Università degli Studi di Bari
Dipartimento di Informatica

LACAM
Knowledge Discovery and Data Engineering

Analisi di Social Media per la Scoperta di Pattern di Interazione

Tesi di Laurea in Metodi Avanzati di Programmazione

Relatori:
dr. Corrado Loglisci
prof. Donato Malerba

Correlatore:
dott. Angelo Impedovo

Laureando:
Donato Meoli

26 Aprile 2018

Obiettivo

- realizzare strumenti computazionali in grado di modellare comunità online
- sintetizzare metodi di analisi che siano in grado di scoprire le interazioni tra i partecipanti

Motivazione

- recente interesse nel monitoraggio dei partecipanti di piattaforme web attraverso le loro relazioni rispetto ad interessi comuni (es. realizzazione di progetti collaborativi)
- indirizzare questo tipo di informazioni per studi sociali rispetto a tecnologie web o social media

Problematiche

- una comunità rappresenta un dominio che evolve nel tempo i cui cambiamenti possono essere molteplici e possono riguardare fattori interni o esterni
- web ricco di User Generated Content (UGC) di tipo non strutturato (es. contenuti di Twitter, StackOverflow, Wikipedia, Reddit, ecc.)
- ambiguità relativa alle informazioni prodotte in linguaggio naturale
- contenuti sintetici, soggetti ad errori ortografici e a linguaggio web (es. lol, asd, ecc.)

Soluzione Computazionale

Scelta del Modello Computazionale

- relazioni non identificate a priori per via delle relazioni sociali che intercorrono tra questi (es. “follow” o “amicizia”)
- previste relazioni tra gli utenti che sono supportate dagli strumenti tecnologici in uso
- rappresentazione a grafo della comunità in cui i nodi rappresentano gli utenti e gli archi corrispondono alle relazioni che intercorrono tra di loro

Costruzione del Grafo (1)

Archi basati su tag sociali:

» **COMMENT_TO** l'utente b commenta il post dell'utente a

» **REPLY_TO** l'utente c risponde al commento dell'utente b sotto il post dell'utente a

» **MENTION_TO** l'utente s menziona l'utente m

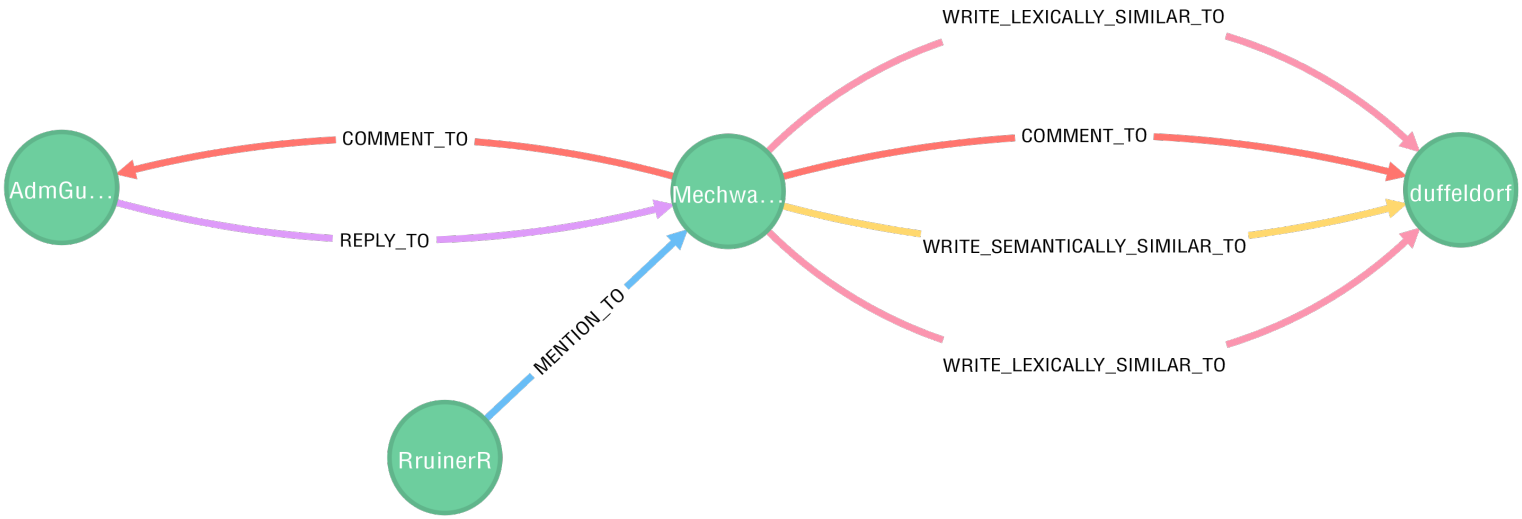
Archi basati sul contenuto:

» **WRITE_LEXICALLY_SIMILAR_TO** dovuta alla similarità lessicale tra i messaggi contenuti nei post di due utenti coinvolti nella stessa discussione

» **WRITE_SEMANTICALLY_SIMILAR_TO** dovuta alla similarità semantica tra i messaggi contenuti nei post di due utenti coinvolti nella stessa discussione

Negli archi basati sul contenuto la direzionalità è determinata dal timestamp.

Costruzione del Grafo (2)



Archi Basati sul Contenuto

» Similarità Lessicale **Lexical Match Algorithm (LMA)** [1], basato sul Vector Space Model (VSM), considera i messaggi che si presuppone possano essere simili

» Similarità Semantica Similarità di **Lin** [2] pesata con i TF-IDF [3] per far si che i termini più frequenti, e quindi semanticamente meno discriminanti, abbiano un'incidenza minore sul valore di similarità finale

---

[1] Fu, Tianjun, Ahmed Abbasi, and Hsinchun Chen. "A hybrid approach to web forum interactional coherence analysis." Journal of the Association for Information Science and Technology 59.8 (2008): 1195-1209.

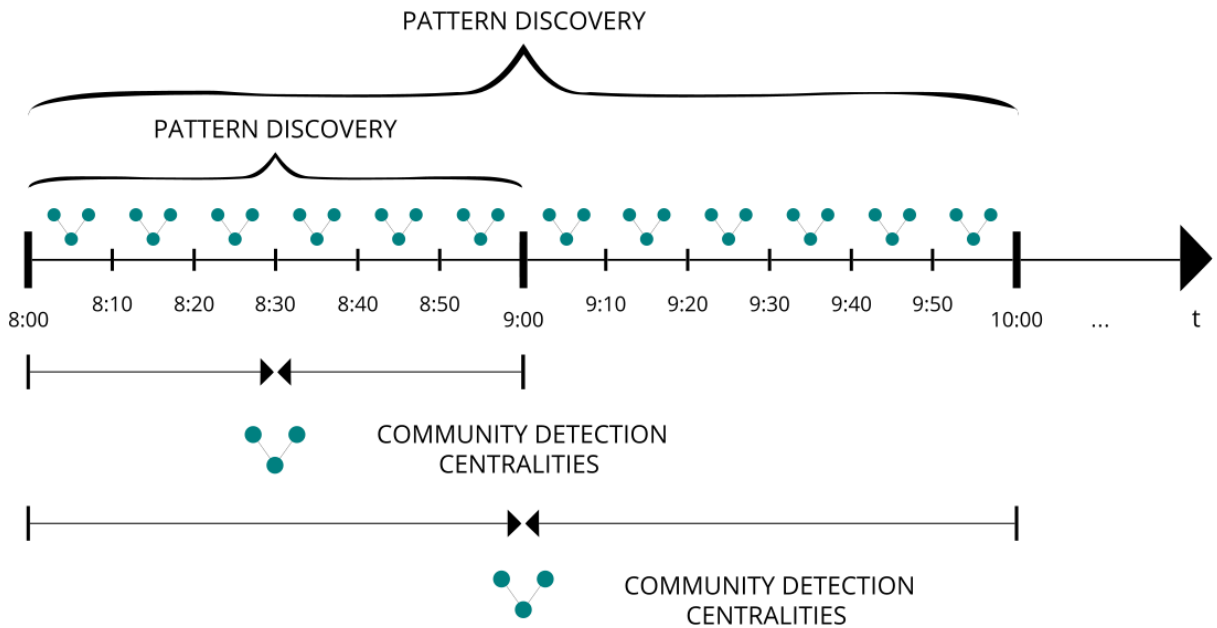
[2] Lin, Dekang. "An information-theoretic definition of similarity." Icml. Vol. 98. No. 1998. 1998.

[3] Term Frequency - Inverse Document Frequency

Scoperta di Interazioni

Analisi di Grafi Temporali

- interazioni scoperte in base a densità e frequenza
- analisi del grafo per time-point al fine di scoprire interazioni tra gli utenti basate su frequenza
- analisi di time-windows di ampiezza incrementale in forma di grafi *cumulativi*, rispetto a quelli di ogni time-point presente nella stessa, al fine di scoprire il ruolo degli utenti e le interazioni tra di essi basate su densità



Ruolo dell'Utente: Indicatori di Centralità

Indicatori nodali definiti in teoria dei grafi:

» **Centralità "Degree"** numero degli archi in cui un nodo è coinvolto facendo distinzione tra quelli in entrata ("in-degree") e quelli in uscita ("out-degree") al fine di mettere in evidenza il ruolo degli utenti come attrattori o mittenti

» **Centralità "Betweenness"** basata sul calcolo dei cammini minimi, utile per trovare i nodi che fungono da tramite da una parte all'altra del grafo

» **Centralità PageRank** determina una stima dell'importanza del nodo all'interno del grafo

Interazioni Basate su Densità: Scoperta di Comunità

- un grafo ha una struttura di comunità se i suoi nodi possono essere raggruppati in modo tale che quelli di ogni comunità siano *densamente* connessi internamente ed abbiano connessioni più *sparse* tra di loro
- metodo di **Louvain** [4] basato sulla massimizzazione della *modularità* di una partizione: funzione euristica che misura la bontà di una comunità ossia la densità dei collegamenti all'interno di essa rispetto ai collegamenti tra le comunità

[4] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.

Interazioni Basate su Frequenza: Scoperta di Sottografi Frequenti

- indagare la presenza di sottografi frequenti rispetto ad una time-window, a partire dai sottografi dei singoli time-point contenuti in essa
- algoritmo **ECLAT** [5] basato su intersezioni insiemistiche e proprità connesse per la scoperta di sottografi frequenti

[5] Zaki, Mohammed Javeed, et al. “New Algorithms for Fast Discovery of Association Rules.” KDD. Vol. 97. 1997.

Esperimenti

Dataset: Reddit, Novembre 2017

Filtraggio preventivo per garantire un dataset formato da utenti particolarmente attivi e da submissions e commenti semanticamente significativi.

- **» Submissions** per cui il numero di commenti è maggiore di 8 e la lunghezza del selftext è maggiore di 168 caratteri
- **» Commenti** per cui la lunghezza del body è maggiore di 170 caratteri
- **» Redditors** che hanno scritto più di 20 submissions e commenti
- **» Subreddits** i 20 migliori tra quelli per cui il numero di submissions che vi appartengono, in seguito al filtraggio definito dai criteri precedenti, risulti maggiore di 77

Setting Sperimentale e Risultati Quantitativi

» **time-windows** da 1 giorno

» **time-points** da 3 ore

	sottografi		comunitá
	2 utenti	3 utenti	
μ	1.476,83	3.126,76	87,97
σ	2.919,28	6.391,88	150,60

Risultati Qualitativi (1)

Redditor `"D2TournamentThreads"`, prima time-window analizzata (01-02/11):

```
{
  "name": "D2TournamentThreads",
  "inDegree": 4,
  "outDegree": 2,
  "betweenness": 4.0,
  "pageRank": 0.39862500000000006,
  "frequentSubgraphs": [
    "{e(D2TournamentThreads, MENTION_TO, 3947977282),
      e(D2TournamentThreads, MENTION_TO, 5874306284)}"
  ]
}
```

Risultati Qualitativi (2)

Redditor "D2TournamentThreads", ultima time-window analizzata (29-30/11):

```
{
  "name": "D2TournamentThreads",
  "inDegree": 497,
  "outDegree": 21,
  "betweenness": 80639.51145093024,
  "pageRank": 46.710466999999994,
  "frequentSubgraphs": [
    "{e(D2TournamentThreads, MENTION_TO, 1107888450),
      e(D2TournamentThreads, MENTION_TO, 5695452105)}",
    "{e(D2TournamentThreads, COMMENT_TO, 1107888450),
      e(D2TournamentThreads, COMMENT_TO, 5874306284)}",
    "{e(D2TournamentThreads, REPLY_TO, 1107888450),
      e(D2TournamentThreads, REPLY_TO, 5695452105),
      e(D2TournamentThreads, REPLY_TO, 5874306284)}"
      . . .
  ]
}
```

Conclusioni

- studio di comunitá online a partire dalla messaggistica intercorsa tra i partecipanti
- soluzione computazionale basata su un modello relazionale e tecniche di data mining

Sviluppi Futuri

- sperimentazione su altri social media
- aggiunta di archi basati sulla similaritá emozionale dei partecipanti in relazione al contenuto dei messaggi scambiati
- analisi delle evoluzioni delle interazioni tra i partecipanti

Grazie per l’Attenzione