ricerca e sviluppo

Aquasearch Portal: una piattaforma di ricerca nel Web

Elisabetta Zuanelli e Mirto Silvio Busico

1. Presentazione

Il portale “Aquasearch” è un sistema di ricerca di contenuti in Internet correlati all'acqua e alla sua gestione. La ricerca per la modellizzazione e lo sviluppo di motori di ricerca innovativi per il recupero dell'informazione in Internet non può confrontarsi in alcun modo con i giganti in materia, primo fra tutti Google.

L'impero dell'informazione globale è un irresistibile successo di chi ha investito, *ante litteram*, nel valore e nel potere della conoscenza globale. Ha affermato Patrick Pichette, *chief financial officer* di Google:

> *We had a great quarter with 27% year-over-year revenue growth. These results demonstrate the value of search and search ads to our users and customers, as well as the extraordinary potential of areas like display and mobile. It's clear that our past investments have been crucial to our success today – which is why we continue to invest for the long term*[1].

L'affermazione del CFO di Google riassume la visione vincente di una nuova generazione di servizi di comunicazione digitale generati da investimenti nelle nuove frontiere di ricerca e sviluppo in Internet.

I siti-sistemi informativi sull'acqua in Internet nelle diverse prospettive (come risorsa naturale, problema ambientale, gestione industriale, documentazione normativa, accordi di programma, ecc.) sono numerosi in ambito internazionale.

Problema generale è l'organizzazione caotica di questa imponente massa di informazioni. Di specifico interesse politico-istituzionale è l'ambiente mediterraneo, ma, più in generale, le aree della terra in fase di sviluppo. La necessità da parte degli operatori politico-istituzionali e industriali di disporre in maniera organica e mirata di informazioni e dati fruibili nei diversi contesti ha dunque condotto a un primo prototipo di piattaforma-motore di ricerca in Internet sui siti dell'acqua.

"Aquasearch" si basa su analisi, filtro, riorganizzazione della massa di informazioni e delle logiche di accesso ai siti e ai portali delle diverse istituzioni/Enti/iniziative presenti in Internet.

Il prototipo di motore di ricerca dispone di un'ontologia di contesto, da espandersi ulteriormente in un'ontologia dell'acqua su base tematica, che orienta e dà accesso diretto ai link dei siti specifici, in ordine ai diversi soggetti, iniziative, documenti, motori di ricerca presenti nei siti web dell'acqua. I contenuti indicizzati e catalogati provengono da siti selezionati accuratamente in base alla loro rilevanza.

A differenza dei normali motori di ricerca che "guardano" al contenuto delle risorse indicizzate, in "Aquasearch" viene gestita una serie di relazioni tra i contenuti e una struttura gerarchica di categorizzazioni. Le ricerche vengono effettuate sulle categorie definite. Per tutte le categorie e sottocategorie, trovate corrispondenti ai criteri di ricerca, vengono presentati i dati che sono in relazione con esse.

---

[1] GOOGLE INC., *Press release announcing the 1st quarter 2011 results*, Mountain View (Calif.), April 14, 2011 = http://investor.google.com/pdf/2011Q1_earnings_google_revised.pdf

Se, per esempio, si effettua una ricerca del termine "*legislation*" vengono trovate le categorie che hanno attinenza con tale parola: "*documents*", "*legislation*", "*European legislation*". Per ognuna di tali categorie vengono estratti i dati che sono stati messi in relazione con esse.

Nella Figura 1, per "*documents*" vengono presentati: il sito di "*ARC – America's River Communities*", l'attività "*CIS – Common Implementation Strategy*", il documento "*Aarhus Convention – Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters*" e altri.

Figura 1. Ricerca di categorie.

Da notare che, per ogni tipo di dato, la presentazione viene effettuata in un'apposita scheda che presenta informazioni differenti; in particolare:

- *institution*: sigla, nome esteso, link alla pagina da cui sono tratte le informazioni, descrizione, link al sito al quale appartiene la pagina originale;
- *activity*: sigla, nome esteso, link alla pagina da cui sono tratte le informazioni, descrizione, link al sito al quale appartiene la pagina originale;
- *document*: sigla, nome esteso, link alla pagina da cui sono tratte le informazioni, descrizione, link al sito al quale appartiene la pagina originale;

- *event*: sigla, nome esteso, link alla pagina da cui sono tratte le informazioni, descrizione, link al sito al quale appartiene la pagina originale;
- *search engine*: sigla, nome esteso, link alla pagina da cui sono tratte le informazioni, descrizione, link al sito al quale appartiene la pagina originale;
- *site*: sigla, nome esteso, link al sito, descrizione, una tabella con i dati, raggruppati per tipo, che sono stati trovati all'interno del sito stesso.

In questo modo si garantisce che la ricerca fornisca una serie di dati riassuntivi, sicuramente rilevanti, che permettono l'accesso diretto alle informazioni originali. Viene anche fornita una ricerca relativa ai soli "motori di ricerca" che sono presenti nei siti. Inoltre è anche possibile ricercare i siti per nome o per URL.

2. Funzionalità

Le funzionalità offerte all'utente sono quelle descritte in Figura 2, vale a dire:

- la ricerca per categorie;
- la selezione dei *search engine*, che possono essere selezionati in base al contenuto del nome o della descrizione, oppure in base alla tipologia: professionali, relativi alle istituzioni o tematici;
- la ricerca dei siti catalogati, che possono essere selezionati in base al contenuto del nome o della descrizione, oppure in base all'URL.

Figura 2. Funzionalità offerte all'utente.

Le ricerche hanno come esito la generazione di “schede”, raggruppate in base alla categoria di appartenenza, che hanno due tipologie di contenuti. Per i siti (Figura 3): sigla, nome esteso, link al sito (pulsante “Site”), descrizione, una tabella collassabile con i dati, raggruppati per tipo, che sono stati trovati all'interno del sito stesso. Il titolo del dato è un link diretto alla pagina di riferimento.

Figura 3. Scheda di sito.

Per tutti gli altri tipi di dato (Figura 4) sigla, nome esteso, link alla pagina da cui sono tratte le informazioni (pulsante “Activity page”), descrizione, link al sito al quale appartiene la pagina originale (pulsante “Site”).

Figura 4. Scheda per altri tipi di dati.

Dietro le quinte, c'è un'interfaccia di gestione (Figura 5) che viene utilizzata dai gestori del database che contiene i dati. A cura di questo personale qualificato nella comprensione e gestione dell'assegnazione alle categorie, vengono effettuati i trattamenti che trasformano i dati grezzi in informazioni a valore aggiunto.

Figura 5. Interfaccia di gestione.

A fronte di una estrazione dei link contenuti nei siti da esaminare, è necessario un lavoro altamente qualificato (Figura 6) per:

- riassumere e sintetizzare le informazioni relative ai link per generare sigla, descrizione breve e descrizione lunga;
- impostare la correlazione tra l’elemento ed il sito di appartenenza;
- assegnare l'elemento a tutte le categorie di appartenenza.

Quest'ultima attività è il punto qualificante di tutto il sistema.

Figura 6. Interfaccia di amministrazione.

3. Architettura

Il sistema evolve continuamente mettendo a disposizione differenti prototipi che offrono funzionalità sempre più complesse. L'architettura qui presentata si riferisce al quarto prototipo (Figura 7).

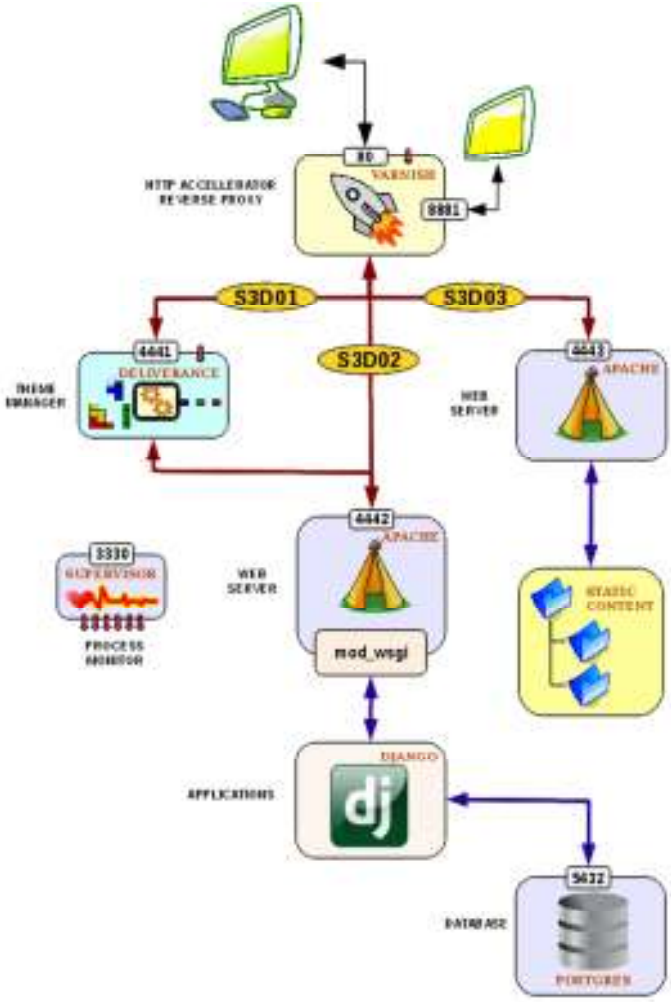

Figura 7. Architettura del prototipo 4.0

Si accede al sistema per mezzo di tre URL differenti:

- l'URL di accesso primario dedicato agli utenti;
- l'URL di accesso dedicato agli amministratori del sistema;
- l'URL di servizio per l'erogazione di contenuti statici.

Il servizio di *front-end* principale viene effettuato con una *web-cache* con funzionalità di reverse *proxy*.

I dati sono gestiti con un database relazionale *Postgres* che viene acceduto e gestito per mezzo di un'applicazione Django. L'applicazione Django offre i suoi servizi per mezzo di un web server Apache che si interfaccia via WSGI.

Questo è l'accesso utilizzato per gli amministratori. L'accesso per gli utenti attraversa un ulteriore strato di software rappresentato da un "motore" dei temi (*deliverance*) che si occupa dell'aspetto grafico del sito.

4. Tecnologie

Le tecnologie utilizzate per questo prototipo sono:

a) **Python**: linguaggio di programmazione per la realizzazione dell'applicazione (http://www.python.org/)

b) **PostgreSQL**: database relazionale che permette una grande scalabilità e possiede estensioni che consentono di ampliarne le finalità di utilizzo come PostGIS, che offre la gestione di dati geografici per applicazioni di *Geographic Information System* (GIS) (http://www.postgresql.org/)

c) **Django framework**: di sviluppo per applicazioni scritte in Python basate su database relazionali (https://www.djangoproject.com/)

d) **Apache Web Server**: espandibile con moduli aggiuntivi (http://httpd.apache.org/)

e) **MOD_WSGI**: modulo di Apache che consente l'esecuzione di applicazioni Python con il protocollo WSGI (http://code.google.com/ p/modwsgi/)

f) **Python-Deliverance**: gestore di temi visuali per siti web (http://pypi.python.org/pypi/Deliverance/0.5.0)

g) **Varnish**: *web-cache* e *reverse proxy* (https://www.varnish-cache.org/)

5. Direzioni future

Aquasearch Portal è un'applicazione in continua evoluzione. Nuove tecnologie vengono valutate, testate e, se adeguate, vengono integrate nel *software* della piattaforma. La piattaforma stessa è progettata per essere modulare e scalabile, per permetterne un impiego su larga scala.

Periodicamente una fase consolidata di sviluppo viene rilasciata per l'accesso pubblico. La fase attuale viene chiamata "quarto prototipo". Le tecnologie che sono in fase di valutazione per il quinto prototipo e le relative funzionalità aggiuntive sono:

a) **OpenId**: autenticazione degli utenti per gestire parti riservate del sito (per esempio per l'offerta di servizi a pagamento). Una parte del sito rimarrà pubblica per consentire la divulgazione. Servizi a valore aggiunto potranno essere offerti ad utenti selezionati.

b) **Merengue**: *Content Management System* (CMS) basato su Django che consente la gestione di contenuti multilingui. Questo consentirà ricerche in lingue differenti: le categorie verranno ricercate utilizzando linguaggi differenti, ma i contenuti trovati saranno gli stessi in quanto la relazione tra dato e categoria

prescinde dal linguaggio. Questo CMS integra anche la gestione di dati geografici.

c) **MapServer**: server di dati geografici che consente la generazione di mappe multistrato che aggregano informazioni di interesse. Ecco un esempio di mappa che riporta informazioni su acqua, città e strade: