Analisi dell'affinità musicale degli artisti di Last.fm

Calogero Giudice
c.giudice@studenti.unipi.it
Student ID: 530155

ABSTRACT

Questo progetto analizza una rete di artisti musicali del social network *last.fm* in base a dei coefficienti di similarità (Sample e Jaccard Ratios) che potrebbero rappresentare delle affinità a livello stilistico o di pubblico. Sono state esaminate sia le **caratteristiche strutturali** della rete, procedendo a svolgere una previsione, in base ai Ratios menzionati in precedenza e al coefficiente di popolarità, su possibili aumenti o cali di popolarità futuri.

Questi risultati possono rappresentare un **modello** che potrebbe riflettersi nello scenario musicale odierno, fornendo nuovi studi e strumenti che possono essere utili a prevedere eventuali nuovi scenari. Infatti, possono esserci artisti emergenti che potrebbero godere di maggiore popolarità futura se ben collegati, con artisti di popolarità alta e collegati con ratio alti; allo stesso tempo, si potrebbero notare alcuni artisti ritenuti popolari che potrebbero avere un trend opposto e quindi in declino.[1]

KEYWORDS

Social Network Analysis, musica, *Last.fm*, affinità, similarità, popolarità

ACM Reference Format:
Calogero Giudice. 2025. Analisi dell'affinità musicale degli artisti di Last.fm. In *Analisi dell'affinità musicale degli artisti di Last.fm.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1]**Project Repositories**
Data Collection: https://github.com/Kalo9603/2024_Giudice/tree/main/data_collection
Network Analysis: https://github.com/Kalo9603/2024_Giudice/tree/main/network_analysis
Open Problem: https://github.com/Kalo9603/2024_Giudice/tree/main/open_problem
Report: https://github.com/Kalo9603/2024_Giudice/tree/main/report

SNA '25, 2024/25, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUZIONE

Nell'epoca dell'affermazione dell'industria musicale digitale, piattaforme come *last.fm* offrono tanti dati sugli artisti musicali, sulle abitudini di ascolto degli utenti e sulle relazioni tra i singoli profili musicali (gli utenti, gli artisti, gli album, etc.). Gli artisti presenti nella piattaforma sono stati analizzati sotto una prospettiva di rete sociale: si possono ipotizzare così degli scenari, dei fenomeni interessanti: la similarità tra artisti, la popolarità degli stessi, l'affermazione di artisti emergenti o l'effetto contrario.

L'analisi parte dalla costruzione di una rete basata su coefficienti di similarità; di questi ne sono stati individuati due:

- Il *Sample Ratio*, ovvero l'intensità del legame tra due artisti in base ai simili condivisi.
- Lo *Jaccard Ratio*, ovvero l'affinità relativa all'insieme totale di simili.

Matematicamente, è molto più probabile che lo *Jaccard* sia inferiore del rispettivo *Sample*, poiché il denominatore del primo è sempre maggiore del secondo (a meno che tali denominatori non coincidono).

In parallelo si è stimata la **popolarità** degli artisti presenti nel dataset: per far ciò ci si è basati sullo **Z-Index**, che mette in relazione il numero di ascoltatori e quello di ascolti. Ciò ha permesso di classificare gli artisti in tre classi: ***mainstream***, ***medio*** ed ***emergente***; a questi ne sono stati affiancati altri tre: ***possibile mainstream***, ***possibile medio*** e ***ritorno emergente*** in base al trend di (de)crescita calcolato.

L'obiettivo principale della seconda parte progetto è quello di identificare le caratteristiche strutturali degli artisti mainstream e di quelli emergenti ed, eventualmente, di prevedere un cambio di status in base a determinate metriche.

2 DATA COLLECTION

La raccolta dei dati da analizzare si è concentrata su tre fasi specifiche:

(1) La generazione di un elenco di artisti in base agli artisti simili;
(2) La creazione di una lista di archi che collegano due artisti se entrambi i coefficienti di similarità superano

la soglia del 50%. Il *Sample Ratio* è calcolato rispetto al numero di simili condivisi, mentre lo *Jaccard Ratio* rispetto all'insieme totale dei simili;

(3) Il successivo accostamento del numero di ascolti, di ascoltatori e il calcolo delle popolarità degli artisti.

Gli elenchi degli artisti e degli artisti simili sono stati ottenuti per mezzo dell'API di *last.fm* e, tramite programmazione ad oggetti, sono state generate le altre liste. Nel complesso vi sono quattro file:

- *artists.csv*, una lista di 28 870 artisti;
- *links.csv*, un elenco di 98 890 archi che collegano 11 745 artisti (il 40.68% del totale). A causa delle lunghe tempistiche è stato elaborato il dataset sul 64.26% dei possibili archi totali (sono stati esaminati più di 267 milioni di possibili archi su oltre 416 milioni totali: $\frac{28\,870 \cdot 28\,869}{2} = 416\,724\,015$);
- *popularity.csv*: la medesima lista di *artists.csv* con inclusi il numero di ascoltatori, quello di ascolti, il coefficiente di popolarità e il suo logaritmo;
- *popularity_z.csv*: la medesima lista di *popularity.csv* a cui si accoda lo Z-Index.

In particolare, i coefficienti di correlazione citati sono calcolati come segue.

Indice	Formula
Sample Ratio	$SR(a,b) = \frac{\lvert S_a \cap S_b \rvert}{k}$
Jaccard Ratio	$JR(a,b) = \frac{\lvert S_a \cap S_b \rvert}{\lvert S_a \cup S_b \rvert}$
Popolarità	$pop_i = \frac{ascolti_i}{ascoltatori_i}$
Popolarità logaritmica	$pop_{\log,i} = \log(pop_i + 1)$
Media	$\mu = \frac{1}{N} \sum_{i=1}^{N} pop_{\log,i}$
Deviazione standard	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (pop_{\log,i} - \mu)^2}$
Z-Index	$Z_i = \frac{pop_{\log,i} - \mu}{\sigma}$

Table 1: Formule utilizzate per la classificazione e l'analisi strutturale

Dati selezionati

Al fine di avere un dataset iniziale di artisti si è pensato, partendo da quattro artisti provenienti da quattro generi musicali diversi (Annalisa per il pop italiano; Giorgio Vanni per le sigle, Katy Perry per la quota internazionale e Ado per il mercato giapponese).

Quindi, partendo da essi, viene creata la lista di artisti simili a cominciare da quello in input: se non è presente, l'artista trovato viene aggiunta in lista. La medesima operazione viene svolta, alla prima iterazione, sui primi quattro artisti; poi, dalla seconda, si considerano anche tutti quelli aggiunti in coda. In questo modo è stato ottenuto un dataset di 28 870 artisti. Relativamente all'ordine di inserimento nel dataset, a ciascun artista è stato assegnato un ID autoincrementale.

Al termine si è passati, basandosi sull'elenco degli artisti, alla generazione del dataset dei *links* del grafo. Per far ciò per ciascun artista della lista, sfruttando le proprietà di un grafo non orientato:

- Si considera l'artista con ID immediatamente successivo;
- Per la coppia in esame si prendono i rispettivi insiemi di artisti simili e si calcolano i ratio, le cui formule sono specificate nel formulario precedente;
- Se entrambi il Sample e lo Jaccard Ratio raggiungono la soglia del 50%, l'arco viene registrato nel dataset;
- Si prosegue il calcolo con l'artista successivo con il medesimo algoritmo fino a raggiungere la fine della lista degli artisti.

In questo modo, ad esempio, il primo nodo (Annalisa) farà 28 869 iterazioni, il secondo (Giorgio Vanni) 28 868, e così via. Quindi gli ultimi nodi faranno molte meno iterazioni rispetto ai primi perché le altre possibili coppie vengono esaminate dai nodi precedenti.

Per migliorare ulteriormente la velocità del processo è stato adottato il *multiprocessing*, in modo che più processi lavorino in parallelo sullo stesso set di artisti. Dati i tempi relativamente lunghi del processo, l'analisi del dataset è iniziato non appena il numero di artisti all'interno dei dataset dei *links* ha raggiunto la soglia richiesta. Al termine sono stati collegati 11 745 artisti con 98 890 archi.

3 NETWORK CHARACTERIZATION

Si esamina qui di seguito il grafo ottenuto dalla fase di Data Collection.

Come anticipato in precedenza, partendo da 28 870 artisti è stato ottenuto un grafo, analizzando il 64,26% dei possibili archi, formato da 11 745 artisti a 98 890 archi.

La densità quindi è molto bassa, ovvero di 1.433×10^{-3}; questo è dovuto principalmente al fatto di aver adottato una soglia minima di Jaccard Ratio più alta. Infatti, matematicamente, la probabilità che lo Jaccard Ratio di un arco abbia valore minore del rispettivo Sample è nettamente maggiore, in quanto il denominatore del rapporto (vedasi la tabella nel

paragrafo precedente) risulta maggiore di un valore fisso come nel Sample.

Riassumendo, le prime generalità del grafo sono le seguenti:

Artisti totali	28 870
Artisti nel grafo	11 745 (40,68%)
Archi	98 890
Densità	1.433×10^{-3}

Table 2: Prime generalità del grafo

Distribuzione di grado

Il grado medio del grafo è di 8.42.
Per calcolarlo sono stati utilizzati, per completezza, due metodi:

- Dividendo il numero di righe del dataset per il numero di artisti connessi;
- Dividendo la somma dei gradi dei singoli artisti per il numero degli stessi.

Nota che l'errore tra i due metodi è dell'ordine di 10^{-4}.
Dal grafico si evince la forte asimmetria del grado medio degli artisti: ve ne sono molti con grado basso e meno con grado più alto.

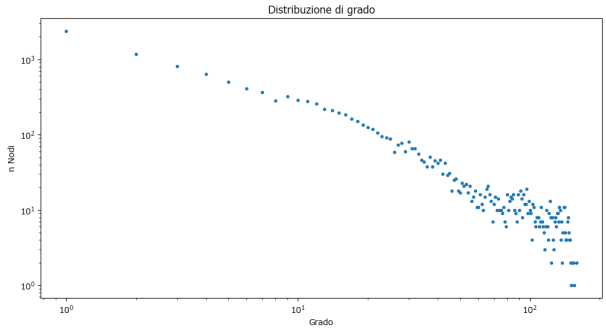

Figure 1: Distribuzione del grado degli artisti

Questa asimmetria dei gradi del grafo si mostra maggiormente dalle distribuzioni CDF e CCDF: la probabilità di trovare un artista con grado elevato decresce.

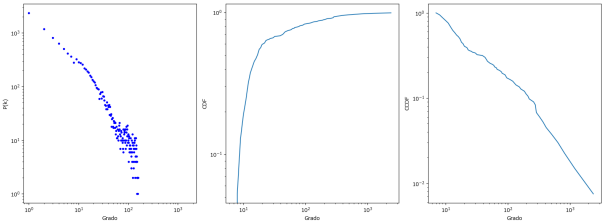

Figure 2: Distribuzione del grado degli artisti - CDF e CCDF

Dato che si tratta di un grafo pesato su due livelli, è interessante mostrare come si comportano i rispettivi pesi con i medesimi grafici. Per cui si illustrano i medesimi grafici ma pesati: qui è evidente come lo Jaccard Ratio risulta più traslato indietro rispetto al Sample corrispondente:

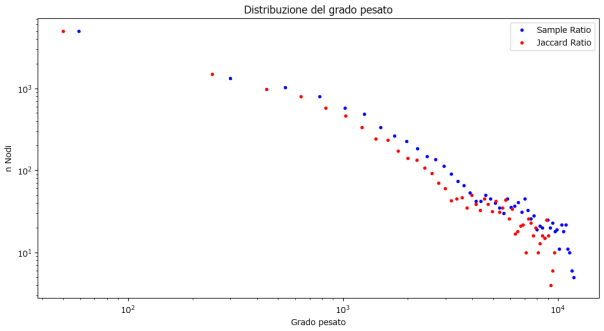

Figure 3: Distribuzione del grado pesato degli artisti

Invece dal grafico corrispondente CDF e CCDF si nota come il range dei gradi pesati Jaccard sia molto più variabile rispetto al Sample. Risulta altresì interessante la continua sovrapposizione di entrambi i gradi nel modello CCDF:

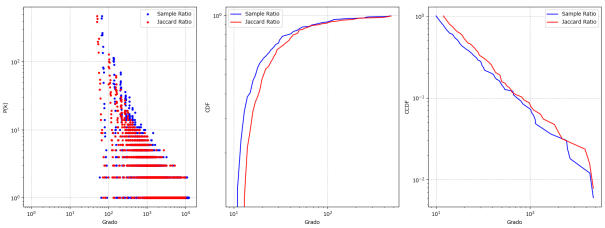

Figure 4: Distribuzione del grado pesato degli artisti - CDF e CCDF