Riassunto Tesi di Laurea

Candidato: Cipriani Simone simone.cipriani@stud.unifi.it
sim.cipr@gmail.com

Relatore: *prof.* Donatella Merlini donatella.merlini@unifi.it

Titolo *(italiano)*: **Analisi con tecniche di *data mining* su dati relativi ad insegnamenti e studenti del Corso di Laurea in informatica**

Titolo *(inglese)*: **Mining and analysis of courses and students data concerning the Computer Science Degree**

La tesi oggetto di questo riassunto riporta una descrizione dell'attività di *data mining* svolta su dati riguardanti studenti e insegnamenti del Corso di Laurea Triennale in informatica.

Sono stati messi a disposizione due insiemi di dati: uno relativo alla carriera universitaria di quattro coorti di studenti iscritti al Corso di Laurea, sotto forma di record anonimi; l'altro riguardante la valutazione dei corsi di insegnamento espressa dagli studenti stessi, in forma aggregata rispetto alle voci del questionario. Su questi dati sono state impiegate delle tecniche di visualizzazione e di *data mining*, al il fine di estrarne qualche informazione utile a migliorare la comprensione di alcuni aspetti del Corso di Laurea.

Per raggiungere questi obiettivi, è stato scelto di usare come *data base management system* il software MongoDB, una tecnologia che segue un paradigma innovativo nell'ambito della gestione di grandi masse di dati, agilmente controllabile tramite degli script in linguaggio *Python* grazie all'apposito driver *pymongo*. Inoltre, è stato utilizzato il software *Weka*, uno strumento *open source* che mette a disposizione delle implementazioni di vari algoritmi utili per il *data mining*. Ci si è inoltre avvalsi di ulteriori strumenti - quali, ad esempio, il linguaggio *R* - per realizzare le tecniche di visualizzazione, oltre a varie *utilities* per gestire ed organizzare il flusso di lavoro.

Sono stati inizialmente esplorati i singoli set di dati, analizzandoli visivamente tramite l'impiego di tecniche di visualizzazione: in questa fase, è stata intuita una correlazione diretta fra la valutazione di un corso da parte degli studenti e le loro performances in esso, intese come voto ottenuto nel relativo esame. Si è quindi provveduto ad effettuare un'operazione di *join* fra le due collezioni a disposizione, aggregando i dati degli studenti rispetto ai risultati negli esami ed incrociandoli con le valutazioni degli insegnamenti, ottenendo così una collezione di record adatta per essere analizzata con tecniche di *data mining*.

Si è riusciti, tramite l'algoritmo *K-Means*, ad ottenere un buon *clustering* sull'insieme di dati risultante dalle operazioni sopra descritte; tale *clusering* li raggruppa in due *cluster*, uno comprendente gli esami considerati migliori e l'altro quelli considerati peggiori rispetto ai seguenti attributi: voto conseguito all'esame, ritardo nel superarlo e valutazione data dagli studenti al corso di insegnamento.

Sono state inoltre ottenute delle regole associative con l'algoritmo *Apriori*, che confermano con un'elevata confidenza la sospettata correlazione fra la valutazione dei corsi e le performance degli studenti in essi.

È stata infine effettuata una ricerca di pattern sequenziali frequenti utilizzando l'algoritmo *GSP*, considerando solamente l'ordine con cui gli studenti superano gli esami del Corso di Laurea. Dai pattern frequenti sono quindi stati estratti gli esami la cui posizione diverge dalla sequenza ideale, confermando che tali esami sono spesso superati dagli studenti dopo altri teoricamente a loro successivi.