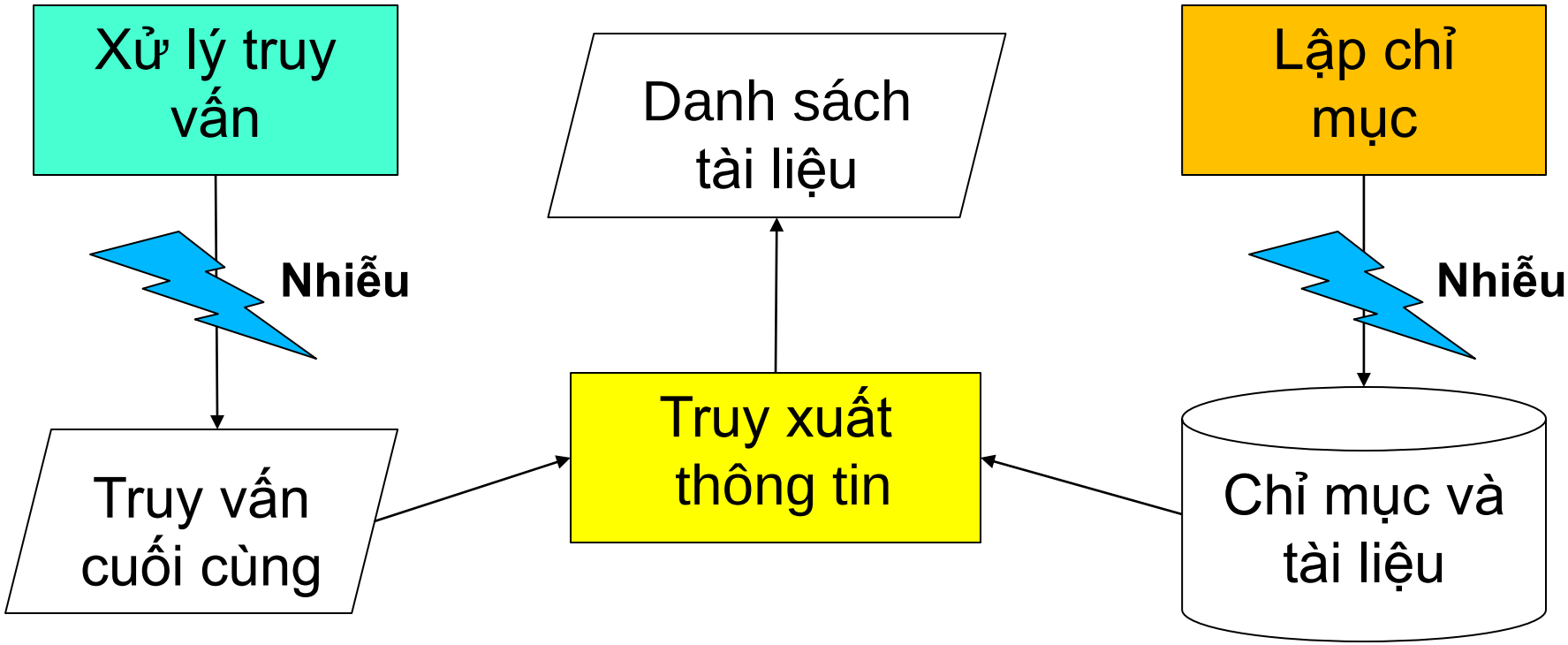
GIỚI THIỆU

❖MÔ HÌNH ĐƠN GIẢN



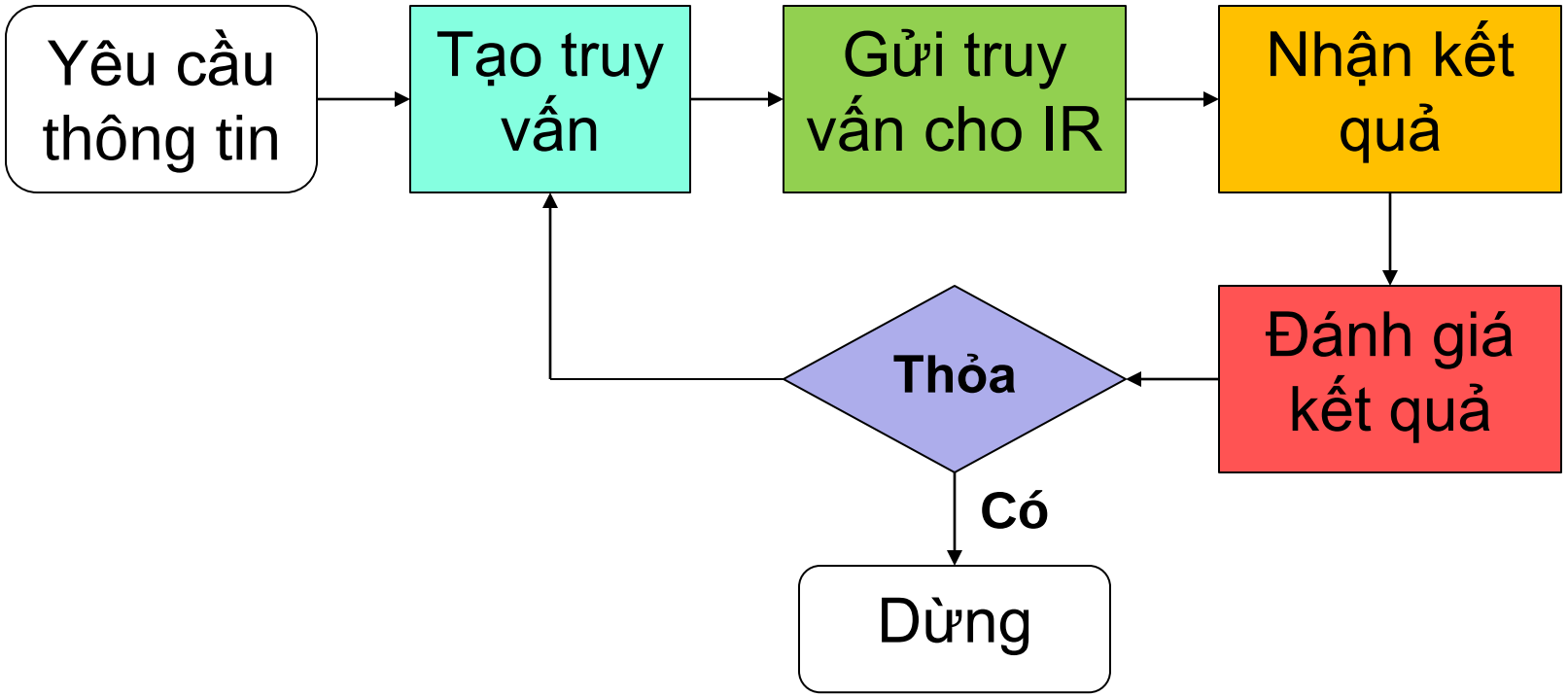
GIỚI THIỆU

- KIẾN TRÚC HỆ THỐNG ĐƠN GIẢN



GIỚI THIỆU

❖SƠ ĐỒ TƯƠNG TÁC VỚI HỆ THỐNG IR



GIỚI THIỆU

❖MÔ HÌNH CHUẨN CỦA IR

Mô hình chuẩn của IR được xây dựng dựa trên các giả thiết sau:

- Mục tiêu: nhằm cực đại hóa đồng thời độ phủ (recall) và độ chính xác (precision) của kết quả tìm kiếm
- Nhu cầu thông tin không thay đổi trong quá trình thực hiện
- Giá trị của mô hình là tập tài liệu được trả về.

GIỚI THIỆU

❖MÔ HÌNH CHUẨN CỦA IR

Một số nhược điểm của mô hình chuẩn:

- Người sử dụng phải thực hiện các việc sau:
  - Đọc tựa của các tài liệu tìm được
  - Đọc tài liệu tìm được
  - Xem lại danh sách các chủ đề hoặc thuật ngữ liên quan
  - Phải duyệt các liên kết.
- Nhiều người sử dụng chỉ muốn trả về một vài tài liệu thực sự cần thiết.

GIỚI THIỆU

❖LỊCH SỬ TRUY XUẤT THÔNG TIN

- Xuất phát từ danh mục các điển tích trong kinh thánh (được thực hiện bởi 500 tu sĩ vào năm 1247)
- Các hệ thống chỉ mục của các bài báo, bài viết từ thế kỷ 17.
- Sau chiến tranh thế giới thứ 2, Cranfield có nghiên cứu đầu tiên về ngôn ngữ chỉ mục và truy xuất thông tin.
- Năm 1951, kết quả nghiên cứu của Bagley cho thấy thời gian tìm kiếm trong tập 50 triệu tài liệu được lập chỉ mục với 30 từ vựng cần 41700 giờ.

GIỚI THIỆU

❖LỊCH SỬ TRUY XUẤT THÔNG TIN

- Năm 1960, Mô hình truy xuất thông tin Boolean được đề xuất.
- Năm 1961, dịch vụ tìm kiếm abstract trong lĩnh vực hóa học trên máy tính được công bố. Hệ thống chỉ mục Medicus của Thư viện y khoa quốc gia được công bố dưới dạng cơ sở dữ liệu có tên MEDLARS
- Năm 1970, tất cả sản phẩm phụ, như hệ thống chỉ mục, các abstract, được xuất bản đều được làm từ máy tính nhờ cơ sở dữ liệu.

GIỚI THIỆU

❖LỊCH SỬ TRUY XUẤT THÔNG TIN

- Năm 1970, hệ thống truy xuất thông tin SMART theo mô hình không gian vector được Salton công bố.
- Năm 1972, Hệ thống truy xuất thông tin pháp lý toàn văn theo mô hình Boolean được công bố với tên LEXIS.
- Những năm 70, truy vấn được phân tích theo xác suất.
- Những năm 80, các vấn đề về tập mờ, logic mờ và suy diễn được áp dụng trong truy xuất thông tin.

GIỚI THIỆU

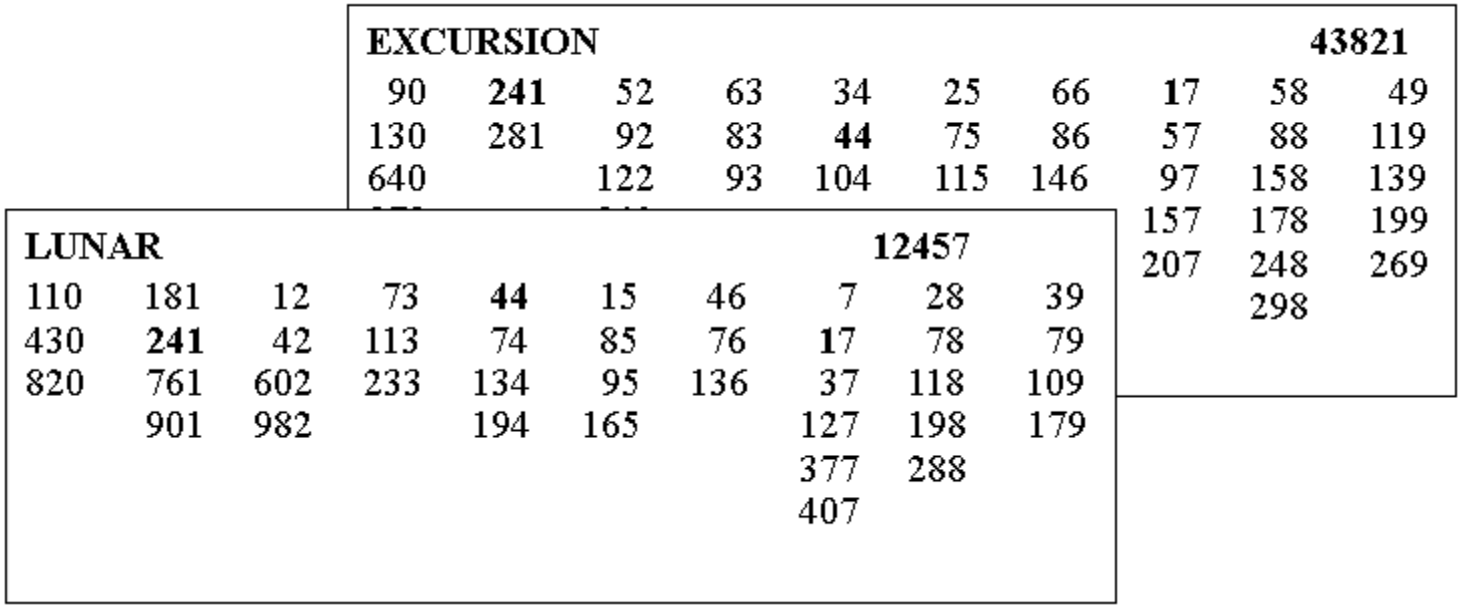
❖LỊCH SỬ TRUY XUẤT THÔNG TIN

- Những năm 90, các vấn đề về mạng neural, mạng suy diễn, chỉ mục ngữ nghĩa tiềm ẩn (Latent semantic index) được nghiên cứu và áp dụng vào truy xuất thông tin.

GIỚI THIỆU

❖LỊCH SỬ TRUY XUẤT THÔNG TIN

Minh họa: thẻ được dùng để tìm tài liệu trong thư viện



Uniterm (Casey, Perry, Berry, Kent – 1958)

GIỚI THIỆU

- MỘT SỐ TẠP CHÍ, KỶ YẾU HỘI NGHỊ

- ACM Transaction on Information Systems
- Am. Society for Information Science Journal
- Document Analysis and IR Proceedings (Las Vegas)
- Information Processing and Management (Pergammon)
- Journal of Document
- SIGIR Conference Proceedings
- TREC Conference Proceedings