UNIVERSITÀ DELLA CALABRIA

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Definizione e implementazione di un algoritmo di influence maximization in reti sociali

Relatori:
Prof. ***Domenico Talia***
Ing. ***Fabrizio Marozzo***

Candidato:
Silvio Mazza
Matricola 195297

Anno Accademico 2019/2020

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

L'introduzione del Web ha rappresentato per l'intera umanità l'ennesima "**rivoluzione**". Si tratta in tali circostanze di rivoluzione digitale, un fenomeno non transitorio che ha dato il via ad un processo di continua crescita.
Negli ultimi anni infatti l'utilizzo dei social network ha modificato il comportamento delle persone, radicandosi nella quotidianità degli utenti che si mostrano sempre più felici ed interessati dalle interazioni digitali che si generano.
I social rappresentano dunque un terreno fertile, dove gli utenti consumano interazioni generando preziose informazioni a partire dalle quali si può estrarre conoscenza.

I Social Scientist approfondiscono lo studio delle reti sociali al fine di comprendere comportamenti che derivano da relazioni di amicizia, gruppi e comunità che si formano o che potrebbero formarsi, mentre gli esperti di marketing analizzano tali informazioni al fine comprendere le possibili strategie da adottare per accrescere il beneficio di una campagna.

La diffusione su larga scala dell'utilizzo dei social, ed in linea generale del web, ha portato in dote quelli che vengono definiti come **Big Data**, ovvero collezioni di dati di grandi dimensioni, provenienti da fonti differenti, che possono essere generati continuamente e con un elevata velocità.

Il processo di analisi e di estrazione di conoscenza da tale immensa mole di dati, di supporto ai processi decisionali, prende il nome di **Big Data Analytics**, le cui applicazioni non hanno il solo scopo di estrarre informazioni per ricavare un profitto diretto, ma possono essere anche usate in contesti come l'analisi dei comportamenti umani.

Al fine di manipolare i Big Data sono necessari strumenti specifici in grado di sopperire alla complessità spaziale e computazionale che caratterizza i dati stessi. Tali strumenti però hanno come caratteristica intrinseca quella che è la difficoltà

di utilizzo, richiedendo un background di competenze informatiche.

Tra i vari aspetti osservabili, la propagazione delle informazioni all'interno delle reti sociali ha mostrato, negli ultimi anni, grande interesse da parte della comunità scientifica. I social scientist hanno studiato meticolosamente le reti in tale ambito, riferendosi però a dati rappresentativi del dominio. Una delle principali motivazioni che ha portato allo studio della propagazione delle informazioni e dei modelli di diffusione è il viral marketing. L'influence maximization è una tecnica per il viral marketing, si tratta del problema di ottimizzazione in cui si vuole identificare un piccolo insieme di nodi, anche detti **influencers**, al fine di propagare un'idea all'interno della rete al maggior numero possibile di utenti. Dunque un problema NP-Hard, in cui le sorgenti di hardness sono due: una prima relativa alla complessità di calcolo dello **spread**, ovvero del numero di utenti che sono influenzati, ed una seconda relativa alla natura combinatoria del problema ovvero l'identificazione delle combinazioni, tra tutte quelle possibili, che massimizzano l'influenza.

Il problema dunque risulta arduo da affrontare ancor di più nel contesto dei Big Data e scegliere la tecnologia adatta non è un compito semplice. Una delle metodologie emergenti è quella BSP, un paradigma di calcolo parallelo che differisce dagli altri framework per aspetti relativi a comunicazione e sincronizzazione. Il framework Apache Hama permette di definire processi sotto tale metodologia sfruttando quello che è il modello **Pregel**, anche detto **Vertex-Centric** in cui vige il "think like a vetex".

L'obiettivo dell'elaborato è quello di analizzare il comportamento della comunità italiana durante l'evento del Referendum Costituzionale del 2016. Analizzare dunque quella che è la propagazione delle informazioni della rete Twitter durante i mesi che portarono a tale evento, ed individuare i principali influencers delle due fazioni coinvolte mediante una metodologia che discende da algoritmi bio-inspired e che prova a migliorare gli aspetti che caratterizzano gli algoritmi ranking proxy based.

Nel primo capitolo l'attenzione sarà incentrata sui social media. Sull'espansione di questi e sull'importanza dei dati che generano, anche detti **Big Social Data**, concentrandosi quindi sul mondo **Big Data**, le sfide di tale e le possibili metodologie da adottare.

Il secondo capitolo descriverà il modello di computazione emergente **Bulk Synchronous Parallel** ed il framework **Apache Hama** che permette di usufruire di tale modello sia in una forma pura, sia mediante **Google Pregel**, quest'ultima

rappresenta una tecnica centrale per l'elaborato.

Il terzo capitolo descrive invece il problema dell'**Influence Maximization**, fornendo dettagli relativi ai principali modelli di propagazione delle informazioni, e alla tassonomia delle moderne tecniche di risoluzione della problematica.

Il quarto capitolo invece descrive la metodologia proposta. Si tratta di una tecnica **bio-inspired**, implementata con lo stile Pregel, ispirata al meccanismo utilizzato dalle api al fine di individuare le fonti di cibo. La metodologia mira a migliorare gli algoritmi di Influence Maximization ranking proxy based, provando a superare le difficoltà che caratterizzano tali.

Il quinto capitolo invece applica tale metodologia ad un caso reale, nello specifico ad un grafo costruito a partire da relazioni su Twitter. Il grafo è costruito analizzando i retweets collezionati sulla base di alcuni hashtag durante i tempi antecedenti al referendum.

Bibliografia

[1] Apache. *Apache Hama*. 2012. URL: https://hama.apache.org/index.html (visitato il 06/03/2020).

[2] Nicola Barbieri, Francesco Bonchi e Giuseppe Manco. "Topic-aware social influence propagation models". In: *Knowledge and information systems* 37.3 (2013), pp. 555–584.

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

[5] 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.

[6] Shishir Bharathi, David Kempe e Mahyar Salek. "Competitive influence maximization in social networks". In: *International workshop on web and internet economics*. Springer. 2007, pp. 306–311.

[7] 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.

[8] Eric Bonabeau et al. *Swarm intelligence: from natural to artificial systems*. 1. Oxford university press, 1999.

[9] Christian Borgs et al. "Maximizing social influence in nearly optimal time". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 946–957.

[10] Arastoo Bozorgi et al. "Community-based influence maximization in social networks under a competitive linear threshold model". In: *Knowledge-Based Systems* 134 (2017), pp. 149–158.

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

[12] Thomas Cheatham et al. "Bulk synchronous parallel computing—a paradigm for transportable software". In: *Tools and Environments for Parallel and Distributed Systems*. Springer, 1996, pp. 61–76.

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

[14] Wei Chen, Laks VS Lakshmanan e Carlos Castillo. "Information and influence propagation in social networks". In: *Synthesis Lectures on Data Management* 5.4 (2013), pp. 1–177.

[15] Wei Chen, Wei Lu e Ning Zhang. "Time-critical influence maximization in social networks with time-delayed diffusion process". In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[16] Wei Chen, Yajun Wang e Siyu Yang. "Efficient influence maximization in social networks". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 199–208.

[17] Wei Chen, Yifei Yuan e Li Zhang. "Scalable influence maximization in social networks under the linear threshold model". In: *2010 IEEE international conference on data mining*. IEEE. 2010, pp. 88–97.

[18] Xiaodong Chen et al. "On influential nodes tracking in dynamic social networks". In: *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM. 2015, pp. 613–621.

[19] Nicholas A Christakis e James H Fowler. "The spread of obesity in a large social network over 32 years". In: *New England journal of medicine* 357.4 (2007), pp. 370–379.

[20] Federico Coro et al. "Exploiting social influence to control elections based on scoring rules". In: *arXiv preprint arXiv:1902.07454* (2019).

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

[22] Pedro Domingos e Matt Richardson. "Mining the network value of customers". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, pp. 57–66.

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

[24] Alexandros V Gerbessiotis e Leslie G Valiant. "Direct bulk-synchronous parallel algorithms". In: *Journal of parallel and distributed computing* 22.2 (1994), pp. 251–267.

[25] Ashish A Golghate e Shailendra W Shende. "Parallel K-means clustering based on hadoop and hama". In: *International Journal of Computing and Technology* 1.3 (2014), pp. 33–37.

[26] Amit Goyal, Wei Lu e Laks VS Lakshmanan. "Simpath: An efficient algorithm for influence maximization under the linear threshold model". In: *2011 IEEE 11th international conference on data mining*. IEEE. 2011, pp. 211–220.

[27] Xinran He et al. "Influence blocking maximization in social networks under the competitive linear threshold model". In: *Proceedings of the 2012 siam international conference on data mining*. SIAM. 2012, pp. 463–474.

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

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

[30] Roberto Interdonato, Chiara Pulice e Andrea Tagarelli. "An Influence Maximization based approach to the Engagement of Silent Users in Online Social Networks." In: *SEBD*. 2017, p. 210.

[31] Kyomin Jung, Wooram Heo e Wei Chen. "Irie: Scalable and robust influence maximization in social networks". In: *2012 IEEE 12th International Conference on Data Mining*. IEEE. 2012, pp. 918–923.

[32] Dervis Karaboga. *An idea based on honey bee swarm for numerical optimization*. Rapp. tecn. Technical report-tr06, Erciyes university, engineering faculty, computer ..., 2005.

[33] David Kempe, Jon Kleinberg e Éva Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 137–146.

[34] 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.

[35] 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.

[36] Jinha Kim, Seung-Keol Kim e Hwanjo Yu. "Scalable and parallelizable processing of influence maximization for large-scale social networks?" In: *2013 IEEE 29th international conference on data engineering (ICDE)*. IEEE. 2013, pp. 266–277.

[37] Jinha Kim, Wonyeol Lee e Hwanjo Yu. "CT-IC: Continuously activated and time-restricted independent cascade model for viral marketing". In: *Knowledge-Based Systems* 62 (2014), pp. 57–68.

[38] Masahiro Kimura e Kazumi Saito. "Tractable models for information diffusion in social networks". In: *European conference on principles of data mining and knowledge discovery*. Springer. 2006, pp. 259–271.

[39] Jure Leskovec et al. "Cost-effective outbreak detection in networks". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 420–429.

[40] Guoliang Li et al. "Efficient location-aware influence maximization". In: *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 2014, pp. 87–98.

[41] Hui Li et al. "Getreal: Towards realistic selection of influence maximization strategies in competitive networks". In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 2015, pp. 1525–1537.

[42] Yuchen Li et al. "Influence maximization on social graphs: A survey". In: *IEEE Transactions on Knowledge and Data Engineering* 30.10 (2018), pp. 1852–1872.

[43] Grzegorz Malewicz et al. "Pregel: a system for large-scale graph processing". In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010, pp. 135–146.

[44] 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.

[45] Mark M Millonas. "Swarms, phase transitions, and collective intelligence". In: *arXiv preprint adap-org/9306002* (1993).

[46] George L Nemhauser, Laurence A Wolsey e Marshall L Fisher. "An analysis of approximations for maximizing submodular set functions—I". In: *Mathematical programming* 14.1 (1978), pp. 265–294.

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

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

[49] Manuel Gomez Rodriguez, David Balduzzi e Bernhard Schölkopf. "Uncovering the temporal dynamics of diffusion networks". In: *arXiv preprint arXiv:1105.0697* (2011).

[50] Kazumi Saito, Ryohei Nakano e Masahiro Kimura. "Prediction of information diffusion probabilities for independent cascade model". In: *International conference on knowledge-based and intelligent information and engineering systems*. Springer. 2008, pp. 67–75.

[51] C Prem Sankar, S Asharaf e K Satheesh Kumar. "Learning from bees: An approach for influence maximization on viral campaigns". In: *PloS one* 11.12 (2016).

[52] Kamran Siddique, Zahid Akhtar e Yangwoo Kim. "Researching Apache Hama: A Pure BSP Computing Framework". In: *Advanced Multimedia and Ubiquitous Engineering*. Springer, 2016, pp. 215–221.

[53] Kamran Siddique et al. "Apache Hama: An emerging bulk synchronous parallel computing framework for big data applications". In: *IEEE Access* 4 (2016), pp. 8879–8887.

[54] Kamran Siddique et al. "Investigating Apache Hama: a bulk synchronous parallel computing framework". In: *The Journal of Supercomputing* 73.9 (2017), pp. 4190–4205.

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

[56] Chi Wang, Wei Chen e Yajun Wang. "Scalable influence maximization for independent cascade model in large-scale social networks". In: *Data Mining and Knowledge Discovery* 25.3 (2012), pp. 545–576.

[57] Yu Wang et al. "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 1039–1048.

[58] Bryan Wilder e Yevgeniy Vorobeychik. "Controlling elections through social influence". In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. International Foundation for Autonomous Agents e Multiagent Systems. 2018, pp. 265–273.

[59] Honglei Zhuang et al. "Influence maximization in dynamic social networks". In: *2013 IEEE 13th International Conference on Data Mining*. IEEE. 2013, pp. 1313–1318.