UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in
INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Progettazione e implementazione di un recommender system basato su grafo per il supporto alle vendite di un grande distributore di componenti elettronici

Design and implementation of a graph-based recommender system supporting sales for a big electronics components distributor

Relatore:
Chiar.mo Prof. ***Domenico Potena***

Laureando:
Alessandrino Manilii

Anno accademico 2021/2022

Indice

Introduzione **5**

1 Il recommender system **9**

1.1 Un po' di storia 9

1.2 Cos'è un recommender system? 12

1.3 Perchè utilizzarli? 13

1.4 Tipi di recommender system 15

1.4.1 Collaborative 16

1.4.2 Content-based 17

1.4.3 Demographic 18

1.4.4 Utility-based 18

1.4.5 Knowledge-based 19

1.4.6 Ibrido 20

2 Contesto applicativo **21**

2.1 Avnet 22

2.2 Il mercato in cui si opera 22

2.3 Sfide progettuali 23

3 Sorgenti informative **25**

3.1 Caratteristiche e valore informativo 25

3.1.1 Design Registrable Management System (DRMS) 26
3.1.2 Quotes 27
3.1.3 Ordini e fatture 27
3.1.4 Schema E-R riepilogativo 28
3.2 Dalle tabelle al grafo 28
3.2.1 Richiami sulla teoria dei grafi 29
3.2.2 Approccio alla costruzione 30
3.2.3 Caratteristiche generali del grafo 32
3.3 Data cleaning 33
3.3.1 Analisi sui clienti 33
3.3.2 Analisi sui progetti 34
3.3.3 Identificatori dei componenti 34

4 Data analysis: duplice approccio **37**
4.1 Obiettivi 37
4.2 Utilizzo diretto del grafo 38
4.2.1 Neighbors 39
4.2.2 Metriche utilizzate 39
Jaccard coefficient 40
Adamic-Adar index 40
4.3 Deep walk 41
4.3.1 Struttura del grafo 42
4.3.2 Il metodo in dettaglio 42
Random walk 43
Commodities e non-commodities 44
Word2Vec 49
4.4 Misura della similarità 50

5 Implementazione e confronto degli approcci **53**

5.1 Ambiente di sviluppo 53

5.2 Implementazione e tempi di esecuzione 54

5.2.1 Utilizzo diretto del grafo 54

5.2.2 Deep Walk 55

5.3 Valutazione 57

5.3.1 Dati di riferimento 57

5.3.2 Risultati 58

6 Conclusioni e sviluppi futuri **61**

Bibliografia **63**

Elenco figure **65**

Introduzione

Il seguente lavoro di tesi presenta il risultato ottenuto a seguito di un periodo di attività lavorativa presso una delle sedi francesi del distributore di componenti elettronici Avnet. La tesi si pone come obiettivo quella di presentare nella sua interezza tutte le fasi affrontate partendo da quelle iniziali di analisi, passando per l'implementazione fino a risultati. L'oggetto principale della discussione è un sistema di raccomandazione, una tecnologia utilizzata in molti ambiti, dallo shopping online ai suggerimenti di contenuti su piattaforme multimediali esprimibile come un algoritmo che utilizza dati sugli interessi e le preferenze degli utenti e, al contempo, sui prodotti stessi, per produrre suggerimenti di consumo il più accuratamente possibile. Dopo un primo esame sui diversi tipi di sistemi di raccomandazione e le loro applicazioni, nonché le sfide e le opportunità associate all'utilizzo di questi sistemi, si passa alla presentazione del contesto aziendale in cui l'attività è stata svolta con l'obiettivo di introdurre l'ambiente in cui il sistema opera e fornire al contempo una panoramica delle varie attività svolte all'interno dell'organizzazione. Presentando nel dettaglio il mercato in cui l'azienda opera, si contestualizzano le sfide ancora aperte e le esigenze che le soluzioni attualmente in uso non soddisfano a pieno. Da qui ha inizio la trattazione più tecnica del progetto presentando le sorgenti dati a disposizione dal punto di vista informativo e descrivendone quindi le caratteristiche quantitative e qualitative. Si trattanno brevemente anche le problematiche correlate al formato in cui i dati si presentano e le relative difficoltà di utilizzo per la

costruzione di modelli di data mining. Quindi si passa a presentare il grafo come nuova struttura tramite cui rappresentare i dati filtrati, motivandone la scelta e riportandone le caratteristiche principali. Si termina questo capitolo discutendo l'attività di data cleaning svolta che risulta propedeutica allo sfruttamento ottimale delle risorse.

Nel quarto capitolo, si tratta invece la fase di analisi che costituisce, con l'implementazione, la fase più corposa del lavoro. Si riprende il discorso sugli obiettivi e le sfide aperte anticipati nel secondo capitolo per descrivere l'analisi effettuata e il duplice approccio di utilizzo del grafo per la generazione di suggerimenti. In particolare, si esplorano:

- un approccio diretto, utilizzando appunto il grafo senza passaggi e strutture intermedie,
- un approccio basato su Deep Walk, in cui si genereranno cammini random di una certa lunghezza e in un certo numero per estrarre informazioni nascoste nella struttura del grafo

descrivendo in dettaglio i principi di funzionamento dei metodi utilizzati assieme al loro significato analitico.

Il successivo capitolo si apre con una parentesi sulle tecnologie utilizzate e sull'ecosistema di strumenti adottati a partire dal servizio cloud, passando per le librerie adottate e le motivazioni alla base delle scelte effettuate, fino agli strumenti di programmazione. Si passa quindi a descrivere l'implementazione dei metodi presentati riportando anche i tempi di esecuzione registrati. Infine si discutono i risultati cercando di valutare le raccomandazioni sulla base di alcuni dati di riferimento e di motivare il perchè si abbiano risultati migliori con un approccio rispetto all'altro per un certo obiettivo di analisi.

Il lavoro si conclude ripercorrendo le sfide affrontate e sottolineando come gli approcci presentati posso considerarsi soluzioni complementari più che alternative,

con l'obiettivo di essere utilizzate insieme, in uno sviluppo futuro del progetto, per migliorare l'efficacia complessiva del sistema. Sempre a proposito degli sviluppi futuri, se ne presentano diversi tutti implicanti un'evoluzione del grafo che ne aumenti la complessità e il carico informativo; si discutono le sfide ancora aperte fornendo quindi spunti di riflessione ed enfatizzando come l'approccio seguito alla costruzione del sistema apra le porte a numerose alternative di utilizzo ancora da esplorare.

Lo scopo di questo lavoro è quindi quello di fornire una panoramica completa del progetto nelle sue varie fasi presentando le problematiche affrontate e contestualizzando la trattazione nello specifico dominio in cui l'azienda opera con la speranza che possa essere occasione di riflessione per la ricerca di soluzioni anche in contesti applicativi differenti.

Il recommender system

In questo primo capitolo si presenterà in linea generale il recommender system come strumento di supporto all'utente contestualizzando l'ambiente e le esigenze che hanno portato alla sua nascita e passando quindi a presentare gli aspetti più tecnici differenziandone le diverse categorie esistenti.

1.1 Un po' di storia

L'idea di sfruttare i calcolatori per consigliare il miglior prodotto ad un utente ha origini lontane. La prima implementazione di cui si ha conoscenza risale al 1979 e conosciuta col nome di Grundy, un sistema bibliotecario programmato per intervistare gli utenti sulle loro preferenze e consigliare di conseguenza dei libri tenendo conto delle informazioni raccolte; in particolare, con un metodo piuttosto primitivo, assegnava l'utente a un gruppo di stereotipi a cui era a sua volta associato un set di libri. Nei primi anni '90 diversi enti di ricerca iniziarono ad interessarsi sempre più al problema dell'information filtering e nacquero, di riflesso, una serie di nuovi progetti. Per citarne alcuni, ad esempio, riscosse successo Tapestry, rivoluzionario sistema di posta e archivio che permetteva di effettuare ricerche in base al contenuto dei documenti e alle reazioni registrate dagli altri utenti, o i recommender systems di casa GroupLens tra cui quello più conosciuto su Usenet, oppure Ringo, un sistema di filtraggio delle informazioni sociali che consentiva agli utenti di formulare

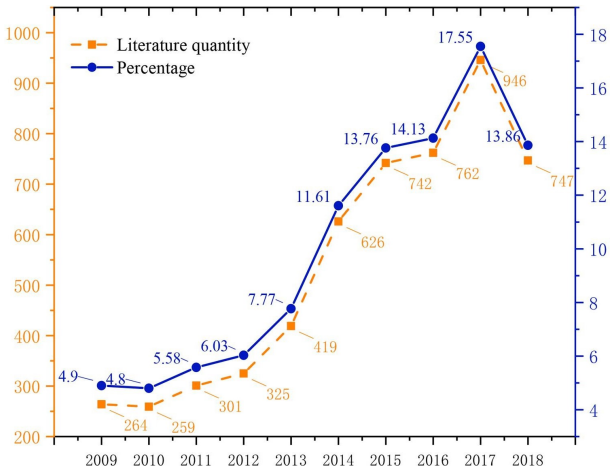

Fig. 1.1: Distribuzione annuale e percentuale annuale di letteratura pubblicata nell'ambito dei recommender system.

raccomandazioni musicali che venivano poi confrontate con le raccomandazioni di altri utenti, presentando quindi all'utente un elenco di brani di suo gradimento. [1]

Con l'espansione capillare di Internet e la crescente necessità di adattarsi al commercio online, i sistemi di raccomandazione hanno subito un rapido sviluppo, diventando ben presto uno strumento necessario per qualsiasi azienda voglia offrire i propri prodotti online, plasmando il più possibile l'offerta sul cliente. Tale espansione risulta fortemente visibile nel grafico in Fig.1.1, che mostra l'evoluzione del numero di pubblicazioni nel campo dei sistemi di raccomandazione, a livello mondiale, nel periodo che va dal 2008 al 2018: le variazioni nel numero di articoli riflette generalmente l'attenzione del mondo della ricerca su argomenti correlati e la velocità di crescita delle conoscenze e, più in generale, può riflettere l'entusiasmo della ricerca sul tema analizzato [2].

Nonostante nell'ultimo periodo considerato dopo lo sviluppo esplosivo del 2017 il numero di articoli pubblicati diminuisca nel 2018, il sistema di raccomandazione è ancora nel boom della ricerca internazionale e il rapido sviluppo della tecnologia dell'intelligenza artificiale, il forte sostegno delle politiche nazionali e gli investimenti senza precedenti nella ricerca e nello sviluppo del settore hanno permesso di

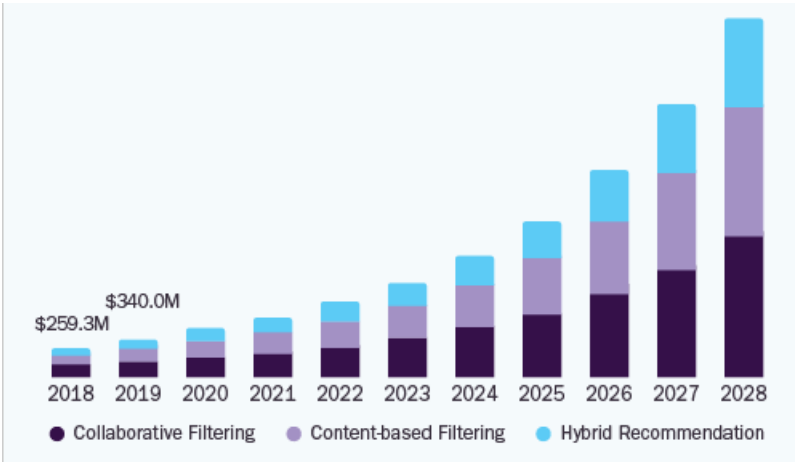

Fig. 1.2: Mercato dei sistemi di raccomandazione negli Stati Uniti, per tipologia, e relativa previsione di crescita - periodo 2021-2028.

mantenere il tema al centro dell'attenzione come uno dei rami più importanti nel campo della ricerca sull'intelligenza artificiale. L'utilizzo di tali tecniche, soprattutto nel campo del machine learning, ha garantito non solo di migliorare in termini di precisione e personalizzazione delle offerte ma ha permesso di adattarsi ai cambiamenti sempre più repentini e radicali nelle abitudini dei consumatori in maniera più o meno automatica. A testimonianza dell'impatto economico che tali strumenti hanno avuto e dell'interesse commerciale per quest'utlimi, sono molteplici i casi di ingenti finanziamenti alla ricerca o per eventi come hackathon e concorsi volti a premiare le capacità di innovazione in tale settore come, ad esempio, il Netflix Prize che nel 2006 assegnava un premio da un milione di dollari allo sviluppatore con la miglior soluzione.

Come mostrato in Fig.1.2, il trend di mercato negli Stati Uniti rispecchia l'interesse economico e l'impatto di tali sistemi sull'economia. Allo stesso modo il mercato globale è stato valutato 1,77 miliardi di dollari nel 2020 e si è prevista un'espansione a un tasso di crescita annuale del 33% dal 2021 al 2028.

1.2 Cos'è un recommender system?

> "A recommender system is a tool that uses active information-filtering techniques to exploit past user behavior to suggest information tailored to an end user's goals." [1]

Scomponendo tale definizione si possono sottolineare i seguenti aspetti:

- è una sottoclasse di sistemi di filtraggio delle informazioni, quindi come tale rappresenta un sistema che rimuove informazioni ridondanti o indesiderate da un flusso di informazioni utilizzando metodi (semi)automatici o computerizzati prima della presentazione a un utente umano e ha come obiettivo principale la gestione del sovraccarico di informazioni e l'aumento del rapporto segnale/rumore semantico;

- si basa su informazioni di comportamento dell'utente ed è questo l'assioma di partenza che rende il risultato adattato il più possibile al target. Talune volte a tali informazioni sono aggiunte ulteriori conoscenze di dominio che mirano a rendere i suggerimenti ancor più efficaci;

- produce consigli in termini di "valutazione" o "preferenza" che un utente potrebbe esprimere per un prodotto, guidandone la scelta con l'obiettivo di ridurre al minimo il gap qualitativo tra domanda e offerta.

Tali sistemi rappresentano uno strumento fondamentale per ottimizzare e adattare l'esperienza del cliente ai suoi interessi in modo da rendere più efficace l'attività di scelta da un lato e quella di offerta dall'altro. Esempi si possono trovare ovunque dalle sezioni di prodotti correlati a partire da un target nei siti e-commerce ai suggerimenti costruiti a monte di una ricerca in piattaforme per l'ascolto e la visione.

1.3 Perchè utilizzarli?

Gli scopi di un sistema di raccomandazione in linea generale sono quelli di aiutare gli utenti a trovare articoli rilevanti da una vasta collezione di possibilità e prevedere quali articoli possano essere di interesse per l'utente fornendogli raccomandazioni personalizzate e riducendo la necessità di cercare manualmente tra un'ampia collezione di articoli. Questo migliora l'esperienza del cliente sia di utilizzo che di acquisto, avendo l'ovvia conseguenza di aumentare i tassi di conversione e delle vendite. All'origine di tali esigenze ci sono diverse ragioni quasi tutte riconducibili al problema dell'information saturation. Oggi l'accesso alle informazioni è ovunque e chiunque può facilmente sperimentare, ad esempio, un senso di smarrimento difronte all'enorme mole di risposte che un motore di ricerca può restituirci a seguito di una query. Sebbene questo, in un primo momento, sembri un aspetto positivo, troppo spesso si è sopraffatti o paralizzati a causa di quello che viene definito sovraccarico di informazioni e nella maggior parte dei casi, il conseguente senso di smarrimento è causato più dalla quantità di informazioni irrilevanti che dalla mole di dati in sé. Il sovraccarico informativo si verifica quando i decisori si trovano di fronte a un livello di informazioni superiore alla loro capacità di elaborazione delle informazioni [3] e tale situazione comporta, al raggiungimento di una soglia più o meno soggettiva, un forte calo della capacità di prendere decisioni e scegliere in funzione dei dati acquisiti.

Più in particolare gli effetti di tale sovrabbondanza sono da ricercare nei disturbi da deficit di attenzione, nella sensazione di sovraccarico (tecno-stress), nella diminuzione della capacità decisionale, nella sensazione di perdita di controllo, nella scarsa soddisfazione e complessivamente si traducono tutti in un'esperienza fallimentare per l'utente a prescindere dal contesto in cui ci si trova, sia esso un sito web piuttosto che una rete sociale o una piattaforma multimediale [4]. A tale fenomeno se

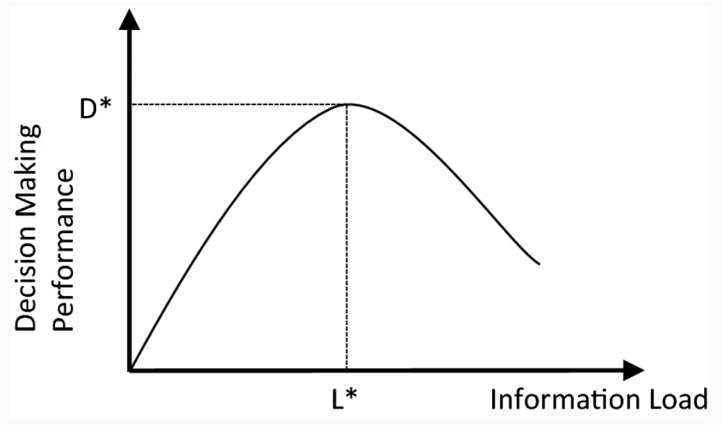

Fig. 1.3: Informazioni e prestazioni decisionali.

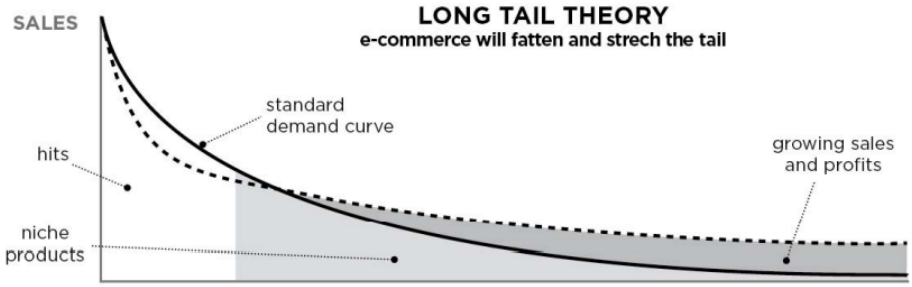

Fig. 1.4: La coda lunga dall'omonima teoria di C.Anderson

ne aggiunge un secondo ad esso strettamente correlato che rappresenta al contrario un elemento di valore se correttamente gestito sia per i clienti, o più in generale gli utenti, che per i soggetti fornitori. Come lo stesso Anderson descrive nel suo libro *The Long Tail: Why the Future of Business is Selling Less of More* [5], l'economia sta progressivamente evolvendo verso una maggiore presenza di nicchie di mercato grazie all'espansione dei mercati digitali. Ciò significa che stiamo passando da una cultura incentrata sui prodotti di massa comunemente definiti mainstream a una maggiore varietà di opzioni disponibili.

Tale fenomeno prende nome dalla nota curva che lo descrive, riportata in Fig.1.4, in cui nella parte sinistra del diagramma troviamo la testa che rappresenta in sostanza le vendite associate ai prodotti maggiormente in voga mentre, nella parte

destra, tutti i prodotti che potremmo definire di nicchia rispetto ai primi. Si può notare come la coda tende virtualmente all'infinito rappresentando una quantità di vendite complessive più che significativa. La coda lunga è la parte di una distribuzione statistica legata alla frequenza più bassa. La forte spinta all'utilizzo di tali sistemi nelle aziende ha le sue radici anche nella naturale spinta alla digitalizzazione da parte delle organizzazioni. La capacità di raccogliere sempre più dati sui clienti ha portato a riconsiderare le risorse a disposizione nelle aziende e a ristrutturare le proprie attività anche in funzione delle evidenze emergenti dalle analisi su tali dati. A tal fine l'utilizzo dei recommender system risulta un'evoluzione naturale verso un cambiamento necessario per rimanere competitivi come organizzazione e fortemente spinto dall'avanzamento tecnologico e conoscitivo nell'ambito del data mining.

1.4 Tipi di recommender system

Le diverse tecniche alla base del funzionamento di tali sistemi si distinguono principalmente in base alla tipologia di dati forniti in input al modello e dei dati contestuali costituenti il background applicativo. In particolare, possiamo distinguere 5 tipologie distinte ed una finale, ibrida, che raccoglie i vantaggi di alcuni tipi [6]. Si presentano di seguito i diversi approcci distinguendoli in funzione dell'insieme I di elementi su cui si potrebbero formulare raccomandazioni, l'insieme U di utenti di cui si conoscono le preferenze, un generico utente u per il quale è necessario generare raccomandazioni e l'elemento i per il quale si vuole prevedere la preferenza di u. Tutti sono accumunati dallo stesso obiettivo di identificare la preferenza prevista di u per i esprimendola in maniera binaria (like/dislike) o su una scala di valori (rating).

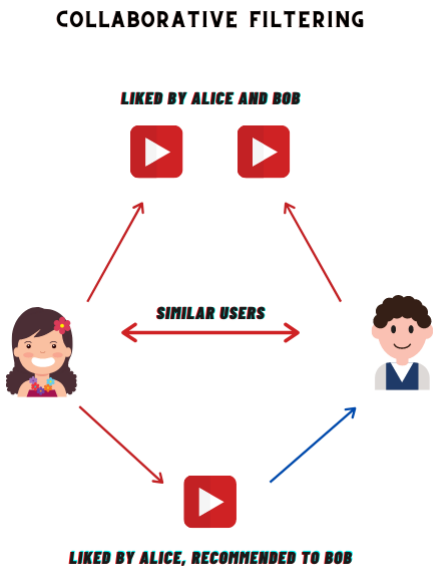

Fig. 1.5: Schema di funzionamento collaborative-filtering.

1.4.1 Collaborative

L'approccio collaborativo consiste nell'identificare utenti in U che risultino simili allo specifico utente selezionato e nell'estrapolare informazioni dalle valutazioni di questi sul prodotto i identificato. L'input del modello coincide con le valutazioni degli oggetti nell'insieme I di prodotti da parte dello specifico utente u, a partite da un background informativo costituito dalle valutazioni dei prodotti nell'insieme I da parte degli utenti in U.

Si aggregano quindi valutazioni o raccomandazioni di oggetti riconoscendo i punti in comune tra gli utenti sulla base delle loro valutazioni e generando nuove raccomandazioni basate sul confronto tra gli utenti. Spesso, a valle di tale processo, è possibile contestualizzare l'output attualizzandolo in base al periodo. Tra i vantaggi: la capacità di identificare nicchie trasversali ai generi, la conoscenza del dominio non necessaria, l'adattività intesa come qualità che migliora nel tempo, il feedback implicito sufficiente; tra gli svantaggi principali invece c'è l'aumento dei nuovi utenti,

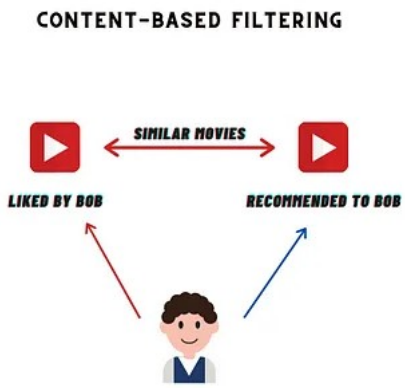

Fig. 1.6: Schema di funzionamento content-based.

il problema del ramp-up di un nuovo articolo, il problema delle "pecore grigie", la dipendenza della qualità da un ampio set di dati storici, il problema della stabilità e della plasticità.

1.4.2 Content-based

L'approccio basato sul contenuto invece si basa sulla generazione di un classificatore che si adatta alle abitudini di valutazione dello specifico utente u per poi utilizzarle sul prodotto i ottenendo di riflesso una valutazione di quell'utente sul prodotto preso in considerazione. In questo caso l'input è dato dalla totalità di valutazioni dei prodotti da parte dell'utente specifico u mentre le informazioni di contesto utilizzabili sono date dalle caratteristiche dei prodotti considerati.

Gli oggetti, infatti, sono definiti dalle loro caratteristiche associate e il processo prevede di apprendere il profilo degli interessi dell'utente in base alle caratteristiche presenti negli oggetti che l'utente ha valutato. Tali modelli sono considerati a lungo termine e vengono aggiornati man mano che si osservano ulteriori prove sulle preferenze dell'utente. I vantaggi di tali sistemi sono la conoscenza del dominio non necessaria, l'adattività come qualità che migliora nel tempo, il feedback implicito sufficiente; tra gli svantaggi invece: l'aumento dei nuovi utenti, la dipendenza della

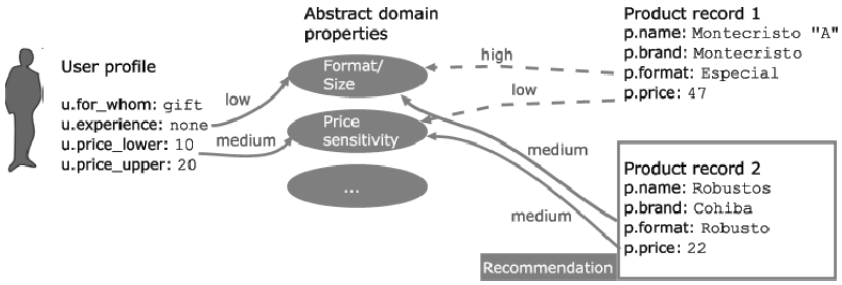

Fig. 1.7: Schema di funzionamento utility-based.

qualità da un ampio set di dati storici, il problema della stabilità e della plasticità.

1.4.3 Demographic

In questa variante dell'approccio collaborativo si identificano utenti che sono demograficamente simili allo specifico utente selezionato e su questa selezione si estrapolano informazioni dalle valutazioni del prodotto i identificato. La peculiarità sta nel categorizzare gli utenti in base agli attributi personali e formulare raccomandazioni basate sulle classi demografiche. I vantaggi di questo approccio sono la capacità di identificare nicchie trasversali ai generi, la conoscenza del dominio non necessaria, l'adattività; tra gli svantaggi invece si ha: l'aumento dei nuovi utenti, il problema delle "pecore grigie", la dipendenza della qualità da un ampio set di dati storici, il problema della stabilità e della plasticità, la necessità di raccogliere dati demografici.

1.4.4 Utility-based

I recommender system basati sull'utilità definiscono una funzione di utilità dei prodotti in I in base alle preferenze dell'utente. L'approccio prevede di applicare tale funzione ai prodotti determinandone una classfica.

Dopo aver identificato le proprietà specifiche di interesse si forniscono suggerimenti basati su un calcolo dell'utilità di ogni oggetto per l'utente tramite una fun-

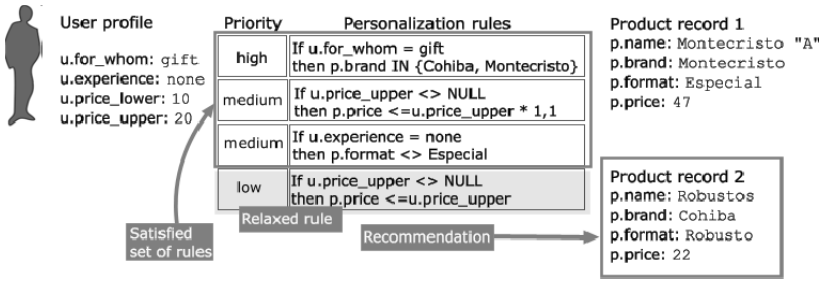

Fig. 1.8: Schema di funzionamento knowledge-based.

zione parametrizzata sulle feature selezionate [7]. A tale scopo è possibile rivisitare il problema nell'ottica dei problemi di soddisfacimento dei vincoli e utilizzare le tecniche a disposizione per la risoluzione di tali problemi al fine di individuare la migliore corrispondenza. A differenza degli approcci precedenti tali modelli non sono generalizzabili a lungo termine sugli utenti e vanno adattati ad ogni utilizzo. Tra i vantaggi di tale approccio: l'assenza di richiesta di ramp-up, la sensibilità ai cambiamenti di preferenze, l'inclusione di caratteristiche non direttamente collegate al prodotto; tuttavia presenta dei limiti dati dalla necessità di ricevere dall'utente la funzione di utilità e dalla staticità della capacità di suggerimento (non apprende).

1.4.5 Knowledge-based

I recommender system basati su conoscenza approcciano il problema basandosi sulla conoscenza funzionale del prodotto: a partire da una descrizione dei bisogni e degli interessi dell'utente, e conoscendo allo stesso tempo come un particolare articolo soddisfa un particolare esigenza permette di correlare gli articoli agli utenti che ne necessitano. I sistemi di questo tipo possono ragionare sulla relazione tra un bisogno e una possibile raccomandazione ma, per far ciò, si necessità di una conoscenza approfondita dei prodotti soprattutto dal punto di vista dei bisogni che questo soddisfa e del tipo di servizio che offre [7]. Anche in tal caso, come per l'approccio utility-based, non esistono modelli adattabili a lungo termine.

Tra i vantaggi: l'assenza di richiesta di ramp-up, la sensibilità ai cambiamenti di preferenze, l'inclusione di caratteristiche non direttamente collegate al prodotto, la possibilità di tracciare una mappa dalle esigenze degli utenti;allo stesso tempo si ha staticità nella capacità di suggerimento (non apprende) e la necessità di conoscenza tecnica del prodotto.

1.4.6 Ibrido

I recommender system ibridi nascono dall'esigenza di alleviare gli svantaggi di alcuni approcci e sfruttare i vantaggi di altri in un unico strumento. Sono la sintesi degli approcci collaborative, content-based e knowledge-based, di cui vengono integrate le funzionalità di maggiore interesse generando un sistema con un metodo paragonabile a quello convenzionale dell'analisi ensemble, in cui la potenza di più tipi di algoritmi di apprendimento automatico viene combinata per creare un modello più robusto. Le modalità con cui si sintetizzano le caratteristiche dei diversi sistemi possono variare e in base a ciò se ne distinguono diversi tipi. I sistemi ponderati, ad esempio, assegnano punteggi alle diverse tecniche di raccomandazione combinate insieme per produrre una singola raccomandazione a partire dal punteggio complessivo. I sistemi basati su commutazione invece passano da una tecnica di raccomandazione all'altra a seconda del contesto applicativo mentre quelli misti producono raccomandazioni provenienti da più raccomandatori diversi e presentate contemporaneamente. Esistono infine approcci che fondono caratteristiche provenienti da sorgenti di raccomandazione diverse in un unico algoritmo di raccomandazione come, ad esempio, i sistemi a cascata in cui un raccomandatore perfeziona le raccomandazioni fornite da un altro prendendo in input l'output del modello a monte [8].

Contesto applicativo

Il seguente elaborato nasce da un lavoro progettuale svolto presso una delle sedi francesi di Avnet, distributore leader nel mercato dei componenti elettronici. Il contesto in cui si opera è quello del mercato dell'elettronica che, come risulta facile immaginare, trova ambiti applicativi tra i settori più disparati dall'automotive al settore industriale, dalla medicina alla difesa militare, e vive al contempo una costante e rapida evoluzione data la diffusione dell'elettronica in ogni settore dell'economia e dalla necessità di assistere l'offerta di prodotti per le più svariate applicazioni con l'obiettivo di garantire soluzioni specifiche e al passo con l'innovazione tecnologica. In particolare, l'azienda opera nel settore con il ruolo di distributore assistendo al contempo i clienti nelle varie fasi della progettazione e supportandone la personalizzazione progettuale a partire da requisiti specifici fino alla consegna del prodotto.

Le attività principali rimangono incentrate tuttavia sul processo di supply chain con l'obiettivo di garantire efficienza nella disponibilità di prodotti e materiali e soddisfazione dei clienti riguardo la redditività complessiva, offrendo canali di acquisizione globali e prevedendo i materiali utilizzati in quantità e tipologia al fine di ridurre il più possibile le tempistiche per i clienti.

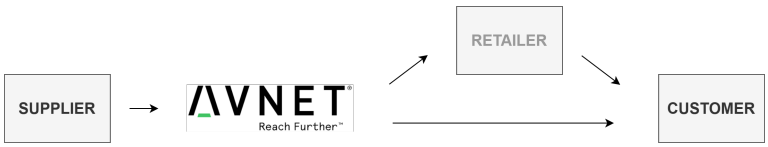

Fig. 2.1: Avnet nella supply chain.

2.1 Avnet

L'azienda nasce nei primi anni '20 quando il fondatore Charles Avnet inizia ad acquistare parti di radio in eccedenza e a venderle al pubblico nei mercati delle città portuali degli Stati Uniti. A metà di quegli anni, quando le radio prodotte in fabbrica cominciano a sostituire i ricambi radiofonici, Avnet modifica la sua linea di distribuzione e inizia a vendere i ricambi a produttori e rivenditori. Durante la Grande Depressione si sposta l'attenzione dalla vendita al dettaglio a quella all'ingrosso e da qui ha inizio la fase di diversificazione, espandendo la propria attività ai kit per autoradio e ai kit di montaggio per automobili. Durante la Seconda Guerra Mondiale Avnet produce antenne per le forze armate statunitensi e, successivamente all'apertura di un secondo impianto di assemblaggio di connettori per l'industria aeronautica, l'azienda viene quotata in borsa. Negli anni '70, Avnet si espande con diverse acquisizioni nei nuovi settori dei semiconduttori, dei relè e dei potenziometri, distributore a valore aggiunto di prodotti informatici e il più grande cliente/partner di IBM. Negli anni a seguire l'azienda ha acquisito oltre 60 società raggiungendo il mercato globale. Attualmente, la divisione EMEA ingloba in se tre principali aziende: Abacus, EBV e Silica; la prima incentrata principalmente sulla vendita di componenti passivi mentre le ultime due operano nel settore dei componenti elettronici attivi. [9]

2.2 Il mercato in cui si opera

Come si evince dalla sintesi sull'evoluzione storica dell'azienda i settori da cui i clienti provengo sono dei più disparati e questo non soltanto per la storia di espansione avvenuta negli anni ma anche per l'utilizzo sempre più massivo dell'elettronica nei più svariati settori dell'economia. Più in particolare l'azienda opera nei seg-

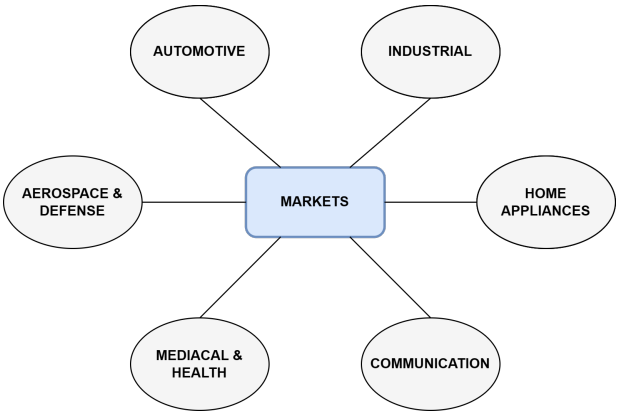

Fig. 2.2: I principali settori nel mercato dell'elettronica.

menti dell'electronics marketing, commercializzando e vendendo semiconduttori, interconnessioni, dispositivi passivi, componenti elettromeccanici, prodotti embedded, e nel segmento delle technology solutions, offrendo prodotti come server, soluzioni di archiviazione dati, software e soluzioni per l'implementazione dei prodotti.

Riguardo il segmento dell'electronics marketing, su cui ci si focalizzerà nel seguente lavoro, è possibile identificare 6 macrosettori che di seguito verranno indicati come markets, e all'interno dei quali poi sussistono svariate sottocategorie indicate invece col nome di applications. La figura 10 rappresenta quanto appena descritto.

2.3 Sfide progettuali

In tale contesto le sfide da affrontare nell'ambito della knowledge discovery e del data mining risultano molteplici e hanno come obiettivo comune quello di supportare al meglio i clienti in fase di acquisto proponendo alternative e prodotti complementari con la maggior precisione possibile. Nello specifico la ricerca di conoscenza nei dati mira a sviluppare meccanismi per il suggerimento di:

- componenti alternativi
- componenti correlati

e tali obiettivi risultano fortemente dipendenti da molteplici fattori inerenti ai clienti, agli ambiti applicativi, alle specifiche tecniche. Un'analisi dei clienti permette infatti di individuare similarità di comportamento nell'acquisto e quindi di raggruppare clienti con la possibilità di utilizzare tale informazione come vincolo nella generazione delle proposte. Riguardo gli ambiti applicativi la capacità di individuare il mercato di afferenza di determinati clienti e componenti determina una precisione maggiore nei suggerimenti sfruttando la conoscenza del settore applicativo per filtrare proposte effettivamente valide e non generate esclusivamente considerando similarità nei comportamenti d'acquisto. Altro fattore chiave per garantire qualità dei suggerimenti è l'aspetto puramente tecnico dei componenti e in tal caso ci si imbatte in una delle problematiche più grandi da gestire dato l'alto livello di granularità nelle specifiche fisiche dei componenti che rende quindi tale informazione poco utilizzabile dagli strumenti classici del data mining.

Sorgenti informative

Nel seguente capitolo si presentano le sorgenti informative a disposizione descrivendone inizialmente il valore informativo e le caratteristiche principali per poi passare ad una trattazione approfondita del grafo generato e delle attività di pulizia effettuate prima e dopo la costruzione.

3.1 Caratteristiche e valore informativo

La sorgente dati a disposizione è rappresentata da una vasta base di dati che raccoglie informazioni generali su clienti, prodotti, applicazioni, fornitori così come informazioni su tutte le attività di interazione con i clienti a partire dalla progettazione fino alla vendita. Oltre alle tabelle relative ai clienti o ai componenti o ai fornitori e che contengono informazioni specifiche sulle diverse entità, sono state selezionate le quattro tabelle che legano effettivamente un cliente a dei componenti e che rappresentano relazioni differenti e con diverso peso a seconda dell'attività di analisi da svolgere. Le tabelle in questione rappresentano:

- progetti registrati
- richieste di preventivo
- ordini
- fatture

3.1.1 Design Registrable Management System (DRMS)

La vendita di componenti agli Original Equipment Manufacturer (OEM), cioè i clienti finale del distributore, richiede una notevole quantità di tempo dalle prime fasi della collaborazione alla progettazione. Ad esempio, un distributore può scoprire che un produttore di computer sta pianificando la produzione di un nuovo server che potrebbe utilizzare uno dei chip utilizzati dal fornitore di componenti. Il distributore passerà mesi a collaborare con l'OEM in modo che quest'ultimo scelga il chip del fornitore per il server. Il distributore, tramite i propri Field Application Engineer (FAE), o collaborando con le case di progettazione, sviluppa soluzioni per i prodotti degli OEM e per proteggersi dalla concorrenza durante questo lungo periodo di progettazione, registra i progetti. Il processo di registrazione segue diverse fasi:

1. si presenta una registrazione per sviluppare un progetto con un componente specifico per un prodotto specifico pianificato dall'OEM;

2. il fornitore del componente valuta se approvare la registrazione;

3. approvazione (o rifiuto) delprogetto da registrare.

In caso di approvazione gli altri distributori non riceveranno incentivi finanziari dal fornitore del componente quindi non competeranno nello sviluppo del progetto registrato sul prodotto in questione. Se il distributore infine esegue il lavoro di progettazione necessario e riesce a convincere l'OEM a utilizzare il componente nel prodotto si aggiornerà lo stato della registrazione come vinta. Tale tabella ha un valore importante poiché, a prescindere dall'esito finale della registrazione che può dipendere da svariati fattori, contiene un'informazione assente nelle altre tre tabelle che è il tipo di application a cui il progetto appartiene, ovvero il sottotipo di segmento di mercato in cui si opera e che, in fase di analisi, aggiunge un aspetto distintivo alla

relazione tra cliente e prodotto. Va sottolineato tuttavia che, nonostante all'interno di uno specifico progetto un prodotto è associato ad una certa applicazione, questo può essere utilizzato in progetti diversi afferenti a settori differenti e quindi non si può inferire nulla sull'applicazione generica di un prodotto. Tra gli attributi di maggiore interesse abbiamo quindi l'identificatore di progetto, il cliente e il fornitore coinvolti, il riferimento temporale di creazione, il tipo di application e naturalmente i componenti.

3.1.2 Quotes

I preventivi rappresentano un impegno formale a fornire determinati prodotti o servizi ad un acquirente. Lo scopo è quello di proteggere il cliente dalle fluttuazioni dei prezzi e per questo motivo un preventivo è valido solo per un certo periodo di tempo. Dal punto di vista del valore che tale dato ha in ottica di analisi in effetti questo rappresenta un interesse del cliente per uno o più prodotti ma senza alcuna informazione correlata sul campo di utilizzo di quel componente o sulla tipologia di progetto.

3.1.3 Ordini e fatture

Nel caso degli ordini e delle fatture, anche se spesso sono fortemente correlati, non necessariamente ad un ordine corrisponde una fattura come nel caso degli ordini annullati o modificati ad esempio. Per tale motivo sono considerati separatamente e hanno valore diverso. Un ordine, infatti, dal punto di vista del valore informativo che questo possiede, rappresenta una relazione molto forte tra il cliente e i componenti dato che l'ordine è la parte conclusiva di un processo di trattativa che, come visto, comprende in alcuni casi anche una fase di consulenza per la progettazione e che assicura il reale interesse da parte del cliente. Per le fatture valgono le stesse

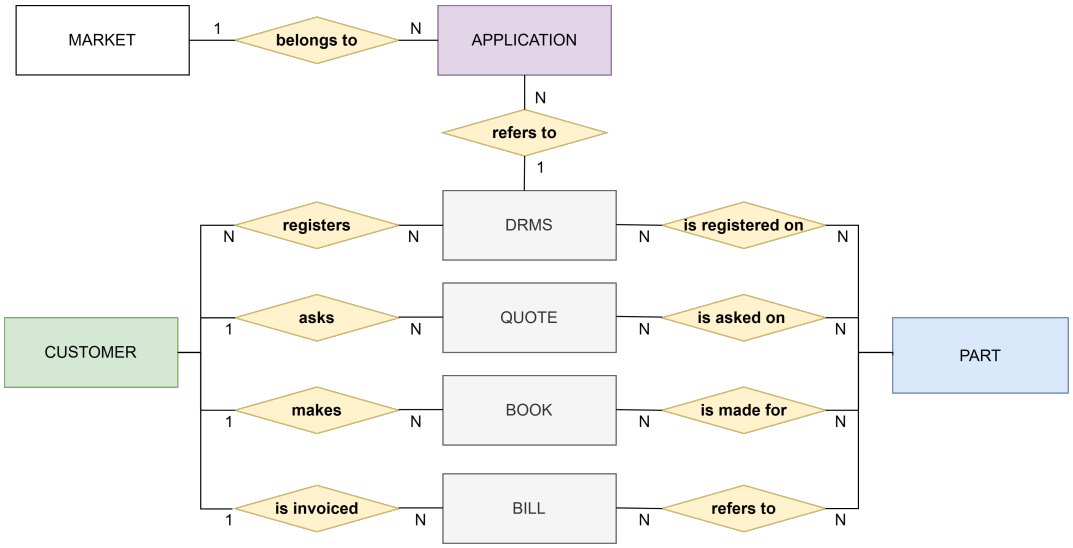

Fig. 3.1: Schema E-R parziale delle sorgenti informative.

considerazioni.

3.1.4 Schema E-R riepilogativo

Si riporta in Fig.3.1 lo schema parziale Entity-Relationship in cui sono considerate esclusivamente le entità e le relazioni di interesse.

Tramite lo schema risulta più facile visualizzare come le informazioni presentate nei sottoparagrafi precedenti siano effettivamente collegate tra loro avendo quindi una visione d'insieme del contesto informativo a disposizione.

3.2 Dalle tabelle al grafo

A partire dai dati a disposizione in forma tabellare si è costruito un grafo che contenesse le relazioni descritte nel paragrafo precedente.

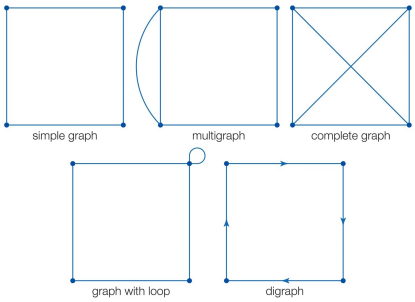

Fig. 3.2: Tipologie di grafo.

3.2.1 Richiami sulla teoria dei grafi

Un grafo è una collezione di vertici e archi. In matematica, il termine "collezione" è generalmente usato per indicare un multiinsieme, cioè un insieme in cui l'ordine è ignorato ma la molteplicità è significativa. I vertici, o nodi, rappresentano generalmente l'entità o le entità fondamentali in un certo contesto mentre gli archi rappresentano una relazione tra questi. A seconda di alcune caratteristiche riguardanti gli archi è possibile distinguere grafi: diretti (o indiretti), semplici (o multigrafi), connessi, completi. Un grafo è diretto se ogni arco ha una specifica direzione ed è possibile distinguere un nodo di partenza e uno di arrivo. Si parla di grafi semplici invece quando ogni coppia di nodi è connessa al più da un arco mentre si ha un multigrafo quando è ammesso più di un arco tra la stessa coppia di nodi. Si parla di grafi connessi quando tutti i nodi risultano connessi tra loro direttamente o per mezzo di una successione di altri archi mentre un grafo si dice completo quando ogni coppia di vertici risulta direttamente connessa.

Esistono infine due proprietà principali di un grafo: i gradi e i cammini. Il grado rappresenta il numero di vertici a cui un particolare vertice è connesso mentre il cammino si riferisce ai diversi percorsi che possono essere intrapresi per iniziare da un vertice e terminare in un altro vertice o nello stesso vertice. Tale breve richiamo permette di motivare con chiarezza alcune scelte adottate nella costruzione del grafo

in funzione degli obiettivi preposti.

3.2.2 Approccio alla costruzione

Si è scelto di seguire un approccio che inglobasse in un unico grafo tutte le informazioni di interesse e che contenesse tutte le tipologie di nodi anziché produrre diversi grafi sconnessi tra loro in cui si consideravano, per coppie di tipologia di nodo, solo i tipi selezionati. L'approccio seguito risulta sempre riconducibile all'alternativa frammentata filtrando i tipi di nodo da considerare ma permette allo stesso tempo, rispetto all'altra, di sfruttare una topologia più complessa, soprattutto nella generazione dei cammini. Un nodo può rappresentare quindi:

1. un componente;
2. un cliente;
3. un'applicazione.

mentre, riguardo gli archi, questi sono generati sempre a partire da progetti registrati nel DRMS, piuttosto che da preventivi, ordini o fatture. La generazione degli archi rispetta i seguenti vincoli:

1. i componenti appartenenti ad uno stesso progetto, preventivo, ordine o fattura sono collegati tra loro;
2. il cliente associato ad un certo progetto (o preventivo, ordine, fattura) viene collegato a tutti i componenti appartenenti al progetto;
3. l'applicazione a cui il progetto afferisce è collegata ad ogni componente appartenente al progetto.

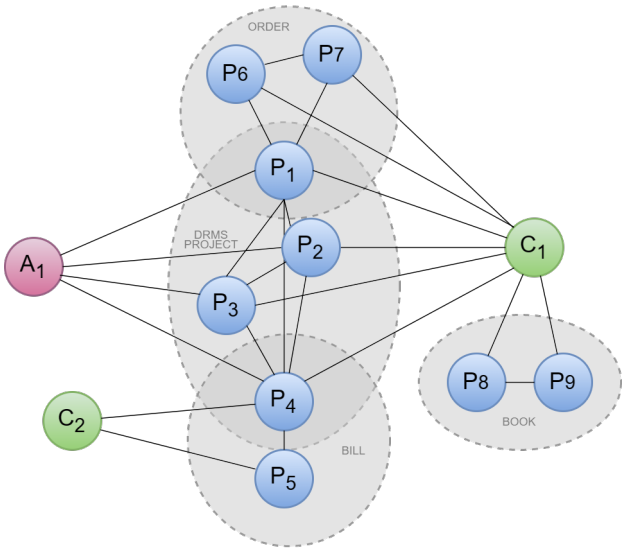

Fig. 3.3: Esempio di possibile interconnessione delle diverse tipologie di nodo.

Si noti che, come è possibile notare anche dallo schema E-R riportato in Fig.3.1, solo nel caso di un progetto registrato nel DRMS è possibile conoscere il tipo di applicazione. Si è costruita quindi la rappresentazione tabellare degli archi del grafo a partire dalle quattro tabelle discusse nel paragrafo 3.1 e secondo i vincoli e le regole appena descritte. La tabella risultante è data da quattro colonne: due che descrivono i nodi che l'arco collega, una colonna *Source* che descrive da quale tabella proviene la relazione tra i due nodi e una quarta colonna *Description* che esprime la natura dei due nodi; ogni nodo può essere un componente (part) etichettato con P, un cliente etichettato con C o una application etichettata con A, per cui i valori che l'attributo può assumere sono tutte le possibili combinazioni di cardinalità 2 dell'insieme A,C,P escluse le combinazioni AA e CC non possibili per costruzione. In Fig.3.3 si riporta un generico esempio esplicativo della struttura del grafo ottenuta a seguito della costruzione appena descritta.

Il grafo risultante è in particolare un grafo eterogeneo non orientato. Il vantaggio ottenuto dalla sintesi delle varie tabelle nel grafo è proprio quella di avere in unica struttura e correlati tra loro i dati di interesse, aggiungendo relazioni tra oggetti

Nodi	212.881
Archi	11.139.695
Componenti connesse	182
Densità	0,000495

Table 3.1: Caratteristiche principali del grafo.

precedentemente scollegati.

3.2.3 Caratteristiche generali del grafo

Il grafo analizzato è stato generato dopo aver selezionato una finestra temporale e i dati relativi a una delle tre società afferenti alla divisione EMEA, in particolare una di quelle focalizzate sulla vendita di componenti attivi. L'approccio seguito per l'identificazione di un arco temporale è stato quello di fissare inizialmente un limite massimo di 4 anni fiscali, data l'obsolescenza dei componenti dopo quest'arco temporale e analizzare il conteggio dei componenti connessi all'aumentare del periodo selezionato. Dopo aver notato un andamento costante nell'evoluzione deli componenti connessi a 3 anni dalla data di inizio si è scelto questo periodo.

Nella Tab.3.1 sono riportate le metriche di base del grafo. Riguardo il numero di componenti connesse, dopo aver studiato le dimensioni di ogni componente in diversi contesti, è emerso come fosse sempre presente una componente contenente quasi la totalità dei nodi del grafo sconnesso e una serie di componenti molto più piccole e isolate. Si è deciso quindi di considerare la componente connessa più grande dato il divario significativo in termini di cardinalità tra i componenti e la necessità di analisi basata anche sull'utilizzo di cammini.

Nella Tab.3.2 si riportano le metriche di base del componente connesso più grande.

Nodi	212.397
Archi	11.138.992
Densità	0,000493

Table 3.2: Caratteristiche principali del componente connesso più grande.

3.3 Data cleaning

All'attività di generazione effettiva del grafo è seguita un'analisi dettagliata dei dati considerati al fine di eliminare rumore nelle sorgenti informative e migliorare la qualità dei dati.

3.3.1 Analisi sui clienti

Nel contesto applicativo di riferimento in cui è possibile avere tipi di clienti profondamente differenti sia in termini di mercato che di ruolo ricoperto all'interno di questo, è necessario filtrare quelle tipologie su cui non c'è interesse a produrre suggerimenti. In particolare, è necessario identificare quelle società comunemente identificate con l'acronimo EMS, ovvero Electronics Manufacturing Services, che rappresentano realtà aziendali che si occupano prevalentemente di collaudare, distribuire, riparare e assemblare componenti elettronici per i produttori (OEM). Questi clienti introducono rumore nei dati poiché acquistano prodotti in quantità e tipologie anomale dato che gli obiettivi sono spesso quelli di rivenderli o utilizzarli per assemblare sistemi per molteplici clienti che operano in settori spesso differenti. In particolare, l'identificazione e l'eliminazione degli EMS è avvenuta:

1. tramite le informazioni a disposizione sui clienti

2. analizzando il grafo

Riguardo il primo punto si sono individuati classificazioni specifiche dei clienti con ruolo di subcontractor, broker, EMS, e così via. A proposito del secondo approccio invece si è utilizzata la struttura del grafo andando a studiare il grado dei nodi clienti e analizzando quelli sopra una certa soglia. A partire da questi si è effettuata una selezione manuale cercando informazioni maggiori su quei clienti sia internamente all'azienda che online. Riguardo l'idea di utilizzare il conteggio degli archi per ogni cliente si ipotizza che questo tipo specifico di utente abbia un numero di connessioni molto elevato dato che, dai comportamenti di acquisto, emerge come spesso acquisti in grande quantità una grande varietà di prodotti.

3.3.2 Analisi sui progetti

Riguardo ai progetti registrati nel DRMS durante il periodo selezionato si è notato come alcuni di questi abbiano associati più clienti e nello specifico:

- il 78% ha un solo cliente
- il 16% ne ha 2
- il 5% ne ha 3
- l'1% più di 3

In totale quindi circa il 22% dei progetti ha più di un cliente associato e per questi specifici progetti si è deciso di verificare se i clienti in questione fossero anche associati a progetti con un solo cliente. In caso contrario sono stati eliminati verificando prima quale fosse l'impatto in termini di dati persi riguardo preventivi, ordini e fatture.

3.3.3 Identificatori dei componenti

I dati relativi al DRMS risultano essere i più incompleti e sporchi poichè inseriti per la maggior parte manualmente dai FAE (Field Application Engineer) o dagli

AM (Account Manager). Questo comporta inconsistenza sia all'interno delle stesse tabelle che tra tabelle differenti. Il problema da gestire riguarda in particolare l'attributo utilizzato per identificare i componenti; questo presenta la seguente struttura:

codice produttore (3 lettere) + *codice identificativo componente*

e il secondo addendo varia in funzione delle caratteristiche del componente. Per tale ragione, nel caso in cui si voglia identificare una serie di componenti piuttosto che un singolo specifico, in fase di registrazione nel DRMS, un'utente (FAE/AM) può inserire identificativi composti in parte anche da asterischi, esprimendo una certa genericità nella scelta. In tali casi, tuttavia, essendo l'identificativo parzialmente valorizzato risulta impossibile utilizzare l'informazione per effettuare eventuali join sull'identificatore. Per evitare di perdere tutti i record di questo tipo, dopo aver ripulito gli identificatori parziali da ulteriori caratteri speciali o da formati particolari, si è effettuato una join condizionata su un costrutto like con la tabella dei componenti. Si è raggruppato il risultato per identificatore parziale ottenendo per la maggior parte dei record un mapping su un singolo componente nella tabella dei materiali, ovvero la tabella contenente ogni dettaglio sui componenti con il relativo identificatore completo, mentre per i casi con più match è stato scelto di utilizzare l'edit distance come discriminante tra le alternative possibili. In particolare, è stata utilizzata la distanza di Levenshtein, calcolata tra ogni alternativa e l'identificatore completo, che esprime la differenza tra due stringhe in funzione del numero minimo di operazioni di inserimento, cancellazione e sostituzione che le rende uguali.

Data analysis: duplice approccio

Nel capitolo di seguito si presentano i maggiori obiettivi di analisi e i due approcci seguiti per raggiungerli, descrivendone nel dettaglio i principi di funzionamento e il significato analitico nel contesto di riferimento.

4.1 Obiettivi

Come brevemente anticipato nel paragrafo 2.3 sulle sfide progettuali, il recommender system permette di assistere le figure preposte alle vendite supportandole nella ricerca di:

- prodotti simili
- prodotti correlati

ad uno prodotto sorgente di riferimento. In particolare, riguardo la ricerca di prodotti simili, da qui in avanti indicata come cross reference, questa si declina cercando le possibili alternative ad un prodotto di partenza tali per cui vengano conservate il più possibile le caratteristiche tecniche e l'ambiente di utilizzo. Quest'ultima caratteristica risulta fondamentale dato che spesso, anche se a livello macroscopico la funzione del componente è la stessa, cambia l'ambito di applicazione e di conseguenza cambiano le specifiche del prodotto. Si pensi, ad esempio, ad un sensore di temperatura: se usato nella realizzazione di un termostato per casa o all'interno

di un meccanismo di sicurezza per un aereo, anche se ha la stessa funzione, avrà specifiche e livello di complessità ben diverse. Alcune volte ma non sempre una caratteristica distintiva utilizzabile per distinguere i contesti applicativi di un prodotto è il prezzo ma questo può dipendere da una serie di ulteriori fattori non trascurabili che lo rendono difficile da utilizzare per questo scopo. A proposito della ricerca di prodotti correlati invece, da qui in avanti indicata come cross selling, questa si focalizza sull'individuazione di componenti che possono essere utilizzate insieme e che rappresentano entità complementari all'interno di uno stesso progetto. Un esempio fortemente esplicativo si ha nel caso in cui si stia progettando un sistema che utilizzi un microcontrollore a cui è necessario connettere almeno una memoria e un alimentatore; in tal caso risulta ovvio il suggerimento degli altri due componenti dato che si sa che questi sono funzionalmente necessari all'interno di un progetto che utilizzi il microcontrollore. Tuttavia, nella maggior parte dei casi non è così facile e immediato trovare delle correlazioni a partire da un certo componente. La difficoltà, come per il caso precedente, è aumentata dal livello di specificità richiesto che rende due prodotti correlati dal punto di vista funzionale a livello macroscopico ma praticamente non utilizzabili insieme dati vincoli sulle specifiche tecniche che li rende incompatibili. Si presentano di seguito i due diversi approcci all'utilizzo del grafo che sono stati seguiti per proporre soluzioni alle due problematiche appena descritte.

4.2 Utilizzo diretto del grafo

Questo primo approccio è basato sull'utilizzo diretto del grafo. Consiste nel calcolare delle metriche di similarità tra nodi tutte basate sull'analisi dei vicini del nodo considerato. Dopo alcune considerazioni sulla struttura del grafo, si passa alla descrizione delle caratteristiche misurate e di ciò che rappresentano, approfondendo il

loro significato nel contesto d'interesse.

4.2.1 Neighbors

Dato un nodo, i vicini di questo sono tutti i nodi ad esso direttamente collegati. A seconda del numero di archi da attraversare per raggiungere, col percorso più breve possibile, un vicino, si distingue l'ordine del vicinato. Nel grafo analizzato, come descritto in dettaglio nel paragrafo 3.2 sull'approccio alla costruzione, dato un nodo, i suoi vicini possono essere:

- componenti
- clienti
- applicazioni

In particolare, nel caso dei componenti, dati due componenti connessi deve esistere almeno un progetto, un preventivo, un ordine o una fattura in cui questi componenti compaiono insieme. Riguardo ai clienti, sempre per costruzione, una connessione tra cliente e componente dipende dall'esistenza di almeno uno dei quattro documenti che li collega. Un'applicazione infine è presente tra i vicini se esiste almeno un progetto registrato nel DRMS di quel tipo specifico che contiene il componente. Generalmente quelle basate sui vicini sono tipologie di analisi più qualitative che quantitative dato che si tiene conto dell'identità dei nodi a differenza dell'analisi sulle centralità che è costruita sul numero di connessioni.

4.2.2 Metriche utilizzate

Le metriche di similarità che sono state considerate sono il coefficiente di Jaccard e l'indice Adamic-Adar che risultano le misure maggiormente utilizzate nell'ambito

dell'information retrival [10]. Entrambe sono basate sul metodo dei Common Neighbours e mentre la prima è più in generale una misura statistica la seconda è una metrica introdotta per la predizione di collegamenti all'interno di un grafo [11].

Jaccard coefficient

Come anticipato, il coefficiente di Jaccard è una msiura statistica utilizzata per confrontare la somiglianza tra insiemi di campioni. Si indica come *J(x, y)* dove x e y rappresentano due nodi di una rete. Nella previsione di similarità, tutti i vicini di un nodo vengono considerati come un insieme e il valore del coefficiente viene calcolato come l'intersezione e l'unione dei due insiemi. Considerato con N il numero di nodi di cui calcolare la similarità e con k il grado medio di un nodo, la complessità del metodo è pari a $\Theta\left(N^2 * k^2\right)$ [11]. Si riporta di seguito la formula del coefficiente:

$$J(x, y) = \frac{\|N(x) \cap N(y)\|}{\|N(x) \cup N(y)\|} \tag{4.1}$$

indicando con *N(x)* l'insieme dei vicini del nodo x. Considerati gli obiettivi di analisi, si è calcolato tale coefficiente per ogni coppia di nodi associati a componenti di tipo non-commodity evitando di eseguire il calcolo per ogni coppia di nodi nel grafo.

Adamic-Adar index

È una misura inizialmente progettata per misurare la relazione tra siti web biografici. Il meccanismo di funzionamento prevede che più è alto il grado di un nodo comune al vicinato dei due nodi analizzati, più basso è il punteggio che gli viene assegnato. Pertanto, il vicino comune di una coppia di nodi con pochi vicini contribuisce maggiormente al valore finale dell'indice rispetto ai nodi con un numero elevato di vicini. In una rete sociale il significato di tale misura è che, se un conoscente comune ha più

amici, allora è meno probabile che presenti i due conoscenti rispetto al caso in cui abbia pochi amici. Riguardo al contesto specifico del recommender system, in base alle specifiche di costruzione del grafo, tale misura di similarità attribuisce quindi più importanza a componenti che sono presenti in un unico progetto, o equivalentemente ordine o fattura, rispetto a quelli connessi a molteplici progetti. Come la Jaccard similarity, è un metodo che si basa sul vicinato comune. Considerato sempre N il numero di nodi di cui calcolare la similarità e k il grado medio di un nodo, la complessità è $\Theta\left(N^2\right)$ [11]. Si riporta di seguito la formula dell'indice:

$$Adamic - Adar(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{log(N(u))} \tag{4.2}$$

con x e y i nodi di cui interessa calcolare la similarità, *N(x)* l'insieme dei vicini di x ed u un generico vicino comune a x e y. Anche in questo caso si è calcolata si è calcolato tale coefficiente per ogni coppia di nodi di tipo non-commodity.

4.3 Deep walk

Il secondo approccio seguito si basa invece su un metodo di machine learning noto come Deep Walk. Dopo aver richiamato la struttura del grafo, si passa a presentare nel dettaglio il meccanismo di funzionamento del metodo e le diverse fasi di implementazione seguite. Si discute il processo di costruzione dei cammini e, in relazione a questo, la distinzione tra componenti di base indicati come commodity e componenti specifici denominati non-commodity descrivendo il metodo di filtraggio utilizzato per differenziarli nel dataset.

4.3.1 Struttura del grafo

Come anticipato nel paragrafo 3.2 riguardo le caratteristiche principali del grafo, si è osservato come questo presentasse diverse componenti connesse delle quali una estremamente più grande delle altre e contenente il 99,7% dei nodi nel grafo originario. Si è deciso di considerare per l'analisi la componente connessa più grande per diverse ragioni:

- la necessità di utilizzare un grafo connesso per l'approccio Deep Walk
- la possibilità di confrontare i due approcci sullo stesso input
- la differenza approssimativamente nulla tra il contenuto informativo del grafo originario e della componente connessa più grande

4.3.2 Il metodo in dettaglio

Il Deep Walk [12] è un metodo di apprendimento non supervisionato che sfrutta tecniche di Natural Language Processing (NLP) per l'analisi di una rete. Le tecniche di NLP utilizzate nel metodo si sono dimostrate particolarmente efficaci per problemi di processamento del linguaggio naturale e, per tale motivo, il metodo si propone di ricondurre l'attività di analisi e apprendimento di una rete ad un problema di processamento del linguaggio. Più in dettaglio, l'algoritmo apprende ciò che è definito come social representation [12], ovvero una rappresentazione di un nodo rispetto gli altri nodi del grafo, modellando un flusso di cammini casuali (random walk). Le rappresentazioni sociali sono definite caratteristiche latenti dei vertici poiché non esplicitamente osservabili nel grafo e tendono a catturare la somiglianza di vicinato e l'appartenenza alla comunità. I nodi espressi attraverso una serie di random walk vengono codificati in uno spazio vettoriale continuo di dimensione ridotta rispetto la lunghezza dei cammini, ottenendo quindi una rappresentazione vettoriale per ogni

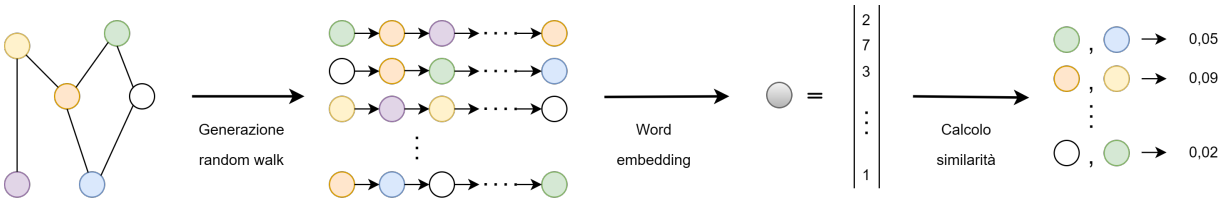

Fig. 4.1: Schema di funzionamento dell'analisi basata su Deep Walk.

nodo attraverso un processo di word embedding. Le due fasi principali in cui il metodo si articola sono:

1. generazione dei cammini

2. rappresentazione vettoriale a partire dai cammini

In Fig.4.1 si è sintetizzato il processo contestualizzandolo nel caso di interesse ovvero per il calcolo della similarità tra nodi.

Random walk

Un random walk è un cammino che inizia da un vertice e ad ogni step si sposta, con una scelta causale, su uno dei vicini del nodo precedentemente attraversato. Normalmente quando il grafo non è pesato, il vertice verso cui si muove il cammino è scelto uniformemente a caso tra i vicini del vertice attuale. Al contrario, quando il grafo presenta archi a cui è associato un peso, si sposta verso un vicino con una probabilità proporzionale al peso dell'arco corrispondente. Dalla precedente definizione si può quindi immaginare come sia possibile rappresentare un cammino come una sequenza di nodi in cui ogni nodo è legato al precedente e al successivo poiché esiste un legame diretto tra questi. In particolare, questi vengono generati in maniera molto semplice: partendo dal nodo radice, si sceglie casualmente un vicino di quel nodo e lo si aggiunge al percorso, poi si sceglie casualmente un vicino del nodo appena scelto e si continua con il cammino fino a quando non si è raggiunta la

lunghezza del cammino desiderata. Alla base del metodo c'è quindi la generazione di random walk per ogni nodo di cui si voglia ottenere una rappresentazione vettoriale. Tale processo è legato alla scelta di due parametri:

1. la lunghezza del cammino ovvero il numero s di nodi attraversati
2. il numero w di cammini per nodo

La scelta dei parametri risulta non ovvia e attualmente non esistono indicazioni che correlino alcune caratteristiche del grafo utilizzato con i valori dei parametri. Un'alternativa, nel caso in cui si avessero a disposizione dei valori attendibili di riferimento, potrebbe essere l'utilizzo di una grid search per identificare dei valori che minimizzano l'errore ma in questo caso non si hanno a disposizione dati etichettati con precisione e utili a tale scopo. Si è cercato allora di correlare i valori dei due parametri alle caratteristiche del grafo e solo nel caso del numero di cammini per nodo si è scelto il grado medio dei vertici ovvero il numero medio di vicini con l'idea di considerare mediamente tutti i vicini del nodo radice. Nel caso invece della lunghezza del cammino è stato mantenuto il valore di default utilizzato nella maggior parte delle implementazioni dell'algoritmo [13] [14].

Commodities e non-commodities

É necessario a questo punto introdurre una distinzione nelle componenti per descrivere, con maggior precisione, gli obiettivi di analisi e i vincoli introdotti nella generazione dei cammini. Si indicano come non-commodities quelle componenti come microcontrollori, memorie, Field Programmable Gate Array (FPGA), alimentatori, batterie, amplificatori che rappresentano elementi complessi di progettazione, con utilizzi specifici e che risultano di maggiore interesse poiché garantiscono un ritorno economico maggiore dato il costo più elevato. Al contrario si considerano

Material group	**Descrizione**
ANA101	General Purpouse Amplifier
MEM104	Flash Memory
OPT300	Transistor and Photovoltaic output- Photocouplers
...	...

Table 4.1: Esempi di *material group* con relativa descrizione.

commodities tutte quelle componenti come diodi, sensori, tiristori, transistor, che costituiscono elementi di base per la progettazione, godono di ampio utilizzo in tutti i progetti e hanno costi ridotti date le funzionalità elementari. Considerando allora i dati a disposizione è possibile ottenere un'informazione codificata sulla tassonomia di un componente a partire dall'attributo material group. Questo è dato da una sigla di tre lettere che identifica la macro-famiglia di componenti e da un codice alfanumerico che distingue sottocategorie all'interno della famiglia. Alcuni esempi sono riportati nella Tab.4.1 con la relativa descrizione funzionale del codice.

Tornando alla necessità di distinguere macroscopicamente i componenti di progettazione di base da quelli più specifici e complessi, l'utilizzo di questa informazione a tale scopo è risultata poco praticabile dato l'elevato numero di material group esistenti e l'assenza di una descrizione precisa come nel caso dei gruppi riportati nella Tab.4.1. Si è reso quindi necessario trovare un metodo alternativo per distinguere i prodotti di tipo commodity dalle non-commodities. A tal fine è stato utilizzato direttamente il grafo e in particolar il grado dei nodi. Considerando che:

- per costruzione, ogni componente di un progetto risulta collegata a tutti gli altri componenti nello stesso progetto,

- una commodity è presente quasi nella totalità dei progetti essendo un elemento di base di progettazione,

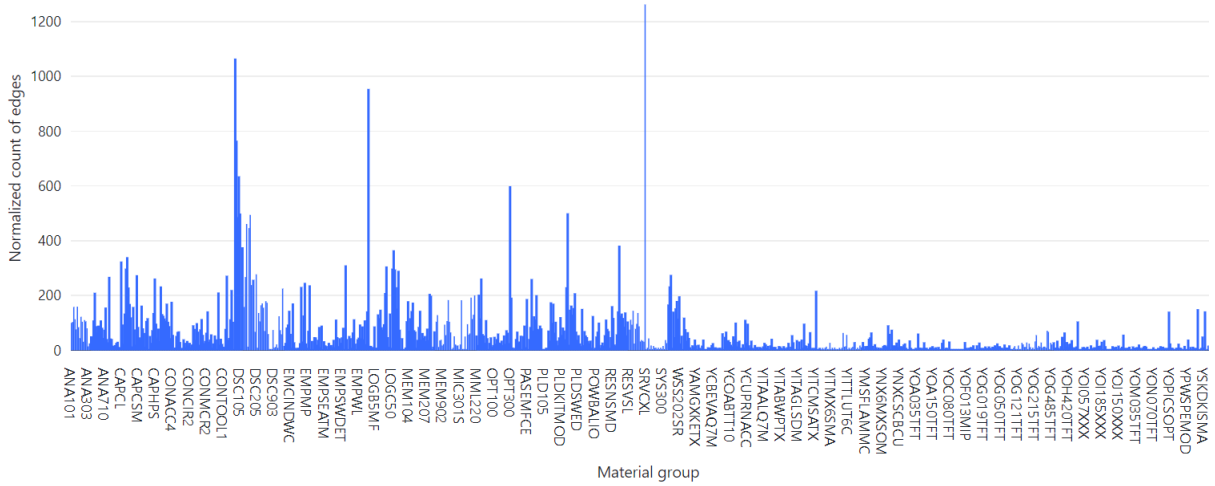

Fig. 4.2: Conteggio archi per material group normalizzato rispetto cardinalità gruppo.

le componenti di tipo commodity risultano estremamente connesse e di conseguenza, i relativi nodi, con un grado elevato. Si è quindi considerato il conteggio totale di archi per material group (Fig.4.2) con l'idea di separare i gruppi corrispondenti a valori elevati e quindi i material group di tipo commodty da quelli con conteggio inferiore. Per rendere tale misura indipendente dalla cardinalità dei gruppi ed evitare di avere quindi valori più alti per alcuni material group solo per l'elevato numero di componenti in essi, si è normalizzato il conteggio rispetto alla cardinalità del gruppo.

Al fine di determinare una soglia indicativa per categorizzare ogni material group si è generato allora un istogramma sul conteggio normalizzato e si è cercato di individuare un valore che segnasse il confine tra valori elevati e valori mediamente bassi, separando quindi tassonomie di tipo commodity dalle non-commodity. Osservando l'istogramma si può notare come la prima colonna corrispondente ai material group con conteggio normalizzato inferiore o uguale a 30 presenta un conteggio nettamente più elevato rispetto alla restante parte dell'istogramma. I material group appartenenti alla prima colonna possono quindi essere associati a componenti non-commodity secondo le considerazioni precedentemente fatte mentre i restanti gruppi,

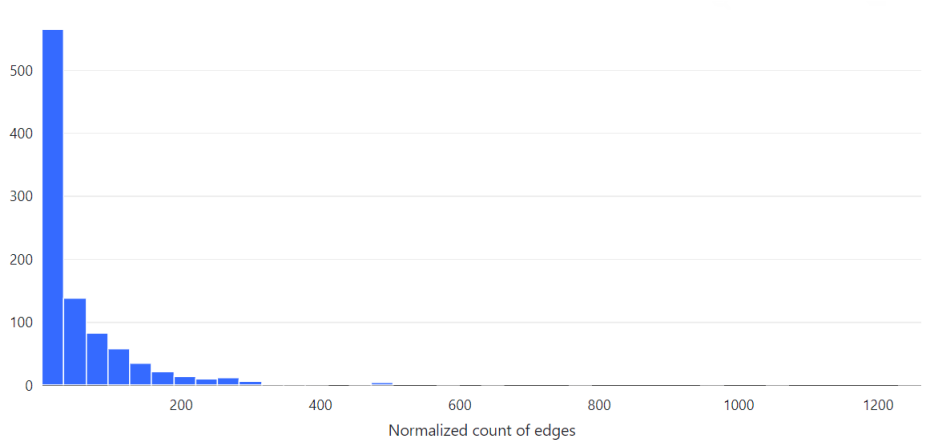

Fig. 4.3: Istogramma conteggio normalizzato degli archi per material group.

associati a conteggi più elevati, sono da considerarsi commodities.

Ricordando come gli obiettivi di utilizzo del recommender system siano incentrati sulla ricerca di prodotti simili e correlati, la differenziazione tra commodity e non-commodity ha permesso di concentrarsi su una specifica tipologia di prodotto, ovvero le non-commodity, di maggiore interesse e di generare cammini random ma con vincoli diversi in funzione della ricerca di alternative o correlazioni. Riguardo al secondo punto, più in particolare, si è deciso di introdurre un vincolo sulla probabilità di attraversare una commodity o no. Per costruzione infatti, come precedentemente descritto, i nodi commodity hanno generalmente un grado maggiore rispetto alle non-commodity e quindi la probabilità di finire su una commodity a partire da un qualsiasi nodo, in fase di costruzione di un cammino, è più alta. In aggiunta a questo è necessario introdurre due ulteriori considerazioni sui componenti commodity sulla base della loro natura di elementi di base di progettazione. In quanto tali, infatti:

1. sono quasi sempre presenti tra i componenti di un progetto o di un ordine e quindi condivisi tra progetti diversi;

2. uno specifico componente di tipo commodity ha bassissime probabilità di trovarsi nello stesso progetto, ordine, o ancor più preventivo, a cui è associata

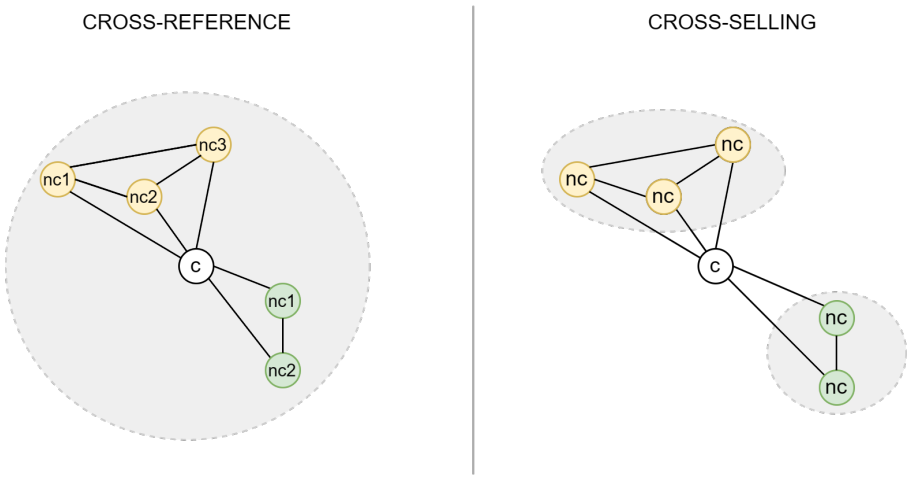

Fig. 4.4: Aree di interesse nella generazione dei cammini.

un'altra commodity con stesse funzionalità.

Sulla base del primo punto allora l'attraversamento di una commodity aumenta le probabilità di spostarsi da un progetto ad un altro dato che sono spesso condivise e rappresentante da nodi altamente connessi; per questo nella creazione dei cammini utilizzati nella ricerca di prodotti simili si è introdotto un vincolo sulla probabilità di attraversare un nodo commodity pari al 10%. Al contrario, per i random walk utilizzati per il cross-selling, ovvero la ricerca di prodotti correlati, c'è maggiore interesse a percorre i nodi di uno stesso progetto e, per questo, si è utilizzato un vincolo sulla probabilità di attraversare una commodity dell'1%, diminuendo così la probabilità di spostarsi tra i nodi di un nuovo progetto. Nella Fig.4.4 si riporta un esempio che descrive l'approccio utilizzato evidenziando le regioni in cui si vincola maggiormente la scelta dei nodi nella costruzione dei cammini.

Si noti come il vincolo introdotto è espresso sempre sulla scelta casuale dei nodi per questo la generazione random dei cammini è garantita. In conclusione, quindi, a seguito delle considerazioni sui parametri descritte nel sottoparagrafo precedente, si sono generati per ogni nodo 30 cammini di lunghezza fissa pari a 80.

Word2Vec

Nella seconda fase del metodo si genera, per ogni nodo, una rappresentazione vettoriale utilizzando i random walk generati. I cammini, infatti, in quanto sequenze di identificatori di nodi, sono associabili a vere e proprie frasi e permettono di essere analizzati utilizzando tecniche di Natural Language Processing per arrivare ad una rappresentazione vettoriale delle parole e quindi, in questo contesto particolare, dei nodi. Continuando l'analogia tra frasi e cammini, come il language modeling stima la probabilità che sia presente una specifica parola da una sequenza di parole in un corpus, allo stesso modo per il grafo si ha una stima della probabilità di osservare il vertice v, dati tutti i vertici precedenti visitati nella passeggiata casuale. In aggiunta a quanto appena detto il modello deve apprendere una rappresentazione vettoriale, non solo una distribuzione di probabilità di co-occorrenza delle parole, e per questo viene definita una funzione di mapping che rappresenta ogni parola (nel nostro caso ogni vertice del grafo) w definita dal prodotto di una matrice di parametri liberi $[V] \times d$, dove V è l'insieme di tutte le parole considerate, noto come vocabolario mentre d è un vettore di parametri indefiniti di dimensione pari a quella dell'embedding [12]. In particolare, grazie all'introduzione del modello Word2Vec e alla sua specifica implementazione data dal modello Skip-Gram [15], è stato invertito il problema della predizione che non utilizza più il contesto per prevedere una parola mancante ma una parola per prevedere il contesto. In aggiunta a ciò, con un impatto importante sull'applicabilità del metodo ai random walk, nel modello proposto in [15] si è eliminato il vincolo sull'ordine delle parole nel contesto, rendendo molto più applicabile il modello ai random walk.

Nella Fig.4.5 è rappresentata una struttura sintetica dell'architettura del modello. In conclusione, il modello Skip-Gram massimizza la probabilità di co-occorrenza tra le parole che compaiono all'interno della finestra w in una frase [14], restituendo

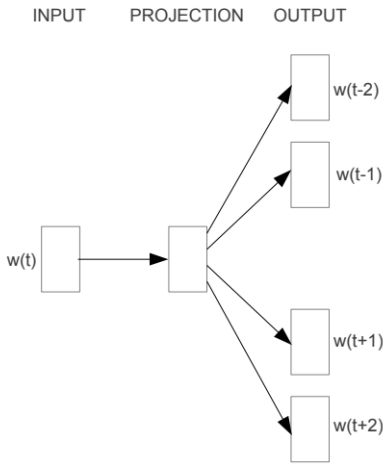

Fig. 4.5: Architettura del modello Skip-Gram.

la rappresentazione vettoriale delle parole in questo caso specifico rappresentate dagli identificatori dei nodi. Va sottolineato che un ulteriore vantaggio di tale modello in questo contesto è dato dalla possibilità di passare in input al modello per l'apprendimento un vertice alla volta, eventualmente ripetendo molteplici volte tale step con frasi differenti, o analogamente random walk multipli, per massimizzare la rappresentazione contestuale della parola o, equivalentemente, del vicinato di un nodo.

4.4 Misura della similarità

Ricordando l'obiettivo di partenza di calcolare la similarità tra due nodi, a partire dalla rappresentazione di ogni nodo come un vettore, si è considerato il calcolo della similarità tramite la cosine similarity.

Tale misura esprime il grado di somiglianza tra due vettori non nulli in funzione della distanza tra di essi nello spazio, la quale può essere espressa dal coseno dell'angolo generato tra i due vettori. I possibili valori che la misura assume sono

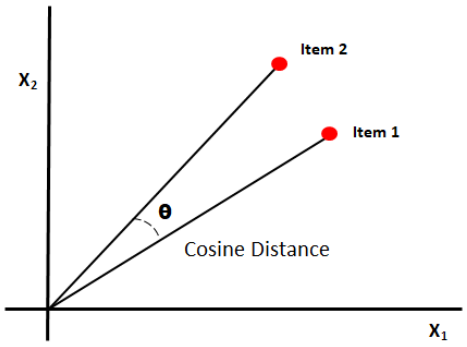

Fig. 4.6: Rappresentazione grafica della coseno-distanza.

compresi nell'intervallo [-1,1].

Implementazione e confronto degli approcci

In questo capitolo si analizzano le principali differenze dei due approcci descrivendo le implementazioni, analizzando i tempi di esecuzione e proponendo una valutazione parziale dei suggerimenti a partire dalle similarità calcolate. Tale confronto non ha tuttavia l'obiettivo di scegliere l'utilizzo di uno o l'altro approccio ma di valutarli nell'ottica di un utilizzo congiunto in sviluppi futuri.

5.1 Ambiente di sviluppo

Per analizzare e porre a confronto i due metodi si introduce l'ambiente di sviluppo utilizzato per l'implementazione. Lo sviluppo è avvenuto interamente sulla piattaforma Databricks, utilizzando l'engine Spark tramite l'API PySpark che espone i suoi servizi in Python. In particolare, tra i moduli di alto livello inclusi in Spark, si è utilizzato:

- Spark SQL, per la gestione delle tabelle come dataframe, che ha permesso di sfruttare le relative ottimizzazioni nella gestione delle query e nel calcolo delle funzioni di aggregazione
- MLlib, la libreria per il machine learning di Spark, che ha permesso di sfruttare il calcolo parallelo in fase di allenamento del modello Word2Vec

Riguardo le risorse di calcolo a disposizione invece si è utilizzato un cluster con 4

worker ciascuno equipaggiato con processore 4-cores da 2.5 GHz più 8GB di memoria RAM e un driver con processore 4-cores da 2.5 GHz e 14 GB di RAM.

5.2 Implementazione e tempi di esecuzione

5.2.1 Utilizzo diretto del grafo

Riguardo il primo approccio relativo al calcolo diretto delle metriche di similarità sul grafo, non avendo a disposizione un'implementazione delle metriche se non tramite la libreria NetworkX che non permette di sfruttare il calcolo parallelo e quindi le risorse a disposizione, si è implementato il calcolo delle metriche lavorando sui dataframe in modo da poter sfruttare il parallelismo. In particolare, per entrambe le metriche, si è costruito un dataframe che memorizza, per ogni nodo non-commodity, i suoi vicini a partire dal dataframe contenente gli archi del grafo, già presentato nel paragrafo 3.2. Nel caso del Jaccard coefficient, dopo aver selezionato dal dataframe relativo ai nodi e ai rispettivi vicini, i soli nodi non-commodity, si è calcolato il prodotto cartesiano ottenendo un dataframe con tutte le coppie di nodi non-commodity e i relativi vicini. A quest'ultimo si è quindi applicato, per ogni riga, il calcolo del coefficiente di Jaccard sfruttando lo sviluppo descritto nella formula 5.1. Il calcolo per circa 85.000 nodi non-commodity ha impiegato complessivamente 1,5 ore.

$$J(x,y) = \frac{\|N(x) \cap N(y)\|}{\|N(x) \cup N(y)\|} = \frac{\|N(x) \cap N(y)\|}{\|N(x)\| + \|N(y)\| - \|N(x) \cap N(y)\|} \tag{5.1}$$

Nel caso dell'indice Adamic-Adar invece, si è costruito, a partire dal dataframe sui vicini, un nuovo dataframe contenente per ogni nodo un dizionario contenente i vicini come chiavi e il grado di questi come valore. Si è quindi calcolato il prodotto

cartesiano, come nel caso precedente, e si infine calcolato l'indice applicando la funzione che lo implementa, ad ogni riga. In questo caso il calcolo dell'indice sullo stesso insieme di nodi non-commodity ha impiegato complessivamente 2 ore.

È importante sottolineare come, al fine di produrre un suggerimento, il valore di similarità è utilizzato congiuntamente all'informazione sul material group del componente, ovvero sulla sua natura macroscopica. Nello specifico:

- per il cross-referencing (ricerca di prodotti simili) si impone un vincolo di uguaglianza tra il material group della sorgente e quello dei target
- per il cross-selling (ricerca di prodotti correlati) si impone che il material group dei componenti target sia differente da quello della sorgente

5.2.2 Deep Walk

Anche nel caso del secondo approccio per sfruttare a pieno le risorse a disposizione si è implementata la generazione dei random walk utilizzando direttamente i dataframe di Spark. Di seguito si riportano gli step che descrivono il meccanismo di generazione dei random walks che è stato utilizzato:

1. costruzione di un dataframe W che contiene sulle righe ogni nodo duplicato un numero di volte pari al numero di cammini da generare per singolo nodo
2. costruzione di un secondo dataframe N, a partire da quello contenente gli archi del grafo, che memorizza per ogni nodo una lista di vicini di tipo non-commodity e un'altra per i vicini di tipo commodity
3. equi-join tra i due dataframe W e N appena costruiti definito sulla colonna dei nodi (NODE_ID)

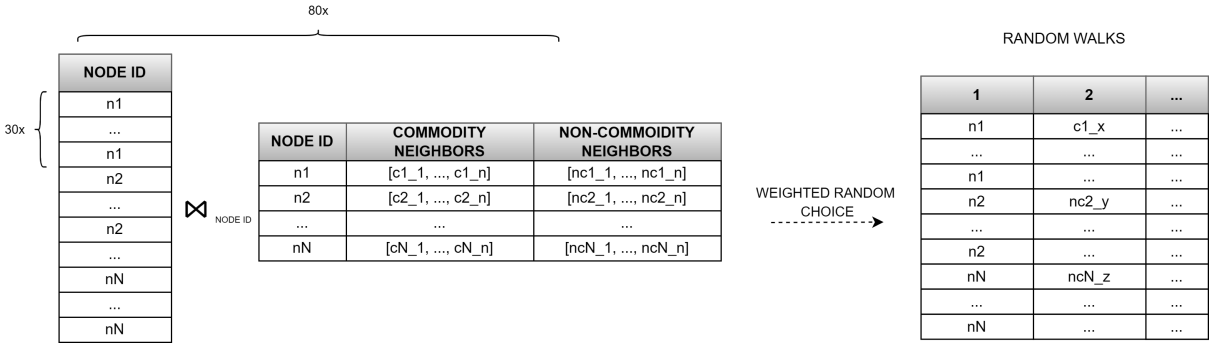

Fig. 5.1: Schema di funzionamento dell'approccio basato su Deep Walk.

4. scelta causale di un nodo all'interno di una delle due possibili liste e sostituzione della lista con tale valore: tale step definisce quindi ogni volta il passo successivo per ogni cammino

5. iterazione degli step 3-4 richiamando la join sulla nuova colonna di W appena valorizzata

Il metodo è stato quindi iterato un numero di volte pari alla lunghezza desiderata dei random walk – 1. In Fig. 5.1 si riporta un'immagine che descrive graficamente il meccanismo.

È necessario precisare, come descritto nel paragrafo 4.3 nella sezione relativa ai random walk, che la scelta tra le due liste allo step 4 dipende dalle probabilità definite in funzione degli obiettivi di suggerimento. Come conseguenza di ciò i random walk generati avranno probabilità differenti di attraversare commodities a seconda che il loro utilizzo sia destinato al cross-referencing piuttosto che al cross-selling. Ogni riga del dataframe risultante è stata quindi raggruppata in un unico array che definisce un certo random walk. Tale implementazione è possibile data la costruzione specifica del grafo che garantisce la presenza di un arco tra ogni coppia di vicini. I random walk così generati costituiscono quindi l'input del modello Word2Vec. Nel dettaglio è stata utilizzata l'implementazione del modello fornita dalla libreria MLlib di Spark. I valori di similarità sono infine stati ottenuti calcolando la cosine similarity

sulle rappresentazioni vettoriali in output. Riguardo i tempi di esecuzione, la fase più dispendiosa dal punto di vista del tempo di esecuzione è sicuramente quella di generazione dei random walk che ha impiegato circa 5 ore mentre l'allenamento del modello è durato circa 1,5 ore. Una nota conclusiva vuole sottolineare come in entrambi gli approcci un'eventuale modifica del grafo dovuta all'inserimento di un nuovo nodo e i relativi archi o l'aggiunta di archi tra nodi già presenti implica:

- nel caso delle metriche direttamente calcolate sul grafo, l'aggiornamento dei dataframe utilizzati per il calcolo modificando solo le righe interessate;
- nel caso dell'approccio DeepWalk, la generazione di nuovi cammini per il nodo inserito e gli eventuali nodi ad esso collegati.

In generale, tuttavia, la necessità di aggiornamento del sistema è strettamente correlata alla quantità di nuove informazioni da integrare per cui è possibile fare valutazioni quantitative o fissare archi temporali periodici in cui mantenere il sistema invariato.

5.3 Valutazione

Nonostante non si abbiano a disposizione dati sufficienti per generare un modello supervisionato è possibile utilizzare alcune informazioni per effettuare una prima valutazione sulle similarità calcolate e quindi sui suggerimenti che il sistema fornisce.

5.3.1 Dati di riferimento

Sia per il cross-referencing che per il cross-selling è stato possibile ottenere informazioni di riferimento. In particolare, riguardo i prodotti simili si ha a disposizione una tabella manualmente popolata da FAE e AM che collega alcuni prodotti considerati intercambiabili e che esprime su una scala da 1 a 4 la qualità delle com-

Metodo	x-ref	x-sell
Jaccard	0.17	0.25
Adamic-Adar	0.33	0.24
Deep Walk	0.2	0.11

Table 5.1: Mean Absolute Error per cross-referencing e cross-selling.

binazioni. Sui prodotti correlati invece si è fatto riferimento alle winning combos ovvero combinazioni di prodotti suggerite da fornitori sulla base delle conoscenze tecniche da questi possedute; si tratta di veri e propri progetti di esperti che mostrano prodotti complementari appartenenti a svariate categorie di prodotti come l'elaborazione embedded, l'analogico, l'alimentazione, la connettività.

5.3.2 Risultati

Per entrambi gli approcci la valutazione avviene tramite un insieme di suggerimenti di riferimento per poter esprimere una valutazione approssimativa dei risultati. Per valutare le raccomandazioni di prodotti simili si sono utilizzati circa 30 suggerimenti di riferimento a disposizione tra quelli con qualità massima. Nel caso dei prodotti correlati si è utilizzato anche qui un set di circa 30 winning combos. Si è quindi calcolato il Jaccard coefficient, l'Adamic-Adar Index e la coseno-similarità per tutti i casi di riferimento appena descritti e si è quindi misurato l'errore medio assoluto (MAE) commesso supponendo una similarità pari a 1 per i dati di riferimento. I risultati ottenuti sono riportati nella Tab.5.1. Infine, nel caso dell'approccio Deep Walk, è stato riportato anche l'istogramma dei valori di similarità generato considerando il 30% della totalità dei valori calcolati.

Da una prima valutazione dei risultati emerge come l'approccio DeepWalk trovi relazioni di correlazione tra prodotti in maniera più precisa rispetto ai risultati ottenuti per la ricerca di prodotti simili, anche rispetto alle altre metriche. Dalle

considerazioni fatte sulla tipologia di prodotti considerati e il livello di dettaglio delle specifiche tecniche necessarie per differenziarli, i risultati rientrano nelle aspettative formulate inizialmente e, tra diverse motivazioni, ciò potrebbe essere soprattutto dovuto alla struttura dei cammini determinata anche dai vincoli introdotti sull'attraversamento delle commodities. Nel caso delle metriche direttamente calcolate sul grafo si nota invece come il Jaccard coefficient trovi risultati migliori per la ricerca di prodotti simili e sia più preciso delle similarità trovate tramite l'Adamic-Adar index. In particolare, per quest'ultimo, dalle considerazioni fatte nel paragrafo 4.3 sulla misura rispetto alla contestualizzazione del meccanismo di funzionamento al grafo di riferimento, ci si aspetta che la cross-sell presenti risultati migliori e i valori sull'errore sembrano confermarlo. Si può notare inoltre come per entrambi gli obiettivi di analisi il coefficiente di Jaccard è più preciso in un caso, e quasi alla pari nell'altro, dell'Adamic-Adar index e ciò può essere dovuto alla natura stessa delle metriche che nel primo caso è più qualitativa dato che si tiene maggiormente in considerazione l'entità stessa dei nodi mentre nel secondo caso si focalizza maggiormente su alcuni di questi sulla base di una valutazione quantitativa dei vicini. Si riportano in Fig.5.2 e Fig.5.3 gli istogrammi relativi ai valori di coseno-similarità calcolati sul 30% dei nodi nella componente connessa considerata. In generale risulta esserci una buonan capacità di distinzione del sistema tra i componenti dato che la distribuzione dei valori di similarità generati per il cross-referencing tendono ad essere mediamente nulli discostandosi dai valori per i casi di riferimento e mostrando, in particolare, una deviazione standard minore rispetto alla distribuzione dei valori generati per il cross-selling che sono mediamente più alti e con una deviazione standard maggiore.

È necessario, tuttavia, precisare come queste valutazioni siano state effettuate su risultati che potrebbero essere dipendenti dai dati di riferimento considerati e che, per questo, rappresentano solo un primo tentativo di verifica che dovrà essere

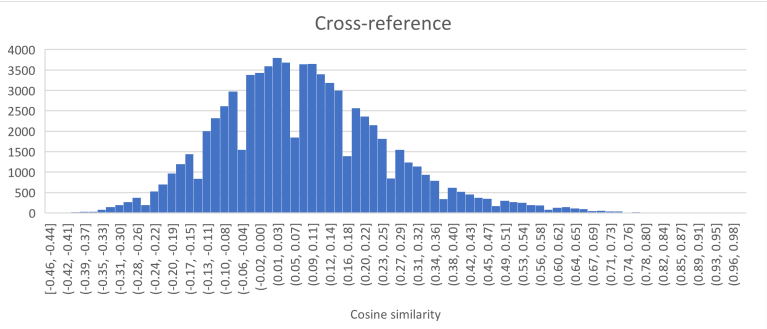

Fig. 5.2: Istogramma coseno-similarità per X-ref.

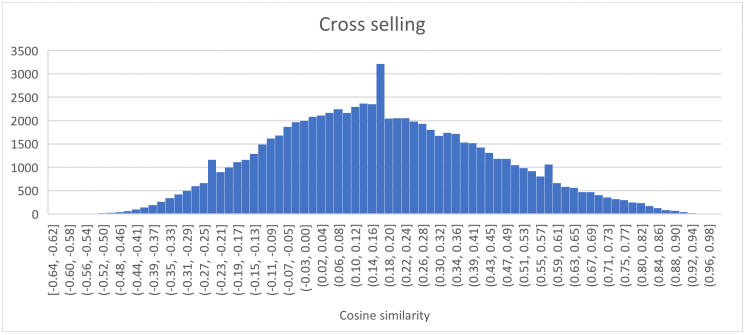

Fig. 5.3: Istogramma coseno-similarità per X-sell.

confermato da ulteriori misure di errore calcolabili quando si avranno a disposizione raccomandazioni di riferimento maggiormente valide per qualità e soprattutto per quantità.

Conclusioni e sviluppi futuri

Il lavoro presentato ha permesso di elaborare un primo approccio alternativo a quello attualmente utilizzato per il supportare alle vendite. Tale necessità di studio è nata da considerazioni sul comportamento dei metodi in uso, come l'algoritmo Apriori per la generazione di regole di associazione, che in un contesto applicativo come quello di interesse risultano non particolarmente adatti sia per la natura dei meccanismi di vendita che per la granularità elevata delle specifiche tecniche dei componenti che rendono complessa la loro differenziazione. Il grafo rappresenta in questo un approccio alternativo che mira a rappresentare informazioni nascoste sulle relazioni tra componenti, tenendo in considerazione anche le informazioni sui clienti e sui settori applicativi dei componenti. Le sfide maggiori affrontate durante la progettazione e la realizzazione hanno riguardato soprattutto la fase iniziale di comprensione dei dati per poter strutturare il grafo e tradurre le informazioni contenute in formato tabellare in modo da rispondere al meglio agli obiettivi di analisi, e la fase di implementazione in cui si è dovuto gestire il calcolo delle similarità sfruttando al meglio le risorse computazionali a disposizione. Per questo sono state implementate in particolare delle soluzioni basate sul calcolo tramite dataframe Spark che hanno permesso di velocizzare e rendere fattibile il calcolo delle metriche presentate. Il risultato del lavoro svolto è un sistema che produce dei suggerimenti riguardo i due principali obiettivi di analisi cioè la ricerca di prodotti simili e correlati. I suggerimenti possono essere generati utilizzando entrambi gli approcci o favorendone uno specifico

in funzione dell'obiettivo (similarità o correlazione). La scelta può dipendere anche dalle prime considerazioni di valutazione fatte nel capitolo precedente. Come anticipato tuttavia quello descritto può essere considerato come un lavoro iniziale che struttura il problema e crea una base di partenza con diverse prospettive di sviluppo. Grazie all'eterogeneità del grafo è possibile, infatti, sviluppare tecniche di clustering dei clienti sulla base della similarità tra nodi customer e andando a generare specifici random walk per tale obiettivo. Anche a proposito del meccanismo di generazione dei cammini è possibile sperimentare soluzioni alternative che dipendono dall'introduzione di nuove caratteristiche nel grafo come la definizione di un peso o di una direzione per gli archi. Più in dettaglio è possibile definire, ad esempio, un peso sugli archi in funzione del numero di occorrenze delle coppie di material group, tra loro collegate, normalizzato rispetto la cardinalità dei gruppi; tale caratterisitica aggiuntiva permetterebbe di generare diversamente le raccomandazioni di prodotti correlati. Sul possibile orientamento degli archi tra le alternative possibili si potrebbe invece definire un verso in funzione del fatto che ci si sposti da un componente con prezzo inferiore ad uno con prezzo superiore introducendo quindi nuovi possibili vincoli nella creazione dei cammini. Per concludere uno sviluppo futuro del progetto potrebbe considerare metriche aggiuntive per il calcolo della similarità e l'utilizzo di una rete neurale artificiale che prenda in input le metriche calcolate su coppie di nodi restituendo in output la probabilità di correlazione tra due nodi in funzione delle varie metriche considerate; questo nell'ipotesi di disporre di dati di riferimento da utilizzare come ground truth.

Bibliografia

[1] Z. Dong, Z.Wang, J.Xu, R.Tang, J.Wen (2022) *A Brief History of Recommender Systems*,

[2] B. Shao, X. Li, G. Bian, Expert Systems with Applications, Volume 165 (2021) *A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph*,

[3] P.G. Roetzel, Bus Res 12, 479–522 (2019) *Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development*,

[4] H. Schroder, M.J. Driver, S. Streufert, New York: Holt, Rinehart and Winston (1967) *Human information processing*,

[5] C. Anderson, Hachette Books (2006) *The Long Tail: Why the Future of Business is Selling Less of More*,

[6] C. Aggarwal, Springer (2016) *Recommender Systems*,

[7] M. Jessenitschnig, User Model. User-Adap. Inter. 19(1-2), 133-166 *Case-studies on exploiting explicit customer requirements in recommender systems*,

[8] R.Burke (2002) *Hybrid Recommender Systems - Survey and Experiments*,

[9] Avnet global website *https://news.avnet.com/about-us/history*,

[10] G. Salton, M.J. McGill, McGraw-Hill (1983) *Introduction to Modern Information Retrieval*,

[11] F. Gao, K. Musial, C. Cooper, S. Tsoka, Department of Informatics, School of Natural and Mathematical Sciences, King's College London, Strand Campus, London WC2R 2LS, UK (2014) *Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics*,

[12] B. Perozzi, R. Al-Rfou, S. Skiena, KDD.'14 - Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 701–710 (August 2014) *DeepWalk:Online Learning of Social Representations*,

[13] 13. CIKM (2020) *Karate Club: API Oriented Open-source Python Framework for Unsupervised Learning on Graphs*,

[14] B. Perozzi (2013) *Original implementation of the model proposed in [13]*,

[15] T. Mikolov, K. Chen, G. Corrado, J. Dean. CoRR, abs/1301.3781 (2013) *Efficient estimation of word representations in vector space*,

[16] A. Spielman (2018) *Spectral Graph Theory: Random Walks on Graphs*,

[17] Spark 2.2.0 Documentation, https://spark.apache.org/docs/2.2.0/mllib-feature-extraction.html *Feature Extraction and Transformation - RDD-based API*,

[18] P. Dangeti: Packt Publishing, O'Reilly (2017) *Statistics for Machine Learning*,

Elenco figure

1. Il recommender system

1.1. Distribuzione annuale e percentuale annuale di letteratura pubblicata nell'ambito dei recommender system.
Sorgente: *B. Shao, X. Li, G. Bian. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph, Expert Systems with Applications, Volume 165, 2021*

1.2. Mercato dei sistemi di raccomandazione negli Stati Uniti, per tipologia, e relativa previsione di crescita - periodo 2021-2028.
Sorgente: *https://www.grandviewresearch.com/industry-analysis/recommendation-engine-market-report*

1.3. Informazioni e prestazioni decisionali.
Sorgente: *P.G. Roetzel,Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. Bus Res 12, 479–522, 2019*

1.4. La coda lunga dall'omonima teoria di C.Anderson.
Sorgente: *C. Anderson, In The Long Tail: Why the Future of Business Is Selling Less of More, 2006*

1.5. Schema di funzionamento collaborative-filtering.

Sorgente: *A. Anand, User-User Collaborative Filtering For Jokes Recommendation, Towards Data Science*

1.6. Schema di funzionamento content-based.

Sorgente: *A. Anand, User-User Collaborative Filtering For Jokes Recommendation, Towards Data Science*

1.7. Schema di funzionamento utility-based.

Sorgente: *M. Jessenitschnig, Case-studies on exploiting explicit customer requirements in recommender systems. User Model. User-Adap. Inter. 19(1-2), 133-166*

1.8. Schema di funzionamento knowledge-based.

Sorgente: *M. Jessenitschnig, Case-studies on exploiting explicit customer requirements in recommender systems. User Model. User-Adap. Inter. 19(1-2), 133-166*

2. Contesto applicativo

2.1. Avnet nella supply chain.

Sorgente: *Autore*

2.2. I principali settori nel mercato dell'elettronica.

Sorgente: *Autore*

3. Sorgenti informative

3.1. Tipologie di grafo.

Sorgente: *Autore*

3.2. Esempio di possibile interconnessione delle diverse tipologie di nodo.

Sorgente: *Autore*

4. Data analysis: duplice approccio

4.1. Schema di funzionamento dell'analisi basata su Deep Walk.

Sorgente: *Autore*

4.2. Conteggio archi per material group normalizzato rispetto cardinalità gruppo.

Sorgente: *Autore*

4.3. Istogramma conteggio normalizzato degli archi per material group.

Sorgente: *Autore*

4.4. Aree di interesse nella generazione dei cammini.

Sorgente: *Autore*

4.5. Architettura del modello Skip-Gram.

Sorgente: [9]

5. Implementazione e confronto degli approcci

5.1. Schema di funzionamento dell'approccio basato su Deep Walk.

Sorgente: *Autore*

5.2. Istogramma coseno-similarità per X-ref.

Sorgente: *Autore*

5.3. Istogramma coseno-similarità per X-sell.

Sorgente: *Autore*