CLAUDIO - CLassificazione AUtomatica dei Documenti InfOrmatici

Committente: **CNIPA**
Inizio : **Marzo 2008**
Fine: **Settembre 2009**

Obiettivi
Realizzazione di un applicativo software open source capace di operare una classificazione automatica o semi-automatica dei documenti, registrati in un sistema di protocollo informatico, ed integrabile nella maggior parte dei prodotti attualmente in uso nelle Pubbliche Amministrazioni.

Soluzione
Il prototipo software realizzato parte dalla considerazione che esiste in ogni amministrazione un set, più o meno ampio, di documenti correttamente classificati che possono essere utilizzati per addestrare il sistema di classificazione automatico dopo essere stati validati da esperti del dominio. L'esistenza di questo corpus di documenti già classificati offre l'opportunità di costruire un Machine Learned Model (MLM) utilizzato per proporre il codice di classificazione di un documento basandosi su una analisi computazionale del contenuto del campo oggetto inserito da un operatore. La criticità dei sistemi di classificazione automatica risiede nella metodologia di calcolo della similarità tra documenti. Il prototipo sviluppato utilizza il TF-IDF. A ciascuna query, sulla base dei risultati ottenuti, viene attribuito un punteggio (score) di similarità. Il punteggio è calcolato attraverso la seguente formula:

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum_{t \in q} tf(t \in d) \cdot idf(t)^2 \cdot t.getBoost(\,) \cdot norm(t,d)$$

Lo score misura la similarità di q, la query data, con d, un oggetto indicizzato nel serarch engine. Il risultato è costituito dal record associato alla query e da ogni record che contiene la rispettiva classificazione. Il risultato tipico conterrà centinaia di records. I records sono ordinati secondo un metodo del pivot sul valore della classe che serve come chiave primaria. Un punteggio di classificazione è generato dalla media tra il punteggio più alto di similarità e la media geometrica di tutti i punteggi di similarità. La classe che ottiene il punteggio maggiore viene proposta come prima candidata per la possibile classificazione.