La matematica del Web

D.A. Bini
Dipartimento di Matematica
Università di Pisa
www.dm.unipi.it/~bini bini@dm.unipi.it

Pisa 10 Febbraio 2017

Il mondo di Internet

Internet è un mondo altamente matematizzato: molte delle operazioni che svolgiamo quando si naviga, così come l'analisi del comportamento dell'utente del Web, e come l'organizzazione del flusso delle informazioni, sono realizzate grazie a strumenti e teoremi della matematica che sono stati inventati quando internet non esisteva ancora.

In questo incontro, attraverso degli esempi, vorrei darvi un'idea

- di quanta matematica c'è dietro il funzionamento di internet
- di come la ricerca matematica più astratta, fatta per pura curiosità intellettuale, trova applicazione anche a distanza di molti anni
- di come sia capillare, pur a nostra insaputa, il ruolo della matematica nella vita di ogni giorno

Per questo voglio fare una breve premessa

Esistono molti luoghi comuni (e tante barzellette) sulla matematica e i matematici, eccone alcuni:

- è una disciplina fine a sé stessa, quindi abbastanza inutile
- ha rigidità di pensiero
- è ripetitiva
- è arida e incapace di esprimere qualcosa di nuovo

Alcuni ingredienti essenziali del buon matematico

- fantasia, immaginazione, libertà di pensiero non vincolato a rigidi schemi mentali
- attrazione per l'eleganza
- attrazione per le singolarità
- estremo rigore logico

Dal mio punto di vista la matematica è un bellissimo giocattolo che

- non stanca mai né si rompe mai
- permette di creare quasi senza limiti strutture eleganti col massimo rigore logico
- sviluppa il pensiero libero, la creatività e la fantasia
- non ha padroni, è proprietà di tutti
- non ha colorazioni né politiche né religiose
- accomuna popoli di varie etnie
- è in continua evoluzione
- è uno strumento indispensabile per studiare e trattare problemi del mondo reale
- numerosi atti che rientrano nella normalità della vita quotidiana sono possibili solo grazie a metodologie matematiche avanzate

Alcuni esempi di applicazione

- Telefonia mobile, reti wireless
  $\rightarrow$ teoria delle code, equazioni matriciali, matrici di Toeplitz, analisi complessa, serie di Laurent,...
- Tecnologia digitale (suoni, musica, film, foto)
  $\rightarrow$ Trasformata di Fourier, trasformate trigonometriche, trasformate wavelet, codici correttori, filtri, geometria computazionale
- Riconoscimento dei volti, riconoscimento di scritti, di impronte digitali
  $\rightarrow$ Decomposizione ai valori singolari, fattorizzazione di Takagi
- Analisi cliniche (TAC, RNM, Radiografie)
  $\rightarrow$ equazioni integrali, trasformate di Radon, Fourier,...
- Previsioni del tempo, modelli biologici, sport, problemi aerospaziali
  $\rightarrow$ equazioni differenziali
- Sistemi di controllo: volo aereo, guida automatica, droni, armi chirurgiche
  $\rightarrow$ sistemi dinamici

- Progettazione e analisi di strutture discrete (ponti, edifici, torre di Taipei)
  $\rightarrow$ Algebra lineare, autovalori, autovettori
- Robotica industriale, processi industriali, simulazione, CAD
  $\rightarrow$ equazioni algebriche, ottimizzazione, computer algebra, equazioni differenziali
- Analisi economiche, analisi della borsa
  $\rightarrow$ Equazioni differenziali alle derivate parziali, equazioni differenziali stocastiche
- Ottimizzazione dei servizi (turni lavorativi, orari di lavoro, orari dei mezzi di trasporto, pianificazione, traffico stradale)
  $\rightarrow$ Teoria dell'ottimizzazione, ricerca operativa
- Biologia matematica, Sociologia matematica, Psicologia matematica

I problemi del Web

- Analisi delle reti sociali, estensione a reti complesse, comunità sociali fisiche, modelli della biologia (lieviti, muffe)
  $\rightarrow$ autovettori di matrici, funzioni di matrici, spazi di Krylov, decomposizione ai valori singolari,...
- Motori di ricerca (PageRank, Google)
  $\rightarrow$ Teoria di Perron-Frobenius, Algebra lineare, spazi di Krylov, raggio spettrale, metodo delle potenze
- Flusso delle informazioni e dei dati
  $\rightarrow$ teoria delle code, processi stocastici, catene di Markov, equazioni matriciali
- Sicurezza delle transazioni economiche, sicurezza nei processi di identificazione
  $\rightarrow$ crittografia mediante numeri primi, codice RSA, teoria dei numeri
- Analisi del comportamento dell'utente del Web a fini di rendere più efficace la pubblicità
- Valutazione dell'affidabilità di sistemi votato/votante

Il problema del PageRank

La ricerca di informazioni su internet (Google) fornisce spesso liste di molte migliaia di indirizzi

È essenziale ordinare questo mare di informazioni in modo da mettere in testa gli indirizzi delle pagine **"piu significative"**

Generalmente Google svolge bene questa operazione. Ma come fa?

Come possiamo definire una pagina più significativa di un'altra?

Il problema del PageRank

La ricerca di informazioni su internet (Google) fornisce spesso liste di molte migliaia di indirizzi

È essenziale ordinare questo mare di informazioni in modo da mettere in testa gli indirizzi delle pagine **"piu significative"**

Generalmente Google svolge bene questa operazione. Ma come fa?

Come possiamo definire una pagina più significativa di un'altra?

Tanti anni fa l'ordinamento era fatto in base ai contenuti (peraltro con spiacevoli inconvenienti)

La pagina che conteneva più occorrenze della parola cercata era messa in testa alla lista

Ai tempi del film "Titanic" la ricerca di "Leonardo Di Caprio" forniva liste con in testa indirizzi di siti "a luci rosse"

Google

Gli inventori di Google, Sergey Brin e Larry Page, capirono che era necessario stabilire una regola di importanza non vincolata ai contenuti ma legata alle relazioni tra le pagine

Proviamo a descrivere l'idea di questo modello facendo prima dei tentativi di definire l'importanza di una pagina Web

Una pagina è importante se:

- **contiene** link a molte altre pagine?

Google

Gli inventori di Google, Sergey Brin e Larry Page, capirono che era necessario stabilire una regola di importanza non vincolata ai contenuti ma legata alle relazioni tra le pagine

Proviamo a descrivere l'idea di questo modello facendo prima dei tentativi di definire l'importanza di una pagina Web

Una pagina è importante se:

- **contiene** link a molte altre pagine?
- **riceve** link da molte altre pagine?

Google

Gli inventori di Google, Sergey Brin e Larry Page, capirono che era necessario stabilire una regola di importanza non vincolata ai contenuti ma legata alle relazioni tra le pagine

Proviamo a descrivere l'idea di questo modello facendo prima dei tentativi di definire l'importanza di una pagina Web

Una pagina è importante se:

- **contiene** link a molte altre pagine?
- **riceve** link da molte altre pagine?
- **riceve link da molte altre pagine importanti**

Principio

L'importanza di una pagina viene distribuita in parti uguali alle pagine che punta.

L'importanza di una pagina è la somma dalle importanze delle pagine che la puntano

È un po' come nella vita sociale dove una persona che ha **conoscenze strette** con persone importanti acquisice importanza essa stessa

Si può matematizzare questo principio?

Principio

L'importanza di una pagina viene distribuita in parti uguali alle pagine che punta.

L'importanza di una pagina è la somma dalle importanze delle pagine che la puntano

È un po' come nella vita sociale dove una persona che ha **conoscenze strette** con persone importanti acquisice importanza essa stessa

Si può matematizzare questo principio?

Certamente! Ci proviamo subito.

Il modello matematico del PageRank

Denotiamo con n il numero di pagine del web (attualmente $n \approx 10^{10}$)

Numeriamo le pagine da 1 a n

scriviamo $i \to j$ se la pagina i **ha almeno un link** alla pagina j

scriviamo $i \not\to j$ se la pagina i **non ha link** alla pagina j

Costruiamo una tabella di $n \times n$ numeri $a_{i,j}$ per $i, j = 1, 2, 3, \ldots, n$ tale che

$$a_{i,j} = \begin{cases} 1 & \text{se } i \to j \\ 0 & \text{se } i \not\to j \end{cases}$$

Il modello matematico del PageRank

Denotiamo con n il numero di pagine del web (attualmente $n \approx 10^{10}$)

Numeriamo le pagine da 1 a n

scriviamo $i \to j$ se la pagina i **ha almeno un link** alla pagina j

scriviamo $i \not\to j$ se la pagina i **non ha link** alla pagina j

Costruiamo una tabella di $n \times n$ numeri $a_{i,j}$ per $i, j = 1, 2, 3, \ldots, n$ tale che

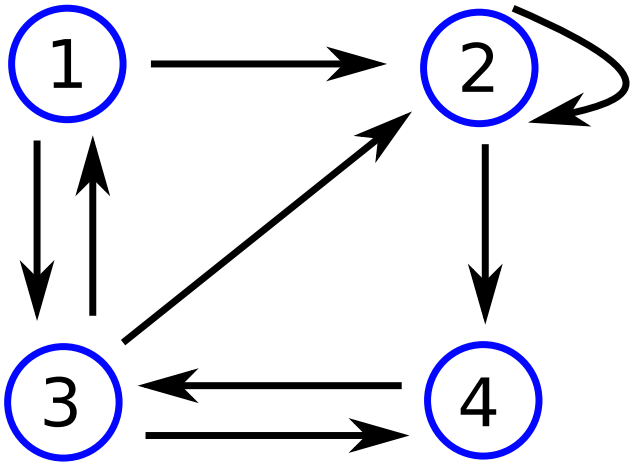
$$a_{i,j} = \begin{cases} 1 & \text{se } i \to j \\ 0 & \text{se } i \not\to j \end{cases}$$

In algebra lineare, tabelle di numeri come questa sono chiamate **matrici**, e descrivono **applicazioni lineari** da uno spazio n-dimensionale in sé

Il concetto di matrice fu introdotto da **Arthur Cayley** nel 1858

Esempio

Una rete formata da 4 pagine numerate da 1 a 4, interconnesse come in figura



è descritta dalla matrice 4×4: $A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

La matrice A è chiamata **matrice di adiacenza**

Ricordiamoci che ogni pagina distribuisce in parti uguali la sua importanza alle altre pagine

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

Per cui le frazioni di importanza che vengono distribuite si ottengono dividendo gli elementi di ogni riga per il numero di uni che sono su quella riga

$$B = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Allora se chiamo x_1, x_2, x_3, x_4 l'importanza delle pagine 1,2,3,4, dalla prima colonna della matrice

$$B = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

ricavo che

$$x_1 = \frac{1}{3}x_3$$

e analogamente dalle altre colonne ottengo

$$x_2 = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3$$

$$x_3 = \frac{1}{2}x_1 + x_4$$

$$x_4 = \frac{1}{2}x_2 + \frac{1}{3}x_3$$

Si ha un sistema lineare omogeneo di 4 equazioni e 4 incognite

$$\begin{cases} x_1 = \frac{1}{3}x_3 \\ x_2 = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 \\ x_3 = \frac{1}{2}x_1 + x_4 \\ x_4 = \frac{1}{2}x_2 + \frac{1}{3}x_3 \end{cases}$$

In generale (supponendo che A non abbia righe nulle) si ha un sistema di n equazioni in n incognite

$$x = \sum_{i=1}^{n} x_i b_{i,j}$$

Alcune domande

- La soluzione esiste
- è unica?
- è non negativa?
- come si può calcolare?
- il fatto che $n \approx 10^{10}$ crea problemi?

Aspetti teorici

La n-upla $(x_1, x_2, \ldots, x_n)$ puó essere vista come un **autovettore** della matrice B, cioè un vettore che viene trasformato in se stesso dall'applicazione lineare associata alla matrice B che corrisponde all'autovalore $\lambda = 1$

Oscar Perron (1880–1975) e Ferdinand Georg Frobenius (1849–1917) dimostrarono numerosi teoremi sugli autovalori di matrici con elementi non negativi che costituiscono la **teoria di Perron-Frobenius**

Applicando questi risultati a B si ha

Se B è non negativa esiste una soluzione $x = (x_1, \ldots, x_n) \geq 0$ tale che $x_1 + x_2 + \cdots + x_n = 1$, ma puó non essere unica

Se B è **positiva** allora x è **unica**, vale $x_1, x_2, \ldots, x_n > 0$, e si hanno altre proprietà interessanti

Conseguenze sul modello

Per garantire l'unicità della soluzione occorre allora ritoccare il modello:

La matrice B viene sostituita dalla matrice

$$H = \gamma B + (1 - \gamma)E$$

dove E è la matrice con tutti elementi uguali a 1 e γ è un numero positivo minore di 1, ma abbastanza vicino a 1

È come dire che l'importanza di ogni pagina non viene distribuita tutta alle altre pagine, ma solo in quantità proporzionale a γ

L'importanza $1 - \gamma$ che avanza viene data a tutte le pagine in parti uguali.

In questo modo non ci possono essere elementi nulli e si può applicare la teoria di Perron-Frobenius

Aspetti computazionali

Ma come si fa a calcolare $x_1, x_2, \ldots, x_n$?

Se guardiamo al problema come a un sistema di equazioni possiamo utilizzare il metodo di eliminazione di Gauss (Johann Carl Friedrich Gauss 1777–1855) che risolve il problema con circa $\frac{2}{3}n^3$ operazioni aritmetiche

Nel nostro caso, assumendo $n = 10^{10}$ dovremmo eseguire circa $\frac{2}{3}10^{30}$ operazioni aritmetiche

Sono tante?

Con un computer portatile, che fa circa 10^9 operazioni al secondo ci vorrebbero $\frac{2}{3}10^{21}$ secondi

Cioè più di **211 miliardi di secoli** probabilmente un tempo superiore alla vita stimata dell'universo!

Usiamo allora un **supercomputer**

Secondo Wikipedia il supercomputer più veloce attualmente è il Sunway TaihuLight a Wuxi in Cina. Esegue 93×10^{15} operazioni al secondo

Impiegherebbe $\frac{2}{3}10^{30}/(93 \times 10^{15}) > 7 \times 10^{12}$ secondi per risolvere il problema del PageRank

Questo tempo corrisponde a più di 2200 secoli!

Ma allora come fa Google a calcolare il vettore PageRank in così poco tempo?

Occorre sfruttare il fatto che la matrice B è molto speciale

Caratteristiche di B

La matrice B che modella il problema del PageRank è molto speciale: la maggior parte dei suoi elementi sono nulli

Infatti le pagine presenti sul Web non hanno link con tutto il mondo e gli elementi non nulli su ogni riga di B sono mediamente meno di una decina

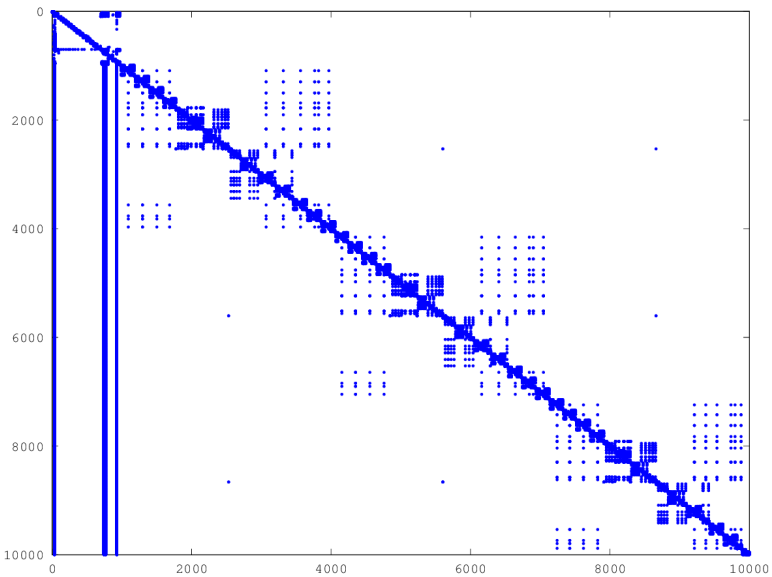


Figura: Matrice di adiacenza della rete web della Berkeley University

Caratteristiche di B

Conseguenza:

Calcolare

$$y_j = x_1 b_{1,j} + x_2 b_{2,j} + \cdots + x_n b_{n,j} \quad \text{per } j = 1, 2, \ldots, n$$

ha un costo proporzionale al numero di elementi non nulli di B

In termini di tempo, questo calcolo costa poco più di minuto con il nostro portatile, e qualche frazione di secondo con un buon server

I metodi iterativi

L'idea di Brin e Page per risolvere l'equazione

$$x_j = x_1 b_{1,j} + x_2 b_{2,j} + \cdots + x_n b_{n,j} \quad \text{per } j = 1, 2, \ldots, n$$

che riscriviamo per comodità come

$$x = Bx$$

è quella di generare una successione definita in questo modo

$$y = Bz, \quad z \leftarrow y$$

a partire da un vettore di z scelto a caso

Per i teoremi di Perron-Frobenius la successione generata in questo modo converge alla soluzione cercata **qualunque** sia il vettore iniziale z da cui siamo partiti purché non nullo

Se ci vogliono, supponiamo, 1000 passi per ottenere una buona approssimazione della soluzione allora il costo complessivo sarà 1000 volte il costo di un solo passo

Sul nostro laptop ci vorranno forse alcuni giorni, su un buon server basteranno poche ore

Google calcola questo vettore una volta al mese

Temi di interesse allo studio:

- Capire da cosa dipende la velocità di convergenza
- Individuare modifiche, o nuovi metodi, che accelerino la convergenza
- Adattare il modello ad altre esigenze

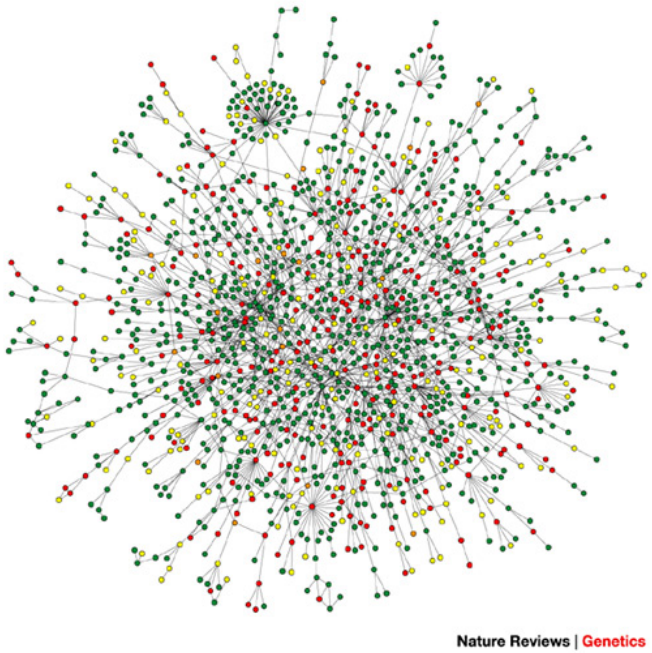
Estrapolazione, spazi di Krylov, metodi dei gradienti, ...

Altri modelli (reti complesse)

La matrice di adiacenza A descrive la **rete complessa** del Web

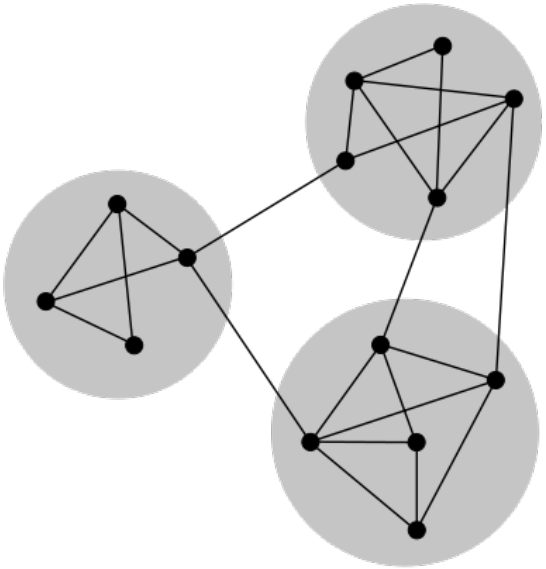
Matrici di adiacenza trovano applicazione in altri contesti

Esistono reti complesse a vari livelli: social networks di internet, reti stradali, comunità sociali, reti di interazioni geni-proteine, reti di distribuzione di energia, reti di collaborazione scientifica, ...



Interazione delle proteine del lievito di birra (Nature Reviews Genetics 5(2):101-13 · March 2004)

Una rete complessa, come può essere una rete sociale (Facebook), è generalmente costituita da **comunità**



(Esempio di 3 comunità preso da Wikipedia)

Il problema di individuare le comunità può essere formulato matematicamente in termini di matrice di adiacenza

Il problema è alquanto complesso ma esistono algoritmi che svolgono automaticamente questa analisi

Altri problemi sono:
Individuare i "nodi più centrali", individuare il tasso di connettività, di vulnerabilità di una rete, la capacità di instradare messaggi

In una rete complessa una infezione (virus) può propagarsi in misura più o meno accentuata in base alle caratteristiche della rete stessa (tasso di infettività)

Queste caratteristiche "fisiche" hanno una **controparte matematica**

In particolare gli autovalori di queste matrici racchiudono molta informazione

Gli elementi diagonali e la loro somma di funzioni particolari di matrici definiscono alcuni di questi parametri

Funzioni di variabile reale possono essere estese a matrici

Esempi:

$$f(x) = x^2 + 2 \qquad f(A) = A^2 + 2I$$

$$f(x) = \frac{1+x}{1-x} \qquad f(A) = (I - A)^{-1}(I + A)$$

$$\exp(x) = 1 + x + \frac{1}{2}x^2 + \cdots + \frac{1}{n!}x^n + \cdots$$

$$\exp(A) = I + A + \frac{1}{2}A^2 + \cdots + \frac{1}{n!}A^n + \cdots$$

Il concetto di funzione di matrice risale alla fine del secolo IX e ha avuto sviluppi nel XX secolo grazie a contributi di molti matematici tra cui alcuni italiani

Buchheim, Cartan, Cayley, Sylvester, Giorgi, Cartan, Fantappié, Cipolla, Poincaré, Schwerdtfeger, Weyr

Attualmente l'interesse è rivolto alla ricerca di metodi (algorimti) per il calcolo di funzioni di matrici con l'analisi delle loro proprietà astratte e della loro efficenza computazionale

Esiste una vasta letteratura a riguardo, in particolare il libro

Nick Higham, "Functions of Matrices", Theory and Computation , SIAM, Philadelphia, 2008.

Altri modelli (sistemi di reputazione)

Il Web permette di dare giudizi su film, libri, album musicali, ed altro

Chi esprime frequentemente il proprio voto può aumentare la propria reputazione e credibilità attraverso valutazioni affidabili, oppure può perdere totalmente credibilità semplicemente dando sempre il massimo o sempre il minimo

Esistono modelli matematici che permettono di valutare il grado di affidabilità di chi vota e allo stesso tempo il valore più corretto dei soggetti valutati.

Il modello è molto simile a quello del PageRank

Altre problematiche riguardano la ricerca di sinonimi che vengono individuati in base alle occorrenze nei testi analizzati e nella mutua vicinanza di parole

Situazioni similari, risolti mediante lo strumento delle matrici, riguardano lo studio della formazione di code nei pacchetti di dati che scorrono su reti complesse (internet)

Ad esempio, il protocollo IEEE 802.11 degli apparati wireless, si basa sulla risoluzione di speciali "catene di Markov"

Nel caso finito si devono risolvere sistemi lineari

Nel caso infinito il problema si riduce a risolvere equazioni dove coefficienti e incognite sono matrici

$$AX^2 + BX + C = 0$$

In certi casi le matrici A, B, C così come X sono infinite

Conclusioni in pillole

- La matematica è una disciplina in continua evoluzione che sviluppa grande creatività e fantasia
- Strumenti matematici anche i più astratti, ottenuti per pura curiosità intellettuale, possono trovare applicazioni inattese
- Internet è un settore importante dove gli strumenti matematici giocano un ruolo fondamentale