Siamo sommersi dai dati.

Ogni giorno viene generata una quantità enorme di dati, anche da azioni della vita quotidiana: dalle applicazioni per smartphone alle carte di credito usate per gli acquisti, dai programmi usati sui computer ai sensori usati nelle infrastrutture intelligenti della città.
Nella stragrande maggioranza dei casi, questa enormità di dati, chiamati Big Data, viaggia attraverso Internet, ed è possibile fruire di queste grandi quantità di informazioni semplicemente esplorando il World Wide Web.
Tuttavia, data l'enormità e l'eterogeneità dei dati che si trovano nel Web, non è possibile utilizzarli direttamente: per farlo, si devono applicare delle metodologie per analizzare ed estrarre l'informazione dal grafo del Web. Non solo, quindi, estrapolare l'informazione da una pagina, ma anche utilizzare la struttura ad hyperlink di cui il World Wide Web si compone.
In quest'ottica, il Data Mining si è evoluto.

0.1 Data Mining

Il Data Mining è l'insieme di tecniche che hanno come obiettivo l'estrazione del sapere o della conoscenza, partendo da grandi quantità di dati. Queste tecniche e metodologie vengono usate sia in ambito industriale che scientifico.
Il termine significa letteralmente "estrazione di dati", la quale si divide in:

- **estrazione**: l'informazione implicita, nascosta o formata da dati strutturati viene estratta per renderla immediatamente usabile;
- **esplorazione ed analisi**: vengono scoperti pattern significativi, per mezzo dei quali si estrae l'informazione significativa.

Con il termine pattern, nel contesto del Data Mining, si intende uno schema, una regolarità, o, in generale, una rappresentazione sintetica dei dati [1].

Natural Language Processing Il Natural Language Processing è il processo di trattamento automatico mediante un calcolatore delle informazioni scritte o parlate in una lingua naturale.
La difficoltà che caratterizza questo processo è l'elevato numero di ambiguità che caratterizza il linguaggio umano: per questo motivo è stato diviso in quattro fasi, o sottoprocessi:

- **analisi lessicale**: in questa fase avviene la scomposizione della sequenza di caratteri, chiamata espressione linguistica, in token (o parole);

- **analisi grammaticale**: in questa fase avviene l'associazione di ciascuna parola ad una parte del discorso;
- **analisi sintattica**: in questa fase avviene il parsing dei token e viene generato un albero di parser (parse tree);
- **analisi semantica**: in questa fase viene assegnato un significato al parse tree, la quale provvede alla disambiguazione l'espressione linguistica, ovvero ad assegnare un significato tra quelli disponibili.

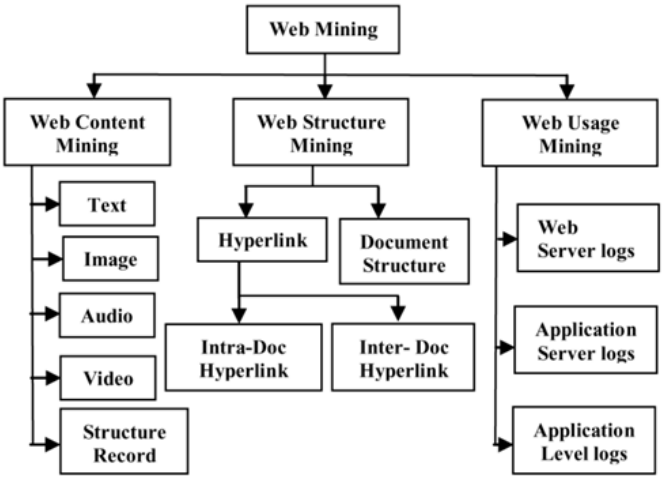
Text Mining Il Text Mining, riferito anche come Text Data Mining o Text Analysis, è l'applicazione delle tecniche e metodologie del Data Mining ai testi. L'obiettivo è simile al Data Mining: estrarre informazioni latenti in documenti e testi analizzando ed esplorando dei pattern significativi.
Utilizzando il Natural Language Processing, è possibile estrarre informazioni incapsulate nei testi, le quali potrebbero essere potenzialmente utili.

0.2 Web Mining

Con l'espressione Web Mining ci si riferisce all'applicazione di procedure analoghe per estrarre automaticamente informazioni dalle risorse presenti nel Web, sia documenti che servizi [2]. In altri termini, è l'applicazione delle procedure di Data Mining per scoprire pattern dal World Wide Web ed estrarre l'informazione. La conoscenza viene estratta dal contenuto, dalla struttura e dall'uso del Web.
In [2] viene spiegato come l'obiettivo del Web Mining viene diviso in vari sotto-obiettivi:

- **Scoperta di risorse**: gli strumenti per la scoperta di risorse, che vengono chiamati Spider, ovvero Web Robot, scandiscono milioni di documenti Web e costruiscono indici di ricerca in base alle parole che si trovano negli stessi;
- **Estrazione di informazioni**: i testi, che sono scritti in linguaggio naturale, vengono trasformati in rappresentazioni strutturate predefinite, dette template, che rappresentano un estratto dell'informazione presente nel testo;
- **Generalizzazione**: i processi di navigazione nel web devono essere generalizzati.

Il Web Mining può essere suddiviso in tre distinte categorie, in base al tipo di dato da estrarre: Web Usage Mining, Web Structure Mining e Web Content Mining.
Di seguito viene presentata la tassionomia delle varie tipologie di Web Mining.



Web Usage Mining Il Web Usage Mining è l'applicazione delle tecniche di Data Mining per la scoperta di pattern e informazioni utili attraverso l'analisi dei log, i quali sono immagazzinati nei Web server o da sistemi che tracciano le attività degli utenti.
L'obiettivo di questo campo è la profilazione dell'utente, ovvero analizzare i suoi comportamenti sul web sia per comprendere quali sono i suoi reali bisogni, sia per offrire dei servizi che possano soddisfare tali necessità e personalizzare l'esperienza Web.
Questo tipo di Data Mining viene usato in disparati campi, che vanno dalle aziende alle agenzie governative: i siti di e-commerce usano questo tipo di tecnologia per presentare all'utente prodotti per i quali potrebbe essere interessato; le agenzia governative, invece, hanno usato il Web Mining per classificare minacce e attentati terroristici.
Alcuni, però, criticano questa tecnologia: il problema etico di cui più si parla è la violazione della privacy.

Web Structure Mining Il Web Structure Mining è un processo di analisi della struttura di un sito web, il quale viene considerato come un grafo i cui nodi sono le pagine e gli archi sono gli hyperlinks tra le pagine. Si divide in:

- Estrazione di schemi dagli hyperlinks, in cui un hyperlink è un arco tra due pagine web;
- Estrazione della struttura del documento, ovvero l'analisi della struttura ad albero basata su HTML ed XML.

Quindi, questo tipo di Web Mining può essere effettuato sia a livello di documento web (intra-pagina), sia a quello di hyperlinks (inter-pagina).
Basata sulla topografia degli hyperlinks, Web Structure Mining può categorizzare le pagine web e generare informazioni come la similarità e le relazioni tra i differenti siti Web [5].
Tra i più importanti algoritmi che appartengono a questa tipologia troviamo Page Rank [4] e HITS [3], i quali sfruttano la struttura ad hyperlink del Web per assegnare un rank alle pagine, ovvero per restituirle in ordine di importanza relativamente ad una determinata query.

Web Content Mining L'ultimo tipo di Web Mining è il Web Content Mining, il quale viene usato per cercare informazioni utili dal contenuto delle pagine Web. Con il termine contenuto ci si riferisce a collezioni di testi, immagini, audio, video, o record strutturati che sono incapsulati in liste e tabelle. Nel campo della ricerca, è stato applicato il Text Mining, che ha permesso di migliorare le attività di mining sui testi grazie al Natural Language Processing.
Per estrarre il sapere da contenuti più complessi, come le immagini, le tecniche di Web Content Mining sono molto limitate[6].

0.3 Rappresentazioni vettoriali di pagine Web

0.3.1 Word2Vec

0.3.2 Line

0.3.3 Doc2Vec

0.3.4 Word space model

0.4 Clustering

Con il termine clustering si intende l'insieme di tecniche che hanno come scopo quello di selezionare e raggruppare, da una collezione di dati, elementi omogenei, avendo come base la somiglianza tra gli stessi. La somiglianza tra

gli elementi è concepita in termini di distanza di uno spazio multidimensionale. La bontà della similarità dipende fortemente dalla funzione che si usa per calcolare la distanza tra gli elementi.

0.4.1 Approcci di clustering

L'operazione di clustering è essenzialmente la creazione di un insieme di clusters, cioè un insieme di insiemi, che generalmente contengono tutti gli elementi iniziali. Si possono usare varie classi di approcci per effettuare clustering su un determinato insieme di dati iniziali. Alcune di queste sono:

- Hard clustering o soft clustering
- Partizionali o gerarchici

Hard clustering e soft clustering Questi algoritmi attuano un approccio secondo cui un elemento può essere assegnato ad un solo cluster o a più cluster. Con hard clustering intendiamo che l'algoritmo assegna un elemento ad uno ed un solo cluster; con soft clustering, invece, l'elemento può essere assegnato a più cluster con gradi di appartenenza diversi.

Clustering partizionale Gli algoritmi di clustering partizionali creano una divisione delle osservazioni minimizzando una certa funzione di costo:

$$\sum_{j=1}^{k} E(C_j) \tag{1}$$

dove k è il numero desiderato di cluster, C_j è il j-esimo cluster ed $E : C \to \mathbb{R}^+$ è la funzione di costo associata al singolo cluster. L'algoritmo più famoso che fa parte di questa categoria è K-Means.

Clustering gerarchico Gli algoritmi facente parti di questa categoria non suddividono lo spazio, bensì costruiscono una gerarchia di cluster. In questa strategia rientrano due sottotipi:

- **Aggregativo**: tale approccio considera n cluster per n elementi, cioè ogni elemento viene considerato un cluster a sè. Successivamente, l'algoritmo unisce tutti i cluster più vicini. Viene anche chiamato bottom-up.
- **Divisivo**: tale approccio ragiona in maniera opposta rispetto al precedente, poichè tutti gli elementi vengono considerati come un unico cluster, e l'algoritmo deve dividere il cluster in insiemi aventi dimensioni inferiori. Questa metodologia viene anche chiamata top-down.

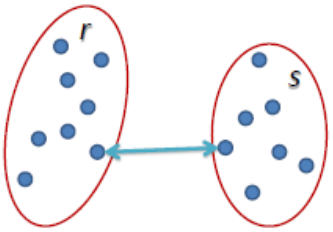
Durante l'aggregazione degli elementi è necessario usare una funzione che permette di calcolare la similarità (o meglio dire la distanza) tra due cluster: questo permette all'algoritmo di unire i cluster simili.

0.4.2 Funzioni (o misure) di distanza

A seconda dell'approccio utilizzato, vi sono delle funzioni (o misure) che permettono di calcolare la distanza tra due cluster. Viene molto usato dagli algoritmi di clustering gerarchico per calcolare la similarià tra i cluster e per unire, eventualmente, i cluster simili. Le funzioni di distanza usate da questo tipo di clustering sono: **single-link proximity**, **average-link proximity**, **complete-link proximity** e la **distanza tra centroidi**.

Single-link proximity Questa funzione calcola la distanza tra due cluster come la distanza minima tra elementi appartenenti a cluster differenti.

$$D(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y) \tag{2}$$

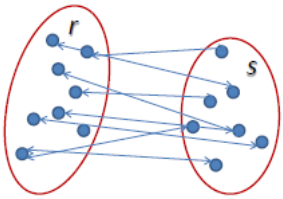


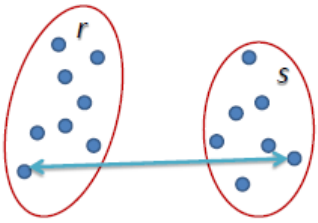
Average-link proximity Questa funzione calcola la distanza tra due cluster come la media delle distanze tra i singoli elementi.

$$D(C_i, C_j) = 1/(|C_i||C_j|) \sum_{x \in C_i, y \in C_j} d(x, y) \tag{3}$$

Complete-link proximity Questa funzione calcola la distanza tra i due cluster considerando la distanza massima tra gli elementi appartenenti ai due cluster.

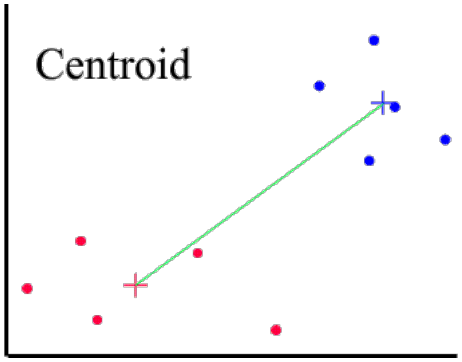
$$D(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x, y) \tag{4}$$





Distanza tra centroidi Questa, invece, è la distanza tra i due cluster prendendo in considerazione i centroidi degli stessi.

$$D(C_i, C_j) = d\left(\hat{c}_i, \hat{c}_j\right) \tag{5}$$



Nei casi precedenti, $d(x, y)$ indica una qualsiasi funzione distanza su uno spazio metrico, le quali possono essere:

- **Distanza euclidea**: chiamata anche norma 2, è la distanza calcolata tra due punti, la quale può essere misurata su uno spazio multidimensionale. Siano $P = (p_1, p_2, ..., p_n)$ e $Q = (q_1, q_2, ..., q_n)$ due punti, la distanza sarà:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{k} (p_k - q_k)^2} \tag{6}$$

- **Distanza di Manhattan**: chiamata anche geometria del taxi o norma 1, è la distanza tra due punti calcolata come la somma del valore assoluto delle differenze delle loro coordinate. Siano $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$ due punti, la distanza sarà:

$$L_1(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2| \tag{7}$$

- **Norma uniforme**

- **Distanza di Mahalanobis**

- **Coseno di similarità**: tecnica euristica usata per misurare la distanza tra due vettori, che viene effettuata calcolando il coseno dell'angolo compresovi, che hanno l'origine coincidente con quello del sistema di assi e passano per i rispettivi elementi. Il valore risultante più sarà vicino ad 1, più i due elementi sono simili tra loro. Siano A e B due vettori di attributi numerici,

$$\cos(\theta) = \frac{AB}{||A||||B||} \tag{8}$$

- **Distanza di Hamming**: misura il numero di sostituzioni necessarie per convertire una stringa nell'altra, oppure può essere vista come un reporting del numero degli errori che hanno trasformato una stringa nell'altra. La distanza di Hamming tra 1011101 e 1001001 è 2; oppure tra 2143896 e 2233796 è 3.

0.4.3 Algoritmi usati

In questa sezione vengono descritti gli algoritmi di clustering che sono stati usati sui dataset della sperimentazione. Questi sono stati usati per verificare la bontà dell'operazione di clustering, partendo da elementi che sono stati appresi mediante algoritmi di apprendimento differenti, la quale è stata analizzata mediante apposite metriche.

K-Means K-Means è un algoritmo di clustering di tipo partizionale, in cui ogni cluster viene identificato mediante un centroide. Si basa sull'algoritmo di Lloyd e consiste in 3 step. Il primo step consiste nella scelta dei centroidi iniziali, i quali saranno K elementi, casuali o usando informazioni euristiche, scelti dal dataset. Successivamente, l'algoritmo assegna per ogni elemento il centrine più vicino e crea nuovi centroidi dalla media di tutti i campioni, assegnati ai centroidi precedenti. Si ripete questa fase finchè l'algoritmo non

converge. Il pregio principale di questo algoritmo è che converge molto velocemente: si è analizzato, infatti, che il numero di iterazioni che l'algoritmo esegue è minore del numero di elementi del dataset. K-means, però, può essere molto lento nel caso peggiore e non garantisce il raggiungimento dell'ottimo globale: la bontà della soluzione dipende dal set di cluster iniziale. Inoltre, un altro svantaggio è che l'algoritmo richiede, in input, il numero dei cluster.

HDBScan HDBScan è un algoritmo di clustering che estende DBScan, rendendolo di tipo gerarchico. Si parte in maniera simile a DBScan: lo spazio viene trasformato a seconda della densità e viene effettuato su di esso una prossimità a single-link. Invece di richiedere come input il parametro ϵ, che viene usato da DBScan per considerare gli elementi del vicinato appartenenti al cluster, viene creato un albero, il quale viene usato per selezionare i cluster più stabili e persistenti. Al posto di ϵ, quindi, viene richiesta la dimensione minima dei cluster per determinare quali gruppi non devono essere considerati come cluster, oppure per dividerli e formare nuovi cluster. Questo algoritmo è molto efficace ed è il più veloce, sia di DBScan che di K-Means.

Bibliografia

[1] Cos'è il data mining. Cineca.

[2] G. Convertino, L. Di Pace, P. Leo, A. Maffione, D. Malerba, and G. Vespucci. Tecniche di web mining per supportare l'attività di navigazione in rete. (2):12.

[3] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, sep 1999.

[4] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[5] N. V. Pardakhe and R. R. Prof. Keole. Analysis of various web page ranking algorithms in web structure mining. 2:6, 2013.

[6] Jaideep Srivastava, Prasanna Desikan, and Vipin Kumar. Chapter 21 - web mining - concepts, applications, and research directions. In *Foundations and Advances in Data Mining*, pages 275–307.