HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 10: Khai phá văn bản và khai phá web

Lê Hải Hà

Nội dung

❶ Khái niệm khai phá văn bản

❷ Quy trình khai phá văn bản

❸ Khai phá Web

Mục tiêu

- Mô tả khai phá văn bản là gì và nhu cầu khai phá văn bản
- Sự khác nhau giữa khai phá văn bản, khai phá web và khai phá dữ liệu
- Hiểu những lĩnh vực ứng dụng khác nhau của khai phá văn bản
- Biết quy trình thực hiện một dự án khai phá văn bản
- Hiểu các phương pháp khác nhau để tạo dữ liệu cấu trúc từ dữ liệu văn bản

Mục tiêu

- Mô tả khai phá web, mục tiêu và lợi ích của nó
- Hiểu 3 nhánh của khai phá web
  - Khai phá nội dung web
  - Khai phá cấu trúc web
  - Khai phá việc sử dụng web
- Hiểu các ứng dụng của 3 mô hình khai phá này

Khái niệm khai phá văn bản

- 85-90% dữ liệu của hầu hết các doanh nghiệp ở dạng phi cấu trúc (thí dụ ở dạng văn bản)
- Dữ liệu phi cấu trúc của các doanh nghiệp tăng gấp đôi sau mỗi 18 tháng
- Khai thác các nguồn thông tin này không phải là một lựa chọn mà là một nhu cầu để duy trì tính cạnh tranh của doanh nghiệp
- Câu trả lời là: khai phá văn bản
  - Tiến trình bán tự động trích chọn tri thức từ các nguồn dữ liệu phi cấu trúc
  - Được biết là khai phá dữ liệu văn bản hay khai phá tri thức trong các CSDL văn bản

Khai phá dữ liệu và khai phá văn bản

- Đều tìm kiếm các mẫu mới và hữu dụng
- Đều là các tiến trình bán tự động
- Khác nhau ở bản chất của dữ liệu:
  - Dữ liệu cấu trúc và dữ liệu phi cấu trúc
  - Dữ liệu cấu trúc: trong các cơ sở dữ liệu
  - Dữ liệu phi cấu trúc: các tài liệu Word, các file PDF, documents, PDF files, các trích đoạn văn bản, các file XML, v.v…
- Khai phá văn bản – trước hết áp đặt một cấu trúc đối với dữ liệu và sau đó khai phá dữ liệu có cấu trúc

Các khái niệm khai phá văn bản

- Các lợi ích của khai phá văn bản đặc biệt nổi bật đối với các môi trường nhiều dữ liệu văn bản
  - Thí dụ: luật (các lệnh tòa án), nghiên cứu khoa học (các bài báo), tài chính (các báo cáo quý), y học (kết luận bệnh án), sinh học (các tương tác phân tử), công nghệ (các file sáng chế), tiếp thị (ý kiến khách hàng), …
- Các bản ghi truyền thông điện tử (thí dụ: Email)
  - Lọc thư rác
  - Mức độ ưu tiên và phân loại Email
  - Tạo phản hồi tự động

Các lĩnh vực ứng dụng khai phá văn bản

- Trích chọn thông tin
- Theo dõi chủ đề
- Tổng hợp/tóm tắt
- Phân lớp
- Phân cụm
- Liên kết khái niệm
- Trả lời câu hỏi

Các thuật ngữ khai phá văn bản

- Dữ liệu cấu trúc và bán cấu trúc
- Ngữ liệu - Corpus (số nhiều là corpora)
- Thuật ngữ - Terms
- Khái niệm - Concepts
- Gốc từ - Stemming
- Từ dừng - Stop words (hay từ nhiễu)
- Đồng nghĩa - Synonyms (và đa nghĩa - polysemes)
- Tokenizing

Các thuật ngữ khai phá văn bản

- Từ điển thuật ngữ - Term dictionary
- Tần suất từ - Word frequency
- Loại từ (danh từ, động từ, …) - Part-of-speech tagging (POS)
- Hình thái học (cấu trúc từ) - Morphology
- Ma trận Term-by-document
  - Ma trận xuất hiện
- Phân rã giá trị đơn - Singular value decomposition
  - Chỉ mục ngữ nghĩa tiềm ẩn

Xử lý ngôn ngữ tự nhiên (NLP)

- Cấu trúc hóa tập văn bản
  - Tiếp cận cũ: bag-of-words
  - Tiếp cận mới: natural language processing
- NLP là …
  - Khái niệm rất quan trọng trong khai phá văn bản
  - Một lĩnh vực con của trí tuệ nhân tạo và ngôn ngữ học tính toán
  - Nghiên cứu việc "hiểu" ngôn ngữ của con người
- Khai phá văn bản dựa trên ngữ nghĩa, cú pháp

Xử lý ngôn ngữ tự nhiên

- Thế nào là "Hiểu" ?
  - Con người hiểu, còn máy tính thì sao?
  - Ngôn ngữ tự nhiên thường mơ hồ và dựa trên ngữ cảnh
  - Hiểu đúng đòi hỏi kiến thức sâu rộng về một chủ đề

  - Liệu máy tính có thể hiểu ngôn ngữ tự nhiên giống như cách chúng ta làm không?

Xử lý ngôn ngữ tự nhiên

- Các thách thức trong NLP
  - Part-of-speech tagging
  - Text segmentation
  - Word sense disambiguation
  - Syntax ambiguity
  - Imperfect or irregular input
  - Speech acts
- Giấc mơ của cộng đồng AI
  - Có các giải thuật có khả năng tự động đọc và lấy tri thức từ văn bản

Xử lý ngôn ngữ tự nhiên

- WordNet
  - Một CSDL các từ tiếng Anh được mã hóa thủ công, các định nghĩa của chúng, các từ đồng nghĩa, và các mối quan hệ ngữ nghĩa khác giữa các tập đồng nghĩa
  - Nguồn chính cho NLP
  - Cần tự động hóa để hoàn thành
- Phân tích cảm xúc - Sentiment Analysis
  - Một kỹ thuật được sử dụng để phát hiện các ý kiến ưa thích hay không ưa thích đối với các sản phẩm hay dịch vụ cụ thể
  - Thí dụ ứng dụng CRM

Các nhóm công việc NLP

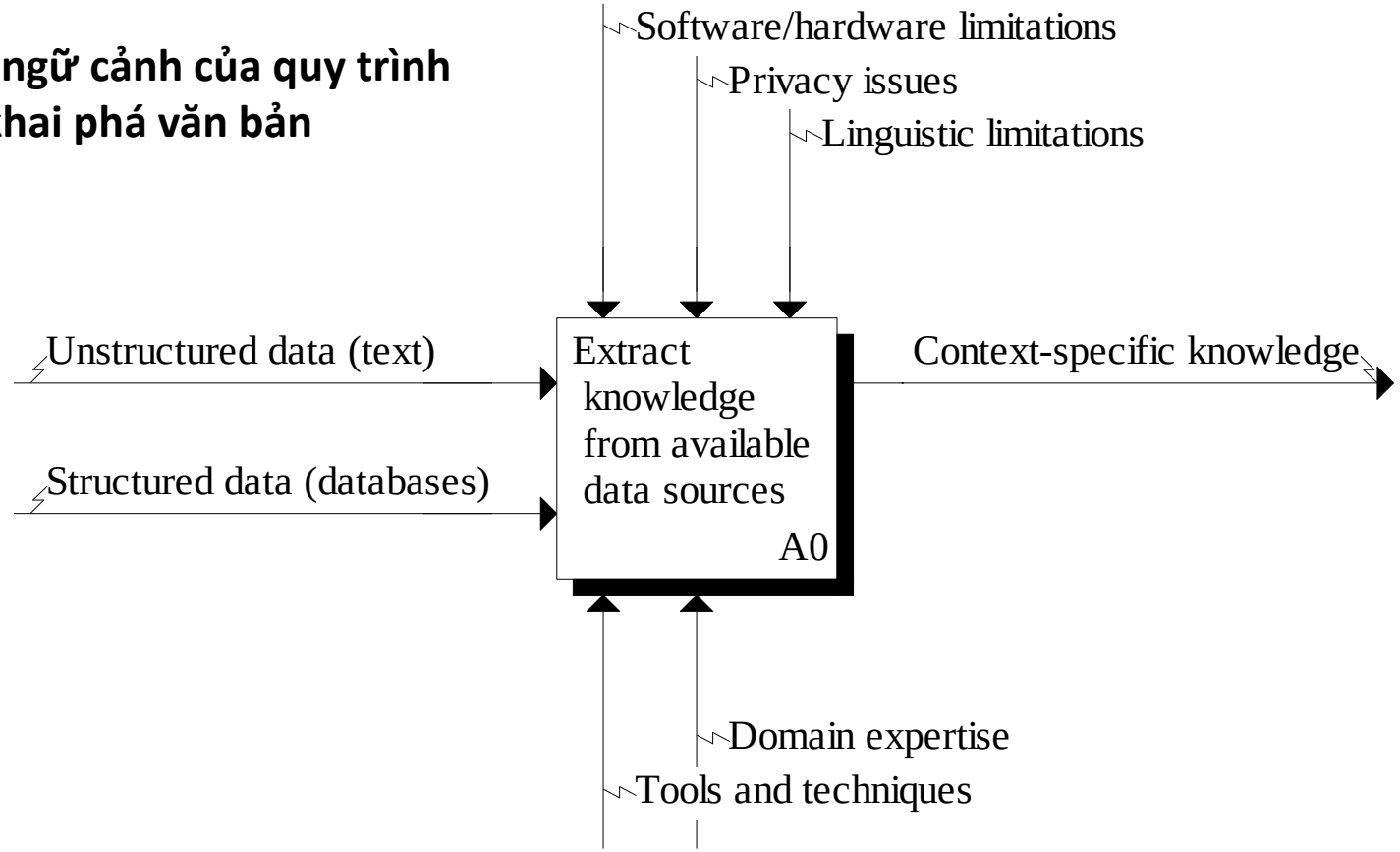
- Information retrieval
- Information extraction
- Named-entity recognition
- Question answering
- Automatic summarization
- Natural language generation and understanding
- Machine translation
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition

Các ứng dụng khai phá văn bản

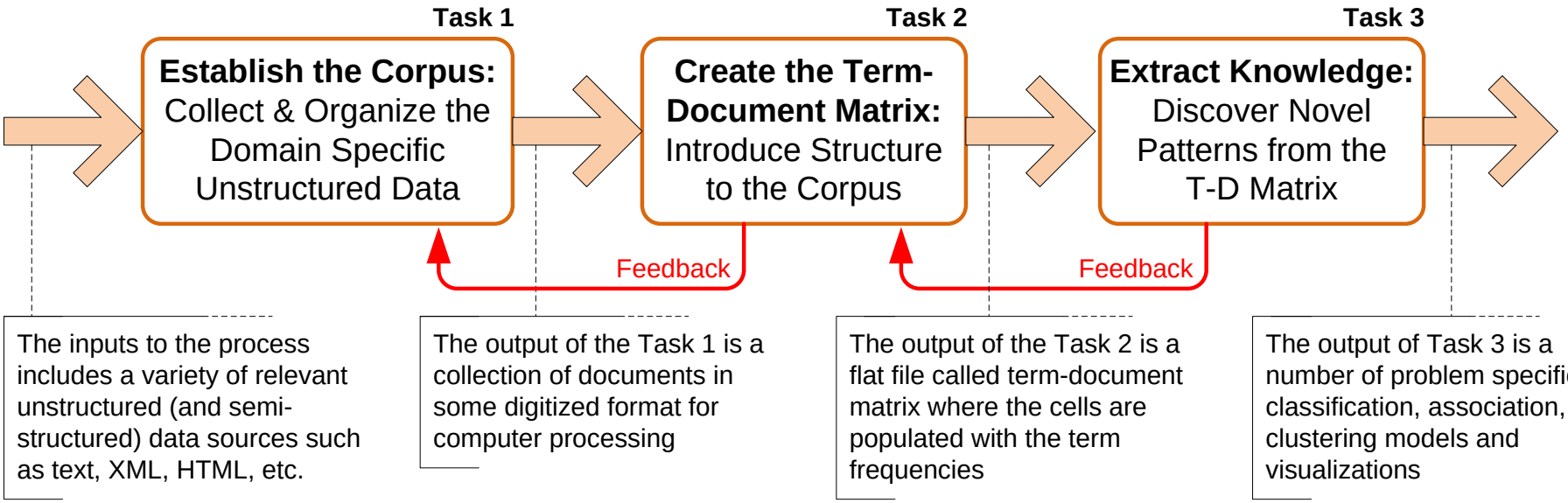
- Các ứng dụng tiếp thị
  - Làm cho các hệ CRM tốt hơn
- Các ứng dụng an toàn bảo mật
  - ECHELON, OASIS
  - Phát hiện lừa đảo
- Y học và sinh học
  - Nhận dạng gen dựa trên tài liệu
- Các ứng dụng trong nghiên cứu/học thuật
  - Phân tích các hướng/luồng nghiên cứu

Quy trình khai phá văn bản

Biểu đồ ngữ cảnh của quy trình khai phá văn bản



Quy trình khai phá văn bản



Quy trình khai phá văn bản 3 bước

Quy trình khai phá văn bản

- Bước 1: Thiết lập dữ liệu/ngữ liệu (corpus)
  - Thu thập tất cả các dữ liệu phi cấu trúc liên quan (thí dụ, các tài liệu văn bản, các file XML, các email, các trang web, các ghi chú ngắn, các bản ghi âm giọng nói, …)
  - Số hóa, chuẩn hóa các dữ liệu thu thập (thí dụ, chuyển tất cả về các file ASCII)
  - Đặt các dữ liệu thu thập ở cùng 1 vị trí (thí dụ trong cùng 1 file, hay là các file riêng biệt trong cùng 1 thư mục)

Quy trình khai phá văn bản

- Bước 2: Tạo ma trận Term–by–Document

Terms / Documents	investment risk	project management	software engineering	development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

Quy trình khai phá văn bản

- Bước 2 (tiếp): Tạo ma trận Term–by–Document Matrix (TDM).
  - Có nên bao gồm tất cả các thuật ngữ không?
    - Stop words, include words
    - Synonyms, homonyms
    - Stemming
  - Các chỉ số (giá trị trong các ô) nên biểu diễn cái gì là tốt nhất?
    - Tần suất, tần suất nhị phân, logarit của tuần suất;
    - Nghịch đảo của tần suất tài liệu

$$idf(i,j) = \begin{cases} 0 & if\ wf_{ij} = 0 \\ (1 + \log(wf_{ij}))\log\frac{N}{df_i} & if\ wf_{ij} \geq 1 \end{cases}$$

Quy trình khai phá văn bản

- Bước 2 (tiếp): Tạo ma trận Term–by–Document.
  - TDM là một ma trận thưa. Ta có thể giảm số chiều của TDM thế nào?
    - Thủ công – một chuyên gia trong lĩnh vực nên xem xét nó
    - Loạt bỏ các thuật ngữ xuất hiện rất ít trong rất ít tài liệu
    - Chuyển đổi ma trận sử dụng phân rã giá trị đơn (SVD)
    - SVD tương tự như phân tích thành phần chính (PCA)

Quy trình khai phá văn bản

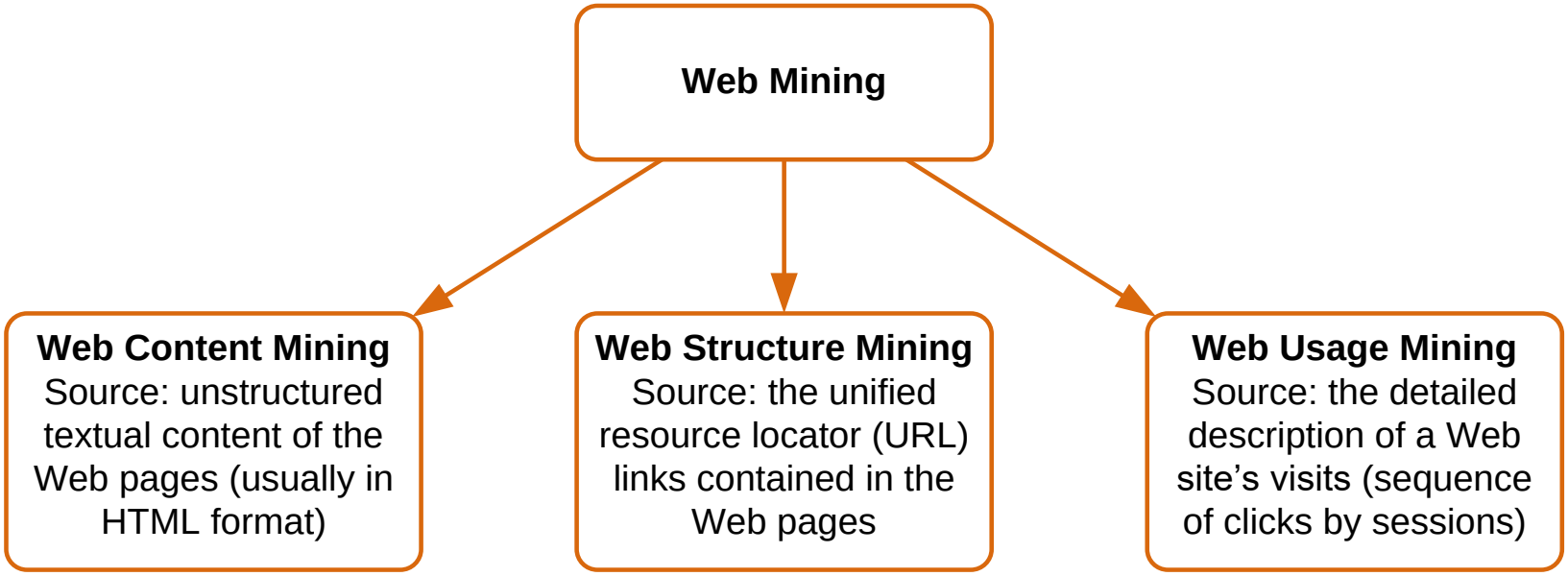
- Bước 3: Trích các mẫu/tri thức
  - Phân lớp (phân lớp văn bản)
  - Phân cụm (các nhóm văn bản một cách tự nhiên)
    - Cải thiện search recall
    - Cải thiện search precision
    - Scatter/gather
    - Query-specific clustering
  - Liên kết
  - Phân tích khuynh hướng

Tổng quan về khai phá web

- Web là kho dữ liệu lớn nhất
- Dữ liệu ở định dạng HTML, XML, text
- Các thách thức (của tiến trình xử lý dữ liệu Web)
  - Web quá lớn để khai phá dữ liệu hiệu quả: *không nghĩ tới việc tạo một DW cho nó*
  - Web quá phức tạp: *không có cấu trúc nhất định*
  - Web quá động: *dữ liệu thường xuyên được cập nhật*
  - Web không xác định một lĩnh vực cụ thể: *rất đa dạng*
  - Web có mọi thứ
- Bổ sung cho Web search engines.
- Cơ hội và thách thức rất lớn!

Khai phá web

- Khai phá Web (hay khai phá dữ liệu Web) là tiến trình khám phá các quan hệ bản chất (thú vị và hữu dụng) từ dữ liệu Web (văn bản, liên kết, hay việc sử dụng)



Khai phá nội dung/cấu trúc web

- Khai phá nội dung văn bản trên web
- Thu thập dữ liệu qua Web crawlers
- Các trang Web chứa các siêu liên kết
  - Authoritative pages
  - Hubs
  - Giải thuật hyperlink-induced topic search (HITS)
- Khai phá cấu trúc Web: trích thông tin hữu ích từ các liên kết được nhúng trong các tài liệu Web

Khai phá sử dụng web

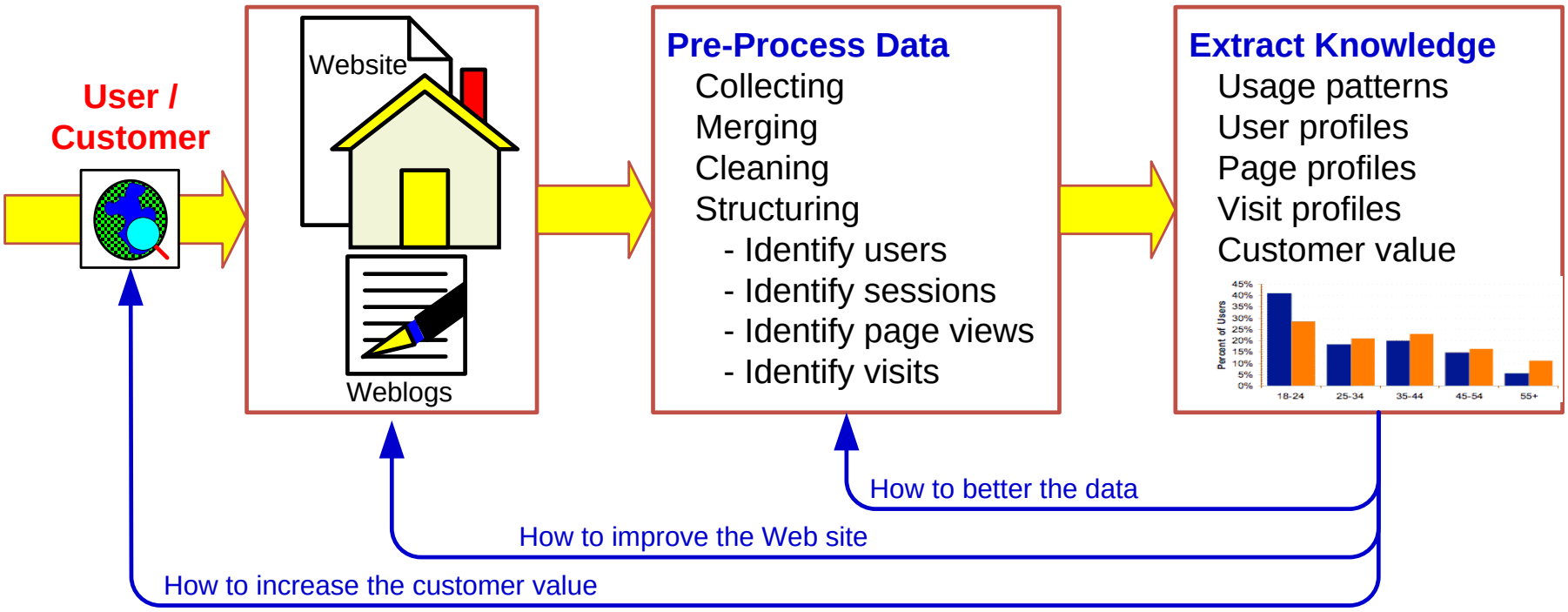
- Trích thông tin từ dữ liệu được tạo ra thông qua việc xem hay giao dịch với trang Web…
  - Dữ liệu được lưu ở server access logs, referrer logs, agent logs, và client-side cookies
  - Các đặc điểm và hồ sơ sử dụng của người dùng
  - Siêu dữ liệu, như thuộc tính của trang, thuộc tính của nội dung, và dữ liệu sử dụng

  => hiểu hành vi người dùng
- Clickstream data
- Clickstream analysis
- Thí dụ quảng cáo trực tuyến cho người du lịch

Khai phá sử dụng web

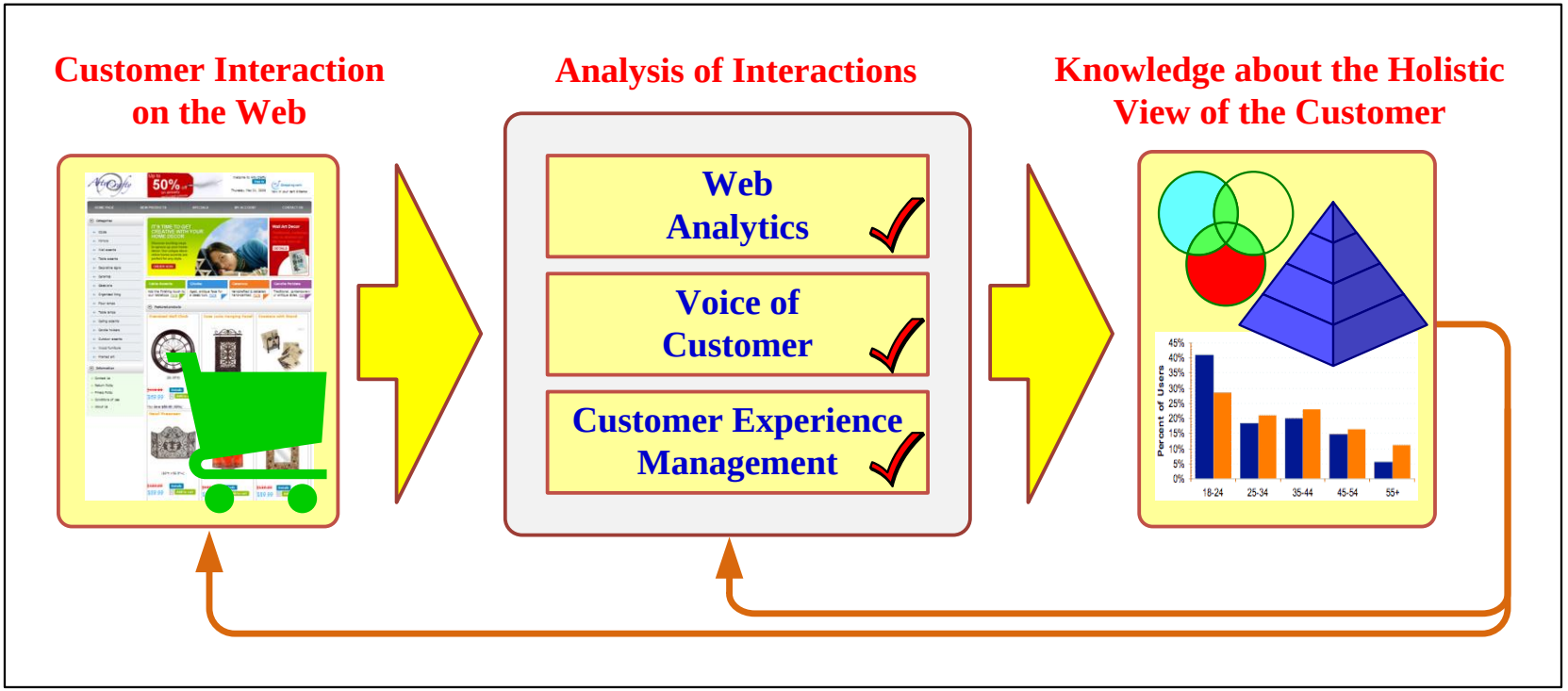
- Các ứng dụng khai phá sử dụng Web
  - Xác định thời gian lưu lại trên trang cụ thể của khách hàng
  - Thiết kế các chiến lược tiếp thị chéo trên các sản phẩm.
  - Đánh giá các chiến dịch khuyến mãi
  - Nhắm mục tiêu quảng cáo điện tử và phiếu thưởng vào các nhóm người dùng dựa trên các mẫu truy cập của họ.
  - Dự đoán hành vi người dùng dựa trên các luật đã học được từ trước và hồ sơ người dùng.
  - Thể hiện thông tin động tới người dùng dựa trên sự quan tâm và hồ sơ người dùng

Khai phá sử dụng web (clickstream analysis)



Các câu chuyện thành công của khai phá web

- Amazon.com, Ask.com, Scholastic.com, …
- Hệ sinh thái tối ưu Website



Các công cụ khai phá web

Product Name	URL
Angoss Knowledge WebMiner	angoss.com
ClickTracks	clicktracks.com
LiveStats from DeepMetrix	deepmetrix.com
Megaputer WebAnalyst	megaputer.com
MicroStrategy Web Traffic Analysis	microstrategy.com
SAS Web Analytics	sas.com
SPSS Web Mining for Clementine	spss.com
WebTrends	webtrends.com
XML Miner	scientio.com