Invididuazione di community di utenti per il turismo utilizzando il clustering congiuntamente al collaborative filtering

**ABSTRACT**

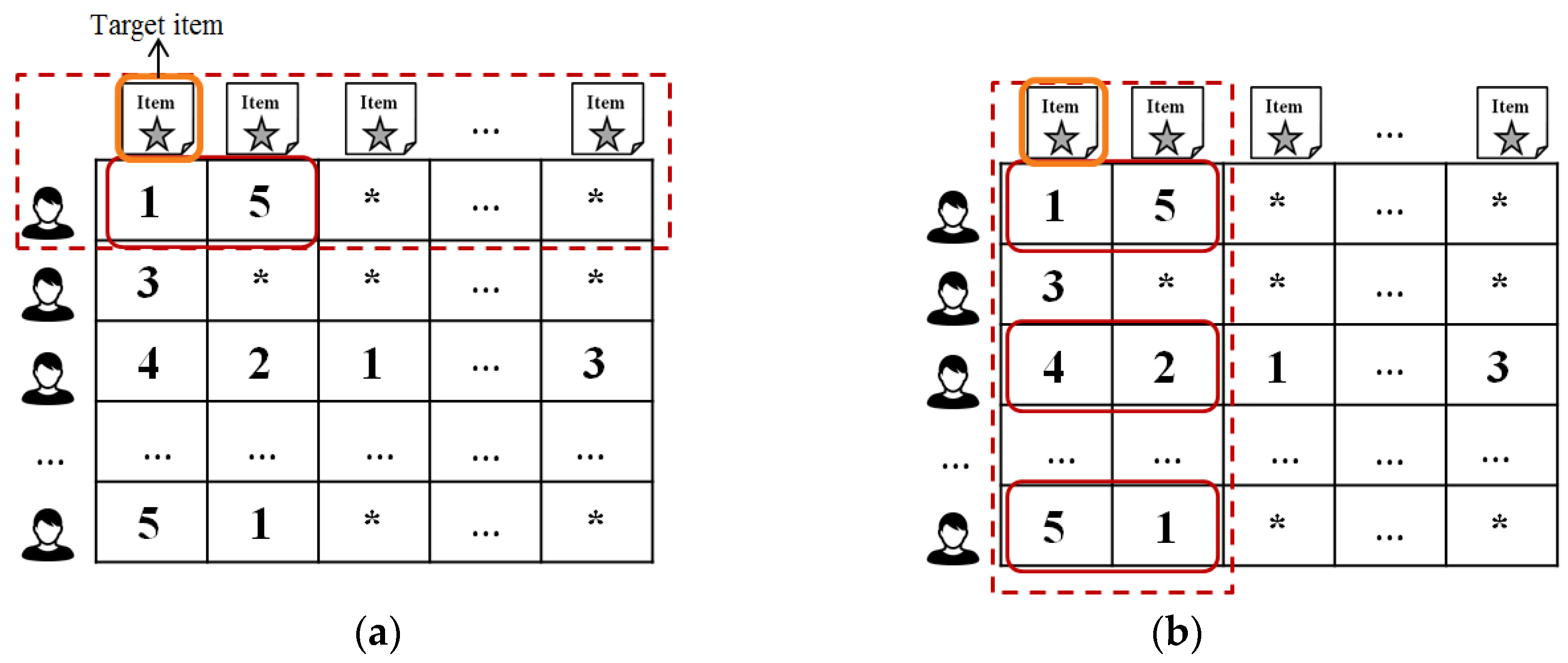
Il turismo è uno dei settori più sviluppati in Italia e uno tra quelli con maggior tasso di crescita, tuttavia molte persone decidono di non partire perché non riescono ad organizzarsi con persone nella propria cerchia di conoscenze, il sistema che proponiamo in questo articolo ha quindi lo scopo di cercare di risolvere il problema per queste persone individuando all’interno del web community di persone con un account Google con gusti simili con l’intenzione di partire cercando di sfruttare il potenziale inespresso di questa cerchia di persone, tuttavia molti utenti non esprimono valutazioni per ogni tipo di attrazione, il sistema tenta quindi di individuare delle community anche dopo aver predetto le valutazioni inespresse e confrontando i risultati con quelli ottenuti non facendo alcuna predizione. Il sistema fa uso di un dataset composto da 5455 utenti con una media di valutazione per ogni tipologia di attrazione e si compone di varie fasi: preprocessamento dei dati, analisi dei dati, individuazione e addestramento del modello predittivo per i voti inespressi, individuazione di community con e senza predizione, visualizzazione mediante grafici dei risultati con e senza predizione dei voti inespressi di ciascun utente, confronto e analisi dei risultati con entrambe le metodologie.

1. Introduzione

Molte persone scelgono di partire con degli sconosciuti sia per mancanza di organizzazione con le persone conosciute, sia per fare nuove conoscenze, tuttavia molti sistemi già presenti non sfruttano a pieno i gusti delle persone basandosi sulle votazioni già effettuate per ciascun luogo e soprattutto non sfruttano a pieno il potenziale che può avere un sistema di raccomandazione per trovare community in modo più preciso. Il Sistema si compone di varie fasi: preprocessamento dei dati eliminando utenti e tipologie di attrazioni per cui non esiste un numero sufficiente di votazioni, analisi dei dati graficandoli, normalizzandoli in quanto non tutti gli utenti esprimono votazioni allo stesso modo, riducendo la dimensionalità del problema individuando delle correlazioni tra le dimensioni (le categorie di attrazioni in questo caso) ed eliminandole, individuazione e addestramento del modello predittivo per i voti inespressi cercando una metrica che misuri bene il grado di similarità tra utenti e items, individuazione di community con e senza predizione, visualizzazione mediante grafici dei risultati con e senza predizione dei voti inespressi di ciascun utente, confronto e analisi dei risultati con entrambe le metodologie utilizzando diversi metodi per l’individuazione di community (clusters).

1. Metodo di ricerca

Un sistema di raccomandazione basato su collaborative filtering è un sistema che tenta di fare delle previsioni sulle votazioni inespresse sulla base dei k utenti o items più simili, con alla base l’idea che utenti che hanno avuto fino ad ora gusti simili su rating già avvenuti, avranno probabilmente gusti simili anche su rating futuri, ciò è alla base dei principali sistemi di raccomandazione di film, musica, etc come Amazon, Netflix etc. Ci sono due tipi di collaborative filtering, uno trovando i K utenti più simili e facendo una media pesata in base alla similarità delle loro valutazioni per l’item non valutato dall’utente per cui si sta facendo una predizione, questo però ha una serie di problemi, come la complessità computazionale, il fatto che la matrice item-user sia una matrice sparsa in quanto ogni utente vota soltanto pochi item etc, oppure trovando gli item più simili e facendo una media pesata sempre in base alla similarità delle votazioni già effettuate dall’utente per i k item più simili (fig a).



Prima di tutto è necessario rappresentare la relazione tra utenti e item sotto forma di una matrice dove le righe rappresentano gli utenti e le colonne gli item, dopo di che si sceglie una metrica per misurare la similarità tra user o tra item, è stato visto empiricamente che per la similarità tra user spesso la metrica migliore è la correlazione di Pearson, mentre tra gli item è la coseno similarità:



Dove zi è l’user/item iesimo e zj e l’user/item jesimo, ru è il rating

Mentre la predizione pesata del rating Rp dell’utente u viene effettuata come segue:

