La ricerca delle informazioni nei siti web di Ateneo con Google Search Appliance

Progetto, implementazione e sviluppi

Il progetto del sistema di ricerca delle informazioni

L'esigenza del sistema di ricerca delle informazioni nei siti web di Ateneo era di individuare una tecnologia che consentisse ai visitatori del portale Unina e ai visitatori dei siti web istituzionali collegati, di effettuare ricerche rapide e complete all'interno del grande volume di informazioni disponibili. Le tecnologie preesistenti all'avvio del progetto, basate su soluzioni software open source, risultavano, infatti, poco efficienti dal punto di vista delle prestazioni oltre che onerose da mantenere e gestire in termini di attività richiesta al personale del supporto informatico.

Definita l'esigenza, nella fase iniziale del progetto sono stati individuati i requisiti di base della soluzione da implementare, requisiti sintetizzati nei seguenti item:

- semplicità e rapidità di implementazione nella fase iniziale;
- impegno minimo, da parte del supporto informatico, per la gestione e la manutenzione del sistema nella fase di esercizio;
- efficienza delle prestazioni, sia in termini di tempo necessario all'indicizzazione che di risposta alle richieste di ricerca;
- possibilità di estendere le funzioni base di ricerca attraverso l'inclusione nei risultati di informazioni provenienti da dati strutturati non presenti in pagine web (database);
- buon rapporto prezzo (iniziale di acquisizione + manutenzione)/prestazioni.

Successivamente sono state analizzate le varie soluzioni disponibili sul mercato che potessero rispondere a tali requisiti: dopo un'analisi delle opzioni disponibili, è stata individuata in Google Search Appliance la soluzione ideale, sia dal punto di vista della semplicità di installazione e personalizzazione che dal punto di vista del rapporto prezzo/prestazioni, oltre al fatto che Google è un brand leader nel settore della ricerca nel web.

Le caratteristiche del motore di ricerca

Il motore di ricerca Google Unina consiste in un sistema dedicato GB1001 (hardware e software) installato nella sala server del CSI, che indicizza in modo completamente automatico le pagine web del portale di Ateneo e le pagine dei siti web istituzionali ad esso collegati. Il server è mostrato in figura 1:

Figura 1 – Unità Google Search Appliance GB1001

L'indicizzazione, eseguita automaticamente ogni notte per non interferire con l'usuale carico diurno dei visitatori dei siti, produce un database di documenti su cui è possibile effettuare ricerche in testo libero attraverso la classica casella "cerca nel sito" presente nella parte superiore di tutte le pagine del portale di Ateneo e nei siti web istituzionali ad esso collegati: il processo che porta alla generazione dell'indice è mostrato schematicamente nella figura 2:

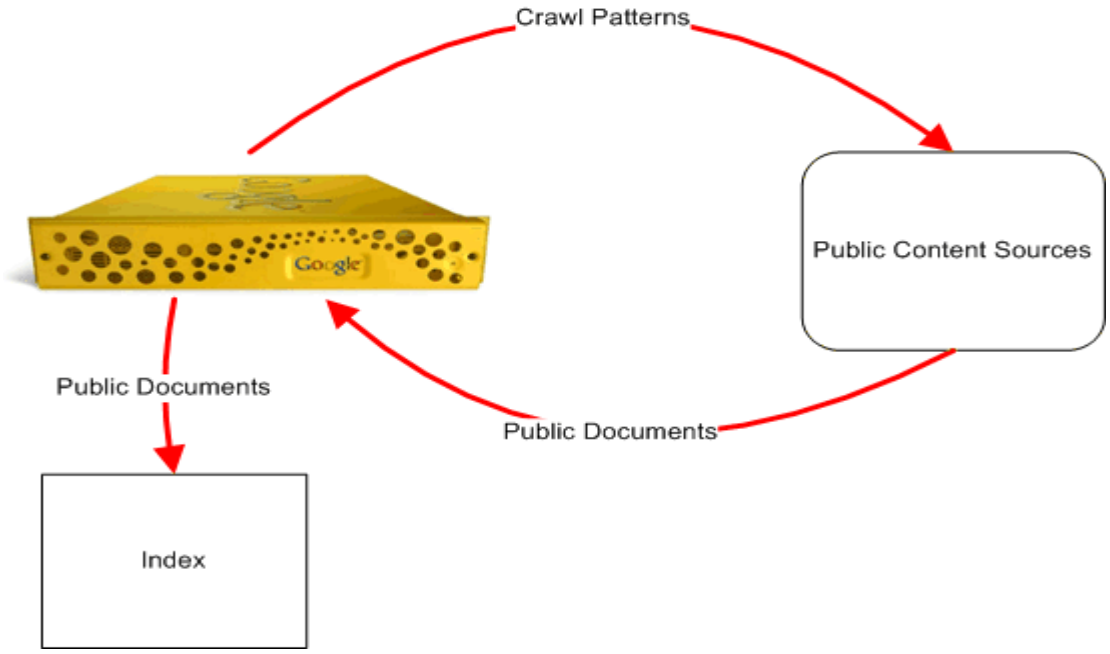

Figura 2 – Generazione dell'indice dei documenti

La capacità di indicizzazione del motore GB1001 di Unina, il più piccolo della serie, è pari a 500.000 *documenti*, dove per *documento* si intende sia la singola pagina web che i veri e propri documenti allegati. Mediante la tecnologia Google $(GSA)^n$, ossia utilizzando più server collegati tra loro, il sistema può scalare fino a diversi miliardi di documenti. Il motore di indicizzazione, quello che in gergo viene chiamato "crawler", usa la stessa tecnologia di ricerca Google per internet e supporta 220 tipi diversi di file sui quali è possibile eseguire ricerche anche per contenuto.

Implementazione

Attualmente nell'indice sono presenti complessivamente circa **279.000** *documenti*, provenienti dal portale Unina di Ateneo, dai siti web di Facoltà, dal sito di web learning Federica e da web docenti. Questi *documenti* sono suddivisi in indici distinti per sito in modo che ad un utente che effettui, ad esempio, una ricerca sul portale Unina siano restituiti solo i risultati provenienti da tale sito. Nel primo semestre 2010, solo sul portale Unina, sono state eseguite 257.550 ricerche la cui distribuzione oraria è mostrata nel grafico di Figura 3:

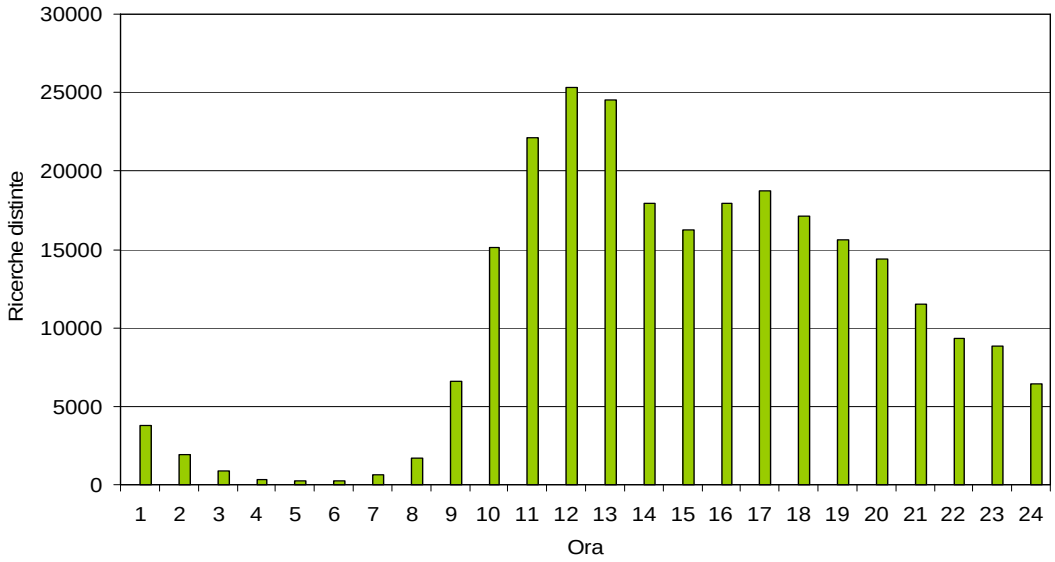

Figura 3 – Distribuzione oraria ricerche nel I semestre 2010 sul portale Unina

Le prime cinque query degli utenti, dove per query si intende tutti i termini che l'utente immette nella casella di ricerca, sempre nel I semestre del 2010 e solo per il portale Unina, sono riportate nella tabella 1 con la relativa ricorrenza:

Query	**Ricorrenza**
esis	8578
part time	3904
web docenti	3314
docenti	1652
esami di stato	1493

Tabella 1 – Prime cinque query nel I semestre 2010 sul portale Unina

Il carico medio del sistema è visualizzato nel grafico di Figura 4, dove l'asse x rappresenta il tempo e l'asse y rappresenta le query ricevute al minuto:

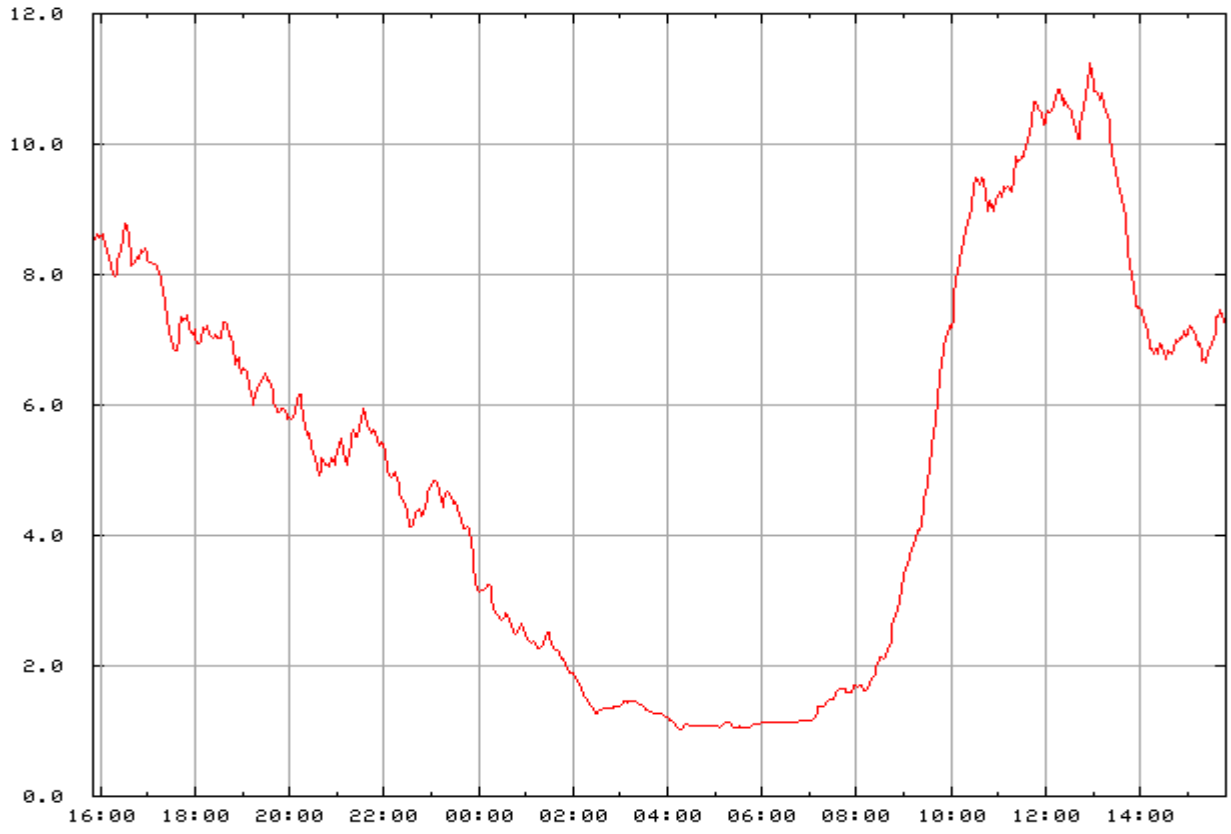

Figura 4 – Carico medio del sistema

Sviluppi

Successivamente all'implementazione iniziale, si è avviato lo sviluppo di nuove funzionalità con l'obiettivo di:

1) migliorare l'esperienza di ricerca degli utenti,
2) estendere i contenuti dell'indice includendo dati provenienti da database di interesse.

Nel primo obiettivo rientrano l'implementazione dei *keymatch* e del protocollo *Open Search Description*.

I *keymatch* consistono in suggerimenti proposti nella pagina dei risultati della ricerca a fronte dell'utilizzo, da parte dell'utente, di determinate parole chiave. Il suggerimento è proposto all'utente con un testo esplicativo e con l'indicazione del link del servizio cercato. Nella successiva figura 5 sono mostrati alcuni esempi di keymatch implementati:

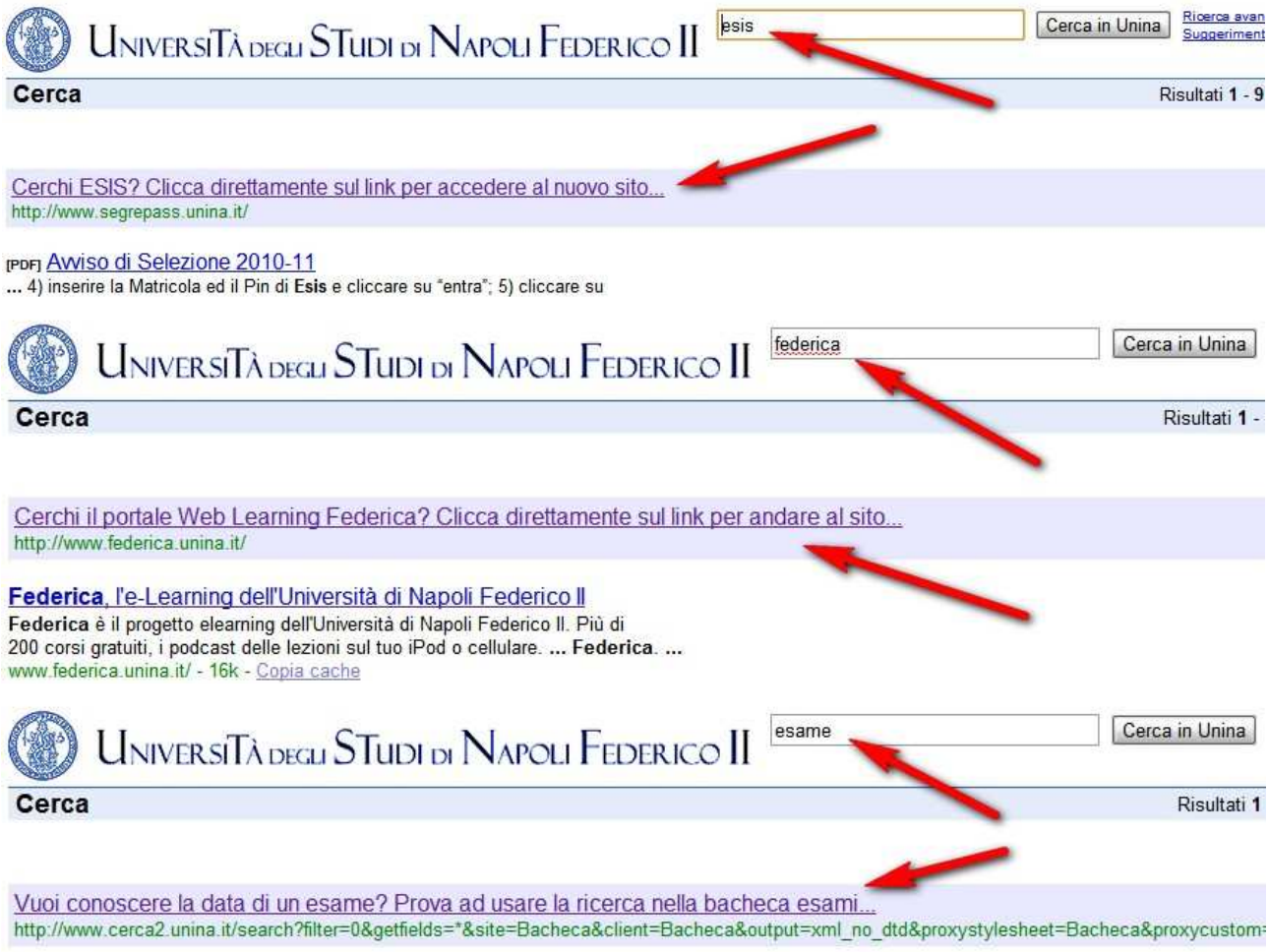

Figura 5 – Esempi di keymatch sul motore di ricerca Unina

L'utilizzo dei keymatch e delle parole chiave utilizzate nelle ricerche è periodicamente monitorato con gli strumenti di reportistica del server Google Search Appliance per verificare l'efficacia dei suggerimenti e per determinare nuovi keymatch utili per le ricerche più frequenti degli utenti.

Open Search Description è invece un protocollo che consente di descrivere un motore di ricerca in modo che il motore possa essere memorizzato, sotto forma di provider di ricerca, direttamente

nel browser dell'utente: ciò consente ai visitatori del portale di memorizzare nel proprio browser il motore di ricerca Google di Unina semplificando e velocizzando l'esperienza di ricerca delle informazioni presenti sul sito. In pratica, cliccando semplicemente su un link che è presentato nella pagina dei risultati della ricerca, l'utente può salvare nell'elenco dei motori di ricerca del suo browser, il motore di ricerca di Unina. Una volta che l'utente abbia effettuato questa operazione, potrà eseguire direttamente delle ricerche sul sito Unina attraverso i box di ricerca del suo browser (Internet Explorer e Firefox) oppure attraverso la barra degli indirizzi di Chrome, semplicemente collegandosi alla rete Internet, avviando il browser e scegliendo come motore di ricerca Google di Unina. Nella figura 6 è mostrato come si presenta il box di ricerca nel browser Firefox dopo che l'utente ha salvato la definizione del motore Google di Unina:

Figura 6 – Google Unina nel box di ricerca del browser Firefox

Le istruzioni su come usufruire di questa funzionalità sono disponibili sul sito web del CSI.

Al secondo obiettivo appartengono la realizzazione della *ricerca nella bacheca esami degli studenti* e la *ricerca nel catalogo dei prodotti della ricerca*. Entrambe queste applicazioni sfruttano la capacità del server Google di indicizzare database relazionali per estrarre dati strutturati da includere nell'indice del server e renderli quindi disponibili per operazioni di ricerca da parte degli utenti. Lo schema di principio di questo meccanismo è mostrato nella successiva figura 7:

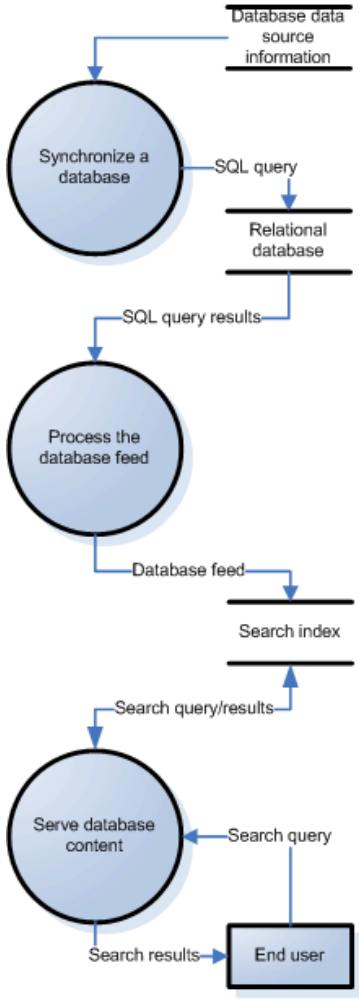

Figura 7 – Indicizzazione di database

Nella fase di sincronizzazione, il server Google accede alle tabelle del database relazionale specificato nel *data source information* e attraverso le query SQL definite dall'amministratore dei dati estrae dal database tutti i record della selezione. Nella seconda fase, i risultati della query sono elaborati dal server Google e trasformati in un flusso dati XML (feed) che è aggiunto all'indice generale del server. Nella terza fase, quando un utente effettua una ricerca sui dati contenuti nell'indice, il server Google elabora come risposta un elenco di link ciascuno dei quali costituisce una query: solo quando l'utente clicca sul link dell'elenco dei risultati della ricerca, la singola query viene eseguita in modo che lo specifico record è estratto dal database e mostrato all'utente.
Il vantaggio di una architettura di questo tipo sono sostanzialmente due:

a) gli utenti possono effettuare ricerche in insiemi di dati strutturati utilizzando semplicemente parole chiave, analogamente a quanto fanno per le usuali ricerche sul web;
b) i tempi di risposta delle ricerche sono ridottissimi ed indipendenti dal database in quanto le ricerche sono effettuate nell'indice del server che è statico ed ottimizzato per tale funzione.

La *ricerca nella bacheca esami degli studenti* indicizza ogni sera tutti gli appelli dei docenti dell'Ateneo presenti nel database Oracle del sistema di gestione delle segreterie studenti (GEDAS). La pagina di ricerca si presenta come mostrato nella successiva figura 8:

Figura 8 – Ricerca nella bacheca esami

E' possibile effettuare ricerche libere, per nome insegnamento, per cognome e/o nome del docente, per intervallo di date oppure con una qualsiasi combinazione di queste chiavi. Specificati i termini della ricerca, i risultati sono visualizzati come mostrato nell'esempio della successiva figura 9 nella quale è stata eseguita una ricerca per cognome:

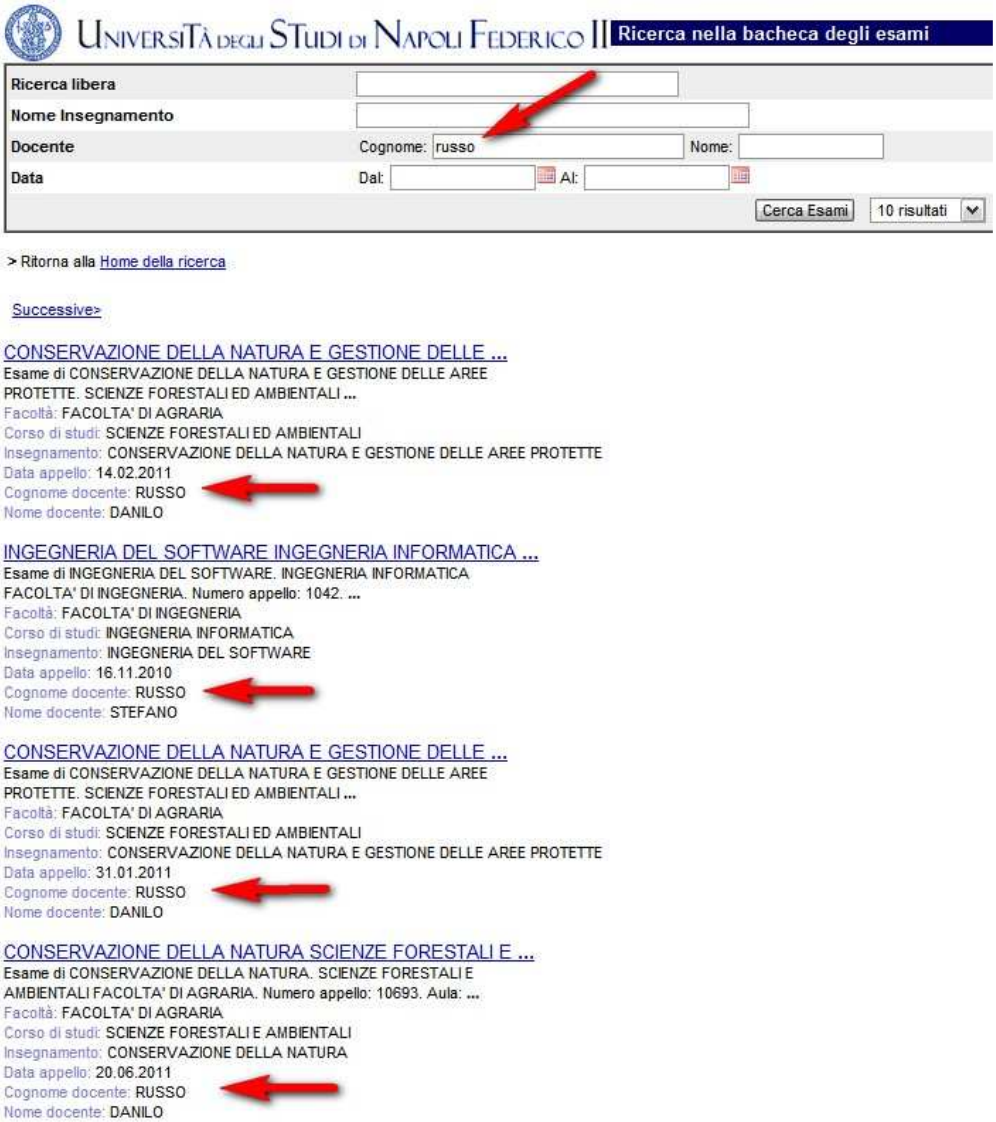

Figura 9 – Esempio risultati ricerca nella bacheca esami

Nella lista dei risultati della ricerca sono visualizzate tutte le informazioni che caratterizzano l'appello, ossia la denominazione completa dell'insegnamento, la facoltà, il nome del corso di studi, la data dell'appello, il cognome ed il nome del docente. Se l'utente clicca su uno dei link indicati nella lista, viene eseguita la query specifica sul database con la visualizzazione di informazioni aggiuntive rispetto a quelle mostrate nella lista (numero appello, aula e ora).
La ricerca nella bacheca esami degli studenti è integrata in web docenti ed è suggerita dall'uso nel portale Unina dei keymatch: *bacheca*, *esame*, *esami*, *appello*, *appelli*, *data appello*, *data esame*, *calendario esami*.

La *ricerca nel catalogo dei prodotti della ricerca* è stata sviluppata per indicizzare i prodotti della ricerca del **Polo delle Scienze e delle Tecnologie dell'Ateneo**. La pagina di ricerca, raggiungibile direttamente dalla home di Unina alla voce "catalogo della ricerca", si presenta come mostrato nella successiva figura 10:

Figura 10 – Ricerca nel catalogo dei prodotti della ricerca del Polo delle Scienze e delle Tecnologie

Diversamente dalla bacheca esami, in questa ricerca esiste un unico campo di ricerca libero a fianco del quale viene mostrato un menu a tendina che consente all'utente di selezionare in quale tipologia di prodotti effettuare la ricerca; selezionata la tipologia di interesse, i risultati della ricerca sono visualizzati come mostrato nell'esempio della successiva figura 11 nella quale è stata eseguita una ricerca per parola chiave *Marrucci* in *Tutte le tipologie* di prodotti della ricerca:

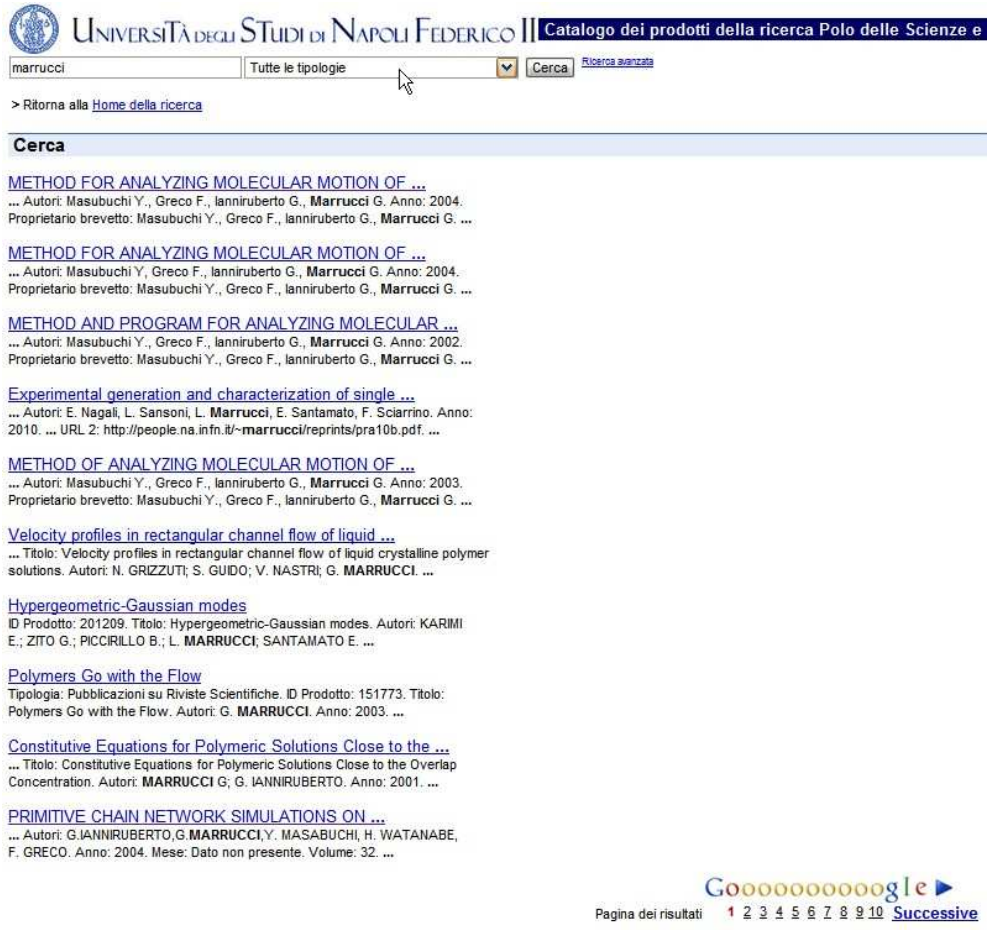

Figura 11 – Esempio pagina di risultati della ricerca nel catalogo dei prodotti della ricerca

Selezionando uno dei link visualizzati nella pagina dei risultati viene mostrato il dettaglio dell'informazione, come mostrato nella successiva figura:

Tipologia: Pubblicazioni su Riviste Scientifiche

ID Prodotto: 370015

Titolo: Photonic quantum information applications of patterned liquid crystals

Autori: L. Marrucci, E. Nagali, F. Sciarrino, L. Sansoni, F. De Martini, B. Piccirillo, E. Karimi, E. Santamato

Anno: 2010

Mese: Dato non presente

Volume: 526

Fascicolo: Dato non presente

Numero pagine: 11

Descrizione Prodotto: In this paper we review recent results we obtained in the field of photonic quantum information that were made possible by transfer of a qubit of quantum information from the spin to the orbital angular momentum of photons and vice versa; (iii) the Hong-Ou-Mandel coal encoded qubits.

Nota: Dato non presente

Codice DOI: 10.1080/15421406.2010.485118

Editore: Dato non presente

Codice Pubmed: Dato non presente

Nome rivista: MOLECULAR CRYSTALS AND LIQUID CRYSTALS

ISSN: 1542-1406

Rilevanza: Dato non presente

URL: http://people.na.infn.it/~marrucci/reprints/mclc10.pdf

URL 2: http://dx.doi.org/10.1080/15421406.2010.485118

URL 3: Dato non presente