Università Politecnica delle Marche

Facoltà Di Ingegneria

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

Individuazione di Pattern of Attention per caratterizzare i bisogni informativi e comunicativi degli utenti del Web

Identification of patterns of attention to characterize the Web users information and communication needs

—

Relatore:

Prof. Alessandro Cucchiarelli

Candidato:

Giacomo Marangoni

Correlatore:

Ing. Christian Morbidoni

Anno Accademico 2015/2016

"L'unico modo per fare un ottimo lavoro è amare quello che fai. Se non hai ancora trovato la cosa che fa per te, continua a cercarla, non fermarti, come capita per le faccende di cuore, saprai di averla trovata non appena ce l'avrai davanti. E, come le grandi storie d'amore, diventerà sempre meglio col passare degli anni. Quindi continua a cercare finchè non l'avrai trovata. Non accontentarti. Sii affamato. Sii folle."

Steve Jobs

Capitolo 1

Introduzione

La crescita sempre più massiva della quantità delle informazioni disponibili e l'aumentata "raggiungibilità" delle stesse per mezzo di risorse Web, ha portato allo sviluppo di metodologie e strumenti che permettono di elaborare i dati e ricavarne informazioni non ovvie, di grande importanza per l'utilizzatore finale, sia esso un ricercatore che studia fenomeni scientifici o sperimentali o il manager di una qualsiasi azienda intenzionata a migliorare i processi decisionali nelle proprie aree di business.
La realizzazione di tali processi di elaborazione è stata resa particolarmente ardua dall'esplosiva crescita delle dimensioni delle basi dati commerciali, governative e scientifiche. In ogni caso, i sistemi di gestione delle informazioni hanno certamente permesso di manipolare i dati immagazzinati in maniera efficace ed efficiente, pur non avendo ancora risolto pienamente il problema di come supportare l'utilizzatore nella "comprensione" e nell'analisi dei dati stessi. L'ambito scientifico a cui ciò si riferisce prende il nome di Data Mining, all'interno del quale il presente lavoro di tesi si colloca, soffermandosi, in particolare, sull'estensione dello stesso al Temporal Mining, cioè a quel processo di estrazione di relazioni non esplicite dai dati che modellano sequenze di eventi. In tale contesto, saranno confrontati ambiti metodi ed algoritmi, prestando particolare attenzione alle tecniche di clustering volte alla definizione di quello che può essere definito

come il problema della Event Detection e dell'elicitazione di Pattern of Attention.
Uno dei principali fattori di criticità riscontrati nel Temporal Data Mining risiede nella scelta di tecniche efficienti per estrarre conoscenza da un gran numero di dati di tipo temporale che sono, per loro natura, complessi da rappresentare e da trattare. A tale proposito, durante gli ultimi decenni, i grandi progressi nel campo dell'hardware, della tecnologia dei database, della grafica, hanno reso possibile la nascita di sistemi potenzialmente in grado di trattare grandi quantità di informazioni complesse e multidimensionali. Il ragionamento temporale e spazio-temporale è, infatti, alla base di molte attività umane. Il problema di trattare con informazioni di questo tipo è diventato un aspetto sempre più rilevante, sul quale si è concentrata gran parte della ricerca scientifica, negli ultimi anni. Nella realtà, infatti, spazio e tempo sono tra loro strettamente interconnessi: la maggior parte delle informazioni che si riferiscono allo spazio attingono anche al tempo. Ci sono, appunto, applicazioni per le quali è assolutamente indispensabile poter utilizzare dati di tipo spazio-temporale. Basti pensare ai sistemi che forniscono grandi benefici in aree quali quelle del monitoraggio ambientale, dei settori amministrativi, dei sistemi di navigazione in real-time, dello scheduling dei trasporti, ecc.
Parallelamente è inoltre possibile definire il concetto di Social Behavior Analysis (SBA) la quale si caratterizza come un sottoinsieme della Business Analysis[1]. La SBA focalizza la sua attenzione sullo studio dei comportamenti degli utenti nel web e sul come e perchè essi interagiscano tra loro o con strumenti quali Social Network, portali eCommerce e giochi online, ecc.
Questo elaborato si propone quindi di fornire un quadro complessivo generale relativo all'area di studi che si occupa di tale problematica, volgendo la sua attenzione sulla caratterizzazione del problema dell'individuazione di classi dinamiche di attenzione collettiva applicata in particolare ai contenuti Web ed ai Social Network.
Tratteremo in dettaglio il problema dell'Information Retrieval e della Temporal Informartion Retrieval per poi focalizzare la nostra attenzione sullo stato dell'arte relativo

[1] È un ambito di ricerca che si occupa di studiare le esigenze legate al commercio e di risolvere problematiche ad esso connesse.

all'Event Detection. Successivamente, proporremo un approccio volto ad elicitare dei pattern di attenzione collettiva applicato al corpus dei ClickLog di Wikipedia andando a definire quello che risulta essere l'elemento di innovatività del nostro lavoro. Esso consiste nel caratterizzare i bisogni informativi degli utenti, in seguito al verificarsi di eventi veicolati dalle più disparate sorgenti comunicative. Per questo utilizzeremo un approccio basato sulla Symbolic Aggregate approXimation per l'estrazione di eventi a partire da cluster ottenuti in seguito all'elaborazione del corpus citato precedentemente. Presenteremo inoltre una metodologia robusta, efficiente e scriptabile per il tuning dei parametri del nostro sistema, il quale ci permetterà di dimostrare sperimentalmente la validità del nostro approccio.
In particolare è possibile suddividere il presente lavoro di tesi come segue:

- **Stato dell'arte sul Temporal Mining**: al suo interno verrà caratterizzata e descritta tale problematica focalizzando l'attenzione sul problema dell'Event Detection. Inoltre verrà presentato il tema dell'Information Retrieval ed della Temporal Information Retrieval;

- **Event Detection basata sulla Symbolic Aggregate approXimation**: descriveremo nel dettaglio gli strumenti utilizzati nell'ambito del nostro lavoro, e caratterizzeremo l'algoritmo SAX* sviluppato presso l'Università "La Sapienza" di Roma. Inoltre introdurremo una nuova variante dell'algoritmo basata su rete semantica NASARI;

- **Elicitazione dei Pattern of Attention dai ClickLog di Wikipedia**: in questa fase caratterizzeremo il corpus dei ClickLog di Wikipedia e la pipeline software sviluppata nell'ambito della nostra ricerca. Tratteremo inoltre la metodologia di valutazione definita per il tuning dei parametri di SAX* NASARI individuando la miglior parametrizzazione per l'Event Detection nell'ambito di eventi di test selezionati nel periodo che va da Giugno 2014 a Settembre 2014;

- **Analisi dei Risultati**: in questa quarta ed ultima parte dell'elaborato, dimostreremo sperimentalmente la validità del nostro, analizzando i risultati per degli eventi rilevanti nel periodo sopra citato ed arrivando ad elicitare pattern di attenzione collettiva derivanti dalla necessità degli utenti del Web di soddisfare i propri bisogni informativi.