

Sintesi del tema di specifico interesse

Import a verifica sintattica e semantica di documenti tecnici in un gestore documentale opensource in cui il dato di ingresso è un documento destrutturato.

L'azienda A si presenta sul mercato come una società di informatica e telecomunicazioni fortemente orientata ai servizi, in grado di proporre soluzioni alle aziende relativamente alle necessità di sviluppo e mantenimento della propria infrastruttura informatica.

Più precisamente l'azienda A opera fondamentalmente in tre settori:

1. realizzazione di impianti di rete (cablaggi, apparati attivi, certificazioni) e fonia;
2. realizzazioni di interconnessioni su area geografica (connessioni WAN, VPN, WiFi ecc.) e relativi aspetti di sicurezza (firewalling) e fonia avanzata su rete a commutazione di pacchetto (VoIP);
3. installazioni di hardware e software per rete locale, quindi: client (PC o ThinPC), server applicativi (W2K3, Linux) e di comunicazioni (proxy server), più una serie di moduli server legati alla comunicazione/condivisione di informazioni (mail server, fax server, application server sullo stile Citrix).

Nel proprio lavoro i tecnici dell'azienda attingono ad una serie di informazioni e di dati tecnici, e parallelamente ne producono di proprie, tanto da necessitare di un sistema di gestione unico che consenta di reperirle in altri momenti, rispetto a quelli in cui si sono utilizzate o prodotte, in modo estremamente semplice e veloce.

A questo scopo è stato adottato in azienda una *Knowledgebase* interna, in quanto alimentata esclusivamente dai propri tecnici, basata sul software opensource OWL (<http://owl.sourceforge.net>).

Il software in questione è basato su WEB, esiste per piattaforma Linux ed è scritto in linguaggio PHP. Il motore dell'applicazione installato è rappresentato dal database relazione opensource MySQL.

Trattandosi di applicazione web il server di base è rappresentato da Apache2. L'applicazione è pubblicata in Internet grazie alla soluzione Access Gateway di Citrix: per questioni di riservatezza e sicurezza la applicazione è stata spostata su HTTPS e collegata per l'accesso alla Active Directory che presiede al dominio Microsoft Interno (autenticazione LDAP).

Oggi il server che esegue l'applicazione è rappresentato da una macchina virtuale, offerta da Vmware server. I dati che i tecnici si trovano a dover inserire sono di varia natura:

- 60% documenti di testo variamente formattati (TXT, DOC, ODT, RTF, PDF);
- 10% schermate di configurazione di moduli software (BMP, TIFF, JPEG);
- 5% diagrammi di rete, tipicamente in formato EDG o PDF;
- 4% schemi forniti dal cliente, tipicamente in PPT o DXF;
- 15% pagine html sotto forma di file unici, ma anche di link contenenti nella pagina ulteriori link ad altre pagine, ricavati dai siti di supporto tecnico dei prodotti;
- 5% immagini fotografiche della installazione;
- 1% documenti video che registrano le operazioni di installazione/configurazione di software).

Ad oggi queste informazioni vengono inserite in una mail che viene spedita ad una casella di posta, sita sul server di posta aziendale, rappresentato da un Microsoft Exchange 2003 server, utilizzato in modalità MAPI dai vari client di rete.

Solo poche regole di formattazione dell'oggetto della mail vengono richieste agli scriventi, mentre la massima libertà viene lasciata per la formulazione del corpo della mail e per la scelta del numero e della tipologia degli allegati. Il messaggio così prodotto viene quindi inviato ad un utente prestabilito del sistema di posta e qui mantenuto, finché un operatore incaricato non provvede, manualmente, ad aprire la mail e dopo averla analizzata provvede ad inserirla nel gestore documentale prima specificato, nella categoria corretta.

Ovviamente le informazioni sono già oggi catalogate secondo criteri che si sono rivelati nel tempo sufficientemente pratici per la consultazione.

L'obiettivo del lavoro è quindi la produzione di un software che automatizzi l'inserimento delle informazioni nella banca dati quindi che realizzi il processo che dalle mail che vengono recapitate nella cassetta postale dedicata (faq) conduce alla immissione in un punto preciso della knowledgebase interna.

Si richiede che il software gestisca una forma base di apprendimento e quindi che incorpori un meccanismo di proposta/approvazione della classificazione, a servizio dell'operatore che oggi effettua la decisione su tutte le informazioni. Priorità assoluta ha il riconoscimento dei documenti testuali, mentre si richiede il solo collegamento dei contenuti multimediali.

Non si esclude la possibilità di integrare il sistema documentale OWL, sempre nel rispetto della licenza GNU che lo caratterizza, con form che guidino il tecnico nella collezione dei dati di interesse.