Challenge 2014 – Sistemi Intelligenti per Internet

Francesco Maria Maglia 407842, Matteo Cannaviccio 403248

Descrizione progetto

L'obiettivo del progetto è di costruire un sistema di raccomandazione che deve essere in grado di suggerire agli utenti dei punti di interesse. Il contesto di esecuzione è quello del location based service Yelp! che fornisce informazioni riguardo punti di interesse. Lo scopo della Challenge è quello di fornire una predizione di rating (da 1 a 5), più precisa possibile, che identifichi il livello di gradimento di un punto di interesse da parte di uno specifico utente.
Il dataset fornito è composto da:

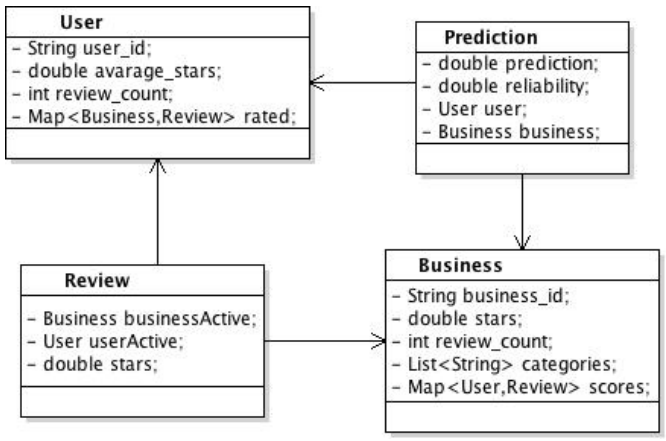
- 11.537 business
- 43.873 user
- 206.786 review

Modello di dominio

L'estrazione dei dati dal dataset consente la creazione di diversi oggetti di dominio. Sono modellati gli utenti (**User**) con i relativi dati (*id, media, review count*) e i business (**Business**) con gli stessi dati (*id, media, review count* e *categorie*). Per i Business vengono estratte le categorie ed inserite in una lista.
Gli oggetti che mettono insieme gli User e i Business sono le Review: un oggetto **Review** ha come parametri un User, un Business e un valore di rating. Inoltre, gli oggetti **Prediction** consentono, in fase di predizione, di fornire un valore di predizione di un User per un Business considerando un valore di "affidabilità" (*reliability*).

Per migliorare l'efficienza del sistema si è scelto di inserire due mappe una per gli User ed una per i Business che consentono di registrare le review di un utente e di un business.

- **Map<Business, Review>** ***rated***
  descrive tutti i rating di un utente per i diversi Business ed è salvata nell'oggetto User;
- **Map<User, Review>** ***scores***
  descrive tutti i rating di un Business effettuati dai diversi utenti ed è salvata nel Business.



La creazione delle mappe (onerosa) viene effettuata in fase di estrazione (e quindi di creazione) consentendo così il rapido accesso ai dati in momenti successivi di esecuzione.

Tecniche utilizzate

UserBased CF

E' il primo approccio che abbiamo eseguito e si basa sull'idea che utenti con gusti simili in passato mantengano gli stessi gusti anche in futuro. L'obiettivo è di calcolare la similarità tra gli user, trovare gli N user più simili ed in base ai rating di questi calcolare una predizione per l'user in esame sul nuovo business.

Similarità
Il primo passaggio dell'algoritmo consiste nel trovare gli user simili all'user attivo di cui predire il rating.

```
calculateUserSimilarity(activeUser)
	for each currentUser that has rated the activeBusiness
		businessCoRated = list of business co-rated between activeUser and currentUser
		if (businessCoRated.size() >= NUMBER_CORATED_USER)
			sim = calculateSimilarity(activeUser,currentUser)
		if (sim > THRESHOLD_SIM_USER)
			similarityUser.put(currentUser, sim);
	return similarityUser;
```

Il controllo degli user simili viene effettuato tra l'user attivo e gli altri user che hanno espresso un rating per il business attivo in modo da consentire la successiva predizione. Se il numero di business co-rated tra l'user attivo ed un altro è maggiore di una soglia (`NUMBER_CORATED_USER`) viene calcolata la similarità. Se il valore di similarità supera una soglia minima (`THRESHOLD_SIM_USER`) viene aggiunta alla mappa `similarityUser` che verrà utilizzata per la predizione. Aumentando le soglie, diminuiscono le possibilità di poter effettuare la predizione ma ne aumenta la precisione. Se la

mappa non viene popolata il rating non è predicibile tramite questa tecnica e si ricorre ad un diverso approccio che sarà spiegato in seguito.
La similarità fra due user è calcolata utilizzando la formula di Pearson, che è espressa con un valore compreso tra [-1,1], indicante quanto due utenti sono simili sulla base dei rating espressi per gli stessi business.

Predizione

$$pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)}$$

La predizione di rating è calcolata utilizzando la formula in figura. Per ogni rating espresso dall'utente simile viene sottratta la sua media in modo da normalizzare i risultati al comportamento dell'utente.
L'implementazione scelta permette di creare una predizione che sia composta dal valore predetto più un valore di affidabilità. Questo valore, che sarà utilizzato per combinare le due tecniche, è calcolato come la similarità media degli user simili:

$$\frac{\sum_{b \in N} sim\,(a,b)}{|N|}$$

ItemBased CF

In questo secondo caso l'obiettivo è stato di calcolare le predizioni utilizzando la similarità tra i business anziché la similarità tra gli user. Vengono calcolati gli N business più simili sulla base dei rating espressi dagli utenti e su questi viene effettuata la predizione considerando i rating che gli user simili hanno espresso per il business in questione.
Le similarità sono migliorate leggermente rispetto agli user grazie all'utilizzo di un valore di similarità offerto dalle categorie.

Similarità

```
calculateBusinessSimilarity(activeBusiness)
	for each currentBusiness that has rated by activeUser
		userCoRated = list of user which have rated both activeBusiness and currentBusiness
		if (userCoRated.size() >= NUMBER_CORATED_BUSINESS)
			sim = calculateSimilarity(activeBusiness, currentBusiness)
		if (simContent() > 0.4)
			sim = sim * 1.1;
		if (sim > THRESHOLD_SIM_BUSINESS)
			similarityBusiness.put(currentBusiness, sim);
	return similarityBusiness;
```

Si tratta dell'implementazione inversa alla precedente in cui viene creata una mappa di similarità dei Business (`similarityBusiness`) utilizzata per la predizione.
Per calcolare la similarità viene utilizzata la formula di Adjusted Cosine Similarity, espressa tramite un valore compreso tra [-1,1]. Il valore ottenuto può essere migliorato tramite la similarità espressa dalle categorie.

Similarità Categorie

Si tratta di una misura di similarità calcolata sui dati a disposizione per i Business. Il calcolo è effettuato tramite la formula di Jaccard Similarity per le categorie dei punti di interesse. Il rapporto di similarità è espresso dal numero di categorie in comune diviso il numero di categorie totali. Per esempio, se il business i ha categorie [Restaurant, Cafè] e il business j ha categorie [Cafè, French] allora il valore di similarità sarà 1/3.

Predizione

$$pred(u,p) = \frac{\sum_{i \in ratedItems(u)} sim(i,p) * r_{u,i}}{\sum_{i \in ratedItems(u)} sim(i,p)}$$

La predizione del rating è calcolata tramite la formula espressa in figura. Ogni rating espresso dell'user attivo all'item i simile viene moltiplicato e normalizzato per il valore di similarità. Anche in questo caso la predizione fornita comprenderà anche un valore di affidabilità calcolato come nel caso precedente.

NaiveBased

Quando non è possibile calcolare la predizione vengono fatte considerazioni di carattere generale riguardo il dataset fornito. In particolare, la predizione viene calcolata sommando tre valori: il primo esprime la media generale per tutti gli user, il secondo la differenza tra la media dell'user specifico e la media generale (user-bias) ed il terzo la differenza tra la media del business specifico e la media generale (business-bias).

$$b_{ui} = \mu + \beta_u + \beta_i$$

ContentBased (CategoryPrediction)

Un valore semplice di predizione è effettuato prendendo in considerazione la media dei rating espressi dall'utente per punti di interesse molto simili al business da predire. Due business sono simili fra loro quando hanno almeno due categorie in comune. In questo caso viene calcolata la media dei rating espressa e tenuta in considerazione durante la fase di combinazione delle predizioni ottenute da ciascun predittore.

Combiner

La predizione finale viene calcolata dal Combiner, il quale accorpa le predizioni in base ai valori di affidabilità di ognuna. La predizione globale è ottenuta dalla seguente formula dove P_x è la predizione del predittore x e R_x è la relativa affidabilità.

$$\frac{P_u \cdot R_u + P_b \cdot R_b + P_c \cdot R_c}{R_u + R_b + R_c}$$

Quando non è possibile calcolare la CategoryPrediction il valore di affidabilità sarà 0 in modo da annullare il fattore relativo alle categorie.

Test effettuati

In base ai test effettuati si è visto che, se vengono fissate soglie che non compromettono troppo l'efficienza, il numero di predizioni che è possibile effettuare con le tecniche di UserBasedCF ed ItemBasedCF è molto basso. Questo avviene perché il dataset è sparso e, anche se è possibile calcolare buone similarità, risulta impossibile applicare la formula per la predizione per le seguenti condizioni:

- Nel caso di ItemBasedCF l'user attivo deve aver espresso una valutazione per tutti i business che sono risultati simili al business da valutare.
- Nel caso di UserBasedCF ogni user risultato simile deve aver espresso una valutazione per il business attivo.

Durante lo studio del modello sono stati eseguiti alcuni esperimenti, fra i quali, quello di sostituire con un *default voting* (avg user/business o avg dataset) il voto mancante (nei casi spiegati sopra). Questo esperimento però non ha portato a grandi miglioramenti dal punto di vista dell'efficacia ma bensì un peggioramento dal punto di vista dell'efficienza. (AVG MAE: 0.77, AVG TIME: 20 p/s)

Per migliorare quindi l'efficienza si è stabilito di effettuare il controllo sui simili (item e user) partendo dai rating che erano stati espressi dall'user o sul business a seconda del caso. In questo modo si ottiene un notevole incremento dell'efficienza. (AVG MAE: 0.74, AVG TIME: 200 p/s)

I test effettuati si sono concentrati, inoltre, sulle migliori soglie con cui testare il modello. Il numero di item/user co-rated è stato fissato a 3, un numero che consente di effettuare predizioni ma nello stesso tempo offre un minimo di informazione utile. La soglia per la similarità è stata fissata a 0.6.

```
USER ---> NAIF: 3296 USERbased: 4684
ITEM ---> NAIF: 3677 ITEMbased: 4303
--------------------------------------------------------------------------------
*** TOTALI *** : 7980
CF: 3651, 46 % *** MIXED CF & NAIF: 1685, 21 % *** PURE NAIF: 2644, 33 %
--------------------------------------------------------------------------------
*** ERRORI *** : 58 % del totale delle predizioni
--------------------------------------------------------------------------------
ERRORI di 1: 3667
ERRORI CF: 1716, 47 % di quelli che vengono predetti CF
ERRORI MIXED CF & NAIF: 738, 44 % di quelli che vengono predetti Misti
ERRORI PURE NAIF: 1213, 46 % di quelli che vengono predetti NAIF
--------------------------------------------------------------------------------
Altri ERRORI: 954
ERRORI CF: 404, 11 % di quelli che vengono predetti CF
ERRORI MIXED CF & NAIF: 738, 16 % di quelli che vengono predetti Misti
ERRORI PURE NAIF: 283, 11 % di quelli che vengono predetti NAIF
END
MAE: 0.7160401002506266
```