Text Mining and Social Media Mining

ThS. Lê Nhật Tùng

Tài liệu tham khảo

1. Manning, C., & Schutze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

2. Feldman, R., & Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

3. Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Pearson Prentice Hall.

4. Clark, A., Fox, C., & Lappin, S. (2010). The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell.

5. Francis, L., & Flynn, M. (2010). Text Mining Handbook. Casualty Actuarial Society, E-Forum.

6. McDonald, D., & Kelly, U. (2012). The Value and Benefits of Text Mining. Jisc.

7. Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). Taming Text: How to Find, Organize, and Manipulate It. Manning Publications.

8. Grimes, S. (2014). Text Analytics 2014: User Perspectives on Solutions and Providers. Alta Plana.

Tài liệu tham khảo

1. Bird, S., Klein, E., & Loper, E. (2010). Natural Language Processing with Python. O'Reilly Media.

[http://www.nltk.org/book_1ed/]

2. Aggarwal, C. (2011). Social Network Data Analytics. Springer.

[excerpts http://www.charuaggarwal.net/socialintro.pdf]

3. Aggarwal, C., & Zhai, C.X. (2012). Mining Text Data. Springer.

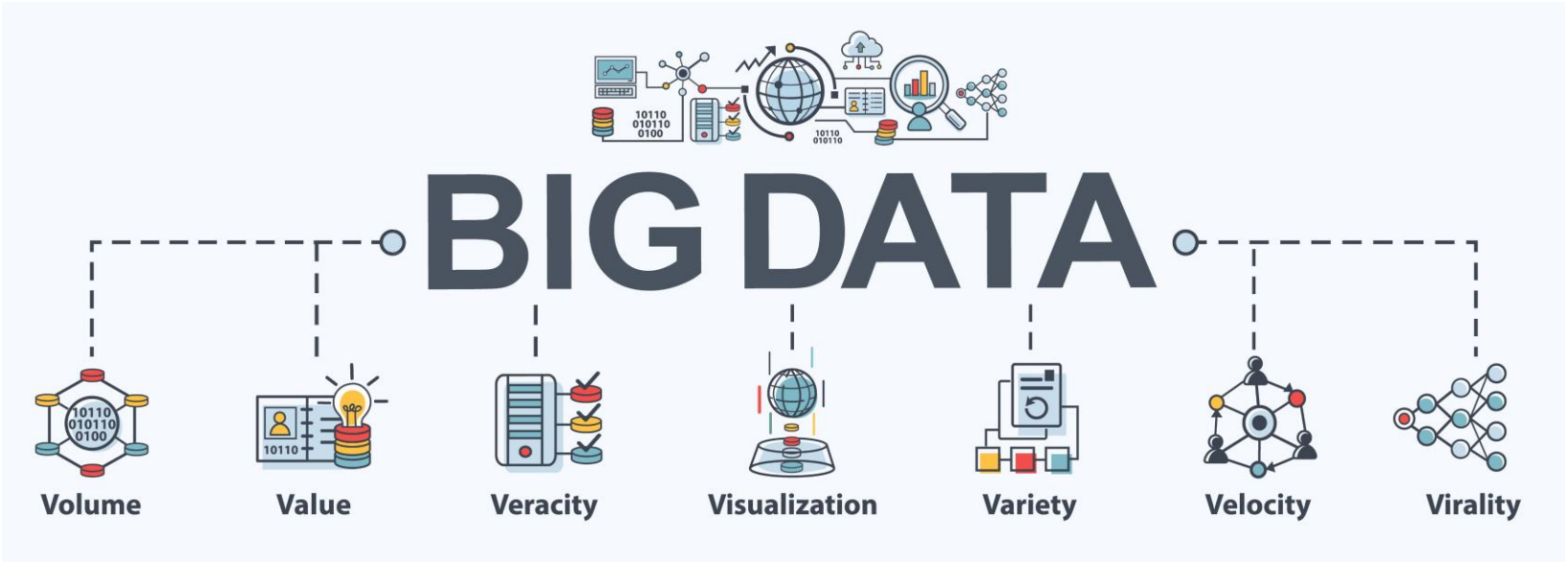
[excerpts http://www.charuaggarwal.net/text-content.pdf]

4. Silge, J., & Robinson, D. (2022). Text Mining with R. O'Reilly Media.

[https://www.tidytextmining.com/]

5. Irizarry, R. A. (2022). Introduction to Data Science: Data Analysis and Prediction Algorithms with R. CRC Press.

[https://rafalab.github.io/dsbook/]



Dữ liệu lớn trong thời đại số

- **Khối lượng ấn phẩm đồ sộ:** sách, tạp chí, báo cáo bài báo khoa học và tài liệu chuyên ngành tư liệu và văn bản đa dạng.
- **Số hóa:** Chi phí sao chép thông tin giảm, dễ dàng lưu trữ và truy xuất, thuận lợi cho việc xử lý và phân tích.
- **Internet - kênh phân phối chủ đạo:** chia sẻ thông tin tức thời, kết nối và trao đổi không giới hạn, mạng lưới tri thức toàn cầu, mạng xã hội phát triển và bùng nổ.

Ví dụ về tri thức của ngành công nghệ sinh học

1. Vấn đề tiếp cận thông tin

- 80% kiến thức chuyên ngành chỉ tồn tại trong các bài báo khoa học

2. Giới hạn khả năng tiếp thu của con người

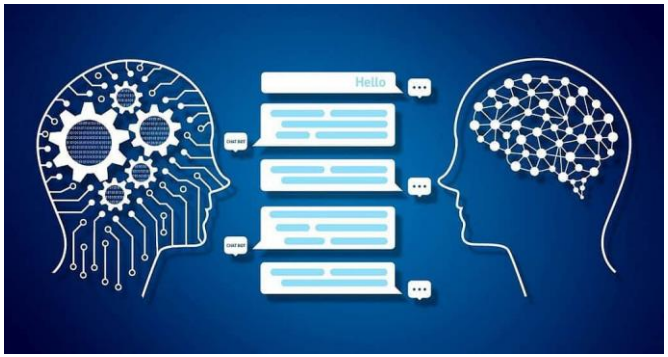
- Trung bình đọc 60 bài báo/tuần
- Chỉ 10% trong số đó thực sự liên quan
- Tốc độ tiếp thu: 6 bài/tuần (khoảng 300 bài/năm)

3. Tốc độ tăng trưởng thông tin

- Riêng cơ sở dữ liệu MedLine: bổ sung 10.000 tóm tắt bài báo mới mỗi tháng

Đặc điểm của Ngôn ngữ Tự nhiên

1. **Định nghĩa và Phát triển**
   - Sản phẩm của quá trình phát triển lịch sử
   - Khác biệt so với ngôn ngữ nhân tạo
2. **Đặc trưng cấu trúc**
   - Cú pháp phức tạp
   - Chứa nhiều tính đa nghĩa
   - Liên tục thay đổi và phát triển
3. **Tính chất đặc thù**
   - Đòi hỏi hiểu biết về thế giới xung quanh
   - Là phương tiện truyền tải tri thức
   - Công cụ giao tiếp và trao đổi thông tin của con người



So sánh khả năng đọc hiểu văn bản: Con người và Máy tính

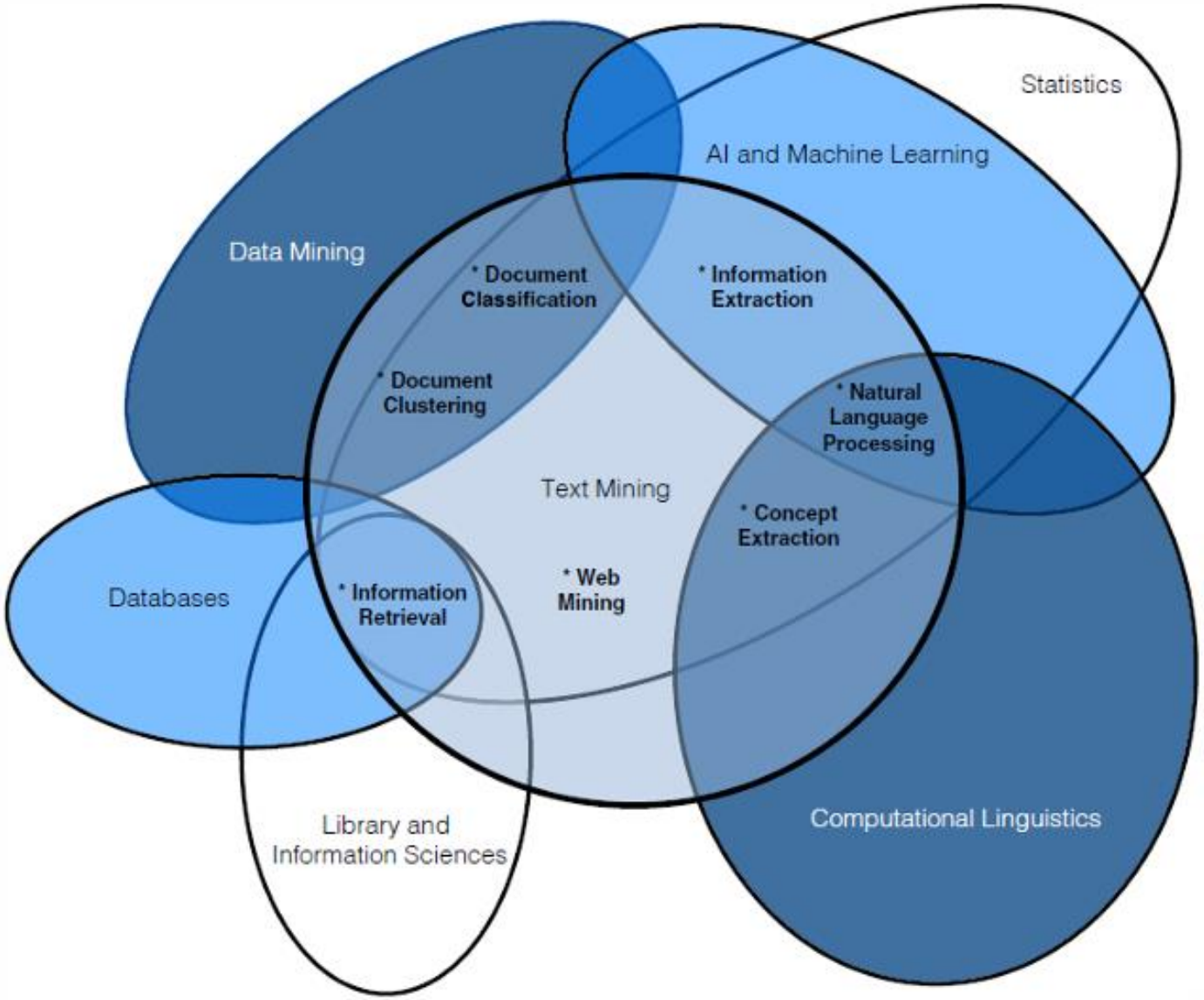
Con người	Máy tính
• Độ chính xác cao • Phạm vi hiểu biết rộng • Phân tích theo trình tự câu • Mức độ hiểu sâu sắc • Nắm bắt ngữ cảnh toàn diện • Xử lý một ngôn ngữ mỗi thời điểm • Tốc độ xử lý chậm	• Độ chính xác thấp (nhiễu) • Phạm vi hiểu biết hạn chế • Chủ yếu phân tích từ điển-kho ngữ liệu • Mức độ hiểu hời hợt • Tuân theo quy tắc suy luận cứng nhắc • Xử lý nhiều ngôn ngữ đồng thời • Tốc độ xử lý nhanh

Phương pháp phân tích dữ liệu phi cấu trúc

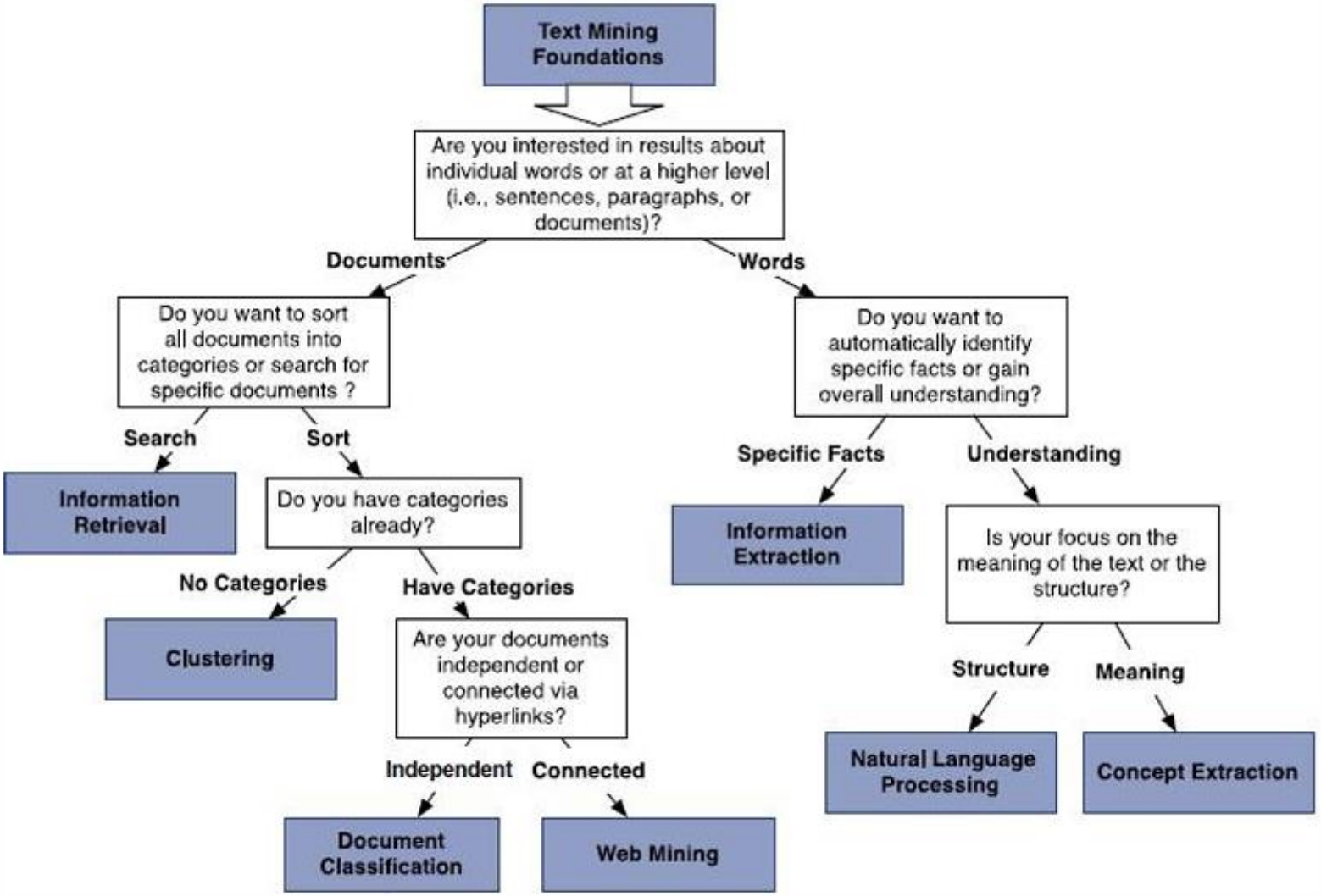
	QUALITATIVE ANALYSIS	QUANTITATIVE CONTENT ANALYSIS	TEXT MINING
METHOD	Manual reading and coding of documents	Dictionaries of words, phrases, patterns, rules	Statistical analysis and data mining techniques
TIME REQUIREMENT	Very High (weeks / months)	Medium (days) Decreasing	Low (seconds / minutes)
VALIDITY	Potentially High (high subjectivity)	Medium to High	Medium to High (sometimes low)
RELIABILITY (REPEATABILITY)	May be problematic	Perfect	Perfect

Khai phá văn bản (Text Mining)

- **Định nghĩa**
  - Quá trình tự động thu thập kiến thức quan trọng và hữu ích
  - Khám phá thông tin mới từ bộ sưu tập tài liệu văn bản
  - Tập trung vào kiến thức chưa được biết đến trước đây
- **Các thuật ngữ tương đương**
  - Khai phá dữ liệu văn bản (Text Data Mining)
  - Phân tích văn bản (Text Analytics)
  - Khám phá tri thức trong văn bản (Knowledge Discovery in Text - KDT)
  - Phân tích văn bản thông minh (Intelligent Text Analysis)



Miner et.al 2012



Miner et.al 2012


TEXT ANALYTICS
Text Mining
Information Retrieval
Document Matching
Search Optimization
Inverted Index
Information Extraction
Entity Extraction
Co-reference
Relationship Exraction
Concept Extraction
Colocations
Word Association
Sentiment Analysis
Natural Language Processing
Tokenization
Lemmatization
Part-of-Speech Tagging
Web Mining
Web Content Mining
Web Analytics
Web Structure Analysis
Classification
Document Ranking
Alert Detection
Document Categorization
Clustering
Document Similarity
Document Clustering

Phân tích văn bản (Text Analytics)

1. Khai phá văn bản (Text Mining)

Trích xuất thông tin (Information Extraction)

- Trích xuất thực thể (Entity Extraction)
- Đồng tham chiếu (Co-reference)
- Trích xuất quan hệ (Relationship Extraction)
- Định vị đồng xuất hiện (Colocations)

Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

- Phân tách từ (Tokenization)
- Chuẩn hóa từ (Lemmatization)
- Gán nhãn từ loại (Part-of-Speech Tagging)

Trích xuất khái niệm (Concept Extraction)

- Phân tích cảm xúc (Sentiment Analysis)
- Kết hợp từ (Word Association)

Phân tích văn bản (Text Analytics)

2. Phân loại và Phân cụm (Classification & Clustering)

- Phân loại tài liệu (Document Categorization)
- Phát hiện cảnh báo (Alert Detection)
- Xếp hạng tài liệu (Document Ranking)
- Độ tương đồng tài liệu (Document Similarity)
- Phân cụm tài liệu (Document Clustering)

Phân tích văn bản (Text Analytics)

3. Khai phá Web (Web Mining)

- Phân tích cấu trúc Web (Web Structure Analysis)
- Phân tích Web (Web Analytics)
- Khai phá nội dung Web (Web Content Mining)

4. Truy xuất thông tin (Information Retrieval)

- Tìm kiếm tối ưu (Search Optimization)
- Chỉ mục đảo (Inverted Index)
- Khớp tài liệu (Document Matching)

Ứng dụng Khai phá văn bản

Quản lý quan hệ khách hàng (CRM)

Nguồn dữ liệu:

- Khiếu nại của khách hàng
- Ý kiến phản hồi
- Dữ liệu tổng đài

Mục tiêu:

- Nâng cao chất lượng sản phẩm và dịch vụ
- Quản lý sản phẩm hiệu quả
- Tự động hóa các hoạt động CRM

Ứng dụng Khai phá văn bản

Tài chính và tuân thủ pháp lý

Nguồn dữ liệu:

- Báo cáo và tin tức tài chính
- Tài liệu doanh nghiệp
- Đăng ký kinh doanh

Mục tiêu:

- Phát hiện gian lận
- Ngăn chặn rửa tiền
- Báo cáo các giao dịch bất hợp pháp

Ứng dụng Khai phá văn bản

An toàn công cộng

Nguồn dữ liệu:

- Báo cáo và hồ sơ không lưu
- Hồ sơ cảnh sát
- Hồ sơ y tế

Mục tiêu:

- Xác định tốt hơn các nguyên nhân để tránh sai sót trong tương lai

Ứng dụng Khai phá văn bản

Quản lý chăm sóc sức khỏe

Nguồn dữ liệu:

- Thử nghiệm lâm sàng
- Hồ sơ bệnh nhân
- Quy định pháp lý và y tế
- Các bài báo y khoa

Mục tiêu:

- Cải thiện chẩn đoán và điều trị
- Thúc đẩy dịch vụ chất lượng cao
- Kiểm soát chi phí
- Thiết kế thuốc

Ứng dụng Khai phá văn bản

Tình báo và chống khủng bố

Nguồn dữ liệu:

- Ghi chú và báo cáo điều tra
- Tài liệu thu giữ

Mục tiêu:

- Phân tích mạng lưới tổ chức nguy hiểm
- Nhận diện mô hình hành vi
- Phát hiện mô hình tấn công
- Phát triển chiến lược đối phó

Các ứng dụng Khai phá văn bản và tỷ lệ sử dụng

Ứng dụng phổ biến (>80%)

- Phân tích dự đoán (Predictive Analytics): 90%
- Ứng dụng tìm kiếm (Search & Search-based Apps): 86%
- Phân tích thông tin kinh doanh (Business Intelligence): 84%
- Phân tích phản hồi khách hàng (Voice of the Customer): 82%
- Hỗ trợ quyết định và quản lý tri thức (Decision Support, Knowledge Management): 81%

Ứng dụng trung bình (70-79%)

- Mạng xã hội (Social Media): 75%
- Dữ liệu lớn khác (Big Data - other): 70%

Ứng dụng cơ bản (<70%)

- Tổng đài và hỗ trợ kỹ thuật (Call Center, Tech Support): 63%
- Tài chính (Finance): 61%
- Rủi ro, tuân thủ và quản trị (Risk, Compliance, Governance): 61%
- Bảo mật và phát hiện gian lận (Security, Fraud Detection): 54%


Nstein
Powering Digital Publishing
Powerset
SAP
fast
A Microsoft Subsidiary
temis
CLEARFOREST
Autonomy
THOMSON REUTERS
SPSS
QL2
manageable data | on demand
Lexalytics
an infonic company
INQUIRA
clarabridge
UIMA
sas
recommind
attensity
IBM
Progress
EasyAsk
ENDECA
net'ēmic
Understand the Internet
ATTIVIO
SYBASE
Lucene
Megaputer
Baynote
Recommendations
WolframAlpha
computational knowledge engine
Vivísimo
[Search Done Right]
LEXIMANCER

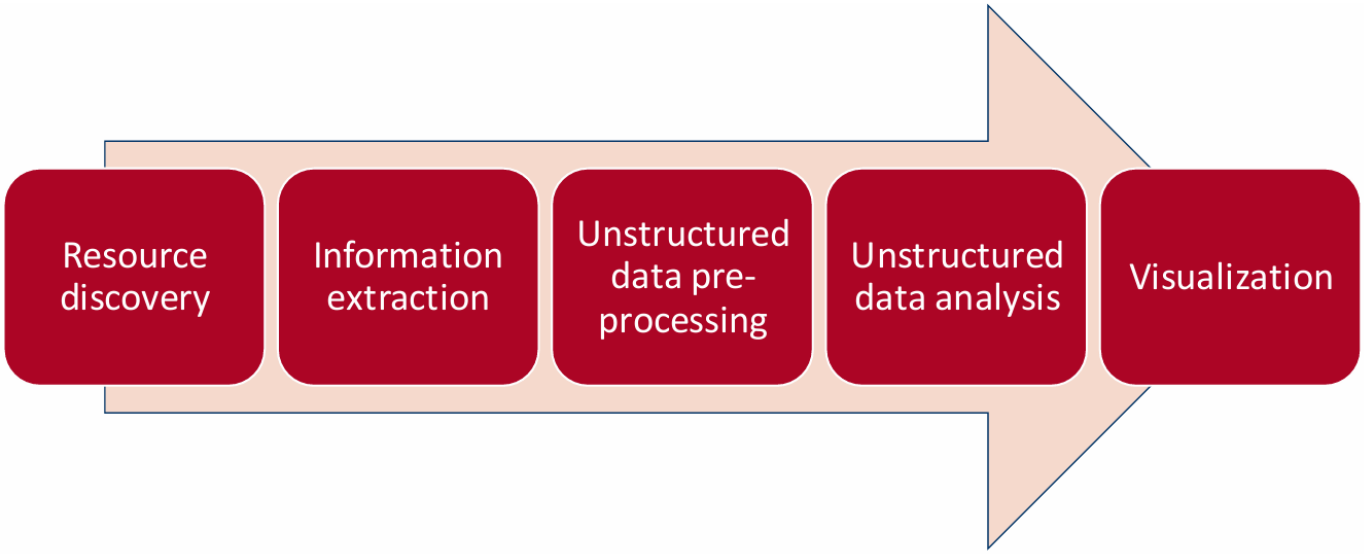
Khai phá Web (Web Mining)

- **Định nghĩa**
  - Quá trình tự động tìm kiếm và trích xuất thông tin từ nguồn trực tuyến
  - Tập trung vào dữ liệu "ẩn" trong các tài liệu siêu văn bản (hypertext)
- **World Wide Web:**
  - Nguồn thông tin công khai lớn nhất thế giới

Khai phá Web (Web Mining)

- **World Wide Web là nguồn thông tin và dữ liệu công khai lớn nhất trên Trái đất, bao gồm:**
- **Đa dạng nguồn nội dung**
  - Các cổng thông tin đa dạng với tin tức, bài báo và tập tin
  - Diễn đàn và blog chứa bình luận, đánh giá và xếp hạng
  - Trang web của các cơ quan công quyền với dữ liệu và thống kê
  - Các cửa hàng trực tuyến cung cấp thông tin sản phẩm và danh mục
  - Dịch vụ điện tử như ngân hàng, viễn thông và học trực tuyến
- **Đặc trưng cấu trúc**
  - Là tập hợp khổng lồ các tài liệu phi cấu trúc được liên kết với nhau qua hệ thống siêu liên kết.

Khai phá Web (Web Mining)



Khai phá Web (Web Mining) gồm 3 loại:

1. **Khai phá nội dung Web (Web Content Mining)**

- Trích xuất thông tin từ các tài nguyên Web như: văn bản, hình vẽ, số liệu, âm thanh và video, v.v.

2. **Khai phá cấu trúc Web (Web Structure Mining)**

- Phân tích cấu trúc các liên kết trong tài nguyên Web như: siêu liên kết (hyperlinks), đường dẫn (links), đánh dấu trang (bookmarks), v.v.

3. **Khai phá sử dụng Web (Web Usage Mining)**

- Phân tích cách thức người dùng sử dụng Web thông qua: nhật ký máy chủ (server logs), nhận dạng người dùng (user identification), v.v.

Khai phá Web bao gồm các ứng dụng sau:

1. **Nhận diện các chủ đề được người dùng đề cập trên một trang web cụ thể**

- Phân tích và xác định những nội dung, vấn đề mà người dùng thường trao đổi/thảo luận trên website

2. **Lọc trang web theo thông tin và mẫu liên kết**

- Sàng lọc các trang web dựa trên nội dung thông tin và cách thức liên kết giữa chúng

3. **Bảo mật Internet:**

- Giám sát trò chuyện (đối với trẻ em)
- Nhận diện thư rác trong email

Khai phá ý kiến (Opinion Mining):

- Nhiều trang web chứa số lượng lớn bình luận và đánh giá của khách hàng
- Có quá nhiều thông tin và bạn không muốn dành nhiều thời gian đọc đánh giá sách hơn là đọc chính cuốn sách đó
- Các công ty gặp khó khăn trong việc theo dõi tất cả các đánh giá xuất hiện trên web về sản phẩm của họ

knowledge available to humanity is in 90% textual!

Text mining techniques

Kỹ thuật khai phá văn bản (Text Mining):

1. **Tần suất từ (Word Frequency)**

- Đo lường các từ hoặc khái niệm xuất hiện thường xuyên nhất trong một văn bản; giúp phân tích các từ hoặc cụm từ mà khách hàng sử dụng nhiều nhất

2. **Liên kết từ (Word Association)**

- Một hình thức phân tích nội dung dữ liệu văn bản để tìm kiếm mối quan hệ giữa các thuật ngữ

3. **Kết hợp từ (Collocation)**

- Xác định cấu trúc ngữ nghĩa ẩn bằng cách đếm các cụm từ đôi (bigrams) và cụm từ ba (trigrams) như một từ/thuật ngữ đơn lẻ

Kỹ thuật khai phá văn bản

1. **Trích xuất từ khóa (Keyword Extraction):**

- Từ khóa là những thuật ngữ quan trọng nhất trong văn bản, những từ tóm tắt nội dung văn bản dưới dạng danh sách
- Trích xuất từ khóa có thể được sử dụng để lập chỉ mục dữ liệu phục vụ tìm kiếm và tạo đám mây từ (biểu diễn trực quan dữ liệu văn bản)

2. **Nhận dạng thực thể (Entity Recognition):**

- Nhận dạng thực thể có tên (NER) tìm kiếm các thực thể như người, công ty hoặc địa điểm tồn tại trong văn bản

3. **Phân loại văn bản (Text Classification):**

- Quá trình gán nhãn hoặc danh mục đã định trước cho nội dung văn bản phi cấu trúc
- Nhiệm vụ phân loại văn bản phổ biến nhất: phân tích cảm xúc

4. **Phân tích cảm xúc (Sentiment Analysis):**

- Quy trình tự động hiểu ý định cảm xúc của từ ngữ để suy luận xem một đoạn văn bản là tích cực, tiêu cực hay trung tính

Kỹ thuật khai phá văn bản

1. **Mô hình hóa chủ đề (Topic Modeling):**

- Là phương pháp phân loại văn bản không giám sát, tìm các nhóm tự nhiên của các mục (tương tự như phân cụm trên dữ liệu số)
- Phương pháp phổ biến để xây dựng mô hình chủ đề: LDA

2. **Phân bố Dirichlet ẩn (Latent Dirichlet Allocation - LDA):**

- Xử lý mỗi tài liệu như một hỗn hợp các chủ đề, và mỗi chủ đề như một hỗn hợp các từ
- Cho phép các tài liệu "chồng chéo" nhau về mặt nội dung, thay vì bị tách thành các nhóm riêng biệt

3. **Phân cụm (Clustering):**

- Phương pháp học máy không giám sát khai thác văn bản và tìm các nhóm tài liệu có nội dung tương tự (tập trung vào một tập các từ tương tự hoặc n-grams)

4. **Trực quan hóa văn bản (Text Visualization):**

- Biểu diễn thông tin văn bản lớn thành bố cục bản đồ trực quan, cung cấp khả năng duyệt nâng cao cùng với tìm kiếm đơn giản
- Các phương pháp trực quan hóa có thể cải thiện và đơn giản hóa việc khám phá thông tin liên quan