

# Machine Learning per Behavior Analytics e Recommender Systems

Massimo Guarascio, Giuseppe Manco, Giuseppe Pirrò, Ettore Ritacco

ICAR-CNR, Via Bucci 8-9c, Rende (CS)

## Abstract

La behavior analytics è un importante tema di ricerca in diversi contesti reali quali la consumer analytics, il social computing, la fraud detection ed il group decision making. Per comportamento si fa riferimento alle azioni ed alle reazioni intraprese da un'entità in risposta a vari stimoli od input. Benché la behavior analytics abbracci un'ampia varietà di domini, in questo documento ci concentreremo principalmente sul tema della *Recommendation*.

## 1 Introduzione e background

Il recente avanzamento tecnologico nel campo dei sistemi di stoccaggio e gestione dati ha permesso di collezionare enormi quantità di dati comportamentali che possono essere opportunamente impiegati per apprendere nuovi modelli matematici in grado di analizzare, comprendere e predire le azioni intraprese da un'entità di interesse. Ciò include stream di eventi, connessioni in una social network, abitudini di acquisto ed informazioni sulle opinioni degli utenti. La comprensione delle dinamiche strutturali, temporali e tematiche del comportamento di un utente può fornire da un lato dettagli sui complessi pattern che governano il processo di propagazione dell'informazione e dall'altro può essere utilizzata per predire eventi futuri. Ciò è cruciale in diversi scenari reali quali: medical events (ad es. l'analisi della sequenza infortunio, visite, diagnosi e cure), consumer behavior, wearable data analysis, social media actions (ed es. post, condivisioni, commenti, messaggi, etc.), smart cities e mobility patterns (ad es. traiettorie, utilizzo di mezzi pubblici, etc.).

Profilare accuratamente gli utenti così da fornire raccomandazioni di prodotti che incontrino i loro gusti, predirne le future ricerche, adattare la visualizzazione dei contenuti in base ai loro principali interessi e migliorare in questo modo la loro esperienza nella fruizione dei servizi è un task di interesse per molte grandi compagnie, specialmente se si pensa ad aziende operanti nel campo del delivery di contenuti multimediali (quali ad esempio film, musica, etc.). Ad oggi, strumenti come *ContentWise* (engine usato per profilare utenti di servizi IPTV e suggerire contenuti) sono tra gli strumenti maggiormente impiegati da grandi società quali Sky, Mediaset, RAI, e BBC per migliorare la qualità dei servizi offerti ed incrementare il loro bacino di utenti. Non di meno, i cosiddetti

*assistenti personali digitali* si stanno rapidamente diffondendo e sono sempre più frequentemente sfruttati per suggerire contenuti all'utente. Apple con *Siri*, Microsoft con *Cortana*, Amazon con *Alexa* e Google con *Google Now* sono la dimostrazione di come metodi e tecniche di intelligenza artificiale possano essere usate per interpretare i comandi dell'utente e migliorare la sua esperienza.

La letteratura recente nell'ambito della profilazione e raccomandazione si è concentrata su approcci basati su *Collaborative Filtering*, che consentono tramite l'analisi dello storico delle interazioni dell'utente di estrarre profili comportamentali accurati. Tali approcci si dividono storicamente in "memory-based" e approcci "model-based", la cui differenza consiste essenzialmente nel fatto che il profilo comportamentale sia individuato in base alla "vicinanza" rispetto alla casistica storica, o in base ad un modello matematico che descrive le principali caratteristiche dei profili. Di fatto, i modelli a fattori latenti rappresentano lo stato dell'arte data la loro capacità di modellare le cause ("nascoste") dell'agire dell'utente (ossia preferenze, scelte ed azioni). Recentemente, nuovi approcci basati sull'impiego di architetture di deep learning sono stati proposti integrando e migliorando sensibilmente gli approcci a fattori latenti.

## 2 Linee di ricerca emergenti

Attualmente, i temi di ricerca di maggior interesse nell'ambito dei recommender systems, e più in generale nella behavior analytics, riguardano i seguenti aspetti: *catalog enrichment*, *social profiling*, *improved accuracy* and *assisted search*.

*Catalog enrichment*. Con il termine si intende il processo non banale di aggiungere struttura, contesto e metadati a diverse tipologie di contenuti così da renderli più facilmente fruibili da esseri umani e macchine. Il challenge è l'automazione di tale processo, tramite l'analisi delle sorgenti e la ricerca di correlazioni implicite. A questo proposito, possono essere utilizzate tecniche di machine learning e di intelligenza artificiale. L'inferenza e la strutturazione di *Knowledge Graphs*, nonché la ricostruzione delle relazioni che li caratterizzano [Pirrò, 2019] rappresenta un tema di grande interesse in quest'ambito: la dimensione e l'eterogeneità di questi grafi, che considera varie tipologie di relazione tra entità, rende disponibile una grande quantità di conoscenza.

*Social profiling*. La modellazione del profilo dell'utente basata sull'analisi delle azioni che esibisce all'interno

di una rete sociale rappresenta un'importante filone di ricerca. Azioni quali la condivisione di contenuti o seguitori determinati utenti, possono fornire importanti indicazioni sui gusti dell'utente ed essere sfruttati per ottenere sistemi di raccomandazione più accurati. In quest'ambito, tematiche come *information diffusion* [Barbieri *et al.*, 2013a; Barbieri *et al.*, 2017], (ossia comprendere come l'informazione si propaghi all'interno della rete e scoprire i migliori percorsi identificando i nodi che fungono da hub e sink della rete) e *influence propagation* [Barbieri *et al.*, 2013b; Barbieri *et al.*, 2014a; Barbieri *et al.*, 2016], (ossia l'identificazione dei fattori latenti capaci di spiegare le ragioni per cui un dato soggetto sia suscettibile a determinati argomenti sulla base della sua egonet) sono attualmente oggetto di studio.

*Improved accuracy.* Le tecniche di raccomandazione tradizionali, basate sull'ottimizzazione dell'accuratezza della predizione, tendono ad esprimere modelli poco efficaci nel generare liste di raccomandazioni interessanti per l'utente. Misure quali la *diversity* e la *novelty* [Steck, 2018; Barbieri *et al.*, 2012] sono in grado di migliorare l'esperienza dell'utente su un catalogo di contenuti. Più in generale, rimane un tema aperto il problema di come iniettare una serie di vincoli di dominio [Gori, 2018] nel delivery personalizzato dei contenuti. In particolare, la modellazione di aspetti temporali stanno ricevendo attenzione crescente ed è stato dimostrato che il *session-based profiling* rappresenti un modello più accurato delle preferenze degli utenti [Barbieri *et al.*, 2013c; Manco *et al.*, 2018; Sachdeva *et al.*, 2019].

*Assisted search.* Sistemi per la ricerca assistita sono sempre più frequenti in diversi settori, in particolare nel campo della fruizione di contenuti digitali. Gli assistenti personali digitali (*digital personal assistant*) si stanno diffondendo in maniera pervasiva e sfruttano strumenti sempre più evoluti quali la *Voice Search*. Benché tecniche per il riconoscimento del linguaggio naturale (NLP) abbiano ottenuto recentemente risultati interessanti, molte problematiche rimangono ancora aperte. In particolare, vi è una ricerca sempre crescente verso approcci efficaci e scalabili. Metodi basati su rappresentazioni distribuite e word embedding (es. *word2vec*), deep networks e reinforcement learning sono alcune delle strategie attualmente oggetto di studio per affrontare questo task.

### 3 Competenze e iniziative progettuali

Il Laboratorio di Analitica Avanzata su Dati Complessi (ADALab) dell'ICAR-CNR conduce ricerca nell'area del *Behavior Computing* ossia la modellazione e l'analisi del comportamento e delle interazioni tra entità. In quest'ambito, il laboratorio sviluppa tecniche di artificial intelligence e machine learning che possono essere applicate ad una ampia gamma di tipologie di dati, incluse sequenze, grafi, dati non strutturati, e dati di sensori/stream. I ricercatori afferenti al laboratorio hanno maturato una vasta esperienza sul tema del behavioral modeling, contribuendo in particolare con approcci innovativi basati su metodi probabilistici generativi per il collaborative filtering [Barbieri *et al.*, 2014b]. Su tali tematiche il gruppo di ricerca ha sviluppato inoltre 4 brevetti internazionali ed ha ricevuto nel 2014 lo YAHOO Faculty Research and Engagement Program, per attività di ricerca relativa a tecniche di machine learning per "Detecting, describing

and distinguishing social Vs. topical communities for social advertisement".

Nell'ambito dello sviluppo di sistemi di raccomandazione evoluti l'ICAR-CNR è stato coinvolto nella proposta progettuale *CATCH 4.0*, presentata nel contesto dei PON I&C 2014-2020 FESR del MISE. Uno degli obiettivi principali del progetto è appunto sviluppare strumenti mirati di *advanced content delivery*, in grado da una lato di variegare il contenuto informativo dei cataloghi digitali, arricchendo l'informazione disponibile con altre informazioni reperibili da sorgenti eterogenee o desumibili tramite tecniche di intelligenza artificiale e machine learning, dall'altro di adattare i contenuti da proporre all'utente in base alle sue preferenze contestuali desumibili in maniera implicita (tramite l'analisi del suo storico) ed esplicita (tramite interazione).

L'ICAR-CNR inoltre, in collaborazione con NTT Data, ha partecipato altresì allo sviluppo di un recommender system evoluto chiamato *Advisor Engine* e basato su deep neural networks e sistemi embedded a fattori latenti. Il sistema, che integra al suo interno un per la specifica di vincoli sulla raccomandazione di vincoli, è stato sviluppato per un'azienda leader nel settore della grande distribuzione.

### Riferimenti bibliografici

- [Barbieri *et al.*, 2012] N. Barbieri, et al. Balancing prediction and recommendation accuracy. *SDM*, pp. 1035–1046, 2012.
- [Barbieri *et al.*, 2013a] N. Barbieri, F. Bonchi, G. Manco. Cascade-based community detection. *WSDM*, pp. 33–42, 2013.
- [Barbieri *et al.*, 2013b] N. Barbieri, F. Bonchi, G. Manco. Topic-aware social influence propagation models. *Knowl. Inf. Syst.*, 37(3), 2013.
- [Barbieri *et al.*, 2013c] N. Barbieri, et al. Probabilistic topic models for sequence data. *Machine Learning*, 93(1), 2013.
- [Barbieri *et al.*, 2014a] N. Barbieri, F. Bonchi, G. Manco. Who to follow and why: Link prediction with explanations. In *KDD*, pp. 1266–1275, 2014.
- [Barbieri *et al.*, 2014b] N. Barbieri, G. Manco, E. Ritacco. *Probabilistic Approaches to Recommendations*. Morgan & Claypool, 2014.
- [Barbieri *et al.*, 2016] N. Barbieri, F. Bonchi, G. Manco. Efficient methods for influence-based network-oblivious community detection. *ACM Trans. Intell. Syst. Technol.*, 8(2), 2016.
- [Barbieri *et al.*, 2017] N. Barbieri, G. Manco, E. Ritacco. Survival factorization on diffusion networks. In *ECML PKDD*, pp. 684–700, 2017.
- [Gori, 2018] M. Gori. *Machine Learning: A Constraint-Based Approach*. Morgan Kaufmann, 2018.
- [Manco *et al.*, 2018] G. Manco, G. Pirrò, G. Ritacco. Predicting temporal activation patterns via recurrent neural networks. In *ISMIS*, pp. 347–356, 2018.
- [Pirrò, 2019] G. Pirrò. Building relatedness explanations from knowledge graphs. *Semantic Web Journal*, 2019.
- [Sachdeva *et al.*, 2019] N. Sachdeva, et al. Sequential variational autoencoders for collaborative filtering. In *WSDM*, pp. 600–608, 2019.
- [Steck, 2018] H. Steck. Calibrated recommendations. *RecSys*, pp. 154–162, 2018.