

Università degli Studi di Padova

DIPARTIMENTO DI MATEMATICA "TULLIO LEVI-CIVITA"

CORSO DI LAUREA IN INFORMATICA



**Analisi di motori di ricerca Open Source
per siti web informativi**

Tesi di laurea triennale

Relatore

Prof. Tullio Vardanega

Laureando

Daniel De Gaspari

ANNO ACCADEMICO 2017/2018

placeholder con citazione.

— Oscar Wilde

Sommario

Il presente documento descrive il lavoro svolto durante il periodo di stage dal laureando Daniel De Gaspari, della durata di circa trecento ore, presso l'azienda InfoCamere S.C.p.A. di Padova (PD).

Gli obiettivi da raggiungere erano molteplici.

Lo scopo dello stage era quello di analizzare le caratteristiche dei motori di ricerca [open source](#) nell'ambito dei siti web di tipo informativo.

In primo luogo era richiesto un approfondimento delle caratteristiche istituzionali dei siti web delle Camere di Commercio.

Successivamente, l'azienda richiedeva di analizzare le potenzialità e specificità di motori di ricerca [Sori](#) e [ElasticSearch](#).

Il passo successivo consisteva nel realizzare un prototipo di un sito web in tecnologia [Drupal](#), con i due motori di ricerca precedentemente citati.

Infine, era richiesta una relazione finale delle potenzialità emerse nell'utilizzo dei due motori di ricerca.

I primi due capitoli del presente documento hanno lo scopo di presentare il contesto aziendale in cui è stato sostenuto lo stage e di spiegare come il progetto di stage si renda utile all'interno della strategia aziendale. Il terzo capitolo documenta invece lo svolgimento dello stage descrivendo le attività che sono state portate a termine, i punti salienti del progetto stesso e le principali scelte attuate. Il quarto ed ultimo capitolo presenta infine una valutazione dello svolgimento dello stage rispetto agli obiettivi aziendali e alle conoscenze acquisite dallo studente.

"Citazione"

— Confucius

Ringraziamenti

Ringraziamenti

Padova, Dicembre 2017

Daniel De Gaspari

Indice

1	L'azienda	1
1.1	Il Profilo Aziendale	1
1.2	Organizzazione aziendale	1
1.3	Tecnologie utilizzate	1
1.4	Rapporto con l'innovazione	1
2	Lo stage	3
2.1	Vantaggi aziendali	3
2.2	L'offerta di stage	3
2.2.1	Presentazione del progetto	3
2.2.2	Aspettative aziendali	3
2.2.3	Vincoli	3
2.3	Vantaggi personali	3
3	Resoconto dello stage	5
3.1	Individuazione dei motori di ricerca	5
3.2	Pianificazione	5
3.3	I siti istituzionali delle Camere di Commercio	5
3.3.1	Funzionalità di ricerca attuali	5
3.3.2	Possibile evoluzione	5
3.4	Ricerca nativa Drupal	5
3.4.1	Introduzione a Drupal	5
3.4.2	Ricerca di base e avanzata	5
3.4.3	Ricerca con Search API	6
3.4.4	Considerazioni di Drupal nativo	6
3.5	Ricerca con Solr	6
3.5.1	Introduzione a Solr	6
3.5.2	Principali funzionalità di ricerca	6
3.5.3	Integrazione con Drupal	6
3.6	Ricerca con Elasticsearch	6
3.6.1	Introduzione a Elasticsearch	6
3.6.2	Principali funzionalità di ricerca	6
3.6.3	Integrazione con Drupal	6
3.7	Considerazioni finali sui motori di ricerca esaminati	7

4	Valutazione retrospettiva	9
4.1	Bilancio degli obiettivi	9
4.1.1	Aziendali	9
4.1.2	Personalì	9
4.2	Conoscenze acquisite	9
4.3	Mondo del lavoro e università a confronto	9
	Glossario	11
	Acronimi	13
	Bibliografia	15

Elenco delle figure

Elenco delle tabelle

Capitolo 1

L'azienda

1.1 Il Profilo Aziendale

Questa sezione conterrà il contesto applicativo in cui l'azienda opera.

1.2 Organizzazione aziendale

Questa sezione conterrà l'organizzazione dell'azienda; essendo un'azienda molto strutturata e con molti dipendenti, l'organizzazione aziendale si concentrerà sul contesto in cui ho operato.

1.3 Tecnologie utilizzate

Questa sezione presenterà le tecnologie utilizzate dall'azienda, ristrette all'ambito in cui ho operato.

1.4 Rapporto con l'innovazione

Questa sezione presenterà il rapporto dell'azienda con l'avanzamento tecnologico.

Capitolo 2

Lo stage

2.1 Vantaggi aziendali

Questa sezione presenterà le motivazioni per cui l'azienda ospita stage.

2.2 L'offerta di stage

2.2.1 Presentazione del progetto

Qui verrà presentato il progetto di stage.

2.2.2 Aspettative aziendali

Qui verranno presentate le aspettative aziendali legate all'offerta di stage.

2.2.3 Vincoli

Qui verranno presentati i vincoli legati allo stage.

2.3 Vantaggi personali

In questa sezione descriverò le aspettative personali legate all'accettazione dell'offerta di stage.

Capitolo 3

Resoconto dello stage

3.1 Individuazione dei motori di ricerca

In questa sezione motiverò la selezione di Solr e Elasticsearch come motori di ricerca da studiare.

3.2 Pianificazione

In questa sezione presenterò come è avvenuta la pianificazione dello stage.

3.3 I siti istituzionali delle Camere di Commercio

3.3.1 Funzionalità di ricerca attuali

Questa sezione presenterà la struttura dei siti camerali attualmente in produzione, ponendo l'accento sugli strumenti messi a disposizione all'utente per ritrovare i contenuti in esso presenti.

3.3.2 Possibile evoluzione

Questa sezione presenterà una possibile evoluzione dei siti camerali, contenente funzionalità di ricerca attualmente non presenti.

3.4 Ricerca nativa Drupal

3.4.1 Introduzione a Drupal

Qui verrà introdotto Drupal, spiegando cos'è e come funziona.

3.4.2 Ricerca di base e avanzata

Qui verranno presentate le principali funzionalità offerte dalla prima tipologia di ricerca nativa Drupal.

3.4.3 Ricerca con Search API

Qui verranno presentate le principali funzionalità offerte dalla seconda tipologia di ricerca nativa Drupal.

3.4.4 Considerazioni di Drupal nativo

Questa sezione conterrà conclusioni riguardanti le funzionalità di ricerca offerte globalmente dalla ricerca nativa Drupal.

3.5 Ricerca con Solr

3.5.1 Introduzione a Solr

Qui verrà introdotto Solr, spiegando cos'è e come funziona.

3.5.2 Principali funzionalità di ricerca

Qui verranno presentate le principali funzionalità di ricerca offerte dal motore di ricerca Solr, di possibile interesse per i siti camerali.

3.5.3 Integrazione con Drupal

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal.

Apache Solr

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal mediante il modulo Apache Solr Search.

Search API Solr

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal mediante il modulo Search API Solr Search.

3.6 Ricerca con Elasticsearch

3.6.1 Introduzione a Elasticsearch

Qui verrà introdotto Elasticsearch, spiegando cos'è e come funziona.

3.6.2 Principali funzionalità di ricerca

Qui verranno presentate le principali funzionalità di ricerca offerte dal motore di ricerca Elasticsearch, di possibile interesse per i siti camerali.

3.6.3 Integrazione con Drupal

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Elasticsearch e Drupal.

Search API ElasticSearch

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra ElasticSearch e Drupal mediante il modulo Search API ElasticSearch.

3.7 Considerazioni finali sui motori di ricerca esaminati

Questa sezione conterrà un confronto tra le principali funzionalità, possibilmente di interesse per l'azienda, offerte dalle tecnologie esaminate e quale di queste potrebbe essere la più adatta ai siti camerali.

Capitolo 4

Valutazione retrospettiva

4.1 Bilancio degli obiettivi

4.1.1 Aziendali

Questa sezione descriverà gli obiettivi aziendali soddisfatti, derivanti dallo stage.

4.1.2 Personali

Questa sezione descriverà gli obiettivi personali soddisfatti, derivanti dallo stage.

4.2 Conoscenze acquisite

Questa sezione descriverà le conoscenze acquisite derivanti dallo stage.

4.3 Mondo del lavoro e università a confronto

Questa sezione analizzerà il gap tra gli insegnamenti universitari e il mondo dello stage, specificatamente allo stage svolto.

Glossario

API Insieme di procedure utilizzabili per interfacciarsi con un programma o un sistema informatico in modo standard. Spesso si intendono le librerie software disponibili in un certo linguaggio di programmazione. . 11–13

CMS E' un software per la realizzazione e la gestione di siti dinamici, che possono accrescere e mutare il proprio contenuto continuamente. Un [Content Management System](#) consente al committente del sito di occuparsi direttamente della sua gestione senza intermediari esterni. 11, 13

Drupal Drupal è un [Content Management System](#), rilasciato sotto licenza [open source](#), che permette la creazione di siti Internet, blog e portali, gallerie di immagini, forum di discussione, piattaforme intranet e molto altro. Essa è altresì un'applicazione completamente web based e può quindi essere utilizzata attraverso un semplice browser.
E' interamente sviluppato in [PHP](#) e utilizza come base di dati [MySQL](#) in modo nativo. v, 11

ElasticSearch Piattaforma di ricerca [open source](#), con capacità full text. E' un server di ricerca basato su [Java Lucene](#) e supporta architetture distribuite. Tutte le funzionalità sono nativamente esposte tramite interfaccia [RESTful](#); le informazioni sono invece gestite come documenti [JSON](#). v, 11

HTTP Formato adatto all'interscambio di dati fra applicazioni client-server. 12, 13

Java Linguaggio di programmazione ad alto livello, orientato agli oggetti e a tipizzazione statica, specificatamente progettato per essere il più possibile indipendente dalla piattaforma di esecuzione. 11, 12

Java Lucene API gratuita ed [open source](#) per il reperimento di informazioni, inizialmente implementata in [Java](#). 11, 12

JSON Formato adatto all'interscambio di dati fra applicazioni client-server. 11–13

Open source Software di cui i detentori dei diritti rendono pubblico il codice sorgente, permettendo ad altri programmatori di apportarvi modifiche. Questo meccanismo è regolato tramite l'applicazione di apposite licenze d'uso. v, 11, 12

MySQL Database relazionale largamente diffuso, composto da un client a riga di comando e un server. 11

PHP Linguaggio di scripting interpretato. 11, 12

REST Stile architetturale che offre la possibilità di manipolare rappresentazioni testuali di risorse Web utilizzando un set predefinito di operazioni. 12

RESTful Le applicazioni basate su **REST**, si definiscono RESTful e utilizzano le richieste **HTTP** per inviare i dati (creazione e/o aggiornamento), effettuare query, modificare e cancellare i dati. In definitiva, **REST** utilizza **HTTP** per tutte e quattro le operazioni CRUD (Create / Read / Update / Delete). 11, 12

Servlet Oggetti scritti in linguaggio **Java** che operano all'interno di un server web oppure un server per applicazioni, permettendo la creazione di web applications. 12

Solr Piattaforma di ricerca **open source**. E' scritto in **Java** e viene eseguito come server di ricerca full text indipendente all'interno di un contenitore **Servlet**. Solr usa la libreria di ricerca **Java Lucene** per la ricerca e l'indicizzazione full text e mette a disposizione chiamate **REST** come ad esempio **HTTP/JSON** e **XML API** che rendono semplice la comunicazione. v, 12

XML Metalinguaggio che consente la rappresentazione di documenti e dati strutturati su supporto digitale. 12, 13

Acronimi

API Application Programming Interface. 11

CMS Content Management System. 11

HTTP HyperText Transfer Protocol. 11

JSON JavaScript Object Notation. 11

XML eXtensible Markup Language. 12

Bibliografia