Tesi

Sommario

...

Introduzione

Tocca tutti gli argomenti. Scrivi alla fine.

Problema e obiettivi

# OBIETTIVI GENERALI

L'obiettivo dello stage è stato la realizzazione di un'applicazione web per la raccolta e la fruizione di dati relativi ai bandi pubblici.

L'applicazione ha lo scopo di raccogliere e memorizzare i **bandi delle gare pubbliche** pubblicati dalla **Gazzetta Ufficiale**, per poi notificare gli utenti registrati al sistema della presenza degli stessi, appena sono disponibili; gli utenti possono esprimere delle preferenze sui bandi a cui sono interessati e sono notificati soltanto se i bandi rispecchiano le preferenze.

## Il sistema prevede un meccanismo di registrazione dei nuovi **utenti** con relativo username e password. Ogni utente imposta alcuni parametri che rappresentano le sue **preferenze** riguardo a certi bandi, ovvero gli argomenti che interessano, tramite alcune espressioni da ricercare nel testo del bando.

## Quando dei nuovi bandi sono esposti sulla Gazzetta sono scaricati e memorizzati; il sistema sceglie quali sono gli utenti registrati interessati ai nuovi bandi: analizza il testo del bando e cerca se sono presenti o no le parole chiave definite dagli utenti, poi li notifica.

### Le **notifiche** avvengono tramite servizi Internet: un sito web e eventualmente posta elettronica, se l'utente lo ha specificato. In fase di registrazione l'utente definisce pertanto anche se desidera ricevere le notifiche per e-mail, oltre che visualizzarle sul sito web.

# PROBLEMI PRINCIPALI

Pagine HTML e bandi

La prima operazione da intraprendere al fine di poter disporre dei bandi di gara, è scaricarli. I bandi si possono raggiungere seguendo una serie di link presenti nelle pagine html, a partire dalla **home page** : http://www.gazzettaufficiale.it/ IMMAGINI TIPO ANALISI?.

### In questa pagina si possono trovare i link alle diverse pubblicazioni: scegliendo "**Quinta Serie Speciale - Contratti pubblici**" si può accedere alle pubblicazioni della Gazzetta Ufficiale dell'ultimo mese che contengono dei bandi di gare pubbliche.

### Questo documento espone diverse pubblicazioni e ciascuna di esse contiene molti bandi al suo interno: ai fini dell'applicazione va raccolto ogni singolo bando di ogni pubblicazione presente. Per raggiungere infine i bandi si sceglie una pubblicazione (seguendo il link esposto) e, nella pagina che si presenta, si sceglie di nuovo un link a uno specifico bando.

## La difficolta nello scorrere le varie pagine per raggiungere quella desiderata ("**web crawling**") risiede soprattutto nella parsificazione delle pagine ottenute, per estrarne le informazioni rilevanti. Per esempio, dalla home page delle pubblicazioni è necessario trovare i link alle varie pubblicazioni: per un umano potrebbe risultare semplice in quanto la pagina è renderizzata appositamente per facilitare la visione e la comprensione di ogni suo elemento, ma un'applicazione deve analizzare il solo contenuto HTML della pagina, che consiste in un file di testo poco strutturato.

### Il modo più veloce per parsificare questo file di testo è servirsi di riferimenti quali **id o class dei tag html**, oppure **sottostringhe significative** che rimangono fisse nel tempo e che identificano un dato elemento, e tramite i quali si ha la relativa certezza di trovare l'informazione che si cerca. Per esempio, si può notare che il link a una specifica pubblicazione è sempre situato in un particolare div, e che il link alla pubblicazione inizia sempre con la stessa sottostringa, e il pattern si ripete per ogni link a pubblicazione:

<span class="titolo\_atto">

<a href="/gazzetta/contratti/caricaDettaglio?dataPubblicazioneGazzetta=2017-07-31&numeroGazzetta=87&elenco30giorni=true/gazzetta/contratti/caricaDettaglio?dataPubblicazioneGazzetta=2017-07-31 class="elenco\_ugazzette">

n&#176; &nbsp;&nbsp;87 del 31-07-2017

### In questo caso, "**/gazzetta/contratti/caricaDettaglio"** è la parola chiave che anticipa il link a una pubblicazione, quindi per trovare tutti i link presenti nella pagina si possono ricercare iterativamente queste parole chiave, per poi leggere tutto il link successivo.

## Analogamente si ricavano i link ai singoli bandi all'interno di una specifica pubblicazione.

Un principale difetto di questa tecnica è la sua bassa affidabilità: non è assicurato che la struttura della pagine rimanga sempre costante nel tempo, soprattutto in un ambito specifico come i nomi dei link o gli id o class dei div. Da un altro punto di vista, però, questo modo di operare ha continuato ad essere efficace per mesi ed è comunque sufficiente modificare i nomi delle parole chiave che anticipano il link nel file di configurazione, e l'applicazione potrebbe continuare a funzionare (se le modifiche sono minime).

## In ogni caso, la struttura modulare dell'applicazione permette un'alta manutenibilità, infatti, se nel futuro si riscontrassero cambiamenti nella struttura delle pagine, un semplice aggiornamento ai componenti addetti al web crawling garantirebbe di nuovo il completo funzionamento del software.

Bandi e Oggetti

Una volta ottenuto il testo di un bando, il principale ostacolo nella realizzazione dell'applicazione, che determina anche la qualità dei dati racolti, è stato estrarre le informazioni principali riguardanti i bandi pubblici: più in particolare, il **riconoscimento**, all'interno del testo del bando, **dell'oggetto del bando**, ovvero l'argomento principale, riassumibile in poche righe.

**Esempio:**

Testo del bando:

Bando di gara a procedura aperta ai sensi del D.lgs. 50/2016

SEZIONE I: AMMINISTRAZIONE AGGIUDICATRICE. Arsenale Militare di

Messina dell'Agenzia Industrie Difesa.

SEZIONE II: **OGGETTO DELL'APPALTO. Affidamento dell'appalto di**

**servizi a richiesta e a quantita' indeterminata per servizi di**

**"Carenaggio, pulizia casse e pitturazioni necessarie nave enotria"**

importo base di gara: € 205.000,00 (euro duecentocinquemila/00) al

netto dell'I.V.A. GARA 6811056 - CIG 71611510C1.

SEZIONE IV: PROCEDURA. Termine ultimo presentazione offerte 12 set.

2017 ore 10:00.

SEZIONE VI: ALTRE INFORMAZIONI. Informazioni in merito possono

essere rilevate sul sito web www.aid.difesa.it nell'area "Bandi di

Gara".

Il responsabile unico del procedimento

ten. col. com. Diego Chiappini

TX17BFC13359

In questo caso l'oggetto del bando potrebbe essere: " Affidamento dell'appalto di servizi a richiesta e a quantita' indeterminata per servizi di 'Carenaggio, pulizia casse e pitturazioni necessarie nave enotria'".

### So nota che l'oggetto del bando compare nelle frasi successive ad alcune parole chiave, come "OGGETTO DELL'APPALTO", "OGGETTO DEL BANDO", "SEZIONE II" e altre. Queste espressioni predittive, se presenti nel bando, anticipano un possibile oggetto.

Per poter permettere al sistema di estrarre in autonomia l'oggetto di un bando a partire dal testo completo, è stato usato il software **WEKA**.

**WEKA** è un software di machine learning che è in grado, tramite un training mirato, di riconoscere stringhe e classificarle in categorie definite. Analizzando delle porzioni di testo del bando in cui compaiono le espressioni predittive e sottoponendole in continuazione al riconoscitore WEKA; grazie alla classificazione da lui operata ("oggetto accettabile" / "non accettabile") il sistema decide se l'oggetto del bando appena ricavato è valido oppure se deve continuare ad analizzare il testo in cerca di altri possibili oggetti

### . La continua iterazione di queste operazioni si conclude con l'estrazione di un oggetto valido oppure mancante, se non è presente nessun oggetto nel testo (caso abbastanza frequente) o se il riconoscitore non è stato in grado di accettare un oggetto valido.

Il software è in grado di ricavare un oggetto accettabile in quasi l'80% della totalità dei bandi analizzati.

La precisione e l'affidabilità degli oggetti riconosciuti dipende da due fattori: **il training dell'agente** e **le porzioni di testo** date in input al riconoscitore.

## Il **training** dell'agente consiste nel fornire un file in cui sono presenti molti esempi di stringhe da accettare e di stringhe da rifiutare: a partire dalla base di conoscenza dell'agente così costruita, il riconoscitore sarà in grado di "predire" se le future stringhe in input rientrano in una classe oppure l'altra. Per realizzare un training efficace si sono scelti esempi significativi di bandi con i relativi oggetti e soltanto gli esempi stessi determinano la qualità del training.

### La scelta degli esempi costituenti il training è stata guidata dall'osservazione pratica delle percentuali di bandi validi riconosciuti, al variare degli esempi stessi. Sono state eseguite quindi numerose prove con diversi campioni "casuali" ed è stato ritenuto migliore il file di training con le percentuali maggiori di tutti gli altri.

## Anche le **porzioni di testo** da sottoporre al riconoscitore giocano un ruolo fondamentale nella classificazione: devono rispecchiare il più possibile le frasi degli esempi nel training per avere la certezza che weka riconosca oggetti validi e rifiuti oggetti non validi: purtroppo non è sempre così.

### Quando si incontrano le parole chiave predittive nel testo del bando, si leggono i successivi **n** caratteri, che dovrebbero contenere l'oggetto del bando, e tutta la stringa letta si dà in input al riconoscitore. La misura di **n** è determinante: oggetti troppo corti non contengono abbastanza informazioni e rischiano di essere rifiutati anche se potenzialmente validi, così come oggetti troppo lunghi potrebbero andare oltre alla frase contenente l'oggetto e quindi contenere altre frasi che contaminano l'oggetto.

La scelta del parametro **n** si è basata su osservazioni empiriche ricavate da numerose prove e manipolazioni del testo: per esempio si leggono inizialmente molti caratteri e si sottopongono subito alla classificazione: se l'esito è negativo si cancellano gli utlimi x caratteri dalla stringa e si riprova con la classificazione; quando l'oggetto diventa troppo corto, la frase viene scartata e si riprova con altre porzioni di testo del bando.

# ALTRI PROBLEMI PRINCIPALI? TIPO DATE DEI BANDI? O QUALSIASI COSA PER OCCUPARE SPAZIO

### *Analisi funzionale*

### L'obiettivo del progetto è di raccogliere e memorizzare i **bandi delle gare pubbliche** pubblicati dalla **Gazzetta Ufficiale**, per poi notificare gli utenti registrati al sistema della presenza degli stessi, appena sono disponibili; gli utenti possono esprimere delle preferenze sui bandi a cui sono interessati e sono notificati soltanto se i bandi rispecchiano le preferenze.

### La Gazzetta è un documento ufficiale pubblicato dalla Republica Italiana con cadenza al massimo giornaliera e viene esposta in una pagina web al seguente indirizzo: <http://www.gazzettaufficiale.it/> (**Home Page della Gazzetta**). Le sue pubblicazioni sono identificate da data e un numero, che si azzera con la fine dell'anno. Una pubblicazione espone, oltre a decreti, sentenze e regolamenti, diverse decine di bandi di vario tipo, offerti da enti anche molto diversi quali Regioni, Province, Comuni, Ministeri...

### Le **pubblicazioni** dei bandi pubblici della Gazzetta degli ultimi 30 giorni sono esposte nel sito web della Gazzetta Ufficiale, al seguente URL: <http://www.gazzettaufficiale.it/30giorni/contratti> (**Pagina delle Pubblicazioni di Bandi pubblici**).

### E' possibile raggiungere la pagina delle pubblicazioni dei bandi pubblici nella sezione "Gazzetta Ufficiale" esposta nella home page, in particolare scegliendo "5a Serie Speciale - Contratti Pubblici".

### I **bandi** rappresentano delle gare pubbliche, cioè appalti di lavori o incarichi pubblici offerti dallo Stato e liberamente fruibili da privati o associazioni, previo concorso. Le informazioni principali che riguardano i bandi sono CIG (Codice Identificativo Gara) e oggetto della gara.

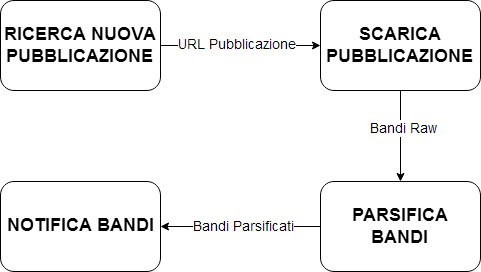
### Per ogni nuovo bando disponibile si memorizzano in modo persistente queste informazioni, per poi notificare gli utenti interessati.

### Il sistema prevede un meccanismo di registrazione dei nuovi **utenti** con relativo username e password. Ogni utente imposta alcuni parametri che rappresentano le sue **preferenze** riguardo a certi bandi.

### Gli utenti possono definire gli argomenti che interessano tramite alcune espressioni da ricercare nel testo del bando. Si definiscono le espressioni che devono comparire nel testo e quelle che invece non devono essere presenti; se poi il bando contiene una o più espressioni della prima categoria ma nessuna della seconda, allora l'utente è notificato.

### Quando dei nuovi bandi sono esposti sulla Gazzetta sono scaricati e memorizzati; il sistema sceglie quali sono gli utenti registrati interessati ai nuovi bandi: analizza il testo del bando e cerca se sono presenti o no le parole chiave definite dagli utenti, poi li notifica.

### Le **notifiche** avvengono tramite servizi Internet: un sito web e eventualmente posta elettronica, se l'utente lo ha specificato. In fase di registrazione l'utente definisce pertanto anche se desidera ricevere le notifiche per e-mail, oltre che visualizzarle sul sito web.

FASI DEL PROCESSO

# RICERCA NUOVA PUBBLICAZIONE

Una volta al giorno si collega al sito web della Gazzetta Ufficiale e cerca se è presente una nuova pubblicazione, controllandone la data e confrontandola con quella dell'ultima pubblicazione memorizzata dal sistema.

In caso ne sia presente una nuova, crea una nuova entry nell'entità Pubblicazione e memorizza le prime informazioni già disponibili (numero, data e URL) . Il processo può così avanzare; nel caso non sia presente nessuna nuova pubblicazione, non sono attivate le fasi successive del processo.

# SCARICA PUBBLICAZIONE

Essendo disponibile una nuova pubblicazione, scarica la pagina corrispondente e analizza la pubblicazione, ricavando tutti i bandi e inserendoli nel database insieme a URL corrispondente, nome dell'ente richiedente e testo per intero. Queste informazioni sono già disponbili all'interno della pubblicazione.

Una volta disponibili tutti i link ai bandi, sono scaricati uno per uno, aggiornando con nuove informazioni la precedente entry creata. La Pubblicazione può ora essere rimossa, in quanto tutti i bandi che conteneva sono stati scaricati.

Questo meccanismo permette di avere un controllo sull'effetivo scaricamento di tutti i bandi della pubblicazione: se tutti i suoi bandi sono stati scaricati, allora la pubblicazione non dovrebbe più essere presente nel database; se invece è presente anche solo un bando di una pubblicazione e la pubblicazione stessa, con relativo stato SCARICATA, significa che c'è stato un malfunzionamento e quindi sarebbe opportuno ripetere tutta l'operazione di scaricamento.

# PARSIFICA BANDI

Una volta disponibile il testo del bando, si parsifica e analizza per cercare di ricavare più informazioni possibili: oggetto del bando, CIG e descrizione; si separano anche logicamente nella struttura di memorizzazione. Queste informazioni, se presenti e se ricavabili, sono memorizzate all'interno dell'entità Bando, pronte per la prossima fase del processo.

Se il Parser non è stato in grado di ricavare le informazioni dal bando, perchè non presenti o male espresse o perchè il documento non era strutturato in un modo riconoscibile, rimane il testo integrale che può comunque essere utilizzato per la fase sucessiva.

# NOTIFICA BANDI

Per ogni utente memorizzato la cui proprietà DT\_NOTIFICA risulta precedente o uguale a DT\_INSERIMENTO (bando), analizza i nuovi bandi disponibili e controlla se rispettano o no le preferenze espresse.

Ad ogni utente sono associate diverse espressioni regolari, tutte in OR fra loro, che devono o non devono comparire nel documento. Per ognuna di queste si tenta il match con l'oggetto (o il testo del documento, se l'oggetto manca): nel caso di espressioni MINUS, il match di una sola di queste comporta la fine immediata del controllo: l'utente non è notificato. Se, a operazione conclusa, una o più stringhe contenute nel bando hanno fatto match con almeno un'espressione regolare PLUS e non con quelle MINUS, l'utente è notificato. In altre parole, tutte le espressioni PLUS in OR fra loro vengono messe in AND con le espressioni di tipo MINUS (in OR fra loro) e tutte insieme vengono processate per il match con il bando, con la differenza che il match con il secondo tipo comporta il "fallimento" del riconoscimento.

In seguito alla notifica di un utente, aggiorna il valore della proprietà DT\_NOTIFICA (nell'entità utente) con la data e l'ora dell'invio della notifica.

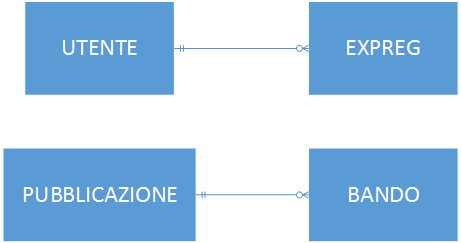
Il controllo di data è necessario in quanto, se il sistema non è stato in grado di notificare l'utente nel giorno stesso in cui è stato aggiunto il bando, l'operazione può essere eseguita in un secondo momento.

Se nessun bando interessa agli utenti registrati (secondo i criteri appena descritti), questa fase si conclude; se invece è presente almeno un bando che rispetta i parametri di uno o più utenti, sono notificati sul sito web. Se l'utente ha scelto di ricevere le notifiche anche tramite e-mail, il sistema provvede ad informare l'utente anche secondo questa modalità

Terminata questa fase, la catena di processi è conclusa e si deve aspettare la disponibilità di una nuova pubblicazione prima che tutto riparta.

Architettura

DATABASE



# PUBBLICAZIONE

Permette di modellare le pubblicazioni della Gazzetta Ufficiale

|  |  |  |  |
| --- | --- | --- | --- |
| Nome Proprietà | Tipo | Null / Extra | Commenti |
| CD\_PUBBLICAZIONE | BIGINT | PRIMARY, AUTO\_INCREMENT | Identificativo della pubblicazione. Popolato automaticamente |
| DT\_INSERIMENTO | DATETIME |  | Data del'inserimento della pubblicazione nel database |
| NM\_PUBBLICAZIONE | INTEGER |  | Numero della pubblicazione |
| STATO | ENUM |  | DA\_SCARICARE, SCARICATA |
| URL | VARCHAR(1000) |  | Indirizzo URL della pubblicazione |

# BANDO

L'entità bando rappresenta un bando di gara pubblica contenuto in una pubblicazione della Gazetta.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome Proprietà | Tipo | Null / Extra | Commenti |
| CD\_BANDO | BIGINT | PRIMARY, AUTO\_INCREMENT | Identificativo del bando, generato in automatico |
| CD\_ESTERNO | VARCHAR (250) | UNIQUE | Codice del bando che si trova nella pubblicazione (TX17BFC9332) |
| CD\_PUBBLICAZIONE | BIGINT | REFERENCES PUBBLICAZIONE (CD\_PUBBLICAZIONE) | Rappresenta la relazione fra BANDO e PUBBLICAZIONE |
| CIG | VARCHAR (250) | Null | Codice Identificativo Gara, riportato sul bando |
| TIPO | VARCHAR (50) | Null | Bando di Gara / Estratto di bando di gara / Bando di gara a procedura aperta |
| TIPORICHIEDENTE | VARCHAR (100) | Null | Ministeri / Regioni / Province / ... |
| NM\_RICHIEDENTE | VARCHAR(1000) | Null | Il nome dell'ente richiedente (Associazione o privato) |
| SCADENZA | DATE | Null | La data di scadenza del bando, riportata sullo stesso |
| OGGETTO | VARCHAR(1000) | Null | Oggetto del bando |
| TESTO | TEXT | Null | Tutto il testo integrale del bando |
| URL | VARCHAR(1000) |  | Indirizzo web identificativo del bando |
| STATO | ENUM |  | DA\_PARSIFICARE, PARSIFICATO |
| DT\_INSERIMENTO | DATETIME |  | Data di inserimento del bando nel database, che corrisponde alla data di inserimento della pubblicazione |

# EXPREG

L'entità gestisce le espressioni regolari associate a ciascun utente, che rappresentano i parametri con cui l'utente esprime il suo interesse in determinati tipi di bando.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome Proprietà | Tipo | Null / Extra | Commenti |
| CD\_EXPREG | BIGINT | PRIMARY, AUTO\_INCREMENT | Identificativo dell'espressione, assegnato in automatico. |
| CD\_UTENTE | BIGINT | REFERENCES UTENTE (CD\_UTENTE) | Rappresenta la relazione fra EXPREG e UTENTE |
| EXPPLUS | VARCHAR(1000) |  | L'espressione regolare vera e propria, di tipo PLUS |
| EXPMINUS | VARCHAR(1000) |  | L'espressione regolare vera e propria, di tipo MINUS |

Le Espressioni Regolari esprimono il fatto che l'utente è o meno interessato in un certo argomento: le espressioni di TIPO PLUS indicano che, se nel bando compare una porzione di testo che fa match con tutte, bisogna notificare l'utente. Quelle di tipo MINUS indicano invece che NON bisogna notificare l'utente se se ne trova almeno una che fa match. La combinazione dei due tipi costituisce l'"argomento" di interesse.

# UTENTE

L'entità utente rappresenta gli account che sono iscritti al sistema e che desiderano ricevere delle notifiche in seguito alla pubblicazione di alcuni tipi di bando.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome Proprietà | Tipo | Null / Extra | Commenti |
| CD\_UTENTE | BIGINT | PRIMARY, AUTO\_INCREMENT | Identificativo dell'utente, assegnato in automatico. |
| USERNAME | VARCHAR (50) | UNIQUE | Pseudonimo (univoco) dell'utente, scelto dallo stesso in fase di registrazione |
| PASSWORD | VARCHAR (255) |  | Password decisa dall'utente in fase di registrazione e di cui si calcola la funzione Hash MD5 |
| EMAIL | VARCHAR(255) |  | Email decisa in registrazione |
| DT\_NOTIFICA | DATETIME |  | Data in cui è avvenuta l'ultima notifica di nuovo bando |

DESCRIZIONE MODELLO A OGGETTI

TOP - DOWN:

prima in generale e poi ogni modulo, dettagliato

Strumenti utilizzati

# STACK TECNOLOGICO E DESCRIZIONI DI OGNI PEZZO

Gli strumenti tecnologici a disposizione sono: Java J2EE, Maven, Spring, Spring MVC, Prime Faces e MySQL.

Per la prima parte dell'applicazione, **GarePubblicheCrawler**, non vi è alcuna parte di presentazione, pertanto è utilzzato Java, accompagnato da Maven per la gestione delle dipendenze, MySQL per la memorizzazione dei dati e eventuali librerie esterne di utilità.

La seconda parte, **GarePubblicheWeb**, è invece un'applicazione web che si occupa della gestione di un sito; è quindi necessario uno stack più ampio:

## Java J2EE, (Maven), MySQL, Spring e Spring MVC per quanto riguarda la parte di back end, di gestione delle pagine dinamiche e del database

## Prime Faces come framework per il front end

# WEKA

http://www.cs.waikato.ac.nz/ml/weka/

Weka è un software di machine learning e data mining sviluppato da "The University of Waikato".

### Tra le sue molte applicazioni (esempi) si trova una funzione di classificazione di testi, utilizzata in questo progetto al fine di riconoscere i gli oggetti dei bandi.