Finding sense in large sets of data

Some powerful (novel) applications of Math and CS

András Benczúr, jr., Lajos Rónyai
MTA SZTAKI, BME

Szeged, June 30, 2006

BusinessWeek online

TOP NEWS | BW MAGAZINE | INVESTING | ASIA | EUROPE

JANUARY 23, 2006

COVER STORY

Math Will Rock Your World

A generation ago, quants turned finance upside down. Now they're mapping out ad campaigns and building new businesses from mountains of personal data

COVER STORY PODCAST

Neal Goldman is a math entrepreneur. He works on Wall Street, where numbers rule. But he's focusing his analytic tools on a different realm altogether: the world of words.

-

State of the Union Address, President Bush, 2006

Our greatest advantage in the world has always been our educated, hardworking, ambitious people – and we're going to keep that edge. Tonight I announce an American Competitiveness Initiative, to encourage innovation throughout our economy, and to give our nation's children a firm grounding in math and science. (Applause.)

. . .

Third, we need to encourage children to take more math and science, and to make sure those courses are rigorous enough to compete with other nations.

Information explosion

- Exlposion in size
  - Hardware speed increases very fast
  - Amount of data grows even faster
  - Speed of external storage devices grows much slower

Information explosion

- Exlposion in size
  - Hardware speed increases very fast
  - Amount of data grows even faster
  - Speed of external storage devices grows much slower
- Inhomogeneity, losing much structure
- Manipulative contents, spam

Information explosion

- Exlposion in size
  - Hardware speed increases very fast
  - Amount of data grows even faster
  - Speed of external storage devices grows much slower
- Inhomogeneity, losing much structure
- Manipulative contents, spam
- Links in networks take center stage
  - Telecommunication, internet, social nets
  - Hyperlinks
  - Small world

Some application areas

- Data mining
  - Knowledge of customers, transaction logs
  - Network security supervision, anomaly detection, alarm management

Some application areas

- Data mining
  - Knowledge of customers, transaction logs
  - Network security supervision, anomaly detection, alarm management
- Exploring the structure of documents
  - Law, law enforcement, monitoring competition, ...

Some application areas

- Data mining
  - Knowledge of customers, transaction logs
  - Network security supervision, anomaly detection, alarm management
- Exploring the structure of documents
  - Law, law enforcement, monitoring competition, . . .
- Intensive use of the internet
  - Web search
  - Content filtering

Some application areas

- Data mining
  - Knowledge of customers, transaction logs
  - Network security supervision, anomaly detection, alarm management
- Exploring the structure of documents
  - Law, law enforcement, monitoring competition, ...
- Intensive use of the internet
  - Web search
  - Content filtering
- Bioinformatics
  - Gigantic problems from genomics
  - Millions of synthesized molecules as potential medications

Some application areas

- Data mining
  - Knowledge of customers, transaction logs
  - Network security supervision, anomaly detection, alarm management
- Exploring the structure of documents
  - Law, law enforcement, monitoring competition, ...
- Intensive use of the internet
  - Web search
  - Content filtering
- Bioinformatics
  - Gigantic problems from genomics
  - Millions of synthesized molecules as potential medications
- Fundamental need for privacy; cryptography

New breed of companies

RSA Security: Founded by the discoverers of RSA crypto algorithms. After a long dry period they flourish from the 90s.

New breed of companies

RSA Security: Founded by the discoverers of RSA crypto algorithms. After a long dry period they flourish from the 90s.

Google: Started as a PhD project at Stanford University – in three years they became one of the flagships of the WWW. At least 5,000 researchers/developers, more than 100,000 computers (estimated).

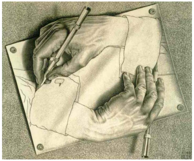
New breed of companies

RSA Security: Founded by the discoverers of RSA crypto algorithms. After a long dry period they flourish from the 90s.

Google: Started as a PhD project at Stanford University – in three years they became one of the flagships of the WWW. At least 5,000 researchers/developers, more than 100,000 computers (estimated).

Akamai: Founded by MIT professors. In the explosive period of the Net (1998–2001) offered infrastructure for fast access. Great survivor of *dotcom* bubble.

Ranking: millions of hits



Idea: use the human judgement and wisdom built into hyperlinks.

- Google: PageRank [Brin, Page 98] the stationary distribution of a stochastic surfer.
  A page is good if it is pointed to by many good pages.
- Teoma: HITS [Kleinberg 98] *authority* and *hub* pages.
  An authority is pointed to by many *hubs*. A hub points to many *authorities*.

Hypertext Induced Topic Selection (HITS)

An authority is pointed to by many *hubs*:

$$au\,(k+1) = h\,(k)A$$

A hub points to many *authorities*:

$$h\,(k+1) = au\,(k+1)A^T$$

$$au\,(k+1) = au\,(1)(A^T A)^k$$

$$h\,(k+1) = h\,(1)(AA^T)^k$$

Hypertext Induced Topic Selection (HITS)

An authority is pointed to by many *hubs*:

$$au\,(k+1) \quad = \quad h\,(k)A$$

A hub points to many *authorities*:

$$h\,(k+1) \quad = \quad au\,(k+1)A^T$$

$$au\,(k+1) \quad = \quad au\,(1)(A^TA)^k = au\,(1)\cdot V \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}^k V^T$$

$$h\,(k+1) \quad = \quad h\,(1)(AA^T)^k$$

Hypertext Induced Topic Selection (HITS)

An authority is pointed to by many *hubs*:

$$au\,(k+1) = h\,(k)A$$

A hub points to many *authorities*:

$$h\,(k+1) = au\,(k+1)A^T$$

$$au\,(k+1) = au\,(1)(A^TA)^k = au\,(1)\cdot V \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}^k V^T$$

$$h\,(k+1) = h\,(1)(AA^T)^k = h\,(1)\cdot U \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}^k U^T$$

Hypertext Induced Topic Selection (HITS)

An authority is pointed to by many *hubs*:

$$au\,(k+1) = h\,(k)A$$

A hub points to many *authorities*:

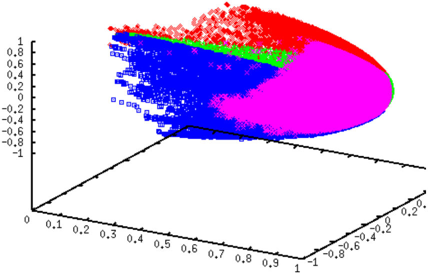
$$h\,(k+1) = au\,(k+1)A^T$$

$$au\,(k+1) = au\,(1)(A^TA)^k = au\,(1)\cdot V \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & 0 & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix}^k V^T$$

$$h\,(k+1) = h\,(1)(AA^T)^k = h\,(1)\cdot U \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & 0 & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix}^k U^T$$

Low rank approximation

A 3D projection of a graph with a "power law".



The grace of geometry: fine and subtle notions of *near* and *far*, similarity, formation of groups (chunks, clusters).

In statistics: principal component analysis. Proven its worth in many traditional applications.

Word–document matrix: geometry of texts

- *document 1*
  Some like it hot, some like it cold,
- *document 2*
  Some like it in the pot,
- *document 3*
  Nine days old.

	cold	days	hot	in	it	like	nine	old	pot	some	the
doc 1	1	0	1	0	2	2	0	0	0	2	0
doc 2	0	0	0	1	1	1	0	0	1	1	1
doc 3	0	1	0	0	0	0	1	1	0	0	0

Two texts are similar, iff their vectors point almost to the same direction (cosine measure).

Latent Semantic Indexing (LSI): synonyms, associative relations

- Topic: *outlaw*.
- Recognize that *fugitive, outcast, exile, pariah, bandit, desperado, brigand, criminal, robber, villain, Robin Hood, Rózsa Sándor, terrorist* are strongly related to that topic.
- Even if *outlaw* is not present in the text!

Latent Semantic Indexing (LSI): Dumais, Deerwester, Berry, Landauer:

- The multitude of ways to express the same content is actually *noise, uncertainity*
- Word-document matrix has way too many degrees of freedom
- Projecting documents into a space of smaller dimension

Analysis of genetical code sequences

- Handling texts at a giga-scale (e.g. complete genom alignment)

Analysis of genetical code sequences

- Handling texts at a giga-scale (e.g. complete genom alignment)
- Fast machines, fast algorithms, sophisticated data structures

Analysis of genetical code sequences

- Handling texts at a giga-scale (e.g. complete genom alignment)
- Fast machines, fast algorithms, sophisticated data structures
- *An important subproblem:* finding a longest increasing subsequence

$$\mathbf{3}, 9, 11, \mathbf{6}, \mathbf{7}, \mathbf{8}, 5, \mathbf{13}, 2, 4, \mathbf{17}$$

  There is a method with complexity $O(n \log n)$.
  Robinson–Schensted algorithm: once pure (high) mathematics...

Reconstructing phylogenetic trees

Interesting new mathematical/computational problems are born.

Saitou–Nei reconstruction method

P-tree: a tree (graph) F with positive edge weights. Weights define a metric on the set of nodes, $d(x, y)$ is the distance of x and y.

Cherry: two leaves of a tree with a common neighbour.

Problem: Suppose we know the values $d(u, v)$ for any pair of *leaves* u, v of F. Find a cherry of F.

Reconstructing phylogenetic trees (contd.)

Saitou N. and *Nei M.* – give a (linear) and conveniently computable function $\delta(x, y)$ of the given distances $d(u, v)$. Let X be the set of leaves of F. For $v \in X$ we set

$$T_v = \sum_{u \in X} d(v, u).$$

For $u, v \in X$ we have

$$\delta(u, v) := d(u, v) - \frac{T_u + T_v}{r - 2}, \text{ where } |X| = r.$$

Theorem

If $x \neq y$ are leaves of F and $\delta(x, y)$ is minimal, then x, y is a cherry.

Many applications (e.g. in CLUSTALW), $\geq 10^4$ citations.

Keeping secrets and mathematics

- Widespread need for secure communication
- The inverted perspective of cryptographers
- Matching pairs of hard and easy problems
- Design and validation of protocols
- Attacks, cryptanalysis

A legendary difficult–easy pair

The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers... Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated. (K. F. Gauß, 1801)

- Only a theoretical problem for a long time
- Decisive turn of events in the 1970-s: secure communication as an ubiquitous demand
- Theory of computation: *easy* and *difficult* problems
- RSA: factoring is difficult, recognizing primes is easy.

Concluding remarks

- Faster innovation at the leading industrial powers
- Hungary and Central Europe: no real high tech companies in the field
- Universities, research institutes harbor these cultures
- Promising example: Morgan Stanley in Budapest
- European hopes