UNIVERSITÀ
DELLA CALABRIA

Dipartimento di Ingegneria Informatica, Modellistica,
Elettronica e Sistemistica

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Analisi di social media per la scoperta automatica delle traiettorie degli utenti

Relatori:
Ing. *Fabrizio Marozzo*
Ing. *Loris Belcastro*

Candidato:
Emanuele Perrella
Matricola 195276

Anno Accademico 2019/2020

Indice

Introduzione **5**

1 Social Media nell'era dei Big Data **7**
1.1 Social Media Analytics 7
1.2 Evoluzione del Web 11
1.3 Dall'evoluzione del web ai social media 14
1.4 Big Data 16
1.4.1 Sfide del mondo Big Data 18
1.4.2 Sorgenti di Big Data 19
1.4.3 Sistemi di elaborazione 21

2 Apache Spark **26**
2.1 Introduzione al modello grafo aciclico 26
2.2 Spark 27
2.2.1 Moduli principali 30
2.2.2 Spark Core 31
2.2.3 Spark Streaming 31
2.2.4 Spark SQL 32
2.2.5 Apache Spark MLlib 33

3 Metodologia **35**
3.1 Keywords Extraction 36
3.1.1 Descrizione del problema 37
3.1.2 Principali metodologie 39
3.1.3 Valutazione generale 41
3.1.4 Metodologia proposta 41
3.1.5 Post elaborazione 42
3.2 RoI Extraction 44
3.2.1 Principali tecniche 45
3.2.2 G-RoI 46
3.2.3 Clustering basato su densità 47

3.2.4 DBSCAN 48
3.2.5 Dominant Set 50
3.2.6 DSets-DBSCAN 53
3.2.7 Una metodologia basata su DBSCAN per l'estrazione delle RoI 54
3.3 Trajectory Extraction 56
3.3.1 Regole associative 57
3.3.2 Approccio algoritmico 59
3.3.3 Apriori 60
3.3.4 FP-Growth 62
3.3.5 Sequential pattern mining e PrefixSpan 62

4 Risultati sperimentali **64**
4.1 Accuratezza nell'estrazione delle keywords 64
4.1.1 Stima delle metriche principali 66
4.2 Accuratezza nell'estrazione delle RoI 73
4.3 Accuratezza nella definizione dei pattern di mobilità 74
4.3.1 Analisi dei frequent itemset 74
4.3.2 Analisi delle traiettorie 78
4.4 Scalabilità 81
4.4.1 Estrazione delle keywords 82
4.4.2 Estrazione delle RoI 85
4.4.3 Estrazione dei pattern di mobilità 89

Conclusioni e sviluppi futuri **93**

Bibliografia **96**

Elenco delle figure

1.1 Differenze principali tra Web 2.0 e 3.0 13
1.2 Esempio grafico di una rete sociale fisica 14
1.3 Modello delle 5 V per i big data [12] 17
1.4 Abitudini degli utenti su Internet [13] 23
1.5 Siti web più visitati nel 2019 [13] 24

2.1 Struttura di Spark. 29
2.2 Workflow relativo all'esecuzione di un'applicazione su Spark. . . . 30
2.3 Struttura di Spark. 30
2.4 Struttura generale di uno stream processing. 32
2.5 Sequenza di Transformers/Estimators in Pipeline. 34

3.1 Workflow implementato. 36
3.2 Esempio di workflow per l'estrazione di parole chiavi. 37
3.3 Pulizia dei dati in maniera esplicativa. 38
3.4 Principali metriche di valutazione. 41
3.5 Con **minPts = 4** i punti rossi sono **core points**, perché l'area che circonda questi in un raggio /*epsilon* contiene 4 punti (compreso il punto stesso). Poiché sono tutti raggiungibili l'uno dall'altro, essi formano un unico cluster. I punti B e C non sono core points, ma sono raggiungibili da A (attraverso altri punti centrali) e quindi appartengono anch'essi al cluster. Il punto N invece è un noise point. 49
3.6 Dominating sets (i vertici di colore rosso). 51
3.7 Analisi del carrello della spesa. 57
3.8 Esempio di database transazionale. 58
3.9 Esempio sul funzionamento algoritmico di apriori. 61
3.10 Esempio di database sequenziale. 63

4.1 Confronto mediante WordCloud relativo ai due approcci implementati. 65
4.2 Andamento della precisione media per metodologia. 69
4.3 Andamento della recall media per metodologia. 70
4.4 Precisione, copertura e F1 a confronto per l'area relativa a Roma. 70

4.5 Un utile confronto per la metrica di valutazione globale. 71
4.6 Andamento della precisione media su un campione di celle. 72
4.7 Andamento della recall media su un campione di celle. 72
4.8 Precisione, copertura e F1 relative all'analisi su Parigi. 73
4.9 Precisione ottenuta sui frequent itemset. 77
4.10 Recall ottenuta sui frequent itemset. 77
4.11 Precisione,recall e F1 a confronto per l'analisi di pattern frequenti su Roma. 78
4.12 Precisione media delle metodologie su un campione di cinquanta utenti. 80
4.13 Recall media delle metodologie su un campione di cinquanta utenti. 80
4.14 Utile confronto tra F1, Recall e Precisione. 81
4.15 Andamento del tempo di esecuzione totale per l'estrazione delle keywords utilizzando TF-IDF. 83
4.16 Andamento del tempo di esecuzione totale per l'estrazione delle keywords utilizzando la metodologia proposta. 84
4.17 Andamento dello speedup ottenuto utilizzando TF-IDF. 85
4.18 Andamento dello speedup ottenuto utilizzando SMA4ADT. 85
4.19 Andamento del turnaround time relativo al processo di RoI Extraction con un numero di punti fissato pari a 5000 87
4.20 Andamento del turnaround time relativo al processo di RoI Extraction con un numero di punti fissato pari a 1000 87
4.21 Andamento dello speedup ottenuto relativo al processo di RoI Extraction con un numero di punti fissato pari a 5000 88
4.22 Andamento dello speedup ottenuto relativo al processo di RoI Extraction con un numero di punti fissato pari a 1000 89
4.23 Andamento del tempo di esecuzione totale per l'estrazione dei pattern di mobilità utilizzando FP-Growth. 90
4.24 Andamento del tempo di esecuzione totale per l'estrazione dei pattern di mobilità utilizzando PrefixSpan. 91
4.25 Andamento dello speedup ottenuto utilizzando FP-Growth. 92
4.26 Andamento dello speedup ottenuto utilizzando PrefixSpan. 92

Elenco delle tabelle

4.1 Accuratezza nell'estrazione delle regioni di interesse. 74
4.2 Esempio del procedimento di conversione in formato transazionale 76
4.3 Turnaround time relativo al processo di estrazione delle keywords. 83
4.4 Speedup relativo al processo di estrazione delle keywords 84
4.5 Turnaround time relativo al processo di RoI Extraction 88
4.6 Speedup relativo al processo di RoI Extraction 89
4.7 Turnaround time del processo di Trajectory Extraction. 91
4.8 Speedup del processo di Trajectory Extraction. 91

Introduzione

Lo sviluppo esponenziale del Web, verificatosi negli ultimi anni, ha visto l'affermarsi di una nuova era sociale, quella dei Social Media.

L'avvento dei social media ha rivoluzionato gli spazi di interazione quotidiana tradizionali, presentandosi come un'opportunità in più che ha inciso molto nella sfera relazionale delle persone. I social media, infatti, hanno avuto una notevole influenza sulla società, modificando abitudini tradizionali e strategie aziendali. Il gigantesco volume di dati prodotto dagli utenti dei social media può essere utilizzato per descrivere eventuali dinamiche e comportamenti nella vita delle persone; l'obiettivo della Social Media Analysis è l'estrazione di informazione utile in diversi contesti applicativi analizzando grandi volumi di dati provenienti da social media.

In molti social media l'analisi di dati geolocalizzati consente di rilevare luoghi geografici rilevanti, denominati Point of Interest (PoI); considerando che un punto di interesse è generalmente identificato dalle coordinate di un singolo punto risulta difficile abbinarlo con le traiettorie degli utenti. A tal proposito è necessario estendere l'oggetto di ricerca ad un area bidimensionale, la regione di interesse (Region of Interest, o in breve RoI) che delimita i confini di un determinato punto di interesse; proprio per questo motivo si parla di *RoI mining* per descrivere tecniche di data mining utilizzate per descrivere il corretto partizionamento dei dati geografici in base ai criteri di interesse.

In questo lavoro di tesi viene proposta una nuova metodologia, denominata **SMA4ADT** (**S**ocial **M**edia **A**nalysis for **A**utomatic **D**etection of user **T**rajectories), che utilizza le informazioni contenute nei dati provenienti dai social media per individuare, con un'elevata accuratezza, le **RoI** e **la traiettorie utente**. Essa è caratterizzata da tre fasi principali:

1. l'estrazione automatica delle keywords, ovvero delle parole chiave, che identificano i luoghi di interesse di cui si vuole calcolare la RoI. Queste chiavi sono usate per raggruppare i dati in base al luogo di riferimento, poichè

ciò consente di aumentare il livello di parallelismo e la scalabilità dei passi successivi del workflow.

2. l'estrazione delle RoI utilizzando un approccio automatico di clustering parallelo, che, partendo dai dati raggruppati dei social media, sfrutta algoritmi di clustering per identificare le RoI in modo efficiente.

3. l'estrazione delle traiettorie generate dagli utenti per scoprire comportamenti e modelli di mobilità delle persone analizzando gli elementi dei social media geotaggati.

Sono stati effettuati diversi esperimenti per stimare il grado di accuratezza e la scalabilità della metodologia proposta su più nodi computazionali; il dataset utilizzato è stato costruito a partire dal social media Flickr su due aree precise, Roma e Parigi considerando nello specifico 21 punti di interesse per area. In particolare, i risultati ottenuti dimostrano che la tecnica proposta garantisce la migliore accuratezza in termini di chiave identificate, definizione delle RoI, e rilevamento delle traiettorie rispetto alle principali tecniche esistenti.

La metodologia proposta ha fornito risultati piuttosto soddisfacenti in tutte le fasi di cui essa si compone; per quanto riguarda l'estrazione delle keywords, i livelli di accuratezza raggiunti dimostrano che mediamente si ottengono keywords più significative per l'area rilevata. Con la fase di estrazione delle regioni di interesse, invece, i livelli di accuratezza raggiunti sono superiori alle altre metodologie presenti in letteratura e in linea con quelli ottenuti con la tecnica **G-RoI**. L'accuratezza delle RoI, infine, è avvalorata anche dalla maggiore precisione ottenuta nella fase di rilevamento delle traiettorie utente, derivate come sequenze di spostamento tra RoI.

Per quanto riguarda l'organizzazione dell'elaborato, nel primo capitolo vengono discussi i principali concetti della Social Media Analytics, dalla sua definizione, fino alla descrizione delle varie fasi di cui si compone. Nel capitolo, si affronta anche il problema dell'analisi dei **Big Data**, evidenziandone le principali sfide e le possibili soluzioni. Nel secondo capitolo viene introdotto il modello di programmazione basato su **Grafo orientato aciclico** ed uno dei framework più utilizzati **Apache Spark** descrivendone funzionamento e principali componenti. Il terzo capitolo descrive nel dettaglio la metodologia proposta e implementata, descrivendo singolarmente ogni task di cui essa si compone. Il quarto e ultimo capitolo, infine, descrive invece i risultati ottenuti in termini di accuratezza e scalabilità del sistema.

Bibliografia

[1] Gohar F Khan. *Seven Layers of Social Media Analytics: Mining Business Insights from Social Media Text, Actions, Networks, Hyperlinks, Apps, Search Engines, and Location Data*. Gohar Feroz Khan, 2015.

[2] Wikipedia. *The Social media analytics*. 2017. URL: https://en.wikipedia.org/wiki/Social_media_analytics.

[3] Loris Belcastro, Fabrizio Marozzo e Domenico Talia. "Programming models and systems for Big Data analysis". In: *International Journal of Parallel, Emergent and Distributed Systems* 34.6 (2019), pp. 632–652.

[4] P Blackshaw e M Nazzaro. *Consumer-Generated Media (CGM) 101: Word-of-mouth in the age of the Web-fortified consumer. Retrieved July 25, 2008*. 2004.

[5] Bruce W Dearstyne. "Blogs, mashups, & wikis: Oh, my!" In: *Information Management* 41.4 (2007), p. 25.

[6] Jan H Kietzmann et al. "Social media? Get serious! Understanding the functional building blocks of social media". In: *Business horizons* 54.3 (2011), pp. 241–251.

[7] Tim Berners-Lee. *Versione italiana dell'intervento di Tim Berners-Lee*. 2008. URL: https://www.globalwebindex.com/ (visitato il 27/11/2008).

[8] J.A. Barnes. *Rete Sociale -Social Network*. Angeli, 1972.

[9] Wikipedia. *Versione italiana dell'intervento di Tim Berners-Lee*. 2008. URL: https://it.wikipedia.org/wiki/Rete_sociale (visitato il 27/11/2008).

[10] Loris Belcastro et al. "Big data analysis on clouds". In: *Handbook of big data technologies*. Springer, 2017, pp. 101–142.

[11] Min Chen, Shiwen Mao e Yunhao Liu. "Big data: A survey". In: *Mobile networks and applications* 19.2 (2014), pp. 171–209.

[12] Anushree Subramaniam. *Edureka*. 2009. URL: https://www.edureka.co/blog/what-is-big-data/ (visitato il 11/03/2020).

[13] We Are Social Inc. *We are social*. 2008. URL: https://wearesocial.com/ (visitato il 07/03/2020).

[14] GlobalWebIndex Inc. *GlobalWebIndex*. 2009. URL: https://www.globalwebindex.com/ (visitato il 07/03/2020).

[15] Gema Bello-Orgaz, Jason J Jung e David Camacho. "Social big data: Recent achievements and new challenges". In: *Information Fusion* 28 (2016), pp. 45–59.

[16] Ekaterina Olshannikova et al. "Conceptualizing big social data". In: *Journal of Big Data* 4.1 (2017), p. 3.

[17] Michael J. Franklin Scott Shenker Ion Stoica Matei Zaharia Mosharaf Chowdhury. "Spark: Cluster Computing with Working Sets". In: *2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'10* (2010).

[18] Xing Xie et al Yu Zheng Lizhu Zhang. "Mining Interesting Locations and Travel Sequences from GPS Trajectories". In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. New York, NY, USA, 2009, 791–800. URL: http://doi.acm.org/10.1145/1526709.1526816.

[19] David J. Crandall et al. "Mapping the World's Photos". In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain: ACM, 2009, pp. 761–770. ISBN: 978-1-60558-487-4.

[20] Yizong Cheng. "Mean shift, mode seeking, and clustering". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 17.8 (1995), pp. 790–799. ISSN: 0162-8828.

[21] Yan-Tao Zheng, Zheng-Jun Zha e Tat-Seng Chua. "Mining travel patterns from geotagged photos". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), p. 56.

[22] Albino Altomare et al. "Trajectory Pattern Mining for Urban Computing in the Cloud". In: *IEEE Transactions on Parallel and Distributed Systems* 28.2 (2017), pp. 586–599. DOI: 10.1109/TPDS.2016.2565480.

[23] Slava Kisilevich, Florian Mansmann e Daniel Keim. "P-DBSCAN: A Density Based Clustering Algorithm for Exploration and Analysis of Attractive Areas Using Collections of Geo-tagged Photos". In: *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. COM.Geo '10. Washington, D.C., USA: ACM, 2010, 38:1–38:4. ISBN: 978-1-4503-0031-5.

[24] Zhijun Yin et al. “Diversified Trajectory Pattern Ranking in Geo-tagged Social Media”. In: SDM. SIAM. 2011, pp. 980–991.

[25] Laura Ferrari et al. “Extracting Urban Patterns from Location-based Social Networks”. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. LBSN ’11. Chicago, Illinois: ACM, 2011, pp. 9–16. ISBN: 978-1-4503-1033-8.

[26] Fosca Giannotti et al. “Trajectory Pattern Mining”. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’07. San Jose, California, USA: ACM, 2007, pp. 330–339. ISBN: 978-1-59593-609-7.

[27] Eugenio Cesario et al. “Analyzing social media data to discover mobility patterns at EXPO 2015: Methodology and results”. In: 2016 International Conference on High Performance Computing and Simulation, HPCS 2016. 2016, pp. 230–237. DOI: 10.1109/HPCSim.2016.7568340.

[28] Jieming Shi et al. “Density-based Place Clustering in Geo-social Networks”. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. Snowbird, Utah, USA: ACM, 2014, pp. 99–110. ISBN: 978-1-4503-2376-5.

[29] C. Bradford Barber, David P. Dobkin e Hannu Huhdanpaa. “The Quickhull Algorithm for Convex Hulls”. In: *ACM Trans. Math. Softw.* 22.4 (dic. 1996), pp. 469–483. ISSN: 0098-3500.

[30] Martin Ester et al. “A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. Portland, Oregon: AAAI Press, 1996, pp. 226–231. URL: http://dl.acm.org/citation.cfm?id=3001460.3001507.

[31] M. Pavan e M. Pelillo. “Dominant Sets and Pairwise Clustering”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.1 (2007), 167–172.

[32] Erich Schubert et al. “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”. In: *ACM Transactions on Database Systems (TODS)* 42.3 (2017), p. 19.

[33] Meeyoung Cha et al. “Measuring user influence in twitter: The million follower fallacy”. In: *fourth international AAAI conference on weblogs and social media*. 2010.

[34] Chaoyi Pang et al. “Dominating sets in directed graphs”. In: *Information Sciences* 180.19 (2010), pp. 3647–3652.

[35] Fabrizio Marozzo e Alessandro Bessi. “Analyzing polarization of social media users and news sites during political campaigns”. In: *Social Network Analysis and Mining* 8.1 (2018), p. 1.

[36] Linton C Freeman e Rosanna Memoli. *Lo sviluppo dell'analisi delle reti sociali: uno studio di sociologia della scienza*. Angeli, 2007.