

Enhancing Data Warehouse management through semi-automatic data integration and complex graph generation

Abstract

L'informazione strategica è una delle risorse principali per molte organizzazioni e, nel prossimo futuro, diventerà sempre più importante per consentire ai Manager di rispondere a domande sulla propria attività, come ad esempio come aumentare la propria redditività. Un processo decisionale adeguato è avvantaggiato da informazioni che sono spesso disperse tra diversi database eterogenei. Tali database possono provenire da diversi sistemi organizzativi e persino da fonti esterne. Di conseguenza, i manager dell'organizzazione devono affrontare il problema dell'integrazione di diversi database da fonti di dati indipendenti contenenti differenze semantiche e nessuna descrizione concettuale specifica o canonica.

I Data Warehouse Systems sono nati per integrare questo tipo di dati eterogenei al fine di essere successivamente estratti e analizzati in base alle esigenze del gestore e ai piani aziendali.

Oltre ad essere difficili e onerosi da progettare, integrare e costruire, i Data Warehouse Systems presentano un altro problema relativo alla difficoltà di rappresentare informazioni multidimensionali tipiche del risultato di operazioni OLAP, come aggregazioni su cubi di dati, estrazione di sotto-cubi o rotazioni del asse dei dati, attraverso viste di facile comprensione. In questa tesi, presento un linguaggio visivo basato su un paradigma logico, denominato linguaggio di Complexity Design (CoDe), e propongo un approccio semi-automatico per supportare il gestore nella generazione automatica di un data warehouse rispondente alle sue specifiche esigenze. In particolare, il manager esprime le sue domande in linguaggio naturale e seleziona le informazioni necessarie tra diverse fonti di dati. I dati selezionati vengono importati dai database, viene eseguito un processo di ottimizzazione dei dati e viene eseguito un algoritmo di estrazione delle regole di associazione per estrarre il modello concettuale sotto i dati. Successivamente, vengono forniti i modelli CoDe e il Data Mart viene generato dai modelli. Infine, il modello CoDe e i dati DW vengono utilizzati per generare un report grafico delle informazioni richieste.

Per mostrare le informazioni che il manager deve prendere in base alle domande strategiche, il rapporto adotta Grafici complessi. In questa tesi, valuto anche l'efficacia

dell'approccio proposto, in termini di comprensibilità della rappresentazione visiva prodotta dei dati estratti dal data warehouse. In particolare, è stata progettata e condotta una valutazione empirica per valutare la comprensione delle rappresentazioni grafiche per consentire decisioni istantanee e informate. Lo studio è stato condotto presso l'Università di Salerno, Italia, e ha coinvolto 47 partecipanti del corso di laurea in Informatica con conoscenze di gestione, sistemi informatici e sistemi avanzati di database e data warehouse). Ai partecipanti è stato chiesto di comprendere la semantica dei dati utilizzando la dashboard tradizionale e diagrammi di grafici complessi.

Lo sforzo richiesto per rispondere a un questionario di valutazione è stato valutato insieme alla comprensione delle rappresentazioni dei dati e il risultato è stato presentato e discusso. I risultati ottenuti suggeriscono che le persone hanno raggiunto una comprensione superiore quando usano i Grafici complessi come quelli prodotti dal nostro approccio rispetto ai grafici standard di Dashboard. Inoltre, l'analisi del tempo necessario per comprendere la semantica dei dati mostra che i partecipanti hanno trascorso molto meno tempo a comprendere la rappresentazione adottando una visualizzazione di Grafici complessi rispetto ai grafici standard di Dashboard. Sia gli utenti esperti che quelli inesperti hanno tratto vantaggio dalla complessa rappresentazione grafica in termini di comprensione e sforzo, mentre i partecipanti più esperti hanno un vantaggio maggiore nel tempo di comprensione. Questa constatazione potrebbe essere molto rilevante per il manager. Infatti, i grafici complessi rappresentano un approccio di visualizzazione efficace che consente una comprensione più rapida e più elevata dei dati, migliorando l'appropriatezza delle decisioni e riducendo lo sforzo decisionale.

Sulla base di questo risultato, è stata presentata una generazione automatica di data warehouse che utilizza grafici complessi per visualizzare i principali fatti che un decisore dovrebbe utilizzare per le decisioni strategiche. In particolare, il linguaggio CoDe è stato adottato per rappresentare tali informazioni.

Nel valutare la generazione automatica del data warehouse, la fase più critica del processo è stata l'integrazione dei dati, a causa della necessità di un importante contributo da parte del gestore. Pertanto, per incoraggiare l'implementazione del processo in un contesto reale, è stato proposto e testato un processo di integrazione dei dati semiautomatico su 2 casi di studio per valutare la bontà dell'approccio. I risultati indicano che, nonostante alcune limitazioni, in uno dei due casi studio abbiamo ottenuto risultati incoraggianti.