Matematica e Mondo Reale: il problema di Google e altre storie

Dario A. Bini

Dipartimento di Matematica, Università di Pisa
www.dm.unipi.it/~bini

6 Febbraio 2007

Matematica: tra Astrazione e Applicazione

Esistono molti luoghi comuni (e tante barzellette) sulla matematica

L'immagine più benevola che si incontra nella mentalità comune è quella di una disciplina fine a sé stessa e per questo assolutamente inutile.

Il matematico è visto come un personaggio strano che vive nel suo mondo fantastico e si occupa di cose strane e di importanza trascurabile per la vita di ogni giorno.

Matematica: tra Astrazione e Applicazione

Alcune delle idee più diffuse e clamorosamente errate sono:

- rigidità del pensiero matematico capace di esprimere solo meccanismi ripetitivi;
- totale mancanza di fantasia;
- aridità e incapacità di esprimere qualcosa di nuovo;
- completa inutilità (a parte i conti della spesa e gli strumenti geometrici elementari) della matematica.

Fantasia e Creatività

Fantasia,
immaginazione,
creatività,
libertà di pensiero,
attrazione per l'eleganza,
attrazione per la singolarità,
rigore logico

sono caratteristiche presenti in chi si occupa di matematica

In matematica si generano idee continuamente nuove e il mondo matematico è per certi versi molto più ricco del mondo reale

Fantasia e Creatività

- **spazi a dimensione elevata**

  $\mathbb{R}$ insieme dei numeri reali $\leftrightarrow$ Retta

  $\mathbb{R}^2$ insieme delle coppie (x_1, x_2) $\leftrightarrow$ Piano

  $\mathbb{R}^3$ insieme delle terne (x_1, x_2, x_3) $\leftrightarrow$ Spazio

  $\mathbb{R}^n$ insieme delle n-uple $(x_1, x_2, \ldots, x_n)$ $\leftrightarrow$ Spazio n-dimensionale

  $X = (x_1, x_2)$ Distanza: $d(X, O) = \sqrt{x_1^2 + x_2^2}$

  $X = (x_1, x_2, x_3)$ Distanza: $d(X, O) = \sqrt{x_1^2 + x_2^2 + x_3^2}$

  $X = (x_1, x_2, \ldots, x_n)$ Distanza: $d(X, O) = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

  Posso costruire spazi tridimensionali ricurvi dentro spazi n-dimensionali ad esempio una “sfera” o un “toro” in $\mathbb{R}^4$.

- **geometrie non euclidee**

Fantasia e Creatività

- **creazione di infiniti non numerabili:** $\aleph_0$ e cardinalità del continuo
- **spazi a dimensione infinita**

  $\mathbb{R}^{\mathbb{N}}$: insieme delle successioni $(x_1, x_2, \ldots)$ Distanza $d(X, 0) = \sqrt{x_1^2 + x_2^2 + \cdots}$

  Attenzione che ci sono oggetti quali $(1, 1, 1, \ldots)$ o $(1, 1/\sqrt{2}, 1/\sqrt{3}, 1/\sqrt{4}, \ldots)$ che hanno distanza dall'origine infinita.
- **spazi funzionali:** con un numero *non numerabile* di dimensioni: $\mathbb{R}^{\mathbb{R}}$

  Insieme di tutte le funzioni da $[a, b] \subset \mathbb{R}$ in $\mathbb{R}$

  distanza $d(0, f) = \left(\int_a^b f(x)^2 dx\right)^{1/2}$

  Si incontrano problemi ancora più sottili

Fantasia e Creatività

- **creazione di infiniti non numerabili:** $\aleph_0$ e cardinalità del continuo
- **spazi a dimensione infinita**

  $\mathbb{R}^{\mathbb{N}}$: insieme delle successioni $(x_1, x_2, \ldots)$ Distanza $d(X, 0) = \sqrt{x_1^2 + x_2^2 + \cdots}$

  Attenzione che ci sono oggetti quali $(1, 1, 1, \ldots)$ o $(1, 1/\sqrt{2}, 1/\sqrt{3}, 1/\sqrt{4}, \ldots)$ che hanno distanza dall'origine infinita.
- **spazi funzionali:** con un numero *non numerabile* di dimensioni: $\mathbb{R}^{\mathbb{R}}$

  Insieme di tutte le funzioni da $[a, b] \subset \mathbb{R}$ in $\mathbb{R}$

  distanza $d(0, f) = \left(\int_a^b f(x)^2 dx\right)^{1/2}$
  Si incontrano problemi ancora più sottili

Fantasia e Creatività

- **creazione di infiniti non numerabili:** $\aleph_0$ e cardinalità del continuo
- **spazi a dimensione infinita**

  $\mathbb{R}^{\mathbb{N}}$: insieme delle successioni $(x_1, x_2, \ldots)$ Distanza $d(X, 0) = \sqrt{x_1^2 + x_2^2 + \cdots}$

  Attenzione che ci sono oggetti quali $(1, 1, 1, \ldots)$ o $(1, 1/\sqrt{2}, 1/\sqrt{3}, 1/\sqrt{4}, \ldots)$ che hanno distanza dall'origine infinita.
- **spazi funzionali:** con un numero *non numerabile* di dimensioni: $\mathbb{R}^{\mathbb{R}}$

  Insieme di tutte le funzioni da $[a, b] \subset \mathbb{R}$ in $\mathbb{R}$

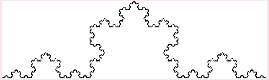
  distanza $d(0, f) = \left(\int_a^b f(x)^2 dx\right)^{1/2}$

  Si incontrano problemi ancora più sottili

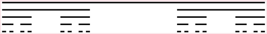
Fantasia e Creatività

- **oggetti geometrici a dimensione frazionaria (frattali)**
  - Le linee nel piano o nello spazio hanno dimensione 1
  - le superfici hanno dimensione 2
  - Esistono oggetti che hanno dimensione frazionaria, ad esempio

    Curva di Koch, dim=$\log 4/\log 3 = 1.2619\ldots$

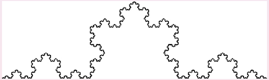
    Insieme di Cantor, dim=$\log 2/\log 3 = 0.6309\ldots$

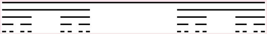
Fantasia e Creatività

- **oggetti geometrici a dimensione frazionaria (frattali)**
  - Le linee nel piano o nello spazio hanno dimensione 1
  - le superfici hanno dimensione 2
  - Esistono oggetti che hanno dimensione frazionaria, ad esempio

    Curva di Koch, dim=$\log 4/\log 3 = 1.2619\ldots$

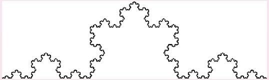
    Insieme di Cantor, dim=$\log 2/\log 3 = 0.6309\ldots$

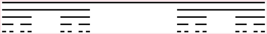
Fantasia e Creatività

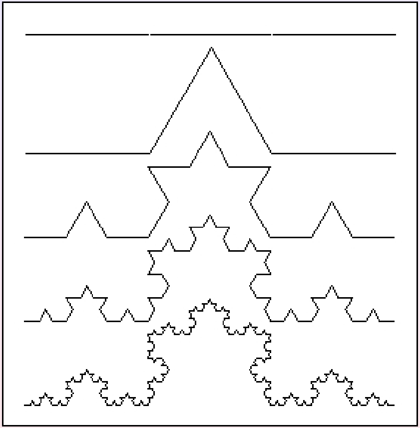
- **oggetti geometrici a dimensione frazionaria (frattali)**
  - Le linee nel piano o nello spazio hanno dimensione 1
  - le superfici hanno dimensione 2
  - Esistono oggetti che hanno dimensione frazionaria, ad esempio

    Curva di Koch, dim$=\log 4/\log 3 = 1.2619\ldots$

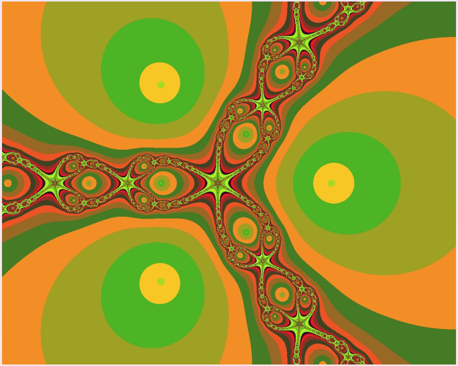
    Insieme di Cantor, dim$=\log 2/\log 3 = 0.6309\ldots$



Fantasia e Creatività

- Frattali più suggestivi si possono costruire con semplici formule algebriche: dall'equazione $x^3 - 1 = 0$ si ottiene



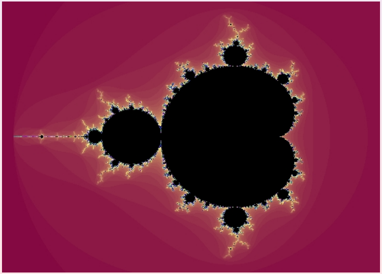
$$\begin{cases} x_{n+1} = \dfrac{2x_n^3 + 1}{3x_n^2} \\ x_0 \end{cases}$$

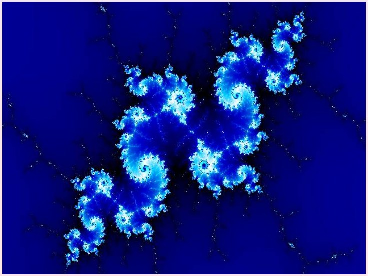
Se la successione $x_0, x_1, x_2, \ldots$ converge, converge ad una soluzione di $x^3 - 1 = 0$

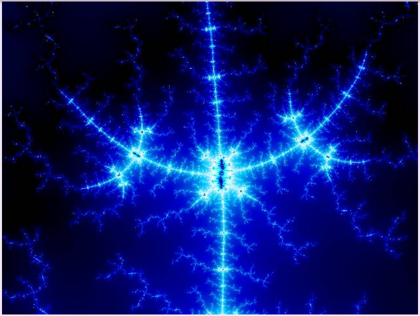
Fantasia e Creatività

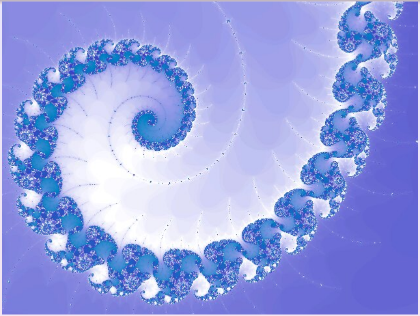
- Insieme di Mandelbrot: insieme dei numeri complessi c per cui non diverge la successione

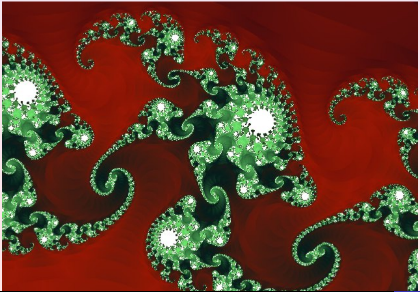
$$\begin{cases} x_{n+1} = x_n^2 + c, n = 0, 1, 2, \dots \\ x_0 = 0 \end{cases}$$

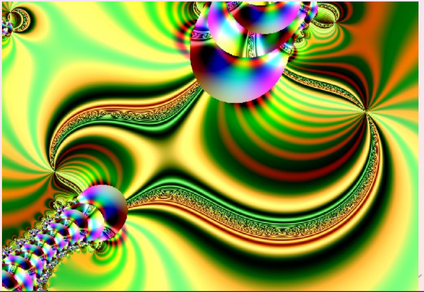










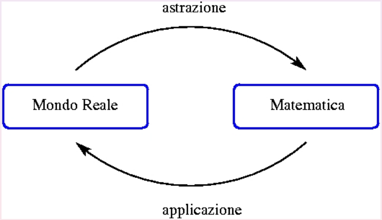


Fantasia e Creatività

- Formulazione di affermazioni che si dimostrano essere non dimostrabili.
- La matematica è forse l'unica disciplina che dimostra la non dimostrabilità delle "verità di fede"

Matematica e Applicazioni

La matematica probabilmente è nata e si è evoluta attraverso una continua interazione col mondo reale in un processo di astrazione e applicazione



Nuove idee matematiche producono strumenti potenti per lo studio di problemi del mondo reale

Nuovi problemi del mondo reale richiedono lo sviluppo di nuove idee matematiche

La matematica intorno a noi

Generalmente il matematico gioca a fare matematica indipendentemente dalle sue possibilità di applicazione, attratto dalla bellezza, dal mistero e dalla sfida posta dai problemi allo studio.

Al giorno d'oggi dovunque ci voltiamo incontriamo inconsapevolmente applicazioni della matematica.

Di fatto la tecnologia avanzata è essenzialmente tecnologia matematica

Senza la ricerca matematica molte delle funzionalità che utilizziamo o che sappiamo essere utilizzate ogni giorno non sarebbero disponibili.

La matematica intorno a noi

Ogni giorno ciascuno di noi fa uso inconsapevole di matematica

- la fotografia digitale: sharpening, deblurring, jpeg...
- la musica digitale: CD, mp3, ipod
- film e tv digitali: alta definizione, dvd, digitale terrestre, digitale satellitare
- telefonia mobile
- gps, cartografia nautica, navigatori satellitari
- volo automatico
- crittografia (bancomat, internet)
- sport
- internet (ricerca di informazioni)

La matematica intorno a noi

- La TAC, la RNM
- modelli cardiaci
- modelli della circolazione sanguigna (aneurismi, ostruzioni)
- modelli epidemiologici
- previsioni del tempo
- indagini statistiche, exit poll
- modelli di code
- analisi di rischio (assicurazioni)
- modelli finanziari

- le armi "intelligenti"
- sistemi antimissile
- aerei killer

Un po' di crittografia: i codici di Giulio Cesare

Idea: si stabilisce una corrispondenza biunivoca tra l'insieme di lettere e sé stesso

NON CAPISCO NIENTE

OPO DBQLTDP OLFOUF

È un gioco da settimana enigmistica decrittare messaggi crittati in questo modo

Codici un po' più astuti

Fase 1: (preparazione) si associa un numero a ciascun carattere alfanumerico, ad esempio il suo codice ASCII (numero compreso tra 0 e 255)

NON CAPISCO NIENTE

78-79-78-32-67-65-80-73-83-67-79-32-78-73-69-78-84-69

Si considera la stringa di numeri come la rappresentazione in base $B = 256$ di un numero intero

$$m = 69 + 84B + 78B^2 + 69B^3 + 73B^4 + \cdots + 78B^{17}$$

Fase 2 (codifica): Il numero m si moltiplica per un numero c (chiave segreta) e si trasmette il risultato $r = cm$

Codici un po' più astuti

Chi riceve il messaggio deve semplicemente dividere r per c per ottenere m; calcolerà poi le cifre di m in base 256 e applicherà i codici ASCII per recuperare il messaggio originale.

Osservazione: In questo modo ad una stessa lettera non corrisponde sempre la stessa codifica.

Codici un po' più astuti

Il codice di crittografia descritto è

- molto semplice
- facilmente implementabile
- più difficile da decrittare di quello cesareo (mescola le lettere)

Ha però un punto debole:

Se il nemico ci cattura qualche messaggio, scopre facilmente (calcolando il massimo comun divisore) che i numeri catturati hanno come comune divisore il numero c.
quindi ricava la chiave!

Il massimo comun divisore si calcola facilmente con l'Algoritmo Euclideo

Un codice molto più astuto: RSA

- Scegliamo p e q numeri primi abbastanza grandi in modo che il numero di cifre di $n = p * q$ sia maggiore del numero di cifre del messaggio m.
- Scegliamo un intero e (esponente della chiave pubblica) coprimo con $f = (p-1)(q-1)$, $1 < e < f$
- Calcoliamo il numero intero d (esponente della chiave privata) tale che $e * d = 1 \bmod f$

Per crittare il messaggio m (numero intero) calcolo $c = m^e \bmod n$.

Per decrittare il messaggio calcolo $c^d \bmod n$

Perché funziona?

Vale

$$c^{d} = (m^{e})^{d} = m^{ed} \mod n.$$

Inoltre, poiché

$$ed = 1 \mod p - 1$$
$$ed = 1 \mod q - 1,$$

dal piccolo teorema di Fermat si ottiene

$$m^{ed} = m \mod p$$
$$m^{ed} = m \mod q.$$

Poiché p e q sono numeri primi, il teorema cinese del resto applicato alle relazioni di sopra dà $m^{ed} = m \mod pq$. quindi $c^{d} = m \mod n$

Il codice è sicuro?

Il codice è sicuro fintanto che i seguenti due problemi rimangono computazionalmente difficili

- Calcolare p e q dato $n = pq$
- calcolare la soluzione dell'equazione $x^e = c \bmod n$, dati e, c, n.

In teoria p e q possono essere calcolati, però i metodi di fattorizzazione al momento conosciuti richiedono l'esecuzione di un numero di operazioni dell'ordine di

2^k $\quad k$: numero di cifre in base 2 di n

Se n è scelto con 500 cifre si hanno circa

$$2^{500} = (2^{10})^{50} > 1000^{50} = 10^{150}$$

operazioni

Secondo voi sono tante?

Il computer più veloce al momento disponibile è il Blue Gene dell'IBM in grado di svolgere 360 teraflops $= 3.6 \times 10^{14}$ al secondo

Esso impiegherebbe $10^{150}/(3.6 \times 10^{14}) > 10^{135}$ secondi per fattorizzare n

cioè 3.17×10^{121} milioni di anni!

Matematica del Web

Internet costituisce una sorgente di problemi matematici di particolare interesse teorico e applicativo

- Page ranking (Google)
- Information retrieval
- Gestione del flusso delle informazioni sulla rete

Motori di ricerca e Page Rank

Due studenti dell'università di Stanford, Sergey Brin e Larry Page hanno fatto la loro fortuna inventando "Google"

Problema: Ordinare le pagine presenti sul web in base alla loro importanza

Come si può definire l'importanza (**page rank**) di una pagina?

Motori di ricerca e Page Rank

Due studenti dell'università di Stanford, Sergey Brin e Larry Page hanno fatto la loro fortuna inventando "Google"

Problema: Ordinare le pagine presenti sul web in base alla loro importanza

Come si può definire l'importanza (**page rank**) di una pagina?

Motori di ricerca e Page Rank

Due studenti dell'università di Stanford, Sergey Brin e Larry Page hanno fatto la loro fortuna inventando "Google"

Problema: Ordinare le pagine presenti sul web in base alla loro importanza

Come si può definire l'importanza (**page rank**) di una pagina?

Importanza di una pagina

Varie proposte

- in base al numero di volte che la parola cercata compare
- in base al numero dei link che da essa partono
- in base al numero dei link che ad essa arrivano
- **in base al numero delle pagine importanti che puntano alla pagina**

Importanza di una pagina

Varie proposte

- in base al numero di volte che la parola cercata compare
- in base al numero dei link che da essa partono
- in base al numero dei link che ad essa arrivano
- **in base al numero delle pagine importanti che puntano alla pagina**

Importanza di una pagina

Varie proposte

- in base al numero di volte che la parola cercata compare
- in base al numero dei link che da essa partono
- in base al numero dei link che ad essa arrivano
- in base al numero delle pagine importanti che puntano alla pagina

Importanza di una pagina

Varie proposte

- in base al numero di volte che la parola cercata compare
- in base al numero dei link che da essa partono
- in base al numero dei link che ad essa arrivano
- **in base al numero delle pagine importanti che puntano alla pagina**

L'idea di Page e Brin

- Ogni pagina ha una sua propria importanza che deriva dalle connessioni (non direttamente dai contenuti)
- L'importanza di una pagina viene trasferita in parti uguali alle pagine che essa punta
- L'importanza di una pagina è data dalla somma delle frazioni di importanza che gli derivano dalle pagine che ad essa puntano

L'idea di Page e Brin

Se il Papa nell'incontro della domenica a Piazza San Pietro dà la sua benedizione al professor Antonello Zibibbo, il professore riceve grande importanza

Se il Papa dà la sua benedizione a tutti i professori del mondo allora il prof. Antonello Zibibbo riceve un'importanza trascurabile

Se il professor Antonello Zibibbo benedice il Papa, quest'ultimo non se ne accorge nemmeno

Modello matematico

Numeriamo le pagine del web da 1 a n

Definiamo la matrice di connettività nel seguente modo:

$$H = \begin{bmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,n} \\ h_{2,1} & h_{2,2} & \dots & h_{2,n} \\ \vdots & \vdots & & \vdots \\ h_{n,1} & h_{n,2} & \dots & h_{n,n} \end{bmatrix}$$

$h_{i,j} = 1$ se c'è un link dalla pagina i alla pagina j
$h_{i,j} = 0$ altrimenti.

Esempio con $n = 4$

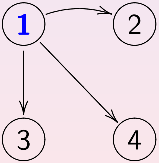
$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

1 2

3 4

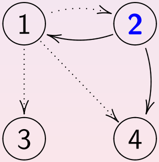
Esempio con $n = 4$

$$H = \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$



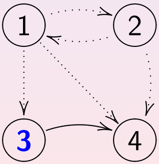
Esempio con $n = 4$

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$



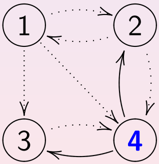
Esempio con $n = 4$

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 0 & 1 & 1 & 0 \end{bmatrix}$$



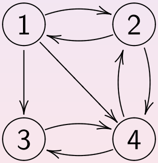
Esempio con $n = 4$

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \end{bmatrix}$$



Esempio con $n = 4$

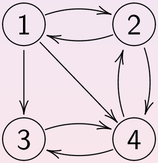
$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$



Sommando i valori sulla riga i si trova il numero di link che partono dalla pagina i. Denotiamo con r_i questo numero

Esempio con $n = 4$

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$



Sommando i valori sulla riga i si trova il numero di link che partono dalla pagina i. Denotiamo con r_i questo numero

Indichiamo con $\mathbf{x}_j$ l'importanza della pagina j

Allora risulta

$$\mathbf{x}_j = h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}, \qquad \text{per } j = 1, 2, \ldots, n.$$

Questo è un **sistema lineare** di n equazioni in n incognite.

Le soluzioni $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ danno il livello di importanza delle singole pagine cioè il **page rank**

Indichiamo con $\mathbf{x}_j$ l'importanza della pagina j

Allora risulta

$$\mathbf{x}_j = h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}, \qquad \text{per } j = 1, 2, \ldots, n.$$

Questo è un **sistema lineare** di n equazioni in n incognite.

Le soluzioni $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ danno il livello di importanza delle singole pagine cioè il **page rank**

Indichiamo con $\mathbf{x}_j$ l'importanza della pagina j

Allora risulta

$$\mathbf{x}_j = h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}, \qquad \text{per } j = 1, 2, \ldots, n.$$

Questo è un **sistema lineare** di n equazioni in n incognite.

Le soluzioni $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ danno il livello di importanza delle singole pagine cioè il **page rank**

L'equazione usata da Google è leggermente diversa

$$\mathbf{x}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

dove d è un parametro fra 0 e 1, di solito viene posto $d = 0.85$

I valori di $\mathbf{x}_j$ sono compresi fra 0 e 1.

Per calcolare il page rank occorre risolvere un sistema di n equazioni ed n incognite

Alla data di oggi ci sono circa $n = 8.5 \times 10^9$ pagine attive

Come si risolve un sistema lineare?

Se $n = 2$ si ha

$$\begin{cases} a\mathbf{x} + b\mathbf{y} = c \\ d\mathbf{x} + e\mathbf{y} = f \end{cases}$$

Il metodo di sostituzione detto anche di Eliminazione Gaussiana si applica in generale a sistemi $n \times n$

Esso richiede circa

$$\frac{2}{3}n^3 + \text{spiccioli}$$

operazioni aritmetiche

Come si risolve un sistema lineare?

Se $n = 8.5$ miliardi il metodo di eliminazione richiede circa

$$\frac{2}{3}(8.5 \times 10^9)^3 \approx 4.1 \times 10^{29}$$

(410 miliardi di miliardi di miliardi) operazioni aritmetiche

sono tante?

Complessità del Page Rank

Anche disponendo del **Blue Gene** dell'IBM, per eseguire 4.1×10^{29} operazioni ci vorrebbero più di 36 milioni di anni

Un tempo "geologico" .... eppure Larry Page e Sergey Brin **calcolano il page rank ogni mese**

come fanno?

Tecnologia hardware vs tecnologia matematica

Anche se la tecnologia fosse in grado di costruire un computer 1000 volte o un milione di volte più veloce non sarebbe possibile risolvere il problema di Google in tempo reale

Solo sviluppando nuovi metodi matematici è possibile risolvere il sistema con tempi di calcolo brevi

L'algoritmo Page Rank

Equazione

$$\mathbf{x}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

Algoritmo

1. Assegna agli $\mathbf{x}_i$ dei valori qualsiasi
2. sostituiscili nella parte destra della formula

$$\mathbf{y}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

3. e ricava i valori di $\mathbf{y}_j$, per $j = 1, 2, \ldots, n$
4. poni $\mathbf{x}_j = \mathbf{y}_j$ per $j = 1, 2, \ldots, n$ e prosegui dal punto 2

L'algoritmo Page Rank

Equazione

$$\mathbf{x}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

Algoritmo

1. Assegna agli $\mathbf{x}_i$ dei valori qualsiasi
2. sostituiscili nella parte destra della formula

$$\mathbf{y}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

3. e ricava i valori di $\mathbf{y}_j$, per $j = 1, 2, \ldots, n$
4. poni $\mathbf{x}_j = \mathbf{y}_j$ per $j = 1, 2, \ldots, n$ e prosegui dal punto 2

L'algoritmo Page Rank

Equazione

$$\mathbf{x}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

Algoritmo

1. Assegna agli $\mathbf{x}_i$ dei valori qualsiasi
2. sostituiscili nella parte destra della formula

$$\mathbf{y}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

3. e ricava i valori di $\mathbf{y}_j$, per $j = 1, 2, \ldots, n$
4. poni $\mathbf{x}_j = \mathbf{y}_j$ per $j = 1, 2, \ldots, n$ e prosegui dal punto 2

L'algoritmo Page Rank

Equazione

$$\mathbf{x}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

Algoritmo

1. Assegna agli $\mathbf{x}_i$ dei valori qualsiasi
2. sostituiscili nella parte destra della formula

$$\mathbf{y}_j = d(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots + h_{n,j}\frac{\mathbf{x}_n}{r_n}) + \frac{1}{n}(1-d),$$

3. e ricava i valori di $\mathbf{y}_j$, per $j = 1, 2, \ldots, n$
4. poni $\mathbf{x}_j = \mathbf{y}_j$ per $j = 1, 2, \ldots, n$ e prosegui dal punto 2

Viene generata una successione di approssimazioni che converge alla soluzione del sistema **qualunque** siano le approssimazioni iniziali.

$$\mathbf{x}_j^{(1)},\ \mathbf{x}_j^{(2)},\ \mathbf{x}_j^{(3)},\ \ldots \quad \longrightarrow \mathbf{x}_j$$

Quanto veloce è la convergenza?

L'errore di approssimazione $e^{(k)} = \max_i |x_i^{(k)} - x_i|$ è tale che

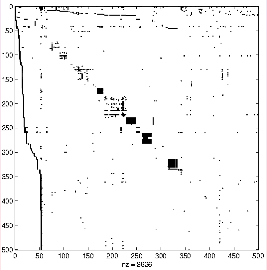
$$e^{(k)} \leq \lambda^k \quad \text{con} \quad 0 < \lambda < 1$$

Purtroppo per valori di d vicini a 1 il valore di λ è molto vicino ad 1 e quindi la convergenza è lenta.

λ coincide col modulo del secondo autovalore più grande in modulo di una opportuna matrice.

Complessità

Per fare un passo dell'algoritmo del Page Rank bisogna eseguire tante moltiplicazioni quanti sono gli elementi non nulli di H e all'incirca altrettante addizioni
Su ogni riga della matrice H ci sono "pochi" elementi diversi da zero.



Se mediamente ci fossero 50 elementi non nulli su ogni riga, un passo del metodo iterativo eseguito col Blue Gene impiegherebbe 2 millesimi di secondo.

Se anche fossero necessari 1000 passi iterativi basterebbero 2 secondi per approssimare la soluzione di Google.

Interpretazione probabilistica

La soluzione $\mathbf{x}_j$ del problema del Page Rank coincide con la probabilità che la pagina j ha di essere visitata da un navigatore virtuale che opera nel seguente modo:

Ad ogni istante (minuto) il navigatore cambia pagina

Per spostarsi da una pagina all'altra il navigatore:

- con probabilità d decide di cliccare su uno dei link presenti sulla pagina corrente scegliendolo a caso con uguale probabilità
- con probabilità $1 - d$ decide di saltare ad un'altra pagina a caso presente sul web scegliendola con uguale probabilità

La probabilità che il navigatore passi dalla pagina i alla pagina j è

$$p_{i,j} = d\frac{h_{i,j}}{r_i} + \frac{1}{n}(1-d)$$

Denotiamo con

$\mathbf{x}_i$ la probabilità che il navigatore stia visitando la pagina i ad un certo istante

$\mathbf{y}_j$ la probabilità che il navigatore stia visitando la pagina j all'istante successivo

Allora vale

$$\mathbf{y}_j = p_{1,j}\mathbf{x}_1 + p_{2,j}\mathbf{x}_2 + \cdots + p_{n,j}\mathbf{x}_n$$

È il sistema di Google!

$$\mathbf{y}_j = d\left(h_{1,j}\frac{\mathbf{x}_1}{r_1} + h_{2,j}\frac{\mathbf{x}_2}{r_2} + \cdots h_{n,j}\frac{\mathbf{x}_n}{r_n}\right) + \frac{1}{n}(1-d)$$

La soluzione $\mathbf{x}_j^{(k)}$ calcolata dopo k passi dell'algoritmo iterativo di Google fornisce la probabilità che il navigatore virtuale ha di trovarsi nella pagina j dopo k istanti

La soluzione $\mathbf{x}_j$ del sistema fornisce la probabilità che il navigatore virtuale ha di trovarsi nella pagina j dopo "infiniti" istanti di navigazione

La teoria delle catene di Markov e la teoria delle matrici non negative forniscono condizioni di esistenza e unicità della soluzione e di convergenza della successione

Problemi allo studio

- trovare modi diversi di costruire successioni che convergono alla soluzione in modo più rapido
  - Successioni di Krylov
  - metodi del gradiente
  - metodi di aggregazione
  - metodi di estrapolazione
- esprimere la soluzione in funzione di d (come serie di potenze) e vedere come si comporta quando d tende a uno
- studiare modelli matematici diversi, ad esempio: il navigatore salta ad una pagina qualsiasi solo se sulla pagina corrente non ci sono link
- studiare il valore di λ

Bibliografia

- A. Langville, C. Meyer, Information Retrieval and Web Search. The Handbook of Linear Algebra. CRC Press, 2006
  http://math.cofc.edu/langvillea/HLA.pdf
- A Arasu, J. Novak, A. Tomkins, J. Tomlin, *Page rank computation and the structure of the Web*. (2002)
  http://www2002.org/CDROM/poster/173.pdf
- A. N. Langville, C. D. Meyer, *A survey of eigenvector methods for Web information retrieval*, SIAM Rev. **47** (2005), pp. 135–161.
  http://math.cofc.edu/langvillea/surveyEVweblRReprint.pdf
- C. Moler, *Numerical computing with Matlab*.
  http://www.mathworks.com/moler sezione 2.11 (2004)
- L. Page, S Brin, R. Motwani, T. Winograd, *The pagerank citation ranking: Bringing order to the Web*,
  http://dbpubs.stanford.edu:8090/pub/1999-66.

Alcuni tools

Qui trovate alcuni tools per giocare col page rank
Gli indirizzi li ho ottenuti cercando sul web "page rank" tools

- http://www.rustybrick.com/pagerank-prediction.php
- http://www.googlerankings.com
- http://www.prchecker.info
- http://www.seochat.com/seo-tools/future-pagerank
- http://www.seoutility.com/
- http://www.markhorrell.com/tools/pagerank.shtml

Alcuni link sui frattali

- Software Xaos per generare frattali, immagini e animazioni
  `http://xaos.sourceforge.net/`
- Software Fractint per generare frattali
  `http://spanky.triumf.ca/www/fractint/fractint.html`