Sự phát triển của ngành khai phá dữ liệu và một vài kết quả nghiên cứu

(Data mining development and some research results)

Hồ Tú Bảo
Japan Advanced Institute of Science and Technology
Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ Việt Nam

Institute of Information Technology
Vietnamese Academy of Science & Technology

SC of PAKDD, PRICAI, ACML

Japan Advanced Institute of Science and Technology

Báo cáo trình bày

➔ Về ngành khai phá dữ liệu: Sự phát triển và mối liên quan với xác suất-thống kê.

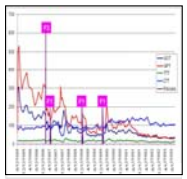
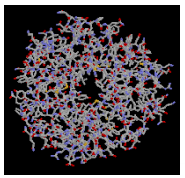
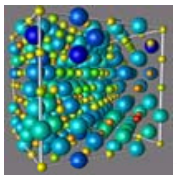
➔ Vài kết quả trong các lĩnh vực:

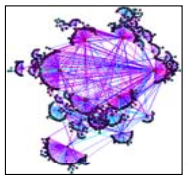
- ❑ Phương pháp kernel
- ❑ Khai phá văn bản
- ❑ Tính toán y-sinh học

What motivated data mining?

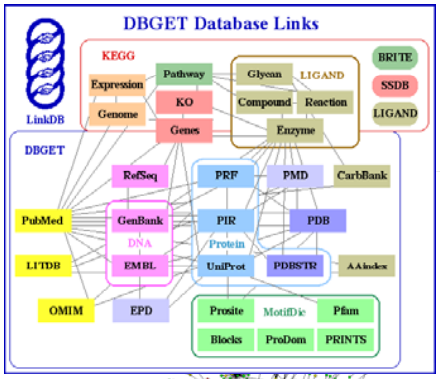
Ta đang sống trong một thời sôi động nhất: Máy tính và mạng máy tính (internet)

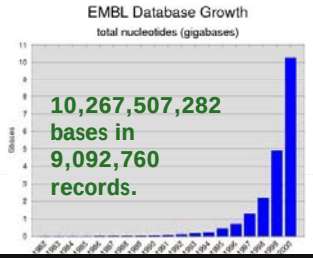
- Rất nhiều dữ liệu hơn bao giờ hết ở quanh ta.
  Chúng được thu thập và lưu trữ trong các CSDL khổng lồ (hàng triệu bản ghi, hàng nghìn thuộc tính).
- Rất nhiều dữ liệu có cấu trúc phức tạp và không ở dạng vectơ (complexly structured data and non-vectorial).

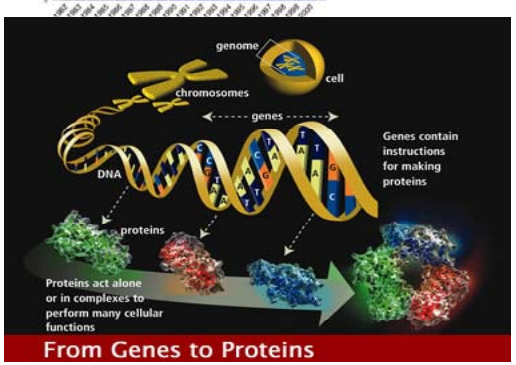






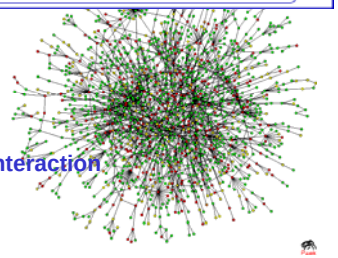
Explosion of biological data







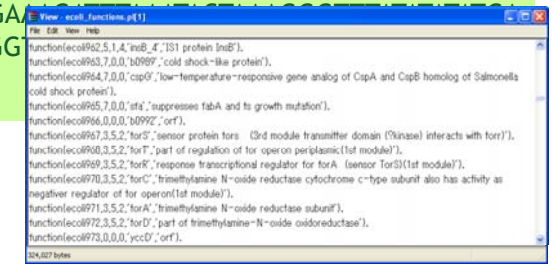
Protein interaction network



How biological data look like?

Đây là một mẩu của chuỗi DNA, toàn bộ chuỗi gồm 1.6 triệu chữ cái. Đoạn DNA này chỉ có 350 chữ cái, 4570 nhỏ hơn chuỗi DNA:

…TACATTAGTTATTACATTGAGAAACTTTATAATTAAAAAAGATTCATGTAAATTTCT
TATTTGTTTATTTAGAGGTTTTAAATTTAATTTCTAAGGGTTTGCTGGTTTCATTGT
TAGAATATTTAACTTAATCAAATTATTTGAATTTTTGAAAATTAGGATTAATTAGGTA
AGTAAATAAAATTTCTCTAACAAATAAGTTAAATTTTTAAATTTAAGGAGATAAAAAT
ACTACTCTGTTTTATTATGGAAAGA
AGTAGTTACCCTTAGAAAAATATGG
TATATTATGT…

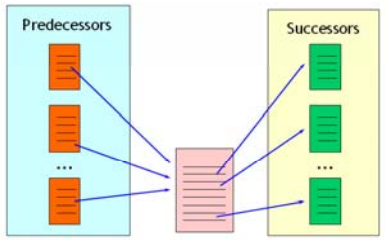


Rất nhiều loại dữ liệu sinh học khác.

Text: huge sources of knowledge

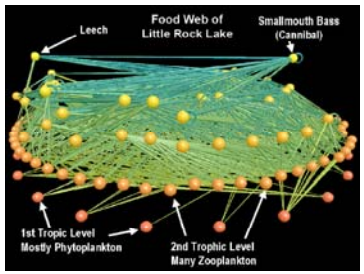
- Khoảng **80%** dữ liệu trên đời không phải dữ liệu số, mà phần lớn ở dạng chữ (source: Oracle Corporation)
- Cơ sở dữ liệu các công bố khoa học
  - Cơ sở dữ liệu ISI (Thomson-Reuteur): SCI/SCIE, SSCI, A&HCI, etc.
  - Cơ sở dữ liệu SCOPUS (Elsevier)
  - Cơ sở dữ liệu Google Scholar (Google)
  - Hơn 100 cơ sở dữ liệu và công cụ khác: arXiv, CiteSeer, ScienceDirect, SciFinder Scholar, MEDLINE, …
  - Hệ thống bằng sáng chế phát minh, chẳng hạn số đăng ký hằng năm trên thế giới là 1.6 triệu và ở Nhật có khoảng 400 nghìn.
- Khai thác các CSDL này thế nào để đánh giá chất lượng các công bố, xu thế trong khoa học, tìm kiếm đúng bài báo hay patent mong muốn, etc.?

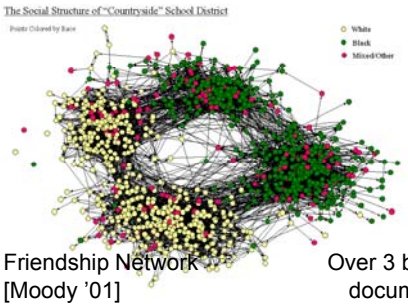
Web link data



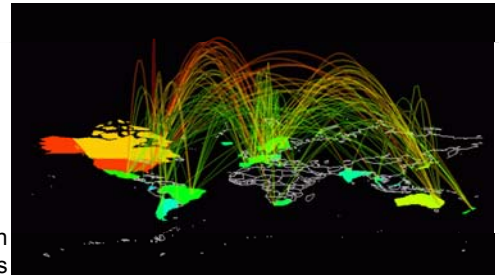
Web link
Web content
Web log

Food Web [Martinez '91]





Friendship Network
[Moody '01]



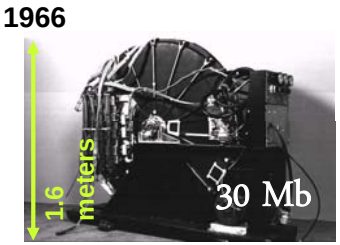
Over 3 billion documents

Multivariate analysis

- Phân tích dữ liệu khám phá (EDA: exploratory data analysis) nhấn mạnh dùng dữ liệu sinh ra các giả thiết.
- Phân tích dữ liệu kiểm định (CDA: confirmatory data analysis) nhấn mạnh kiểm định giả thiết.
  - Factor analysis; Principal component analysis (PCA)
  - Regression analysis
  - Linear discriminant analysis: taxonomy
  - Cluster analysis
- Thấy gì từ các phương pháp truyền thống?
  - Không chạy được/nghèo nàn với dữ liệu rất lớn và phức tạp
  - Các phương pháp cổ điển chỉ phân tích được các tập dữ liệu nhỏ
  - Chi phí để lưu trữ và xử lý dữ liệu là trở ngại lớn trong nhiều thập kỷ.

Multivariate analysis

- Các phương pháp được tạo ra khi các tập dữ liệu nhỏ và vừa đang phổ biến, khi tính toán chỉ thực hiện được trên các máy tính còn yếu.
- Do tốc độ và hiệu quả tính toán được cải thiện, rất nhiều phương pháp phân tích dữ liệu nhiều chiều được tạo ra để giải quyết các bài toán dữ liệu kích thước lớn.
  - ➔ projection pursuit (friedman, 1984)
  - ➔ neural networks
  - ➔ reduced-rank regression
  - ➔ nonlinear manifold learning
  - ➔ independent component analysis
  - ➔ kernel methods, support vector machines
  - ➔ random forests (Breiman, 2001)



What is data mining? Khai phá dữ liệu

"Khám phá các mô hình và mẫu dạng từ các tập dữ liệu lớn thu thập được"

Data mining metaphor:
Extracting ore from rock

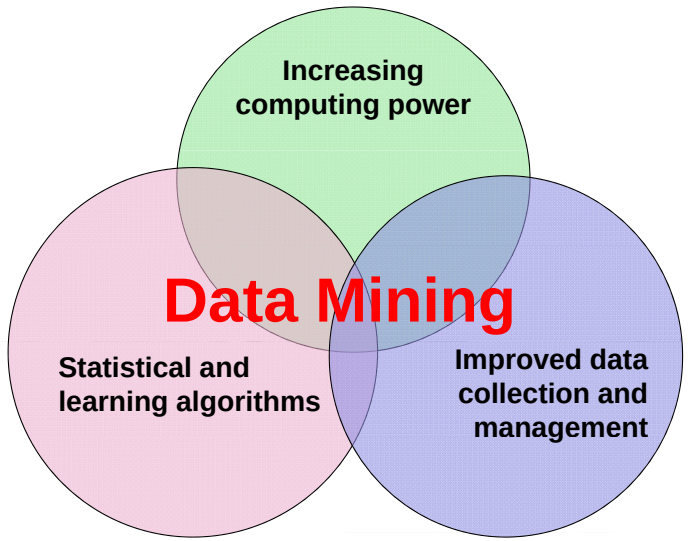
Languages, Representations

Engineering, Data Management

Applications

ACM SIGKDD (1995), ECML/PKDD (1997), PAKDD (1997), IEEE ICDM SIAM DM, …

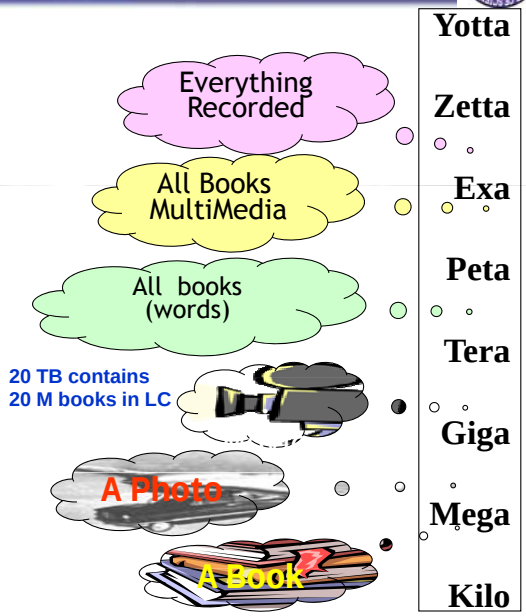
Convergence of three technologies



Improved data collection and management

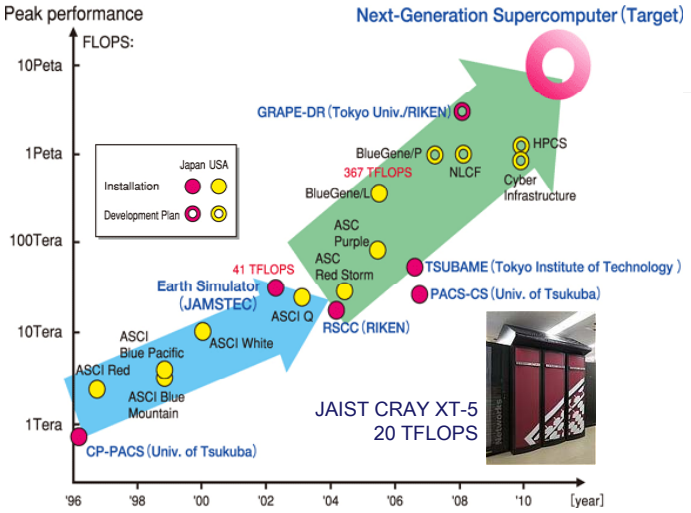
- Mọi thứ trên đời sẽ sớm được biểu diễn và lưu trữ trên máy tính
- Hầu hết các dữ liệu này chẳng bao giờ được chứng ta ngó ngàng tới
- Đâu là công nghệ chủ để đối mặt với nguồn dữ liệu và thông tin khổng lồ này?

["How much information is there?" Adapted from the invited talk of Jim Gray (Microsoft) at KDD'2003]



Increasing computing power

National key project (2007-2012)



ODRA1304

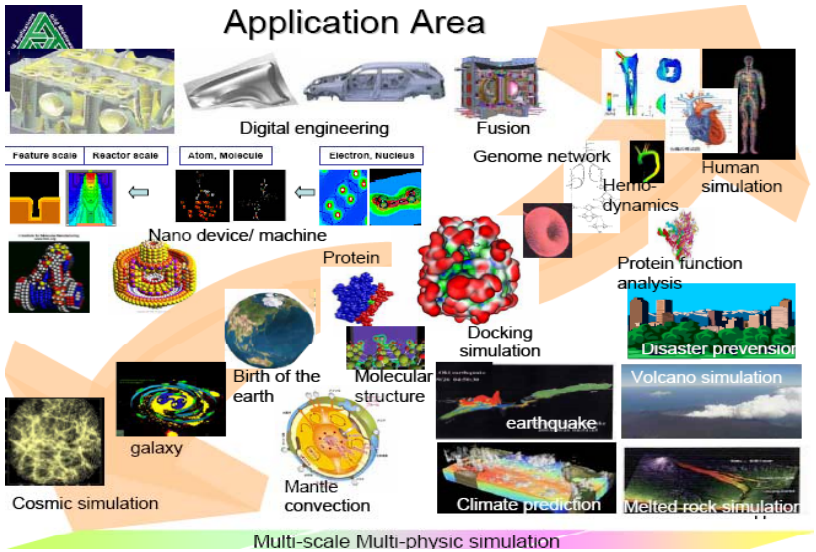
GPGPU

Tesla S20, 2.1 ~ 2.5 TFLOPS, 12,995 USD

II

Cray XT-3, 2.2 TFLOPS

Next generation supercomputer project

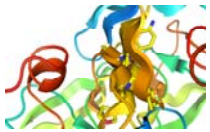


(Kennichi Miura, DEISA Symposium, 5.2007)

Data schemas vs. mining methods

Types of data

- Flat data tables
- Relational databases
- Temporal & spatial data
- Transactional databases
- Multimedia data
- Genome databases
- Materials science data
- Textual data
- Web data
- etc.



Key issue: Model selection

Mining tasks and methods

- **Classification/Prediction**
  - → Decision trees
  - → Neural networks
  - → Rule induction
  - → Support vector machines
  - → Hidden Markov Model
  - → etc.
- **Description**
  - → Association analysis
  - → Clustering
  - → Summarization
  - → Trend detection
  - → etc.

Challenges in data mining

Large data sets (10^6-10^{12} objects) and high dimensionality (10^2-10^3 attributes)
[Problems: efficiency, scalability?]

Different types of data in complex forms (mixed numeric, symbolic, text, image, voice,…)
[Problems: quality, effectiveness?]

Data and knowledge are changing

Human-computer interaction and visualization

10 challenging problems in data mining
(J. IT&DM, Vol.5(4), 2006)

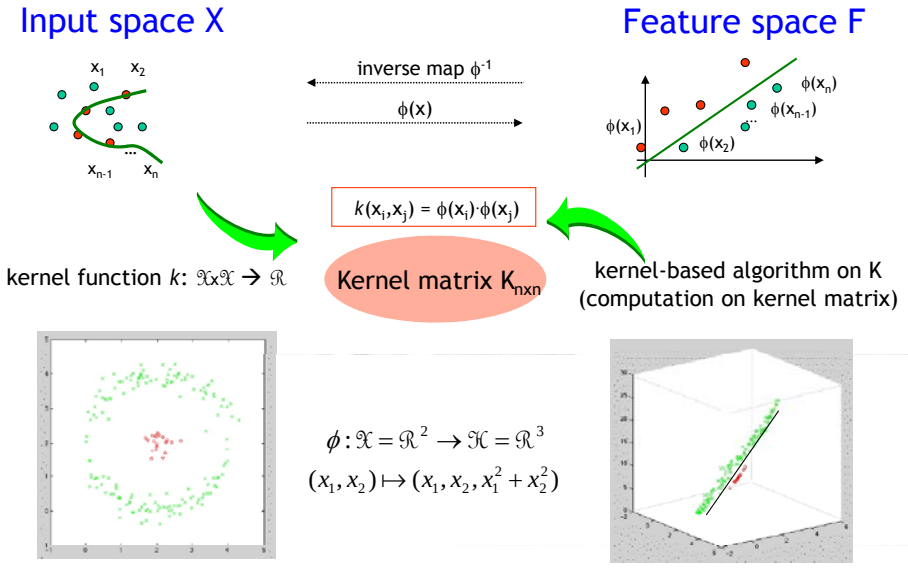
1. Developing a unifying theory of data mining
2. Scaling up for high dimensional data/high speed streams
3. Mining sequence data and time series data
4. Mining complex knowledge from complex data
5. Data mining in a network setting
6. Distributed data mining and mining multi-agent data
7. Data mining for biological and environmental problems
8. Data-mining-process related problems
9. Security, privacy and data integrity
10. Dealing with non-static, imbalanced and cost-sensitive data

The talk aims to …

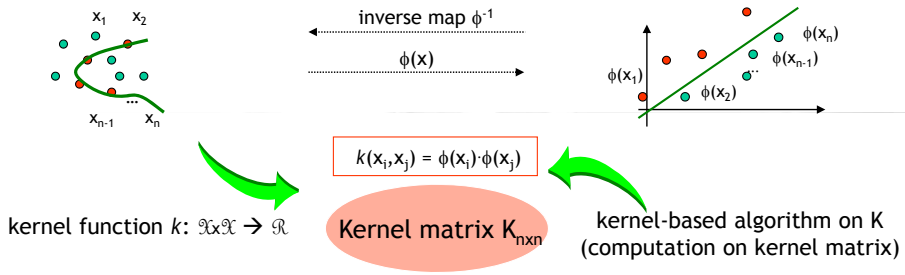
➔ Về ngành khai phá dữ liệu: Sự phát triển và mối liên quan với xác suất-thống kê.

➔ Vài kết quả trong các lĩnh vực:
- ❑ Phương pháp kernel
- ❑ Khai phá văn
- ❑ Tính toán y-sinh học.

Kernel methods: the basic ideas



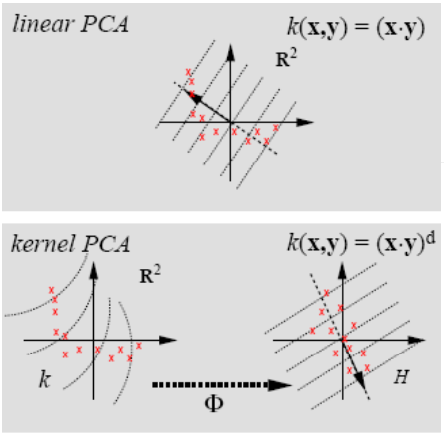
Kernel methods: math background



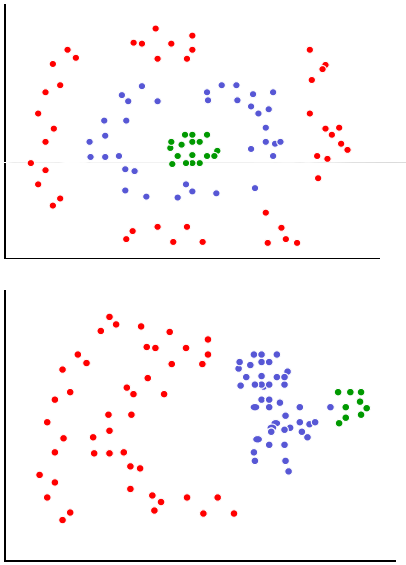
Linear algebra, probability/statistics, functional analysis, optimization

- Mercer theorem: Any positive definite function can be written as an inner product in some feature space.
- Kernel trick: Using kernel matrix instead of inner product in the feature space.
- Representer theorem (Wahba): Every minimizer of $\min_{f\in\mathcal{X}}\{C(f,\{x_i,y_i\})+\Omega(\|f\|_H)$ admits a representation of the form $f(.)=\sum_{i=1}^{m}\alpha_i K(.,x_i)$

Kernel PCA



Sử dụng hàm kernel, các toán tử tuyến tính ban đầu của PCA được thực hiện trong một không gian Hilbert kernel tái tạo với một ánh xạ phi tuyến.



Kernel matrix evaluation

- Chỉ ra độ đo phổ biến và hiệu quả nhất KTA (Kernel Target Alignment, 2001) đánh giá các ma trận kernel có những hạn chế lớn (điều kiện đủ nhưng không cần).
- Đề xuất độ đo mới FSM (Feature Space-based Kernel Matrix Evaluation Measure) dùng phân bố dữ liệu trong không gian feature. FSM hiệu quả, có những tính chất mong muốn.
- Implication of FSM is vast.

Comparing directly matrices in the input spaces

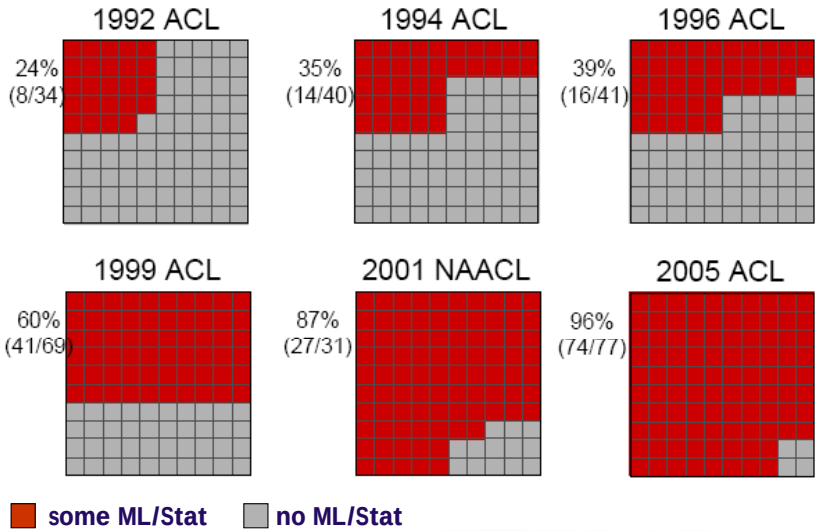
$$KTA(K,y) = \frac{\langle K, y.y^T\rangle_F}{\sqrt{\langle K,K\rangle_F \langle y.y^T, y.y^T\rangle_F}}$$

Comparing data images distributions in the feature space

$$FSM(K,y) = \frac{var_+ + var_-}{\|\phi_- - \phi_+\|}$$

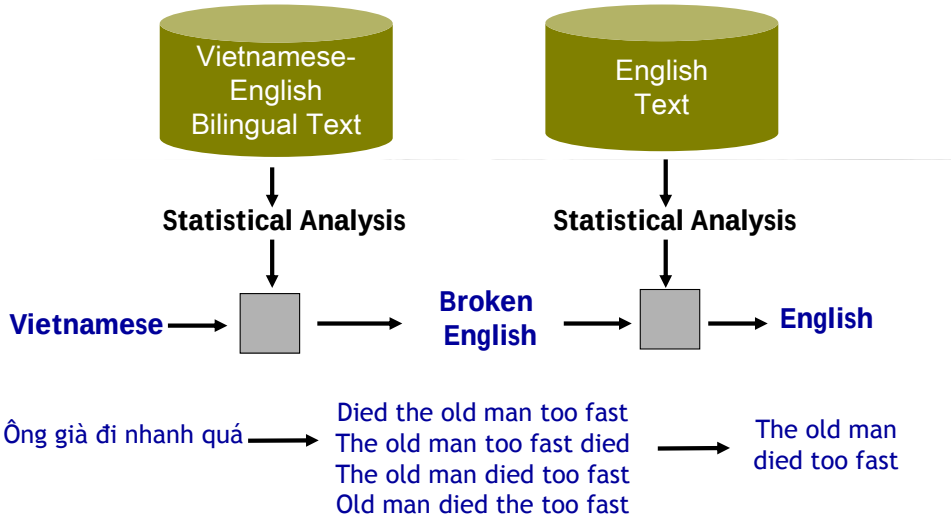
(Nguyen Canh Hao, Ho Tu Bao, Pattern Recognition, 2008)

Statistics-based natural language processing



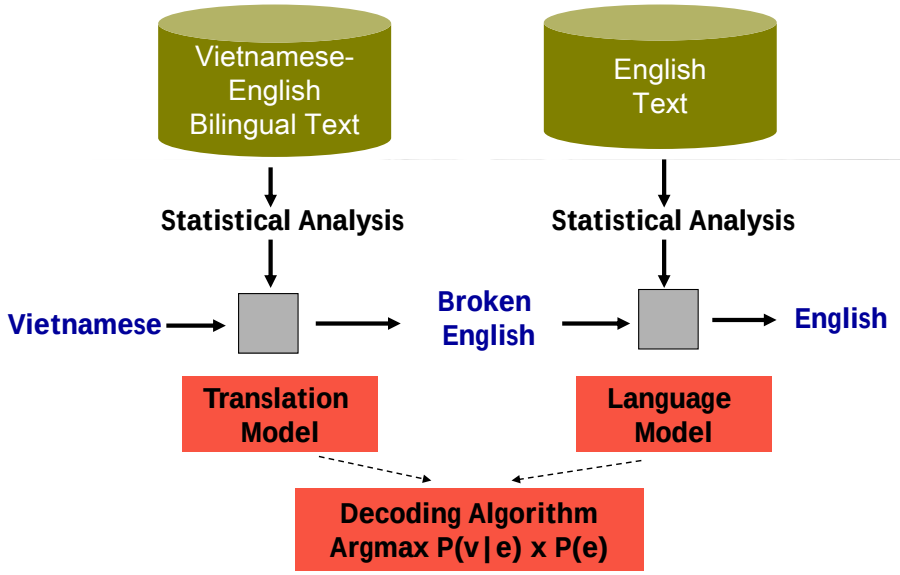
(from Marie Claire talk, ECML/PKDD 2005)

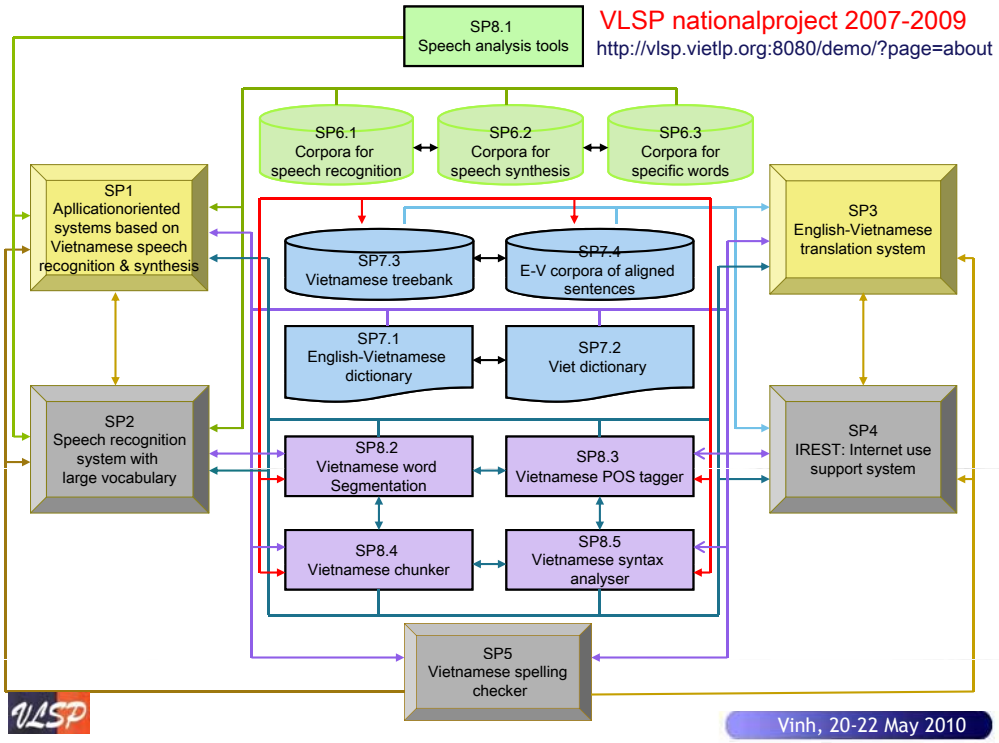
Statistical machine translation



(tutorial on SMT, K. Knight and P. Koehn)

Statistical machine translation





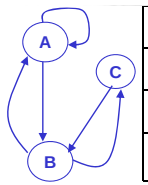
PageRank algorithm (Google)

- Google từ'weather forecast' → 4.2 triệu trang web.
- Làm sao Google biết các trang web quan trọng và liên quan nhất?
- Google gán cho mỗi trang web một con số (PageRank number) tính qua việc giải bài toán

$$Pw = \lambda w$$

- Kính thước (2007): 4.2×10^9

Larry Page, Sergey Brin

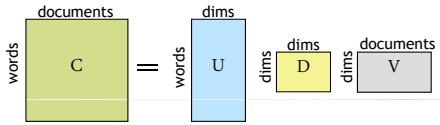


	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

Để máy hiểu được nghĩa các văn bản?

- Tìm tài liệu trên Google liên quan 3 chủ đề "thực phẩm", "mắm tôm", "dịch bệnh".
- Google cho ra rất nhiều tài liệu, với precision và recall thấp.
- Làm sao máy tính hiểu được nội dung văn bản để tìm kiếm cho hiệu quả?
- Thông qua chủ đề của văn bản
- Latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999): Biểu diễn văn bản trong một không gian Euclid, mỗi chiều là một tổ hợp tuyến tính các từ (giống PCA).

Latent semantic analysis



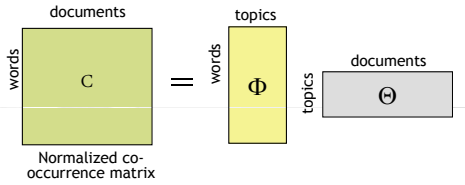
	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

	D1	D2	D3	D4	D5	D6	Q1
dim1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
dim2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

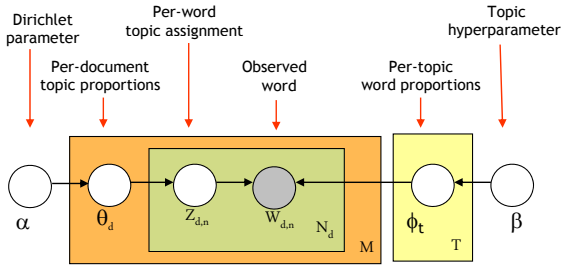
Topic modeling: key ideas

- Topic modeling key idea (LDA, Blei, JMLR 2004)
  - → mỗi văn bản là một mixture của các chủ đề
  - → mỗi chủ đề là một phân bố xác suất trên các từ.
- Thí dụ
  - → "thực phẩm" = {an toàn, rau, thịt, cá, không ngộ độc, không đau bụng …}
  - → "mắm tôm" = {tôm, mặn, đậu phụ, thịt chó, lòng lợn, …}
  - → "dịch bệnh" = {nhiều người, cấp cứu, bệnh viện, thuốc, vacine, mùa hè, … }
  - → D1= {thực phẩm 0.6, mắm tôm 0.35, dịch bệnh 0.8}

Topic modeling



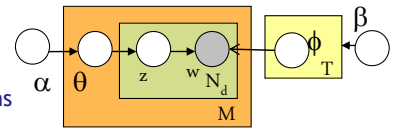
Latent Dirichlet Allocation (LDA)



Latent Dirichlet allocation (LDA) model

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

Dirichlet prior on the per-document topic distributions



$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)\,p(w_n|z_n,\beta)$$

Joint distribution of topic mixture θ, a set of N topic **z**, a set of N words **w**

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)\,p(w_n|z_n,\beta)\right)d^k\theta$$

Marginal distribution of a document by integrating over θ and summing over z

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)\,p(w_{dn}|z_{dn},\beta)\right)d^k\theta_d$$

Probability of collection by product of marginal probabilities of single documents

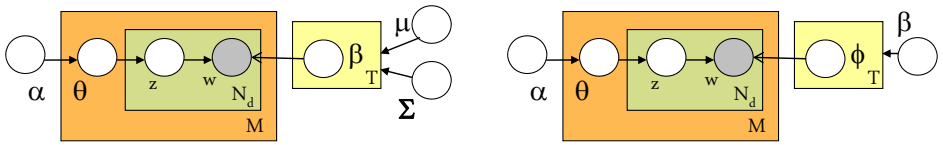
Example of topics learned

- From 16000 documents of AP corpus → 100-topic LDA model.
- Each color codes a different factor from which the word is putatively generated

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Dirichlet-Lognormal (DLN) topic model



Model descrption:

$$\theta \mid \alpha \sim Dirichlet(\alpha)$$
$$\beta \mid \mu, \Sigma \sim Lognormal(\mu, \Sigma)$$
$$z \mid \theta \sim Multinomial(\theta)$$
$$w \mid \beta \sim Multinormial(f(\beta))$$

$$\Pr(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\Sigma}.x_1\ldots x_n} \exp\left\{-\frac{1}{2}(\log x-\mu)^T\Sigma^{-1}(\log x-\mu)\right\}$$
$$where\ \log x = (\log x_1,\ldots,\log x_n)^T$$

Spam classification

Method	DLN	LDA	SVM
Accuracy	**0.5937**	0.4984	0.4945

Predicting crime

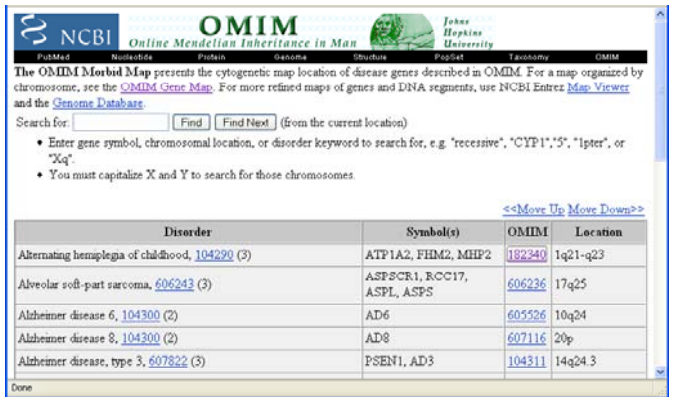
DLN	LDA	SVM
0.2442	0.1035	0.2261

(Than Quang Khoat, Ho Tu Bao, 2010)

Dự đoán genes gây bệnh bằng tính toán

Problem

- Đã biết 3,053 genes trong số 25.000-30.000 genes của người là genes gây bệnh, thông báo trong CSDL OMIM.
- Dự đoán (tính toán) xem có genes nào khác cũng là genes gây bệnh?

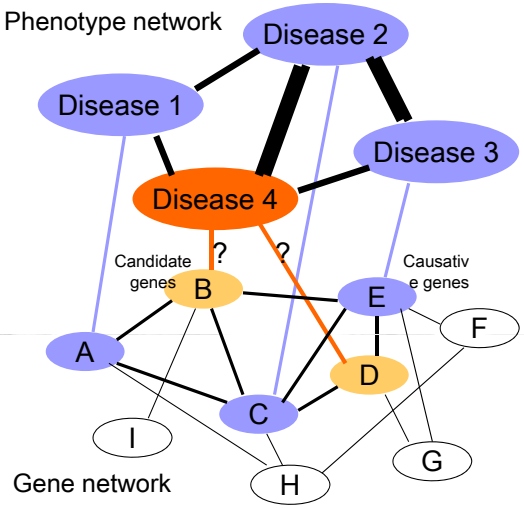


Dự đoán genes gây bệnh

Key assumption

Láng giềng của một gene trong mạng lưới các genes gây bệnh cũng có nhiều khả năng gây ra bệnh đó hoặc bệnh tương tự.

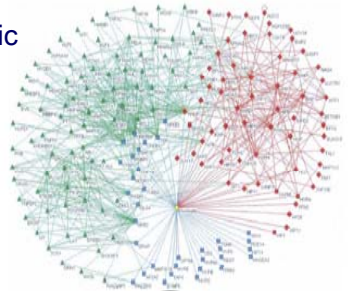
(Goh et al. 2007;
Oti and Brunner 2007).



(Reproduced from www.blackwell-synergycom)

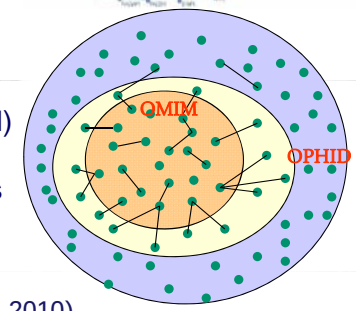
Key idea of the method

1. Phát hiện các sự kiện nền genomic/proteomic của proteins và tương tác protein từ nhiều CSDL → xây dựng mạng tương tác proteins.



the Gene Ontology · InterDom · InterPro · Pfam · prosite · PRINTS · UniProt · mips

2. Phát triển một phương pháp mới về học với dữ liệu một phần có nhãn (semi-supervised) để kiểm chứng giả thuyết cơ bản về genes gây bệnh dựa trên mạng tương tác proteins → đoán 568 genes có khả năng gây bệnh.



.

(Nguyen Thanh Phuong, Ho Tu Bao, AI in Medicine, 2010)

- **Prof. Katsuhiko Takabayashi, Head, Division of Medical Informatics and Management**
- **Prof. Osamu Yokosuka, Prof. Tatsuo Kanda, Department of Internal Medicine, Chiba Univ.**

Take-home message

- Nhiều dữ liệu và nhiều kiểu dữ liệu mới, khả năng tính toán mạnh hơn → cần mô hình và công cụ xác suất-thống kê mới thích hợp.
- Khai phá dữ liệu (data mining) là lĩnh vực mới ra đời từ đòi hỏi thực tế, đang phát triển sôi động → liên quan đến những thay đổi sâu sắc trong xác suất-thống kê.
- Nhận thức được sự thay đổi này và các thách thức lý thuyết cũng như giải pháp thực tiễn cho các nhu cầu.
- Hướng đến những xu thế mới, tới những vấn đề lý thuyết và ứng dụng cần thiết, nhiều ý nghĩa?
  (L. Breiman, D. Han, M. Jordan, …)

- Izenman, A.J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer.

- **http://www.jaist.ac.jp/~bao**