Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

TÀI LIỆU LÝ THUYẾT KHAI THÁC WEB

Chủ đề 4

KHAI THÁC CẤU TRÚC WEB (PHẦN 1)

Giảng viên: ThS. Lê Ngọc Thành
Email: lnthanh@fit.hcmus.edu.vn

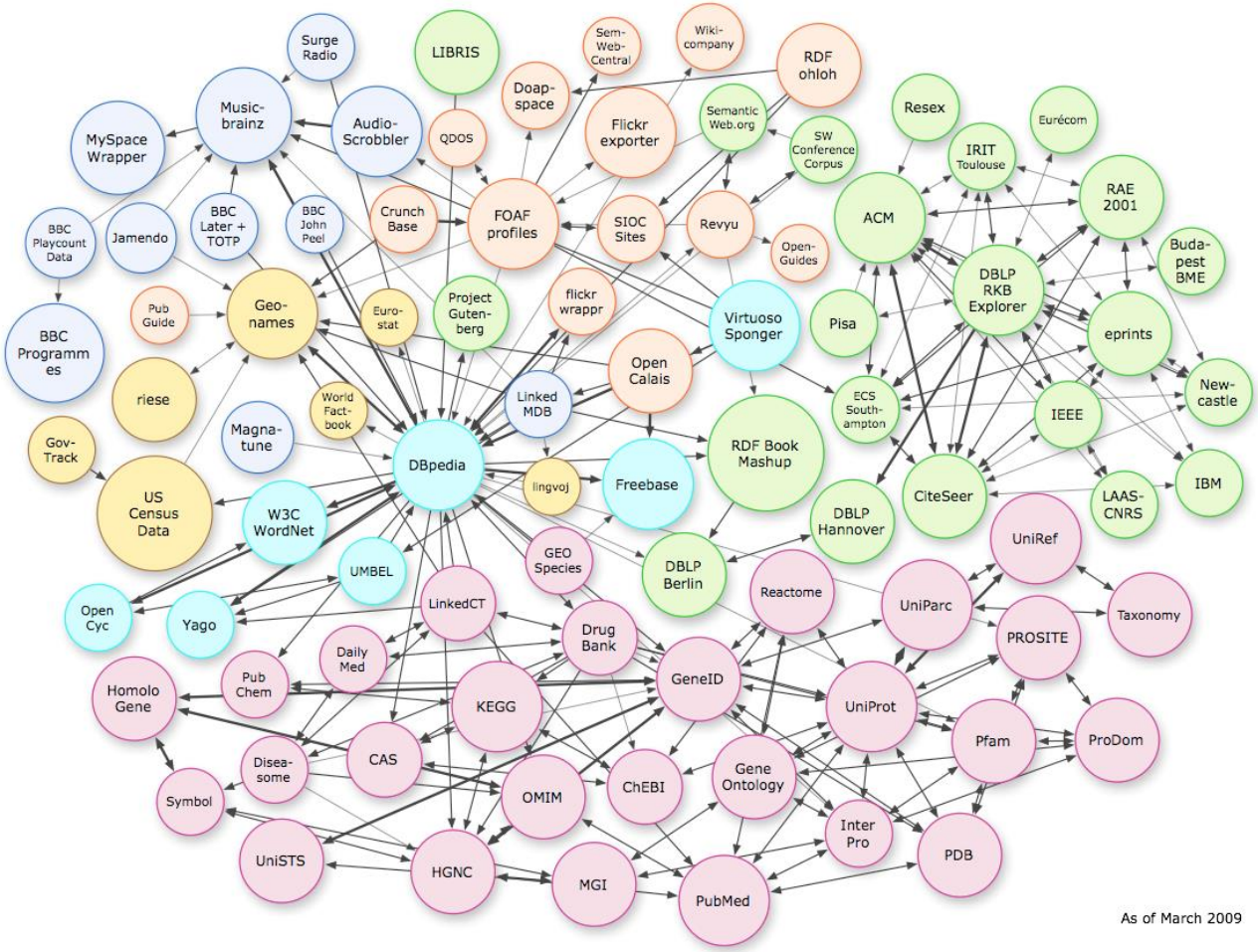
HKII/2011-2012

NỘI DUNG

- Khái niệm khai thác cấu trúc Web
- Đặc điểm của khai thác cấu trúc Web
- Một số cấu trúc Web phổ biến
- Giới thiệu Web crawler
- Giới thiệu một số thuật toán
- Ứng dụng

Khai thác cấu trúc web

- Khai thác cấu trúc Web là khám phá ra dữ liệu hữu ích từ siêu liên kết.



Đặc điểm khai thác cấu trúc Web

- Phát sinh ra một *cấu trúc tóm tắt* về Website và trang Web (Web page).
- Sử dụng *lý thuyết đồ thị* để phân tích node và cấu trúc kết nối của một website.
- Từ đó xác định các tài liệu có độ chính xác cao hơn.

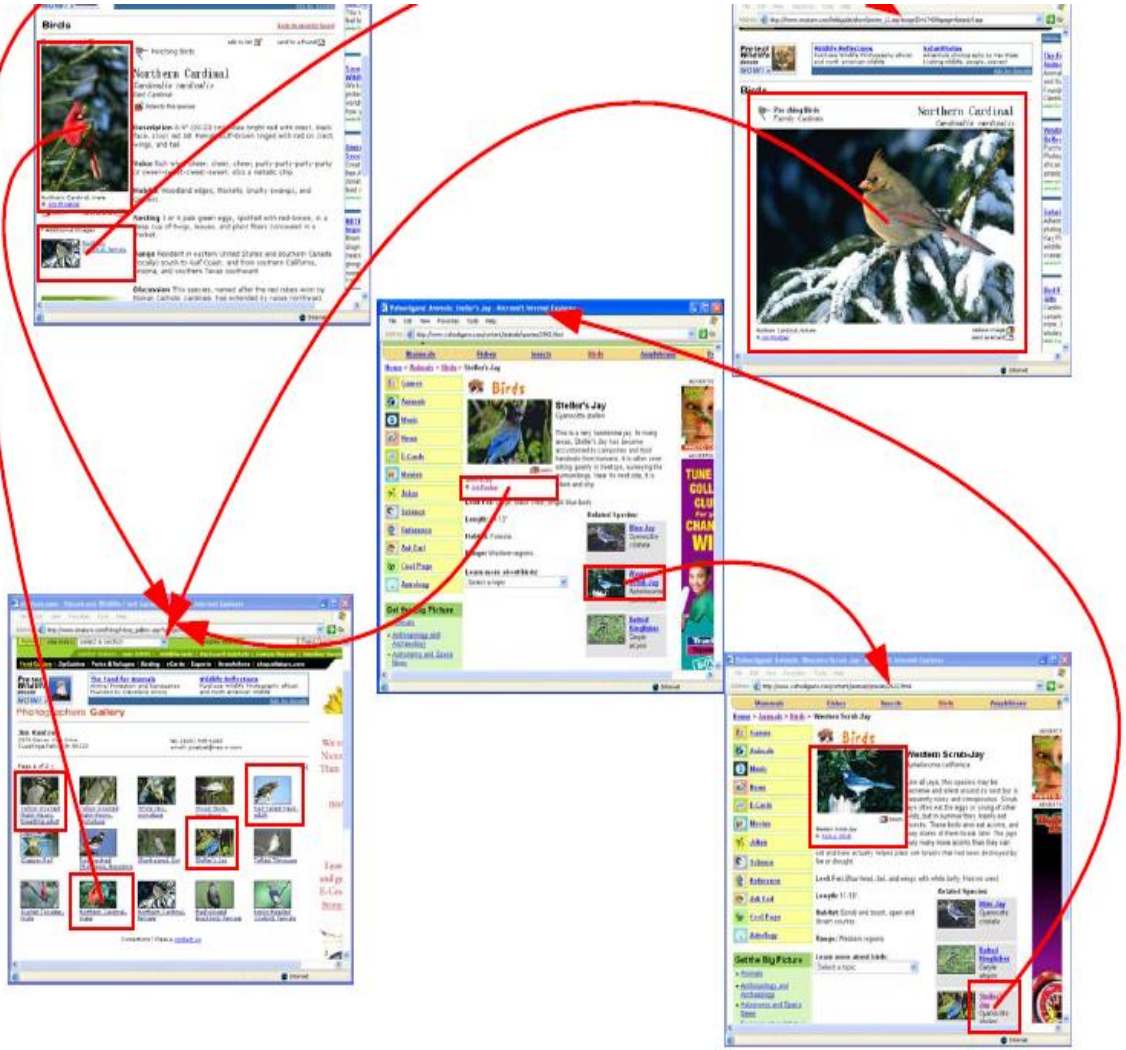
Đặc điểm KTCT Web (tt)

- Khám phá ra *tính tự nhiên của phân tầng liên kết* trong website và cấu trúc của nó.
- Siêu liên kết (*hypelink*) xác định sự chứng thực của tác giả đến các trang web khác.
- Truy vấn thông tin về độ tương quan và chất lượng của trang Web.

Đặc điểm KTCT Web (tt)

- Khai thác có thể được thực thi ở mức tài liệu (*intra-page*) hay ở mức siêu liên kết (*inter-page*)
- Nghiên cứu ở mức siêu liên kết cũng được gọi là phân tích liên kết (*hyperlink analysis*)

Ví dụ về layout và phân tích liên kết cho các hình ảnh Web



Wei-Ying Ma, Microsoft Research Asia

Một số khái niệm Web

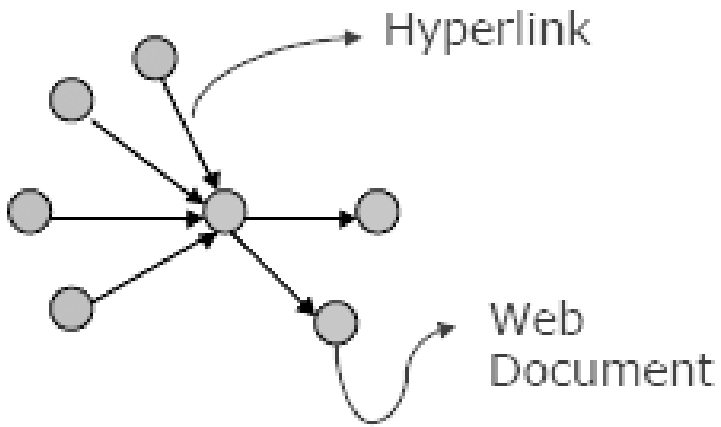
- Một Web là một tập hợp rất lớn của các tài liệu được liên kết với nhau bằng các tham chiếu (*reference*).
- Các tham chiếu được mô tả bằng những siêu từ hay từ neo (*anchor text*) và được nhúng trong HTML

Một số khái niệm Web (tt)

- HTML mô tả như thế nào tài liệu nên được trình diễn trên cửa sổ duyệt.
- URL để xác định tính đơn nhất của website

Các cấu trúc Web phổ biến

- *Cấu trúc đồ thị web*: bao gồm các trang Web là node và siêu liên kết là cạnh kết nối giữa hai trang liên quan.
- Trong khi *truy vấn thông tin* tập trung trên thông tin được cung cấp bởi các chữ trong tài liệu, web còn cung cấp *thông tin thêm* thông qua cách các tài liệu được kết nối với nhau.



Web Graph Structure

Đồ thị cấu trúc Web

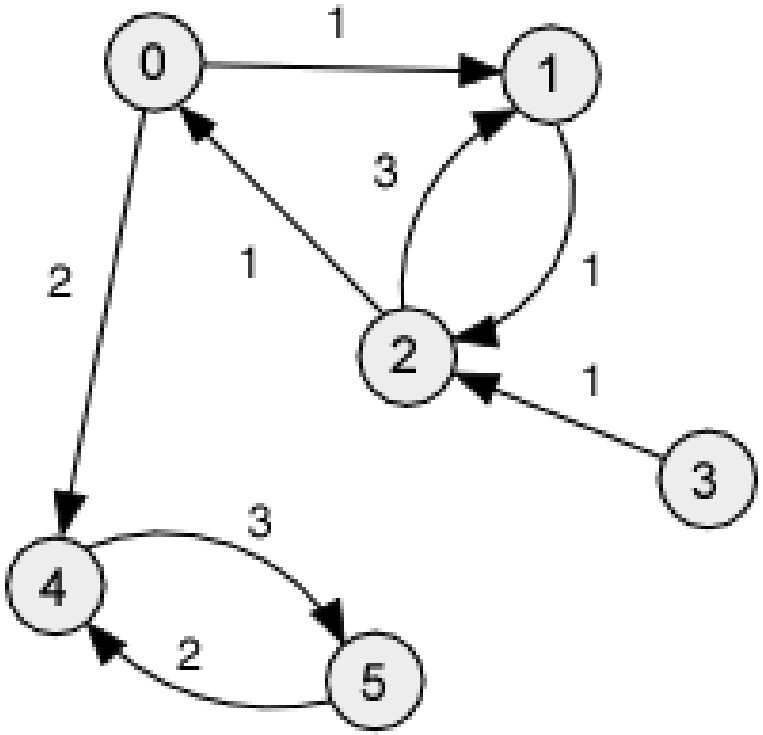
- *Đồ thị Web*: một đồ thị có hướng trong đó:
  - Mỗi node p là một trang web.
  - Cạnh có hướng là siêu liên kết trên Web.
  - Bậc trong của p: là số liên kết phân biệt trỏ đến p.
  - Bậc ngoài của p: là số liên kết phân biệt xuất phát từ p trỏ đến node khác.

Đồ thị cấu trúc Web (tt)

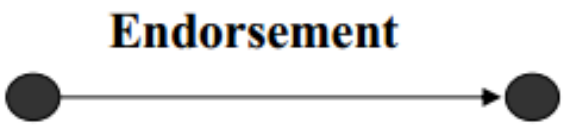
- *Đường đi có hướng*: là chuỗi tuần tự các liên kết từ điểm p đến q.
- *Đường đi ngắn nhất*: là đường đi có số liên kết ít nhất trong số các đường từ p đến q.
- *Đường kính của một đồ thị*: là giá trị lớn nhất của đường đi ngắn nhất giữa tất cả các cặp node trong đồ thị.
- *Khoảng cách kết nối trung bình*: là trung bình chiều dài các đường đi ngắn nhất của các cặp node trong đồ thị.

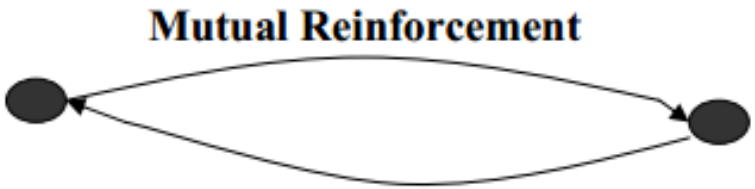
Xác định các khái niệm trên

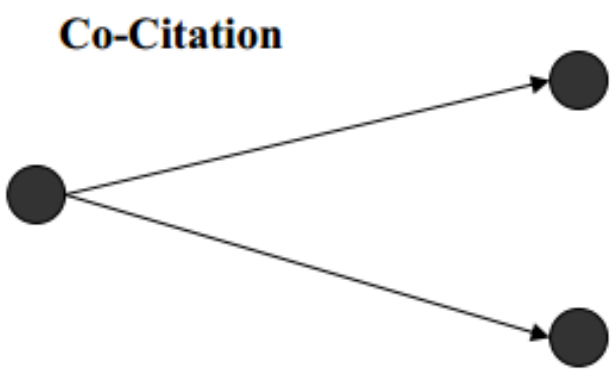
Xác định đỉnh, cạnh, đường đi ngắn nhất, đường kính, khoảng cách kết nối trung bình…

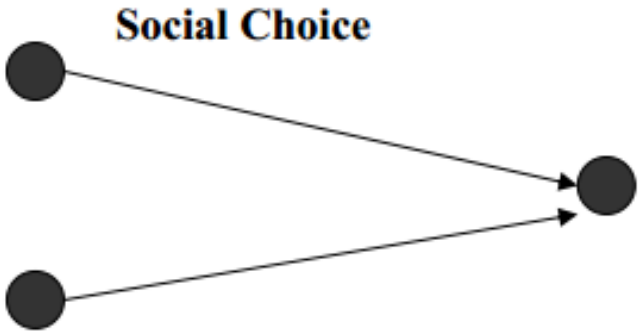


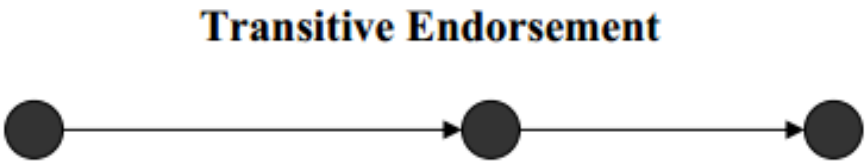
Một số loại liên kết





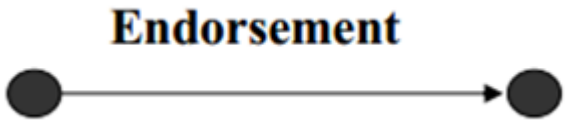




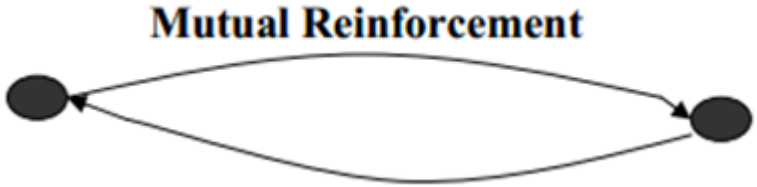


Một số loại LK (tt)

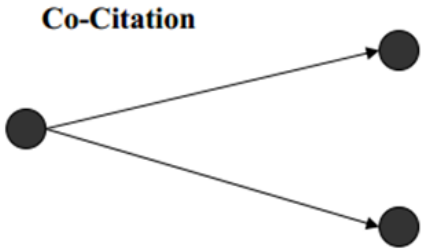
- *Endorsement* là liên kết từ nguồn đến đích thể hiện sự tham chiếu



- *Mutual reinforcement* là liên kết thể hiện sự tham chiếu lẫn nhau nhằm làm tăng cường thông tin từ hai phía.

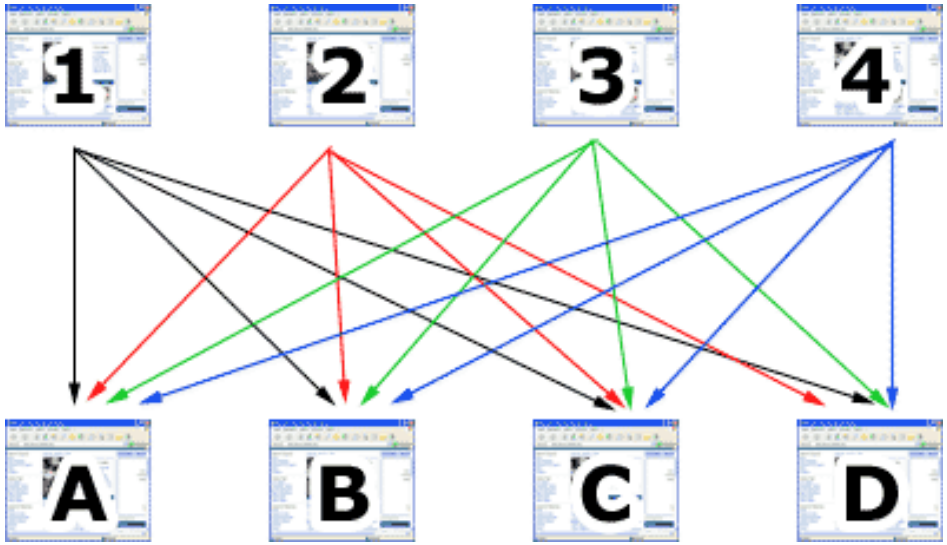


Một số loại LK (tt)



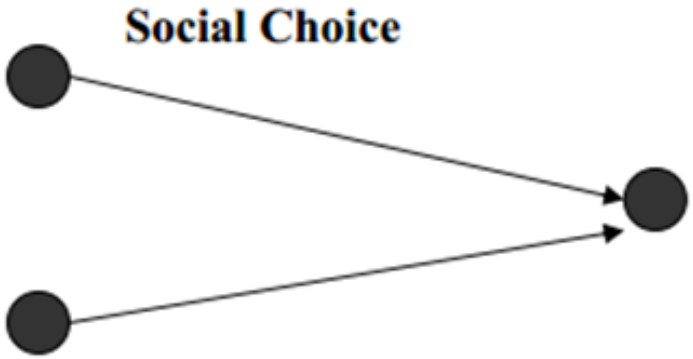
- *Co-citation* là liên kết đồng thời từ một nguồn đến các nguồn khác.

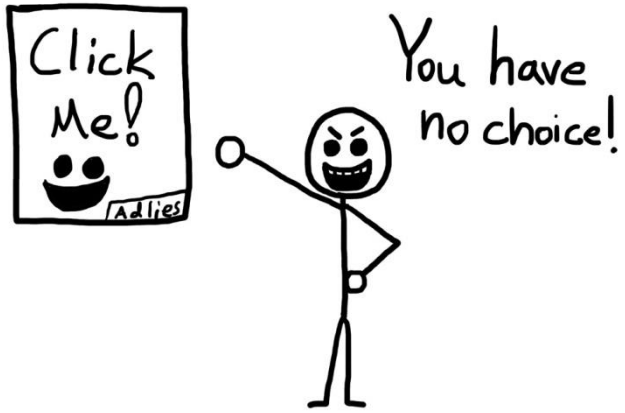
  Ví dụ: website 1,2,3 và 4 liên kết đến website A,B,C,D. Mặc dù A,B,C,D không liên kết với nhau nhưng công cụ tìm kiếm vẫn nghĩ là có bởi có cùng trang web liên kết đến chúng.



Một số loại LK (tt)

- *Social choice* là liên kết thể hiện tính thú vị của cộng đồng thông qua sự bình chọn hay liên kết đến một đối tượng (website).

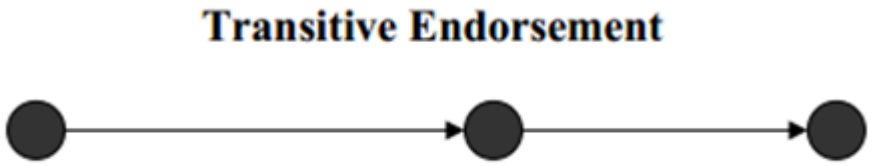




Một số loại LK (tt)

- *Transitive endorsement* là tham chiếu ngầm giữa hai website không có tham chiếu trực tiếp nhưng lại có tham chiếu đến website thứ ba.

  Ví dụ: A tham chiếu B, B cham chiếu C. Như vậy A được gọi là tham chiếu ngầm đến C.

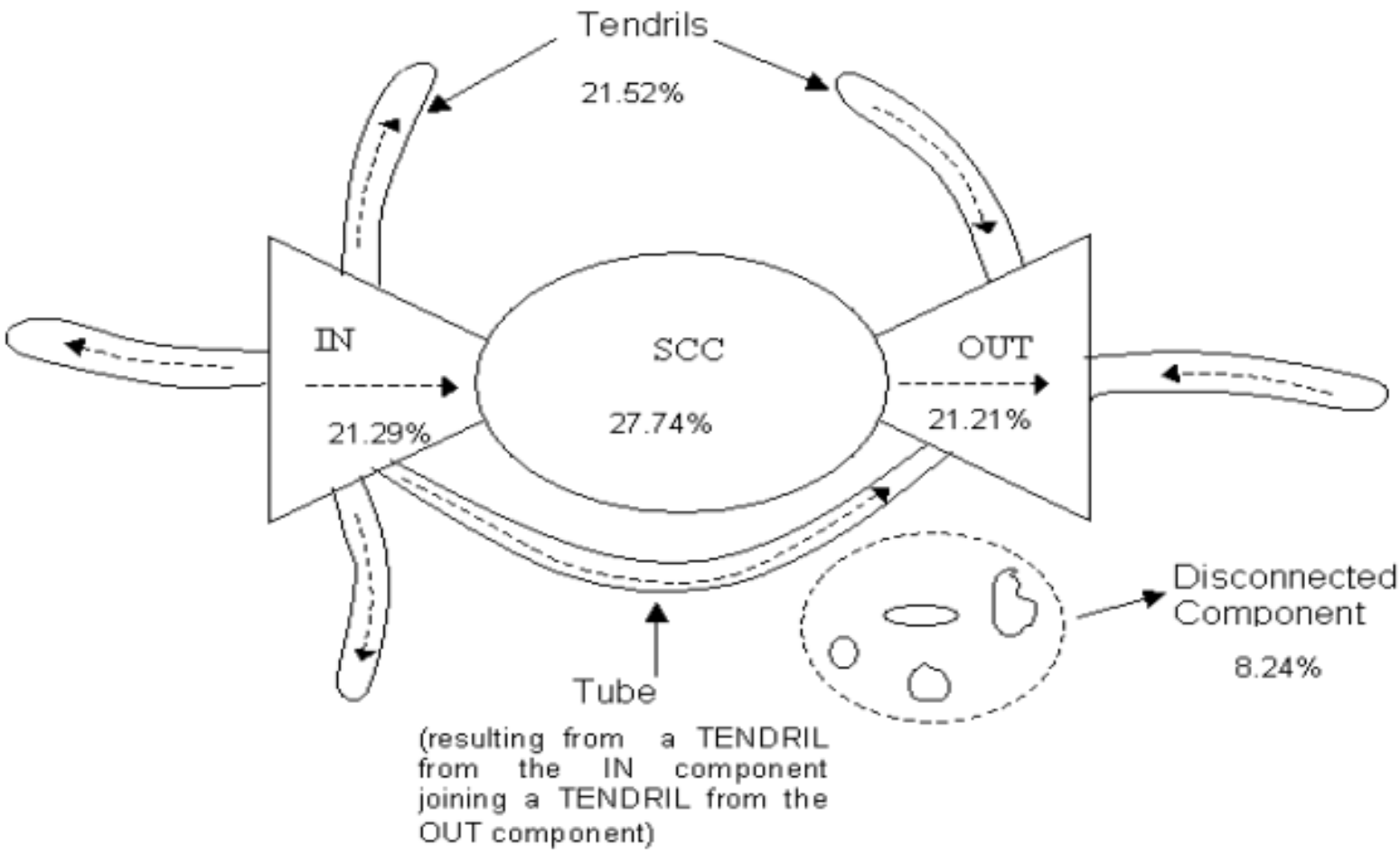


Các bước phân tích siêu liên kết

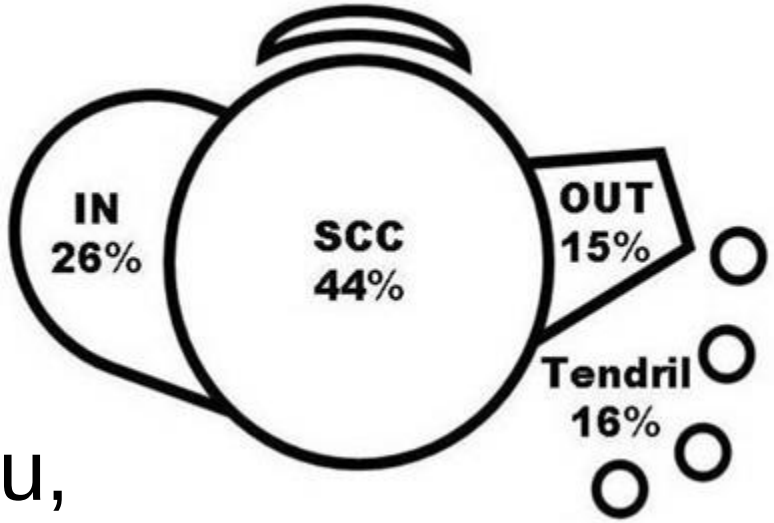
- *Mô hình hóa tri thức (knowledge model)*: thể hiện bên dưới để hình thành nên cơ sở cho ứng dụng với nhiệm vụ cụ thể.
- *Phân tích phạm vi và thuộc tính (analysis scope and properties)*: phân tích phạm vi là xác định nhiệm vụ chỉ liên quan đến một node hay một tập node hay toàn đồ thị. Thuộc tính là đặc trưng của một node đơn hay tập các node hay toàn bộ trang web.
- *Các độ đo và thuật toán*: độ đo là những chuẩn của thuộc tính như chất lượng, độ liên quan hay khoảng cách giữa các node. Thuật toán được thiết kế để tính toán hiệu quả những độ đo này.

Mô hình Bow-Tie của Web

- Một cách nhìn tổng thể về Web

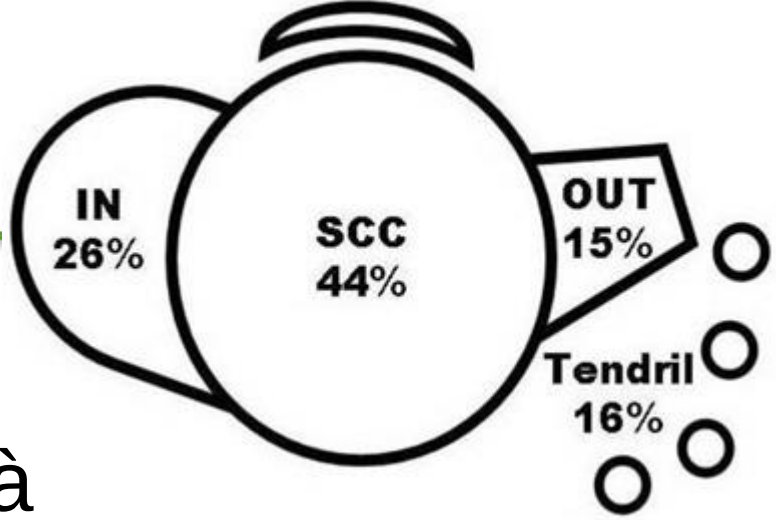


Mô hình …(tt)

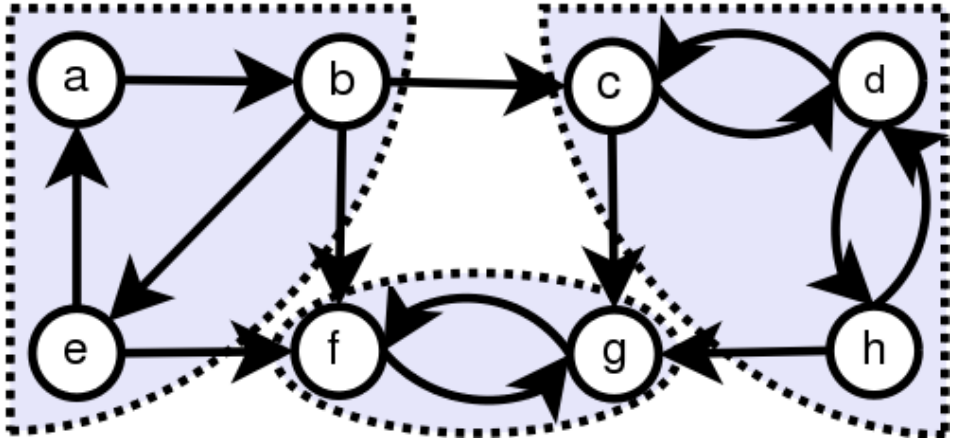


- Khối **SCC** là lõi tương tác lẫn nhau, hay là thành phần kết nối mạnh (**S**trongly **C**onnected **C**omponent).
- Khối **Tendril** chứa các trang không có liên kết đến SCC cũng như không có liên kết từ SCC đến.

Mô hình …(tt)



- Một đồ thị có hướng được gọi là kết nối mạnh nếu mỗi đỉnh trong đồ thị đều có liên kết đến các đỉnh khác.



Đồ thị với các thành phần kết nối mạnh được đánh dấu viền khung

NỘI DUNG

- Khái niệm khai thác cấu trúc Web
- Đặc điểm của khai thác cấu trúc Web
- Một số cấu trúc Web phổ biến
- Giới thiệu Web crawler
- Giới thiệu một số thuật toán
- Ứng dụng

Giới thiệu Web crawler

- *Web crawler* là một bộ công cụ dùng để thu thập tất cả tài liệu web bằng cách duyệt Web có hệ thống và tường tận.
- *Miền của trang web* được bò (crawl) có thể xác định bằng cách sử dụng cấu trúc URL
- Được sử dụng bởi một cổ máy tìm kiếm để cung cấp vị trí truy xuất đến các phiên bản mới nhất của tất cả các trang Web có thể.

Nhắc lại tìm kiếm Web

- Có hai loại của dữ liệu:
  - Được cấu trúc (structured).
  - Chưa được cấu trúc (unstructured)
- Dữ liệu được cấu trúc có những khóa liên hệ với mỗi phần tử dữ liệu mà phản ảnh nội dung của nó.
- Truy xuất dựa trên nội dung đến dữ liệu chưa được cấu trúc mà không cần quan tâm đến ngữ nghĩa của nó là một phương pháp tìm kiếm từ khóa

Nhắc lại tìm kiếm Web (tt)

- Để thuận tiện cho quá trình so khớp từ khóa và tài liệu, một số bước tiền xử lý được thực hiện:
  - Tài liệu được tách từ
  - Kí tự chuyển đổi thành hoa hay thường
  - Các từ được giảm đến thể gốc
  - Stopword thường bỏ đi
  - ….

NỘI DUNG

- Khái niệm khai thác cấu trúc Web
- Đặc điểm của khai thác cấu trúc Web
- Một số cấu trúc Web phổ biến
- Giới thiệu Web crawler
- Giới thiệu một số thuật toán
- Ứng dụng

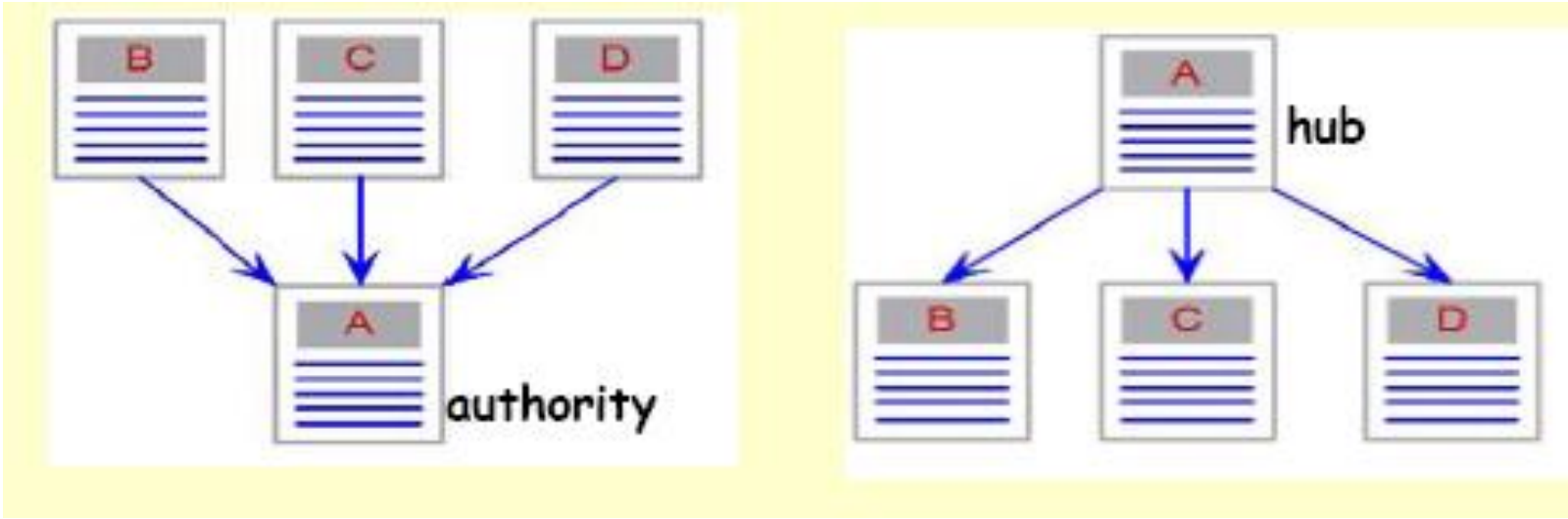
Giới thiệu thuật toán

- Có hai thuật toán chính được sử dụng trong khai thác cấu trúc Web
  - HITS (Hypertext-Induced Topic Search)
  - Thuật toán xếp hạng trang (Page rank)

HITS

- Là thuật toán phân tích liên kết
- Bình chọn cho các trang web
- Được đề xuất bởi Jon Kleinberg
- Xác định hai độ đo cho một trang:
  - Độ *authority*: ước lượng giá trị của nội dung của một trang
  - Độ *hub*: ước lượng giá trị của những liên kết chính nó đến các trang khác

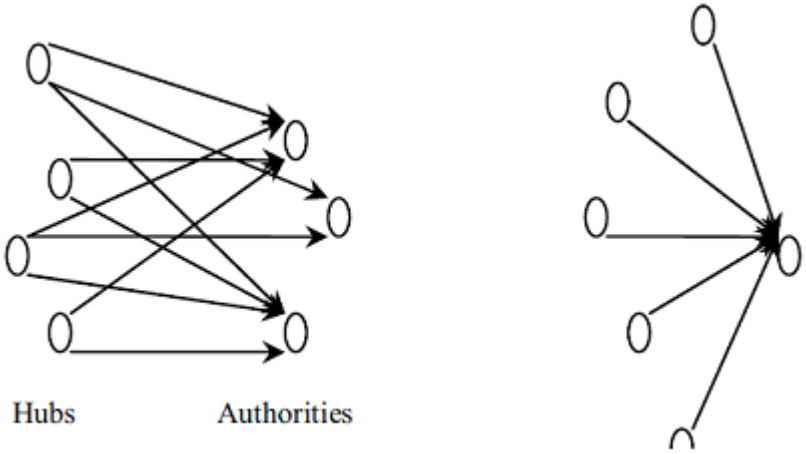
Các hub và authority



- Các trang Hub trỏ những liên kết thú vị đến các authority (các trang liên quan)
- Các authority là mục tiêu của các trang hub

Đặc điểm hub và authority

- Độ đo hub và authority có thể được xác định qua lại lẫn nhau hay có quan hệ hỗ tương.
  - Một hub tốt là một trang trỏ đến nhiều authority tốt; một authority tốt là một trang được trỏ đến bởi rất nhiều hub tốt

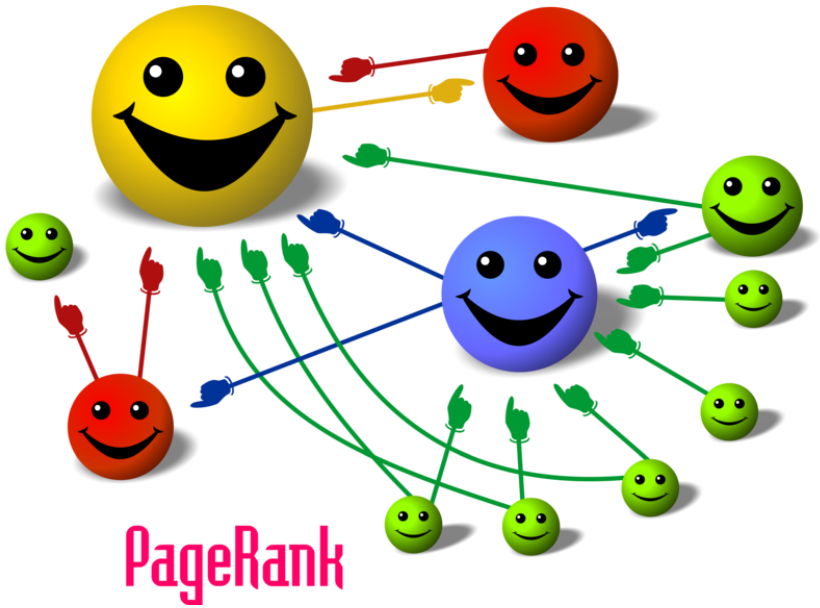


Page rank

- Là thuật toán phân tích liên kết
- Thực hiện gán một trọng số đến mỗi phần tử của một tập siêu liên kết của tài liệu.
- Được kí hiệu là PR(E)

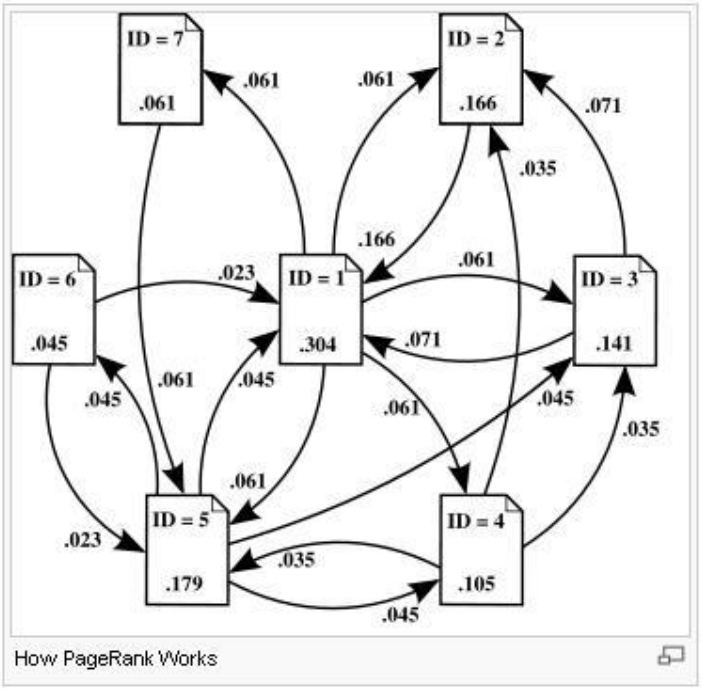
Đặc điểm Page rank

- Dựa trên bản chất *dân chủ đơn nhất*
  - Liên kết từ trang A đến trang B là một phiếu bình chọn (vote) của trang A cho trang B.
  - A xem xét độ quan trọng của chính nó và giúp đỡ để làm B quan trọng.



Đặc điểm Page rank (tt)

- Cũng là một phân bố xác suất thể hiện xác suất mà một click liên kết đến từ bất kì trang cụ thể nào.
  - Hạng của trang là 0.5 hay 50% nói đến cơ may mà một người click vào một liên kết sẽ hướng đến tài liệu là 0.5



NỘI DUNG

- Khái niệm khai thác cấu trúc Web
- Đặc điểm của khai thác cấu trúc Web
- Một số cấu trúc Web phổ biến
- Giới thiệu Web crawler
- Giới thiệu một số thuật toán
- Ứng dụng

Các ứng dụng

- Truy vấn thông tin trong mạng xã hội (social network).
- Tìm ra độ tương quan của mỗi trang Web.
- Đo tính toàn vẹn của Website
- Được sử dụng trong các cổ máy tìm kiếm để tìm ra thông tin liên quan

Tìm ra các trang cùng loại

Jiawei Han

Professor, Department of
Univ. of Illinois at Urbana-
Rm 2132, Siebel Center f
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: hanj[at]cs.uiuc.ed

Ph.D. (1985), Computer

Knowledge Discovery and Data Mining,

Tel: (217) 265-6
Fax: (217) 244-

Gerald DeJong (a.k.a. Mr. EBL)

Professor of Computer Science
Affiliate of the Electrical and Computer Engineering Department

- **Current Research** (Selected Publications)
  - Information Network Analysis and Discov
  - Sequential and Structured Pattern Discove
  - Discovery of the Dynamics of Data Stream
  - Ranking and Multidimensional Analysis in
  - Analysis of Spatiotemporal, Trajectories, a
  - Knowledge Discovery in Cyberphysical Sy
  - Assured Information Sharing Lifecycle (AI
  - Software Bug Detection in Sensor Network
  - CS-BibCube: OLAPing and Analysis of Co
- **Teaching**
  - UIUC CS512: Data Mining: Principles and
  - UIUC CS412: An Introduction to Data Wa
  - UIUC CS591Han Advanced Topics in Data
  - UIUC CS591 Yahoo!-DAIS (Data and Info

I received my Ph.D. i
Shin. I was an Asista
I then joined the Univ
tenure. My interests i
networks, distributed
developing theory, an
systems, motivated b
determinism. Applica
avionics to homeland

Who is Mr.

1995-present Pr
University of Illi
1985 1995 Ass
Engineering, Uni
1981-1985 Assi
1980 Instructor,
1979 Ph.D., Co
1974 B.S., Phys

ENGINEERING AT ILLINOIS

PROFESSOR MICHAEL T. HEATH

AWARDS
BIOGRAPHY
CONTACT
PHOTO GALLERY
PUBLICATIONS
RESEARCH
STUDENTS
TEACHING
CSAR
CSE
HOME

Professor Michael T. Heath is Fulton Watson Copp Chair in the Department of Computer Science at the University of Illinois at Urbana-Champaign, where he is also Director of both the Computational Science and Engineering Program and the Center for Simulation of Advanced Rockets. His research interests are in numerical analysis—particularly numerical linear algebra and optimization—and in parallel computing. He is an ACM Fellow, a SIAM Fellow, and a member of the European Academy of Sciences.

Flickr: Bức ảnh nào giống nhất

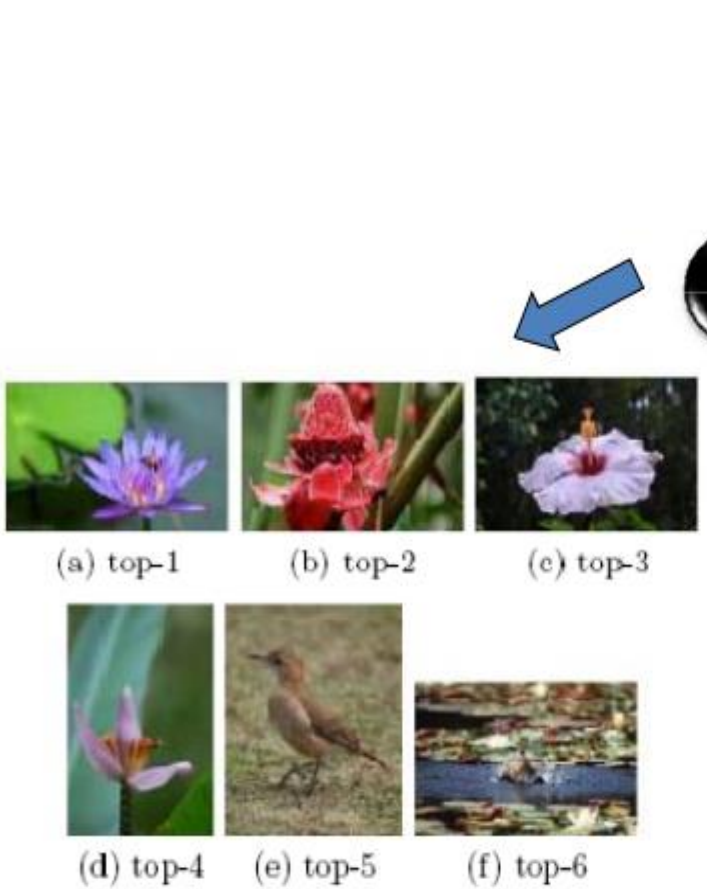


Figure 5: Top-6 images in Flickr network under path schema *ITI*

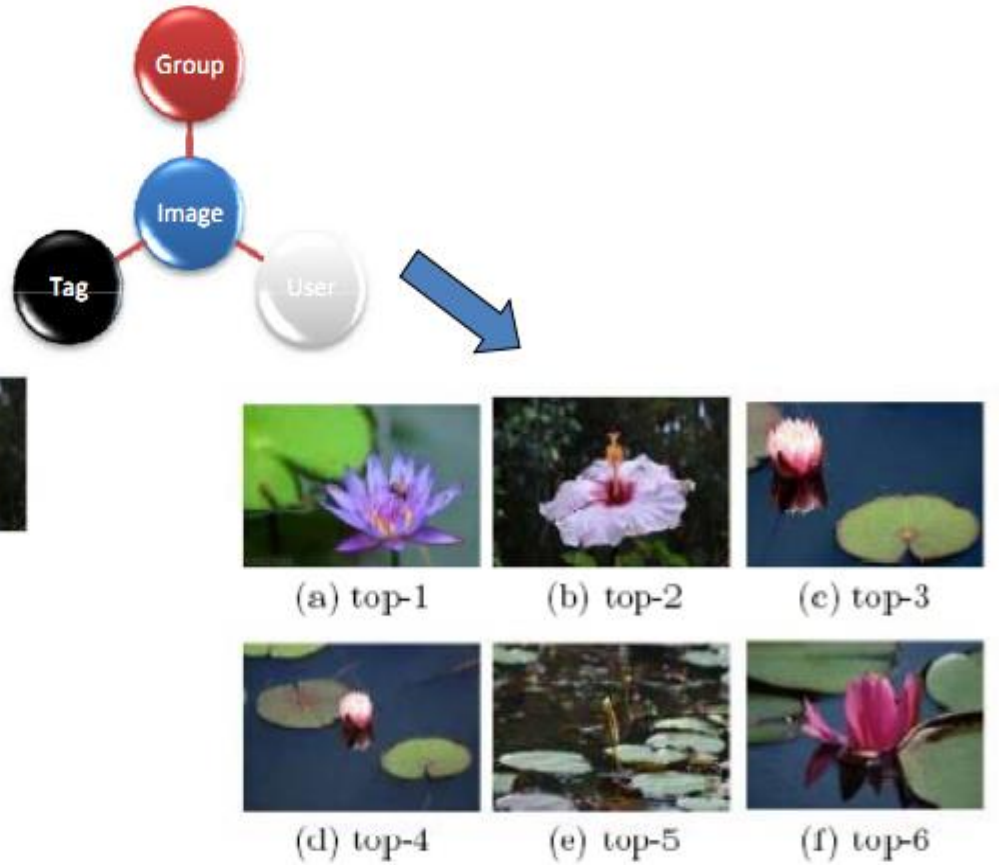
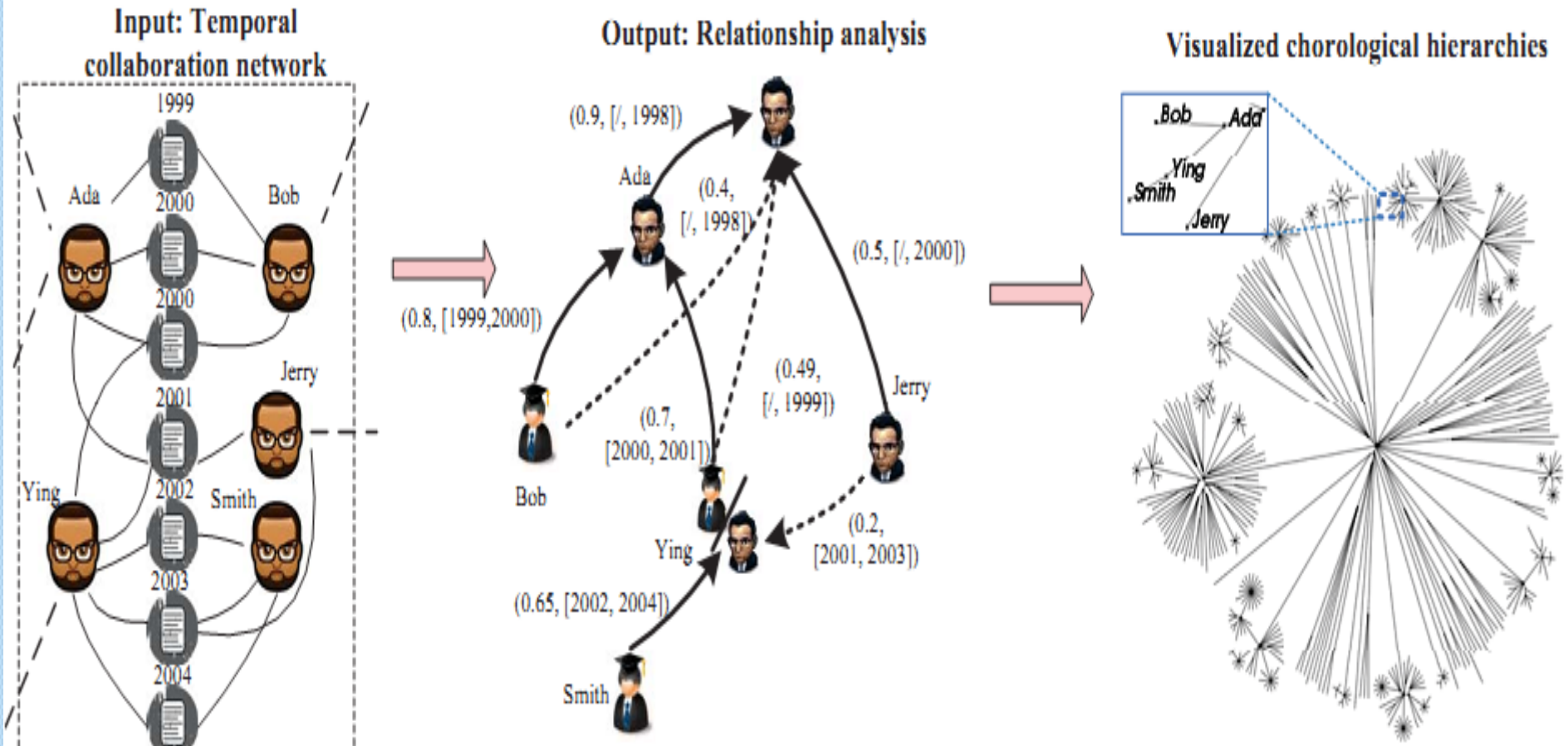
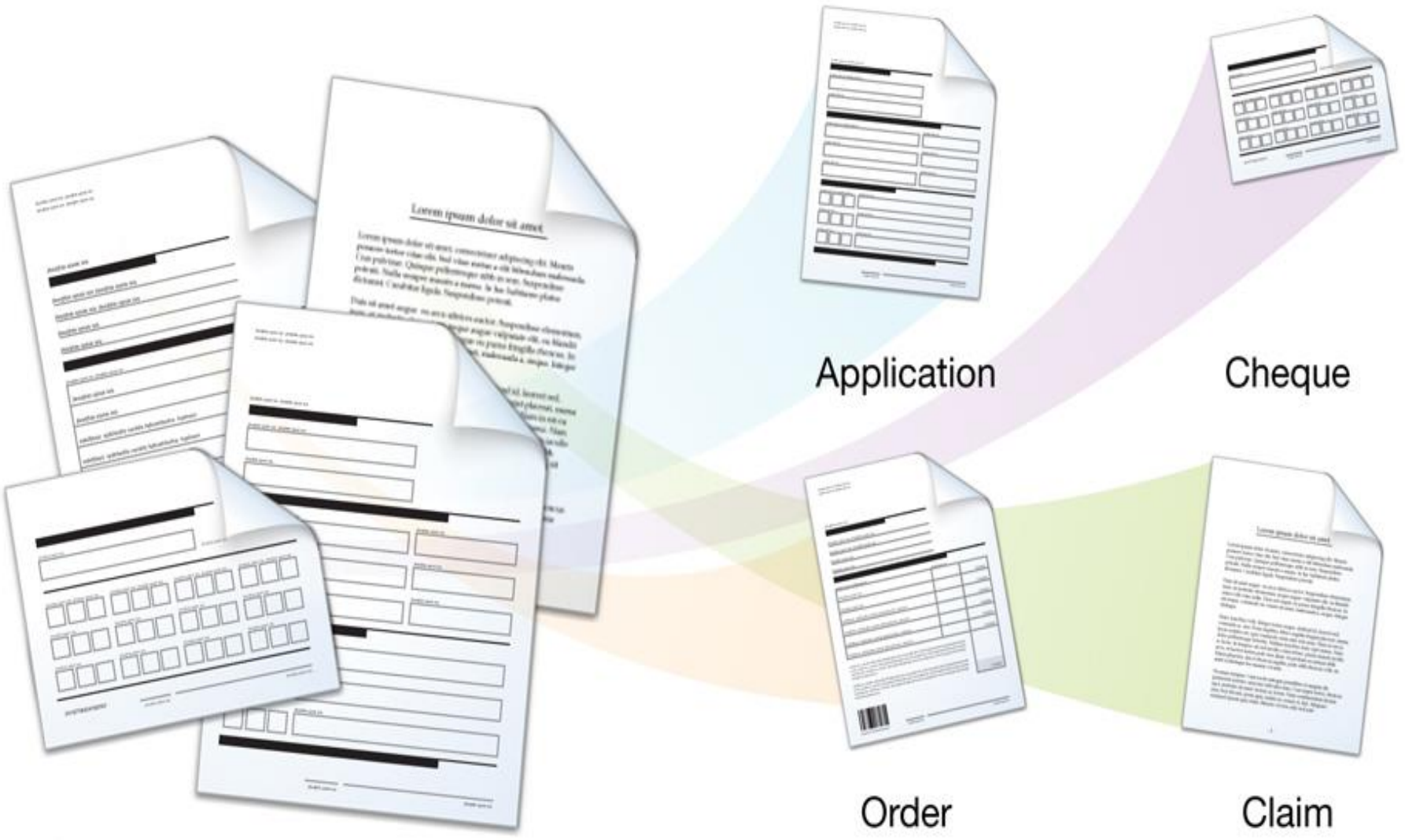


Figure 6: Top-6 images in Flickr network under path schema *ITIGITI*

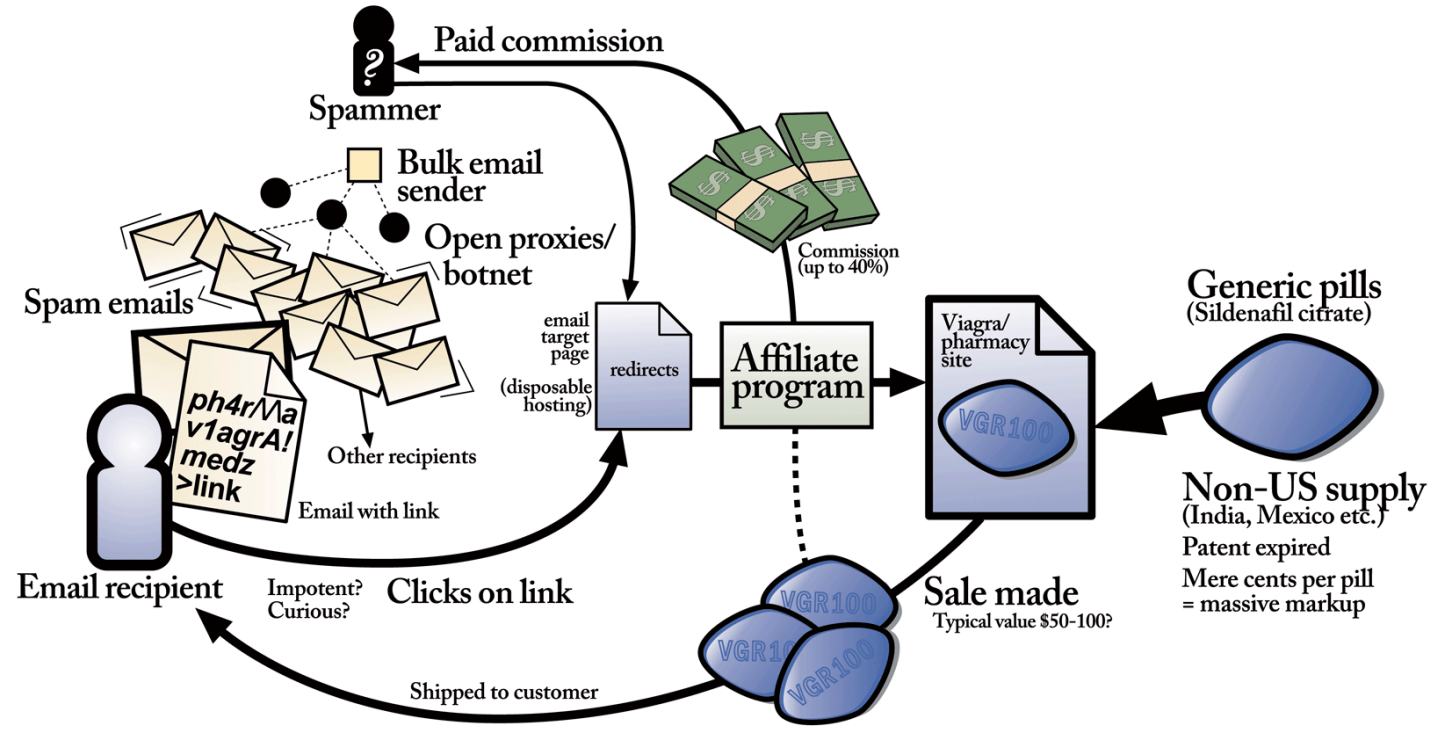
Mạng cộng tác



Phân lớp trang Web



Phát hiện liên kết spam



How Viagra spam works

(modern life is rubbish) http://www.modernlifeisrubbish.co.uk

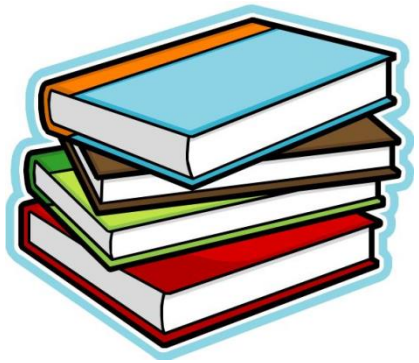
Tóm tắt

- Cỗ máy tìm kiếm sử dụng khai thác cấu trúc Web để tìm ra thông tin.
- Chúng ta có thể tạo ra tri thức mới từ những tri thức sẵn có.
- Khai thác nội dung Web có thể được thêm vào để tăng cường sự thực thi của các cỗ máy tìm kiếm.

TÀI LIỆU THAM KHẢO

- http://www.cs.sunysb.edu/~cse634/spring2007/group3_final.ppt
- http://www.cyberartsweb.org/cpace/ht/lanman/wsm1.htm
- http://www.web-datamining.net/structure/
- http://www.expertsupdates.com/seo-articles/web-mining-12.aspx

KẾT THÚC PHẦN 1



Thông tin cấu trúc Web

- Thông tin mang tính cấu trúc phát sinh từ khai thác cấu trúc Web bao gồm:
  - Thông tin đo lường độ phổ biến của liên kết cục bộ trong các bộ tuần tự Web (Web tuple) trong bảng Web
  - Thông tin đo lường độ phổ biến của các bộ tuần tự Web trong một bảng Web chứa các liên kết bên trong và các liên kết bên trong cùng một tài liệu
  - Thông tin đo độ phổ biến của các bộ tuần tự Web trong bảng Web mà chứa các liên kết toàn cục và liên kết đến trang các website khác
  - Thông tin đo độ phổ biến của các bộ tuần tự web giống hệt nhau mà xuất hiện trong bảng Web hay giữa các bảng Web.