UNIVERSITÀ DEGLI STUDI DI UDINE

Thesis Title

di

Andrea Beggiato

Tesi presentata per il conseguimento
della laurea magistrale

in

Facoltà di scienze MM.FF.NN.
Comunicazione multimediale e tecnologie dell'informazione

Febbraio 2014

Dichiarazioni dell'autore

Io, ANDREA BEGGIATO, dichiaro che la tesi intitolata 'THESIS TITLE' ed il lavoro presentato in essa sono frutto di . Confermo che:

- Questo lavoro stato fatto interamente durante la frequentazione del corso "Comunicazione multimediale e tecnologie dell'informazione" presso questa Universit.
- Qualora sia stata presentata in precedenza qualsiasi parte di questa tesi di laurea presso questa Universit o qualsiasi altra istituzione, questo stato chiaramente affermato.
- Dove ho consultato il lavoro pubblicato di altri, questo sempre chiaramente attribuito.
- Dove ho citato lavoro altrui, la fonte sempre data. Con l'eccezione di tali citazioni, questa tesi interamente lavoro personale.
- Qualora la tesi si basa sul lavoro svolto da me insieme ad altri, espressamente sottolineato ci che stato fatto da altri e quello a cui io ho contribuito.

Firmato:

Data:

"Prediction is very difficult, especially if it's about the future."

Niels Bohr

UNIVERSITÀ DEGLI STUDI DI UDINE

Abstract

Facoltà di scienze MM.FF.NN.

Comunicazione multimediale e tecnologie dell'informazione

Doctor of Philosophy

di Andrea Beggiato

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Riconoscimenti

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

For/Dedicated to/To my...

Chapter 1

Introduzione

Chapter 2

Stato dell'arte

In questo capitolo sarà descritto lo stato dell'arte riguardanti gli studi e le analisi effettuate relative ad informazioni di tipo geolocalizzato riguardanti gli esseri umani; questa tipologia di dati abbraccia diversi ambiti, tra i quali l'ambito sociale e l'ambito scientifico.

L'interesse che gli accademici che si occupano di sociologia e comportamento umano è dovuto principalmente al modo in cui gli spostamenti delle persone condizionino le amicizie e relazioni tra loro, ma soprattutto come l'analisi dei dati sia spesso contrastante con l'impressione che hanno le persone stesse dei loro spostamenti, come esposto da Nathan Eagle ed altri in [1].

Un'altro ambito in cui lo studio di informazioni di tipo geolocalizzato è quello scientifico, dove si può trovare l'applicazione di diverse discipline matematiche per effettuare l'analisi di dati grezzi; tra queste la teoria dei grafi e la statistica sono sicuramente le maggiormente utilizzate. Gli studiosi che si occupano di scienze sono interessati all'aggregazione dei dati grezzi in dati pi strutturati, per poter, ad esempio, essere in grado di predirre sia le posizioni delle persone nel futuro, sia le relazioni di amicizia tra gli utenti nel tempo.
Infine, utilizzando i dati strutturati in maniera adeguata, è possibile creare un

sistema di raccomandazione e stimare la similitude tra utenti basandosi solamente su informazioni di tipo geolocalizzato.

2.1 Ambito sociale

Lo studio di Nathan Eagle ed altri in [1], il quale durato complessivamente nove mesi, si basa su 94 soggetti dello stesso gruppo di lavoro dotati di smartphone che hanno installato al loro interno alcune applicazioni che permettono di registrare ed inviare ad i ricercatori diverse informazioni, tra cui il log delle chiamate, l'identificativo dei dispositivi Bluetooth che sono stati a meno di cinque metri dal soggetto e l'identificativo della cella attraverso la quale lo smartphone riceve il segnale.

Oltre a questi dati analitici, che possiamo definire comportamentali, ad ogni soggetto è stato chiesto di compilare alcuni questionari mensili che mirano a raccogliere informazioni personali riguardanti le relazioni d'amicizia e la durata approssimativa della vicinanza con altri soggetti; i dati emersi da questi questionari vengono definiti autodichiarati.

L'analisi di tutti i dati raccolti divisa in tre fasi:

- analisi della relazione tra dati comportamentali ed autodichiarati
- analisi della presenza di dati comportamentali che caratterizzano l'amicizia

2.1.1 Relazione tra dati comportamentali ed autodichiarati

Negli ultimi trent'anni si è discusso molto sull'affidabilità delle misurazioni esistenti per le relazioni, osservando soprattutto che le osservazioni comportamentali sono debolmente correlate con le interazioni riportate dai soggetti; alcuni studi [2] hanno dimostrato come le persone riescono a ricordare meglio le strutture sociali a lungo termine rispetto a quelle nel breve periodo. Si possono riscontrare due diverse tipologie di bias, uno basato sul ricordo degli eventi recenti denominato *recency bias*, l'altro basato sul ricordo degli eventi pi importanti denominato *salience bias*;

attraverso i dati raccolti si pu quindi assimilare l'effetto di *recency bias* alla quantità di interazioni in un periodo prefissato antecedente al questionario e l'effetto di *salience bias* alla presenza o meno di una relazione di amicizia tra due soggetti.

Attraverso l'incrocio dei dati comportamentali ed autodichiarati è emerso che la maggiorparte della prossimit è misurata è stata invece dichiarata dal soggetto come non vicinanza; inoltre, quando i due dati non erano in contrasto, il tempo di contatto sempre stato sovrastimato, essendo la media dei dati comportamentali di 33 minuti al giorno contro la media dei dati autodichiarati di 87 minuti al giorno. Infine si è osservato che i dati riportati da soggetti che si reputano amici sono molto pi precisi rispetto ai dati riportati da soggetti che non si considerano amici.

2.1.2 Analisi della presenza di dati comportamentali che caratterizzano l'amicizia

Analizzare i dati comportamentali per evidenziare il grado di relazioni tra due soggetti, come l'amicizia, è diverso da misurarne la vicinanza; infatti è plausibile che due persone anche essendo amiche, siano distanti anche per periodi di tempo piuttosto lunghi. Ad ogni modo il contesto, sia spaziale che temporale, pu aiutare a definire alcuni pattern per la predizione delle amicizie, ad esempio l'aver passato con un'altra persona poche ore un sabato sera in un posto diverso dal luogo di lavoro indica una relazione differente rispetto all'aver passato quattro ore nel luogo di lavoro di un mercoledì pomeriggio.

In figura 2.1 è rappresentata graficamente la distribuzione della probabilit di vicinanza sia all'interno del luogo di lavoro, sia all'esterno, tra persone che si reputano amici reciprocamente, persone tra le quali solamente una delle parti si reputa amica dell'altra parte e persone che non si reputano amici a vicenda; si nota immediatamente come la vicinanza è pi probabile tra le prime due categorie di persone, ma, essendo il luogo d'incontro un fattore determinante, si può osservare

come la vicinanza misurata all'esterno del luogo di lavoro sia maggiore per chi si reputa amico reciprocamente.

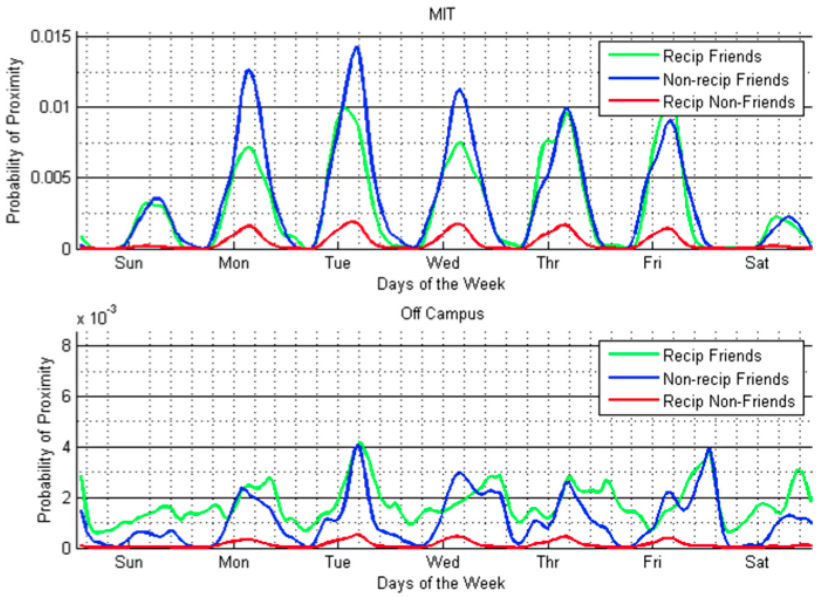

FIGURE 2.1: Distribuzione della vicinanza nel tempo e nello spazio

Nathan Eagle ed altri hanno successivamente classificato la vicinanza in diverse variabili corrispondenti alla vicinanza all'interno o all'esterno del campus, alla vicinanza di giorno o di notte e alla vicinanza nei giorni lavorativi o nei fine settimana; una fattorizzazione di queste variabili ha evidenziato come esistano solamente due fattori discriminanti, ovvero la vicinanza durante le ore del giorno nel luogo di lavoro e la vicinanza nelle ore serali o nei fine settimana all'esterno del campus. Solamente utilizzando il secondo fattore è possibile predirre il 96% percento dei rapporti di non amicizia reciproca ed il 95% percento dei rapporti di amicizia reciproca, potendo quindi tralasciare i dati autodichiarati di amicizia, quest'ultima risulta stimabile correttamente utilizzando solamente i dati comportamentali.

2.2 Ambito scientifico

2.2.1 Predizione delle traiettorie

2.2.2 Similitude tra utenti

Negli ultimi anni gli utenti del web hanno iniziato ad utilizzare tecnologie per il tracciamento dei dati GPS anche per registrare esperienze sportive o viaggi personali, per poi condividerle nel web con altri utenti oppure per visualizzare in una mappa i propri spostamenti. Tuttavia le applicazioni che utilizzano i tracciati GPS utilizzano i dati grezzi, senza focalizzarsi sulla comprensione degli stessi; esistono inoltre applicazioni che cercano di razionalizzare i log derivati dal tracciamento GPS, estrapolando l'attività che un utente sta compiendo in quel momento, come ad esempio una corsa a piedi o in bicicletta, oppure cercando di sintetizzare i dati grezzi in luoghi che un utente ha visitato.

Un ulteriore campo di applicazione che prevede l'utilizzo di tracciati GPS riguarda il confronto tra diversi utenti per calcolare la similitudine tra essi, come hanno provato ad analizzare Quannan Li ed altri in [3]. Secondo la prima legge della geografia ogni cosa è correlata, ma le cose che sono più vicine in termini geografici lo sono di più, quindi la tesi dei ricercatori, basandosi su questa legge, consiste nella similitudine tra due utenti in base alla loro storia in termini di tracciati GPS.

2.2.2.1 Definizioni

Log Gps : Un log GPS è una sequenza di punti GPS, che contengono una coordinata che rappresenta la latitudine, una coordinata che rappresenta la longitudine ed il timestamp della registrazione.

Punto stazionario : Un punto stazionario è una regione geografica dove l'utente rimane per un determinato periodo ti tempo; a differenza di un punto GPS,

un punto stazionario possiede un significato semantico, come ad esempio la casa o l'ufficio,

Storia dei luoghi : Dato un log GPS e i punti stazionari che si possono ricavare, la storia dei luoghi di un utente è rappresentata dalla sequenza dei posti che esso ha visitato, corredati di una data di arrivo ed una data di partenza.

Grafo gerarchico : La similitudine in termini geografici tra due utenti non è binaria, ma deve essere valutata rispetto ad una scala; per questo motivo i ricercatori hanno considerato l'insieme di tutti i punti stazionari di tutti gli utenti come un unico dataset che hanno poi clusterizzato utilizzando la distanza tra i vari punti stazionari; ad ogni step dell'algoritmo di clusterizzazione veniva generato un nuovo livello in cui i cluster rappresentavano aree geografiche pi ampie. Ad ogni livello hanno potuto creare dei grafi diretti per ogni utente, in cui un nodo era rappresentato dal cluster di appartenenza del punto stazionario del relativo utente, mentre un arco veniva tracciato tra due cluster quando la traiettoria dell'utente cambiava area geografica; in questo modo la similitudine tra due utenti è può essere valutata per ogni livello sulla base dei cluster e delle traiettorie tra cluster condivise, risultando in una similitudine più elevata per le condivisioni che avvengono ai livelli più bassi. Questo procedimento è illustrato in 2.2.

2.2.2.2 Processo per il calcolo della similitudine

La struttura del grafo gerarchico di 2.2 è una valida rappresentazione della storia dei luoghi di un utente, contenente quindi sia informazioni di tipo spaziale che temporale; i ricercatori, per poter calcolare la similitudine tra due utenti, hanno innanzitutto cercato i nodi in comune nei rispettivi grafi gerarchici degli utenti e successivamente sono state generate delle sequenze contenenti i nodi trovati, riconducendo il problema della similitudine di due utenti ad un problema di pattern matching tra sequenze di oggetti.

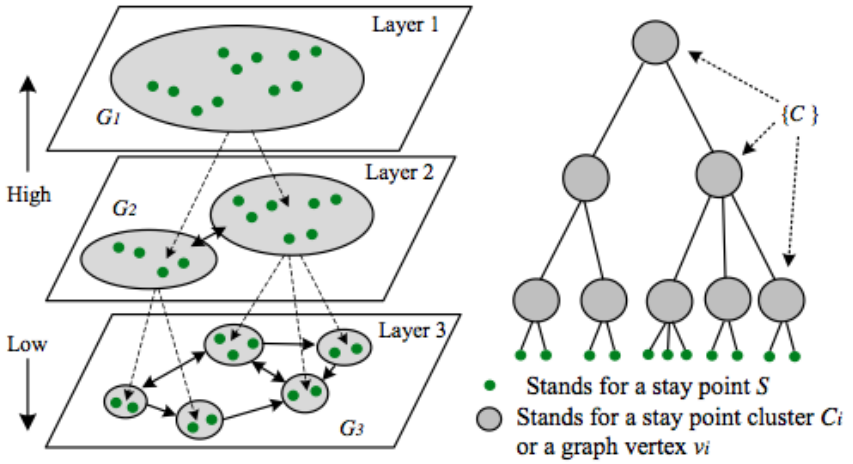

FIGURE 2.2: Rappresentazione del grafo gerarchico basato su cluster di punti stazionari

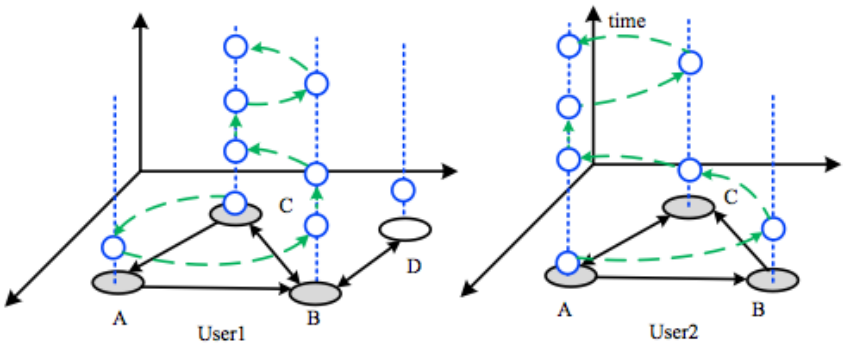

FIGURE 2.3: Rappresentazione delle sequenze di due utenti

In 2.3 sono visualizzati due grafi gerarchici di due utenti distinti allora stesso livello di clusterizzazione; in ognuno dei due grafi sono tracciati in grigio i nodi che corrispondono al cluster, mentre i nodi più colorati in blu e collegati con una linea tratteggiata identificano i punti stazionari visitati dagli utenti in tempi diversi. La freccia tratteggiata in verde, invece, rappresenta la successione temporale in cui gli utenti hanno esplorato i propri punti stazionari.

Si può notare come i due utenti condividano i nodi A, B e C e generino, rispettivamente, le sequenze $< C, A, B, B, C, C, B, C >$ e $< A, B, C, A, A, C, A >$ che per semplicità vengono rappresentate con $< C(1), A(1), B(2), C(2), B(1), C(1) >$ e $< A(1), B(1), C(1), A(2), C(1), A(1) >$ dove il numero tra parentesi che segue il

nome di un nodo rappresenta il numero di volte che un utente ha viaggiato successivamente in punti stazionari appartenenti allo stesso cluster; utilizzando inoltre il tempo di entrata e di uscita da ogni cluster è possibile calcolare l'intervallo di tempo Δt_I tra due elmetti delle sequenze, ottenendo:

User 1: $C(1) \xrightarrow{\Delta t_1} A(1) \xrightarrow{\Delta t_2} B(2) \xrightarrow{\Delta t_3} C(2) \xrightarrow{\Delta t_4} B(1) \xrightarrow{\Delta t_5} C(1)$
User 2: $A(1) \xrightarrow{\Delta t_{1'}} B(1) \xrightarrow{\Delta t_{2'}} C(1) \xrightarrow{\Delta t_{3'}} A(2) \xrightarrow{\Delta t_{4'}} C(1) \xrightarrow{\Delta t_{5'}} A(1)$

In modo generico, quindi, la storia dei luoghi di un utente può essere scritta come:

$$seq =< a_1(k_1) \xrightarrow{\Delta t_1} a_2(k_2) \xrightarrow{\Delta t_2} a_3(k_3) \xrightarrow{\Delta t_3} \ldots >$$

dove $a_i \in V$ è l'identificativo del cluster e k_i è il numero di volte che un utente ha visitato successivamente un punto stazionario all'interno del cluster a_i. In modo analogo si può calcolare il tempo tra due elementi della sequenza come:

$$\Delta t_i = a_{i+1}(0).arvT - a_i(k_i - 1).levT$$

dove $arvT$ identifica il tempo di entrata, $levT$ il tempo di uscita e l'indice all'interno delle parentesi tonda identifica un punto stazionario all'interno del cluster.

Pattern matching Per definire simili due sequenze devono essere soddisfatte le seguenti condizioni:

$$seq_1 =< a_1(k_1) \xrightarrow{\Delta t_1} a_2(k_2) \xrightarrow{\Delta t_2} a_3(k_3) \xrightarrow{\Delta t_3} \ldots a_m(k_m) >$$
$$seq_2 =< b_1(k_{1'}) \xrightarrow{\Delta t_{1'}} b_2(k_{2'}) \xrightarrow{\Delta t_{2'}} b_3(k_{3'}) \xrightarrow{\Delta t_{3'}} \ldots b_m(k'_m) >$$

1. $\forall 1 \leq i \leq m, a_i = b_i$, ovvero i nodi alla stessa posizione devono avere lo stesso identificativo.

2. $\forall 1 \leq i < m, |\Delta t_i - \Delta t_{i'}| \leq t_{th}$, ovvero due utenti devono avere un'intervallo di tempo tra due transizioni minore di una certa soglia.

Utilizzando questa definizione di similitudine tra due sequenze è possibile definire sequenze simili di lunghezza diversa; un esempio di queste definizione si può visualizzare graficamente in 2.4

Per calcolare la similitudine tra due utenti, basandosi sulle definizione precedentemente date, si dovranno tener conto sia delle lunghezze delle diverse sequenze simili tra loro ma anche del livello del grafo gerarchico in cui queste sequenze sono state calcolate; possiamo quindi definire il punteggio di ogni sequenza come:

$$s_{(m)} = \alpha_{(m)} \sum_{i=1}^{m} min(k_i, k_{i'})$$

dove $\alpha_{(m)}$ è un coefficiente che dipende dalla lunghezza m della sequenza e che aumenta il punteggio per le sequenze più lunghe.

La similitude di due utenti ad un certo livello del grafo gerarchico si basa invece su tutte le sequenze simili trovate in quel livello; si può quindi definire:

$$S_t = \frac{1}{N_1 * N_2} \sum_{i=1}^{n} s_i$$

dove s_i è il punteggio della sequenza i-esima calcolata con la formula precedente, n è il numero di sequenze simili tra due utenti trovati nello specifico livello ed N_1 e N_2 sono rispettivamente il numero di punti stazionari dei due utenti. La divisione per questi due fattori è motivata da un problema di bilanciamento tra i dati di due utenti; intuitivamente se non venisse tenuto conto della quantità di dati gli utenti che hanno maggior punti stazionari sarebbero favoriti nel contenere al loro interno sequenze simili di coloro che hanno meno punti stazionari.

Infine i ricercatori hanno definito una misura di similitudine tra utenti attraverso tutti i livelli, definita come la somma pesata attraverso un opportuno coefficiente dipendente dal livello, che a livelli inferiori assume un valore maggiore, di tutti i vari punteggi ottenuti sui livelli del grafo gerarchico.

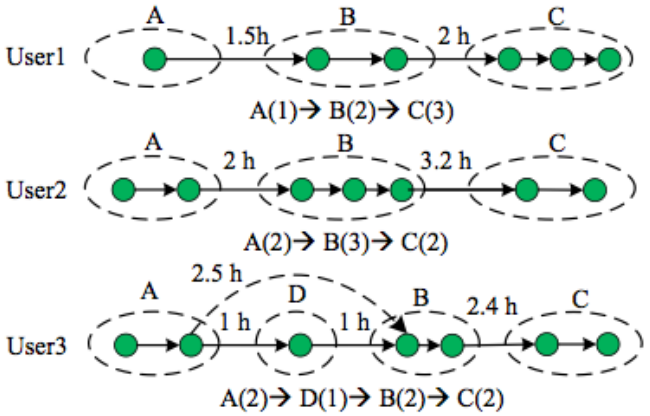

FIGURE 2.4: Similitudine tra sequenze tra tre utenti

2.2.3 Sistema di raccomandazione

Un sistema di raccomandazione è una soluzione che cerca di predirre una valutazione o un insieme di preferenze che un utente attribuirebbe ad un oggetto; esistono diverse metodologie per la costruzione e l'applicazione di un tale sistema divisibili in due categorie principali:

- collaborative filtering
- content-based filtering

2.2.3.1 Collaborative filtering

L'approccio di tipo *collaborative filtering* si basa essenzialmente sul collezionare ed analizzare un grande numero di informazioni sui comportamenti, sulle attività o sulle preferenze degli utenti e di predirre successivamente quello che può interessare ad un utente basandosi sulla similitudine con altri utenti; uno dei principali vantaggi di questa soluzione risiede nel fatto che non è necessario che i contenuti da proporre siano "capibili" dalla macchina, potendo quindi raccomandare elementi complessi quali film, musica o libri.

La raccolta delle preferenze degli utenti può essere effettuata in due modalità, spesso utilizzate congiuntamente, ovvero la raccolta di dati espliciti da parte degli utenti, come ad esempio chiedere all'utente di valutare un oggetto o di scegliere tra due oggetti quali preferisce, e la raccolta di dati impliciti, come ad esempio osservare gli oggetti che un utente visita in un sito di e-commerce oppure memorizzare una lista degli ultimi oggetti acquistati. La similitudine tra gli utenti, invece, viene valutata utilizzando spesso tecniche che appartengono a diverse discipline, come l'algoritmo k-NN per la classificazione di elementi derivato dal *machine learning* oppure come la correlazione di Pearson derivata dalla statistica.

Molti sono i sistemi di *collaborative filtering* presenti, sia commerciali che non commerciali, tra cui i più famosi comprendono il sistema di Amazon.com che propone oggetti che altri utenti hanno acquistato congiuntamente all'oggetto che si desidera acquistare in quel momento, Last.fm che consiglia un brano musicale basandosi sulle abitudini d'ascolto di altri utenti e Facebook che consiglia nuovi possibili amici basandosi sulle connessione tra gli utenti ed i loro amici.

Questo tipo di approccio soffre di alcuni problemi, tra cui i principali si possono riassumere in:

- Cold Start: questi sistemi richiedono un'enorme quantità di dati presente per ogni utente per poter generare raccomandazioni accurate.
- Scalability: nella maggiorparte dei sistemi di raccomandazioni esistono milioni di utenti ed oggetti, necessitando quindi di una grande potenza di calcolo.
- Sparsity: il numero di oggetti che un utente può valutare è spesso elevato, quindi è molto probabile che per ogni oggetto ci siano poche valutazioni da parte degli utenti, anche per l'oggetto che ad esempio ha più vendite in un sito di e-commerce.

2.2.3.2 Content-based filtering

Un'altro approccio comune tra i sistemi di raccomandazione è il *content-based filtering* che si basa sulle descrizione degli oggetti e sul profili delle preferenze degli utenti; in questo tipo di sistema, le parole chiave sono utilizzate per descrivere gli oggetti, mentre i profili delle preferenze degli utenti sono costruite per indicare che tipologia di oggetti l'utente predilige.

Il confronto tra le preferenze degli utenti e le descrizioni degli oggetti è effettuato utilizzando algoritmi di information retrieval basati sugli spazi vettoriali; il sistema scompone le descrizioni degli oggetti in un insieme di attributi discreti che possono essere rappresentati quindi come vettori in modo analogo a come accade per i documenti in un sistema di information retrieval, mentre il profilo dell'utente rappresenta la query di ricerca. Come accade nei sistemi di information retrieval classici, diverse tecniche possono essere utilizzate sia per determinare la rilevanza di un oggetto nei confronti di un utente, tra cui BM25 o le reti neurali, sia per validare e migliorare l'efficacia dell'algoritmo a posteriori proponendo all'utente di valutare l'oggetto reperito dal sistema.

Uno tra i più noti sistemi di raccomandazione basati sul *content-based filtering* è utilizzato da Pandora Radio, che permette ad un utente di ascolta musica con caratteristiche simili ad una canzone che l'utente ha inviato precedentemente e che quindi sarà utilizzata come query di ricerca.

2.2.3.3 Location-content-aware filtering

L'emergere di social network sia basati sugli eventi, come Meetup e DoubanEvent, sia basati sulla posizione geografica, come Foursquare o Gowalla, hanno permesso a Hongzhi Yin ed altri[4] di sperimentare un nuovo sistema di raccomandazione basato su dati geolocalizzati per predirre posti o eventi agli utenti basandosi sugli

interessi personali e sulle preferenze locali, ovvero l'insieme delle preferenze degli utenti in una determinata area.

La realizzazione di un tale sistema comprende ulteriori difficoltà rispetto ai precedenti, in quanto una persona difficilmente visita molti luoghi diversi, causando quindi una grande sparsità nella matrice utenti-posti; inoltre è molto frequente che un utente visiti posti od eventi in un'area circoscritta alla città in cui vive, rendendo quindi molto difficile il suggerimento di buone raccomandazioni in città diverse, data la mancanza di dati storici per quell'utente. Alcune analisi infatti evidenziano come le attività che un utente compie all'esterno dell'area della propria città rappresentino solamente lo 0.47% delle attività complessive, introducendo quindi il problema denominato della *new city*.

Un esempio che rafforza l'intenzione degli autori di creare un nuovo sistema di raccomandazione per questa tipologia di dati può essere il seguente: un utente u è uno shopaholic e tende a visitare spesso il centro commerciale $v^{'}$ nella sua città; v è un centro commerciale molto popolare nella città l_v di cui u è nuovo. Intuitivamente un buon sistema di raccomandazione suggerirebbe v ad u quando quest'ultimo di trova ad l_v, ma un metodo basato sul *collaborative filtering* fallirebbe in quanto ci sarebbero troppi pochi utenti in comune tra v e $v^{'}$ e di conseguenza i vettori di v e $v^{'}$ non sarebbero simili.

Il sistema di raccomandazione creato da Hongzhi Yin ed altri si basa su due modelli, uno offline ed uno online, come rappresentato in 2.5.

Modello offline Il modello offline, denominato LCA-LDA, è un modello generativo basato sulla probabilià che cerca di simulare il processo di decisione umano nella scelta di posti ed eventi; questo modello considera sia gli interessi personali dell'utente, sia le preferenze locali in maniera unificata con la seguente formula:

$$P(v|\theta_u, \theta_{l_u}) = \lambda_u P(v|\theta_u) + (1 - \lambda_u) P(v|\theta_{l_u})$$

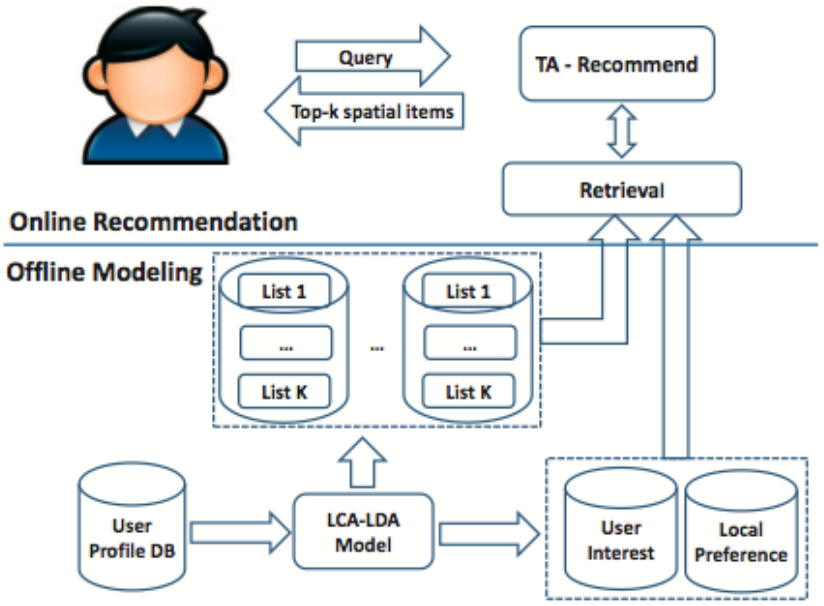

FIGURE 2.5: Rappresentazione del sistema di raccomandazione di Hongzhi Yin ed altri

dove $P(v|\theta_u)$ è la probabilità che il posto v sia generato secondo gli interessi personali dell'utente u, indicati con θ_um e $P(v|\theta_{l_u})$ è la probabilità che l'elemento v sia generato secondo le preferenze locali, indicate con θ_{l_u}. Il parametro λ_u è un moltiplicatore che quantifica l'influenza degli interessi personali rispetto alle preferenze locali per l'utente u.

I valori dei parametri nascosti, ovvero θ_u, θ_{l_u} e λ_u sono inferiti utilizzando diverse iterazione del campionamento di Gibbs

Modello online Una query nel sistema di raccomandazione presentato è composta da due argomenti (u, l_u), ovvero un utente u con una città l_u dove desidera viaggiare. Il risultato è una lista di posti o eventi della città l_u ordinati secondo la probabilità calcolata nel modello offline. Durante l'interrogazione è necessario calcolare il punteggio per ogni posto o evento e successivamente selezionare i miglio k risultati; quando il numero di posti o eventi diventa molto grande, ad esempio

supera il milione, il calcolo dei primi k elementi richiede milioni di operazioni tra vettori.

Per migliorare il tempo di calcolo necessario i ricercatori hanno esteso l'algoritmo basato sulle soglie descritto in [5] cercando di mantenere offline una lista dei posti o degli eventi ordinati per le preferenze locali; in questo modo è sufficiente solamente calcolare le probabilità per gli interessi specifici degli utenti.

Dataset utilizzato DoubanEvent è il più grande social network sugli eventi in Cina, dove un utente può creare un evento specificando cosa, dove e quando l'evento si terrà; successivamente gli altri utenti possono esprimere il loro intento a partecipare o meno all'evento creato. Il dataset utilizzato nell'esperimento di Hongzhi Yin ed altri è estratto da DoubanEvent e contiene 100.000 utenti, 300.000 eventi e 3.500.000 espressioni di partecipazione da parte degli utenti.

Risultati ottenuti Per validare i risultati ottenuti, i ricercatori hanno confrontato la *recall* dei primi k risultati sia in query che contenevano la città propria dell'utente, sia in query che contenevano nuove città. In 2.6 si può notare come il sistema LCA-LDA proposto sia migliore sia nella raccomandazione di posti o eventi nella città di appartenenza dell'utente, sia nel risolvere il problema della *new city*, mostrando una recall di 0.33@10 e di 0.42@20.

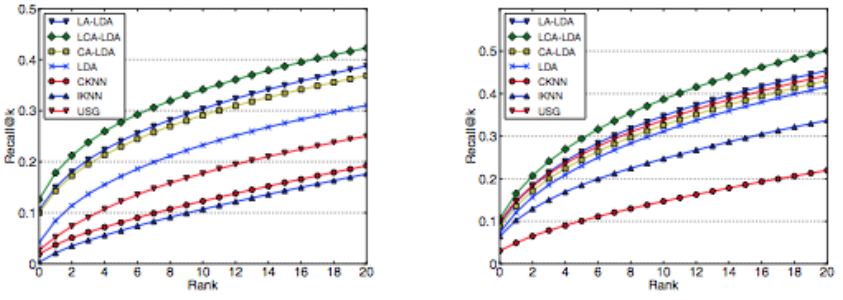

FIGURE 2.6: Risultati della recall@k nelle nuove città (a sx) e nella citt propria dell'utente (a dx)

2.2.4 Predizione delle connessioni tra utenti

Il suggerimento di nuovi amici nei social network utilizza principalmente le relazioni di amicizia tra gli utenti ed i loro amici; con l'introduzione di servizi di geolocalizzazione, quali Facebook Places o Foursquare, le informazioni finora utilizzate possono essere ampliate aggiungendo i posti frequentati dagli utenti. Una caratteristica fondamentale di social network basati sulla posizione geografica consiste nella presenza di milioni di nodi, ma allo stesso tempo una grande sparsità, risultando in una bassa densità di archi tra i vari nodi.

Salvatore Scellato ed altri [6] hanno cercato di utilizzare questa nuova tipologia di informazioni per rispondere alla domanda: *Com'è possibile progettare un sistema di predizione dei collegamenti tra utenti utilizzando i loro check-in?*. Infatti in questi sistemi le interazioni che gli utenti hanno con i luoghi è volontario, ovvero un utente deve effettuare una specifica azione, come ad esempio il click su di un link, per poter registrare la propria posizione presso quel luogo; tale azione è denominata check-in.

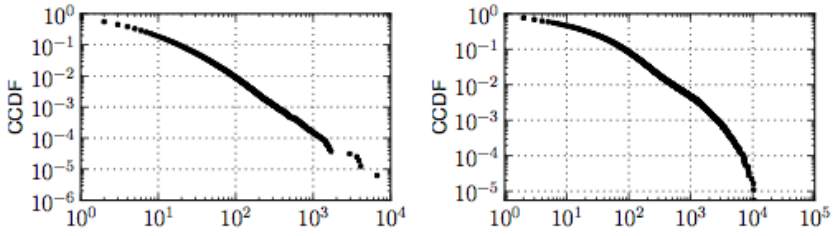

FIGURE 2.7: CCDF del numero di amici (a sx) e del numero di luoghi (a dx) per utente

2.2.4.1 Dataset utilizzato

Gowalla è un servizio di social network creato nel 2008 e permette agli utenti di aggiungere amici e condividere la propria posizione; la relazione di amici è reciproca, ovvero ogni utente deve accettare la relazione di amicizia per perchè

Mese	Utenti	Utenti attivi	Luoghi	Check-in
Maggio	252.000	148.234	958.823	7.475.401
Giugno	291.812	168.925	1,104.771	9.073.157
Luglio	325.025	189.512	1.226.847	10.537.516
Agosto	382.750	216.734	1.421.262	12.846.151

TABLE 2.1: Snapshot Gowalla tra maggio e agosto 2010

Mese	Nodi	Archi	Componente gigante	Grado medio
Maggio	109.045	476.409	102.951	8,73
Giugno	124.190	559.901	117.868	9,01
Luglio	138.387	630.045	131.711	9,10
Agosto	159.391	736.778	152.011	9,24

TABLE 2.2: Proprietà del grafo di Gowalla tra maggio e agosto 2010

sia permessa la condivisione della posizione. I ricercatori hanno scaricato quattro snapshot mensili da Gowalla, tra maggio e agosto del 2010, riassunti in 2.1 dove è possibile notare come il numero di utenti registrati e salito da circa 250 mila a circa 380 mila, risultando però invariata la percentuale di utenti attivi, ovvero coloro che hanno almeno un check-in o un amico.

In 2.2, invece, sono descritte le proprietà del grafo sociale per ogni snapshot, utilizzando solo gli utenti attivi; si nota come esista una componente gigante comprendente circa il 94% per di nodi, e il grado medio è cresciuto da 8,73 a 9,24. In 2.7 si può invece notare come sia la distribuzione del numero di amici, che la distribuzione del numero di luoghi per utente, seguano una power-law con un fattore molto elevato, in quanto solamente l'1% degli utenti possiede più di 100 amici e il 90% degli utenti ha visitano meno di 100 luoghi.

Gli utenti di un social network tendono a non creare relazioni d'amicizia in modo casuale, ma preferiscono altri utenti che sono vicini a loro, sia in termini sociali che attraverso altre dimensioni come la vicinanza geografica o la condivisione di interessi. Anche nel dataset utilizzato dai ricercatori è possibile riscontrare questo fenomeno in quando il numero di nuove relazioni di amicizia che si formano tra i diversi snapshot decresce esponenzialmente aumentando il numero di archi che

dividono due utenti; infatti la probabilità che due utenti con almeno un amico in comune, ovvero a distanza 2, diventino amici nello snapshot successivo è di 10^{-4} ed aumentando la distanza a 3 e a quattro, la probabilità decresce a 10^{-5} e 10^{-6} rispettivamente.

2.2.4.2 Definizioni

Ad ogni modo in un social network basato sulla geolocalizzazione la dimensione sociale non è l'unica da tenere in considerazione, quindi i ricercatori hanno introdotto, oltre al concetto di *friends-of-friends*, il concetto di *place-friends*, ovvero hanno ipotizzato che un utente possa instaurare una relazione d'amicizia con un altro utente anche se l'elemento che hanno in comune è un luogo geografico. A questo proposito sono stati ideati due insiemi di potenziali amici per un dato utente u_i. **Friends-of-friends**

$$S_i^t = \{(u_i, u) : u \in (\bigcup_{u_k \in \Gamma_i^t} \Gamma_k^t) \setminus \Gamma_i^t\}$$

Place-friends

$$P_i^t = \{(u_i, u) : u \in (\bigcup_{m_k \in \Theta_i^t} \Phi_k^t) \setminus \Gamma_i^t\}$$

Il primo insieme identifica tutti gli utenti che condividono almeno un amico, senza però che questo sia direttamente connesso con l'utente in questione; il secondo insieme, invece identifica tutti gli utenti che hanno effettuato almeno un check-in in uno stesso luogo, senza che esista una relazione di amicizia tra i due utenti. Per ogni snapshot è stato creato il bacino delle potenziali nuove connessioni, unendo gli insiemi dei *friends-of-friends* e dei *place-friends*, e verificando nello snapshot successivo il numero di nuovi collegamenti che provengono dal bacino creato; i ricercatori hanno osservato che più di due terzi delle nuove connessioni create tra due snapshot successivi appartenevano al bacino creato, con un contributo del 50%

da parte dell'insieme dei *friends-of-friends*, ma il 30% condividevano sia amici che check-in.

2.2.4.3 Proprietà dei luoghi

Sebbene la condivisione di alcuni luoghi favorisca la creazione di nuove relazioni di amicizia, alcuni luoghi hanno un maggior impatto, e per questo motivo i ricercatori hanno deciso di esplorare le proprietà dei luoghi per differenziare quelli che hanno maggior importanza rispetto a quelli con minor importanza. Intuitivamente un luogo in cui pochi utenti registrano la loro presenza è probabilmente importante per loro, infatti potrebbe essere un'abitazione privata o un ufficio; al contrario, un luogo che ha lo stesso numero di check-in, ma effettuati da molti più utenti, avrà un'importanza minore essendo un posto pubblico come un aeroporto o un luogo turistico.

Seguendo questa logica Scellato e gli altri hanno creato un'unità di misura di entropia per i luoghi: definendo C_k^P il numero totale di check-in che gli utenti hanno effettuato presso il luogo m_k e con $q_{i,k} = c_{i,k}/C_k^P$ la frazione di check-in che un utente u_i ha effettuato presso il luogo m_k rispetto al numero totale di check-in presso il luogo m_k, è possibile creare la distribuzione discreta di probabilità $\{q_{1,k}, ..., 1_{N,k}\}$ che descrive quanto probabile sia un check-in per un utente in un luogo. È possibile quindi definire la misura di entropia di un luogo come:

$$E_k = -\sum_{u_i \in \Phi_k} q_{i,k} \log q_{i,k}$$

ovvero i luoghi che sono visitati da molti utenti, e quindi hanno un maggior valore in termini di entropia, sono meno importanti per la creazione di nuove relazioni di amicizia, come si può notare in 2.8.

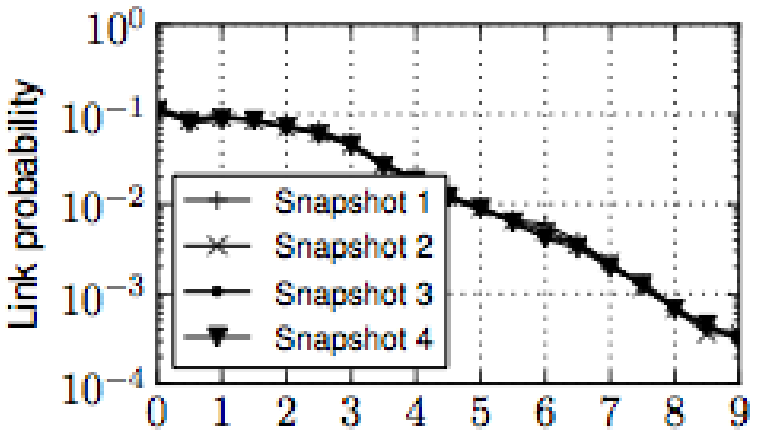

FIGURE 2.8: Probabilità di un nuova amicizia rispetto all'entropia di un luogo

Chapter 3

Setup dell'esperimento

3.1 Datasets

3.1.1 Foursquare

3.1.2 Geolife

Le traiettorie GPS presenti nel dataset sono state collezionate da *Microsoft Research Asia* per la realizzazione del progetto *Geolife*; esse contengono 182 utenti in un periodo di più di cinque anni che spazia dall'aprile del 2007 all'agosto del 2012. Una traiettoria GPS è definita nel dataset come una sequenza di punti contenenti la latitudine, la longitudine, l'altitudine ed il timestamp della registrazione.

3.2 Strumenti utilizzati

3.2.1 Python

3.2.2 Networkx

3.2.3 Google Maps API

Chapter 4

Dettaglio dell'esperimento

4.1 Raccolta dati grezzi

4.1.1 Visualizzazione dati grezzi

4.2 Creazione del grafo

4.2.1 Algoritmo basato sulla densit

4.2.2 Algoritmo basato sulla suddivisione

4.2.3 Algoritmo basato sul tempo e sullo spazio

4.2.4 Creazione del grafo aumentato

4.3 Analisi dei grafi

4.3.1 Analisi dei grafi personali

4.3.1.1 Analisi dei grafi giornalieri

4.3.1.2 Analisi dei grafi periodici

Chapter 5

Discussione dei risultati

Chapter 6

Conclusioni e sviluppi futuri

Bibliography

[1] Alex (Sandy) Pentlandb Nathan Eagle and David Lazerc. Inferring friendship network structure by using mobile phone data. *PNAS 2009*, 2009.

[2] Freeman S Freeman L, Romney A. Cognitive structure and informant accuracy. *Am Anthropol*, (89):310–325, 1987.

[3] Xing Xie Quannan Li, Yu Zheng. Mining user similarity based on location history. *ACM GIS '08*, 2008.

[4] Hongzhi Yin Yizhou Sun Bin Cui Zhiting Hu Ling Chen. Lcars: A location-content-aware recommender system. *KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 221–229, 2013.

[5] M. Naor R. Fagin, A. Lote. Optimal aggregation algorithms for middleware. *PODS*, pages 102–113, 2001.

[6] Salvatore Scellato Anastasios Noulas Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. *KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054, 2011.