CHITECHNOLOGIES®

Università degli Studi Roma Tre

Seminario Sistemi Informativi 2005-2006

Analisi e ingegnerizzazione di un processo di estrazione dati da Web

Alessio Pace

Sommario

1. Contesto di riferimento dello studio
   - il progetto universitario **RoadRunner**
   - la spin-off **CHI Technologies**
2. Studio di caso: **FinFox**
3. Analisi e diagnosi della situazione attuale
4. Necessità di un intervento di ingegnerizzazione
5. Soluzione proposta
6. Indicazioni per lo sviluppo

Contesto di riferimento

I wrapper per le pagine Web

- Il Web è la più grande fonte di informazioni:
  - pagine tuttavia pensate per il consumo da parte di utenti umani attraverso un browser e difficilmente processabili dai programmi che girano sui calcolatori elettronici
  - alto valore aggiunto qualora le informazioni possano essere *estratte* e *conservate* localmente per essere *analizzate*, *rielaborate*, *riproposte* sotto una diversa forma o in congiunzione con altre informazioni
- Per estrarre i dati sono necessari dei programmi chiamati **wrapper** che:
  1. effettuano il parsing di una pagina web (HTML/XHTML)
  2. estraggono i dati di interesse
  3. riversano i dati in un altro formato elettronico (XML, RDBMS, Excel, ..)

Esempio costruzione manuale di wrapper

teamName	*town*
Atalanta	Bergamo
Inter	Milano
Juventus	Torino
Milan	Milano
…	…

```
<html><body>
<h1>Italian Football Teams</h1>
<ul>
<li><b>Atalanta</b> - <i>Bergamo</i> <br>
<li><b>Inter</b> - <i>Milano</i>  <br>
<li><b>Juventus</b> - <i>Torino</i>  <br>
<li><b>Milan</b> - <i>Milano</i>  <br>
</ul>
</body></html>
```

Wrapper Procedure:

Scan document for
While there are more occurrences
  scan until
  extract teamName between and
  scan until
  extract town between and
  output [teamName, town]

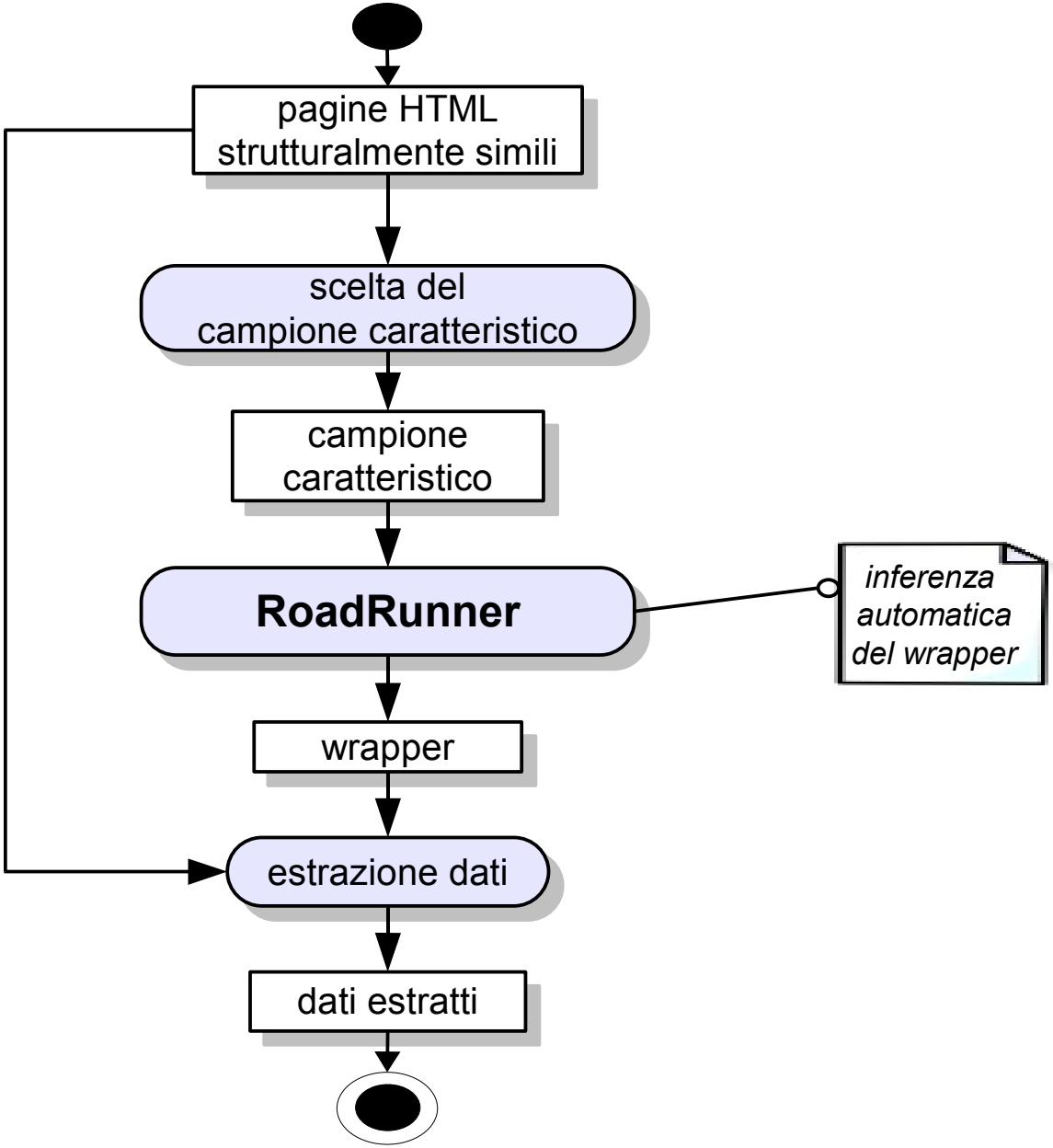
Necessità di automatizzare la generazione dei wrapper

- Limiti e costi dei wrapper costruiti manualmente:
  - procedimento lungo e tedioso
  - soggetto ad errori
    - la struttura della pagine potrebbe essere complessa (tag annidati) e i dati da estrarre molti
  - poco generalizzabile
    - la struttura delle pagine potrebbe essere poco regolare
  - poco robusto
    - piccoli cambiamenti nel tempo della struttura della pagina possono rendere necessaria la riscrittura del wrapper
- Necessità di generare automaticamente i wrapper:
  - abbattimento dei costi di generazione di un wrapper e suo mantenimento (-> lo si rigenera)
  - scalabilità al numero di pagine simili (un wrapper per più pagine “simili”)
  - scalabilità al numero di tipologie di pagine simili (un wrapper per tipologia di pagina) da cui si vogliono estrarre i dati

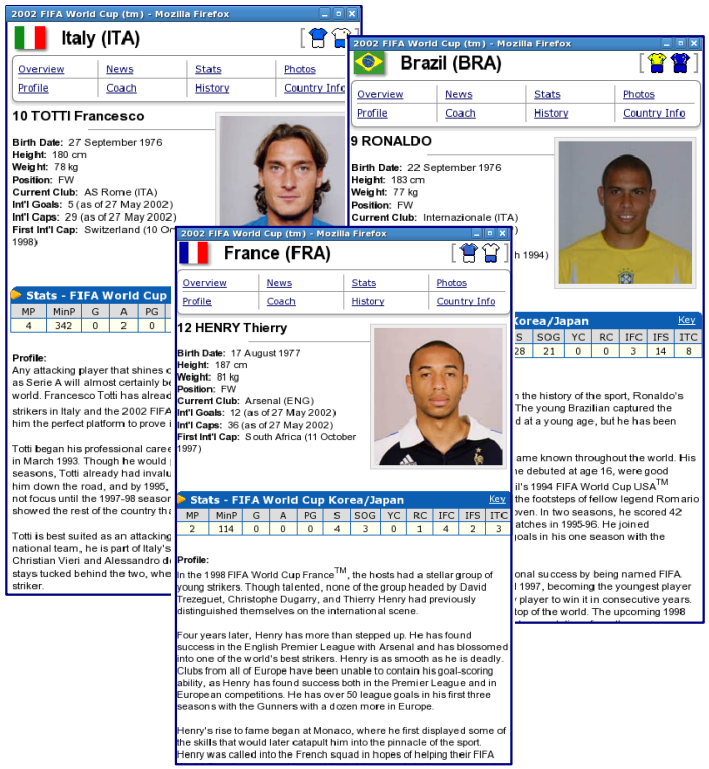
Il progetto RoadRunner

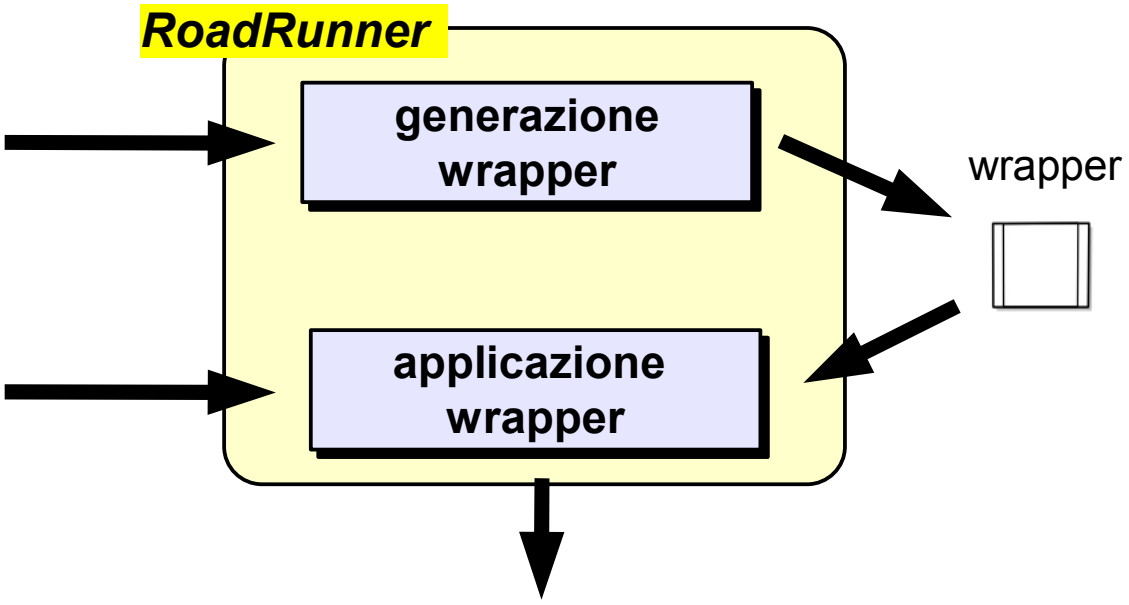
- Generazione automatica di wrapper di pagine web:
  - INPUT: un piccolo *campione caratteristico* preso da un insieme di pagine **strutturalmente simili** (nel seguito: una “classe di pagine”)
  - OUTPUT: un wrapper, descritto come una grammatica regolare per il codice HTML delle pagine
- Applicazione di un wrapper
  - la grammatica viene usata per il parsing delle pagine e l'estrazione dei dati in essa contenuti

RoadRunner: schema di funzionamento



RoadRunner: esempio di funzionamento





_N_	_O_	_P_	_Q_	_R_	_S_	_T_	_U_	_V_	_W_
	10 Totti Francesco	27 September 1976	180cm	78kg	FW	AS Rome (ITA)	5	29	Switzerland (10 October 1998)
	9 Ronaldo	22 September 1976	183cm	77kg	FW	Inter (ITA)	37	57	Argentina (23 March 1994)
	12 Henry Thierry	17 August 1977	187cm	81kg	FW	Arsenal (ENG)	12	36	South Africa (11 October 1997)

classe di pagine:
pagine strutturalmente simili

RoadRunner: esempio dati estratti e salvati in XML

```
<!-- TOTTI -->
   <instance
            source=”file:/mirrors/fifaworldcup.yahoo.com/02/en/t/t/pl/165247/index.html"
            name="index">
      <and>
         <attribute label="_N_"><imageref
            source="http://us.i1.yimg.com/us.yimg.com/i/fifa/gen/tr/pl/s/165247.jpg"/>
         </attribute>
         <attribute label="_O_">10 TOTTI Francesco</attribute>
         <attribute label="_P_">27 September 1976</attribute>
         <attribute label="_Q_">180 cm</attribute>
         <attribute label="_R_">78 kg</attribute>
         <attribute label="_S_">FW</attribute>
         <attribute label="_T_">AS Rome (ITA)</attribute>
         <attribute label="_U_">5</attribute>
         <attribute label="_V_">29</attribute>
         <attribute label="_W_">Switzerland (10 October 1998)</attribute>
        <!-- ..... -->
 </instance>
```

Cosa non fa RoadRunner

- Individuazione e scaricamento delle pagine di interesse
- Scelta del campione caratteristico per la generazione del wrapper:
  - la scelta deve essere fatta esternamente a RoadRunner
- Trasformazione degli schemi dei dati estratti (nel seguito, “mapping”):
  - la *struttura* e le *etichette* delle informazioni estratte non rispecchiano la logica dei dati ma soltanto l'organizzazione fisica con cui i dati vengono pubblicati nella pagina

Motivazioni dello studio condotto

- Per valorizzare in ambito industriale i risultati del progetto RoadRunner è sorta una spin-off universitaria, la **CHI Technologies**, con l'obiettivo di definire ed offrire strumenti e servizi per l'estrazione dati da Web

- La tesi nata in questo contesto ha avuto l'obiettivo di studiare come ingegnerizzare un processo di estrazione dati da Web avente come nucleo RoadRunner (vedi studio di caso)

Studio di caso: FinFox

Studio di caso: descrizione

- Collaboratore e committente:
  - Sincro Consulting S.p.A.
- Input:
  - 2 siti web (http://`www.bluerating.com` , http://`www.morningstar.it`)
  - ~6000 fondi di investimento fra italiani ed esteri per sito
  - 5 pagine HTML collegate per fondo su ogni sito (~30000 pag per sito)
  - ~1GB occupati in locale per sito
- Output richiesto:
  - 1 base di dati con le informazioni *estratte* dai due siti web in un giorno ed *integrate* fra loro

Studio di caso: i dati

- Dati da estrarre:
  - informazioni anagrafiche sugli strumenti finanziari
  - informazioni sui rendimenti
  - informazioni sul rating/ranking dei diversi provider
- Possibilità di effettuare analisi sui dati estratti:
  - per tipologia
  - per mercato
  - per settore industriale
  - per società di gestione

Studio di caso: risorse a disposizione

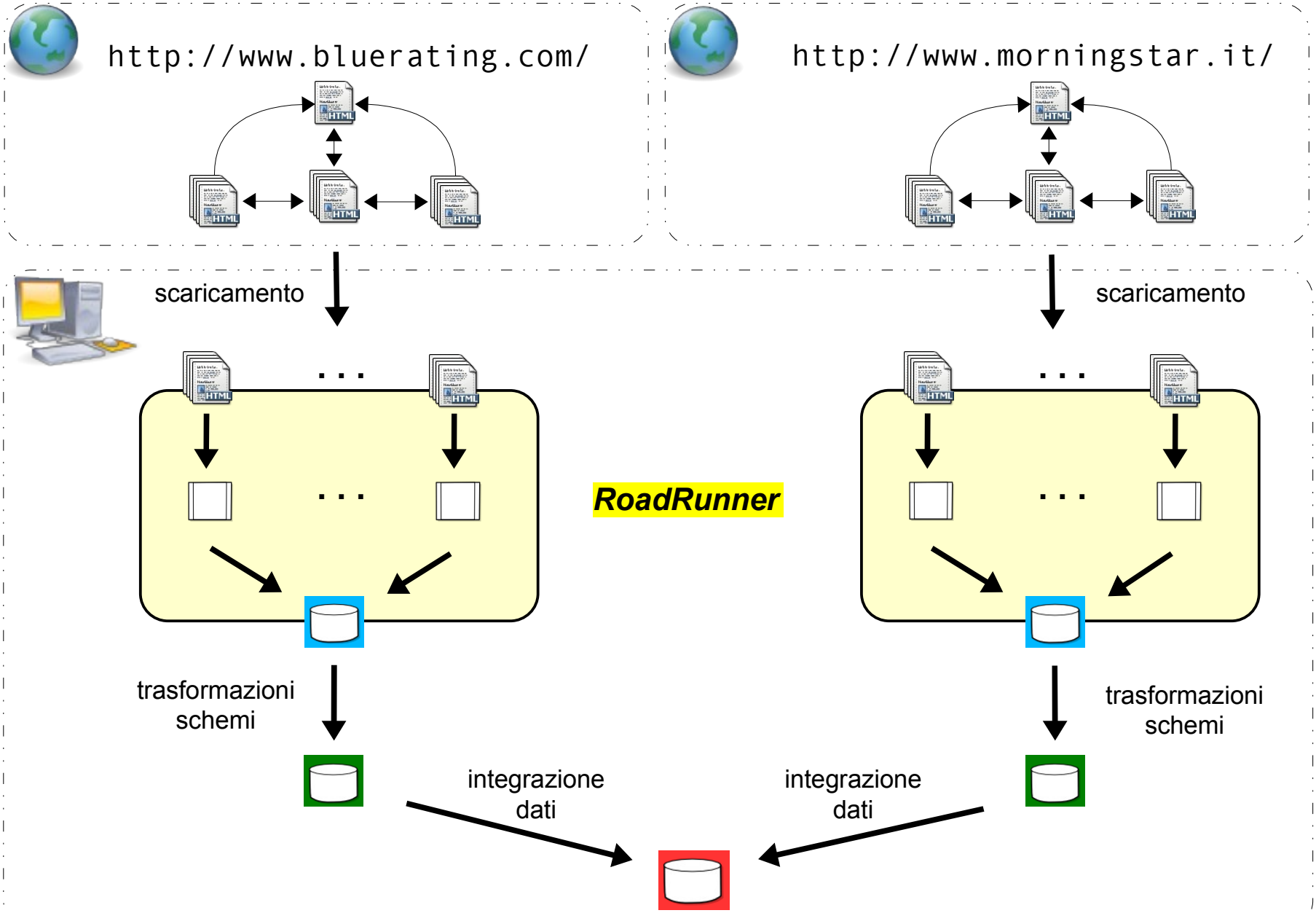
- Vincoli temporali:
  - 2 mesi (gennaio-marzo 2005)
- Risorse umane:
  - 2 tesisti full time per la fornitura dei dati richiesti (Manicardi, Pace)
  - 2 coordinatori (Crescenzi, Merialdo)
  - 1 project manager (da Sincro Consulting)
  - 1 addetto al front end (da Sincro Consulting)
- Strumenti a disposizione:
  - RoadRunner

Studio di caso: approccio adottato

- Cercare le soluzioni realizzative più efficaci nei vincoli e con le risorse a disposizione

- Cercare di mantenere un occhio di riguardo sulla riutilizzabilità in contesti successivi

Le fasi necessarie



Dettagli delle fasi e sottofasi

- Navigazione (*per ciascun sito*)
  - inferenza dell'*algoritmo navigazionale*
  - esecuzione dell'algoritmo navigazionale al fine di scaricare le pagine web di interesse e raggrupparle in classi di pagine
- Generazione dei wrapper (*per ciascun sito*)
  - (sotto-classificazione delle pagine scaricate)
  - generazione dei wrapper con RoadRunner
- Estrazione (*per ciasun sito*)
  - dati estratti e salvati su RDBMS da RoadRunner
- Mapping (*per ciascun sito*)
  - dei dati verso gli schemi relazionali richiesti dal committente
- Integrazione (*fra i due siti*)
  - *record linkage* (*entity matching*) per la corrispondenza dei nomi dei fondi delle due sorgenti informative

Risultato finale: FinFox

FinFox

BR | MS

Home
Anagrafica
Portafoglio
Commissioni
Fin&Stat

Area Anagrafica

Società di gestione: Bipiemme Iniziativa Europa Acc

Gestore: Armando Carcaterra

Categoria Assogestioni: AZIONARI EUROPA

Indirizzo:

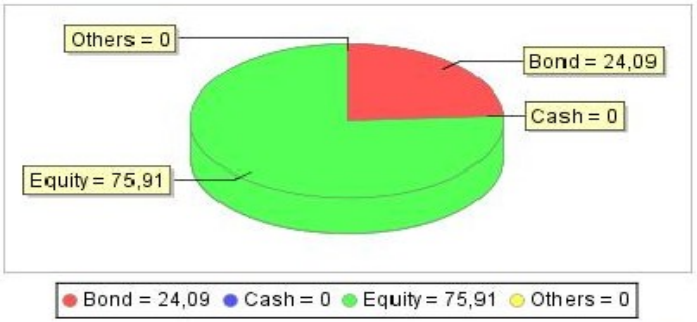
Benchmark di riferimento: 90% MSCI Europe Small Caps in Euro 10% MTS BOT Lordo

Valuta: Eur

Patrimonio: 147,84 milioni di Eur al 31/01/05

Area Portfolio

Portfolio per comparti di investimento



more...

Area Commissioni

Commissioni di ingresso: Min: 0,00%-Max: 0,00%

Commissioni di uscita: 0,00%

Area Fin&Stat

Rendimenti ultimi 3 anni

Analisi e diagnosi della situazione attuale

Analisi

- Demo utile per:
  - individuare le fasi e attività richieste per l'estrazione dati da Web
  - rilevare reali problematiche e requisiti
- Risultati finali soddisfacenti
- Modalità e costi per il raggiungimento dei risultati non soddisfacenti:
  - mancanza di strumenti automatici o semi-automatici (esterni o scritti in casa) per lo svolgimento della maggior parte delle attività
  - ricorso a soluzioni codificate manualmente per l'occasione, poco riutilizzabili

Navigazione: criticità della soluzione

- Costosa:
  - crawler scritti manualmente
- Non scalabile:
  - se il numero di siti richiesti fosse stato 10 e non 2 sarebbe stato necessario visionare manualmente ciascuno dei 10 siti e scrivere per ognuno un crawler specializzato
- Poco manutenibile
  - a lievi cambiamenti nella struttura di un sito corrisponde la necessità di rimettere mano al codice sorgente del suo crawler
- Troppo tecnica:
  - solo gli esperti delle tecnologie utilizzate possono (eventualmente) scrivere/modificare il crawler per un sito
  - di fatto i crawler sono codice usa e getta..

Generazione wrapper: criticità della soluzione

- Costosa:
  - attività di sotto-classificazione svolta manualmente
- Poco manutenibile:
  - al variare delle caratteristiche delle pagine il lavoro andrebbe svolto di nuovo da zero
- Incompleta:
  - solo un cluster per classe di pagina fu individuato (utilizzo di circa il 1% dei fondi di ciascun sito)

Mapping: criticità della soluzione

- Costosa:
  - trasformazioni strutturali eseguite tramite codice Groovy/SQL scritto manualmente per la specifica situazione
- Soggetta ad errori:
  - scrivendo a mano le trasformazioni strutturali è facile incorrere in errori (es: scambiare un attributo per un altro)
- Poco manutenibile:
  - al cambiare dei wrapper (es: se cambiano le pagine) bisogna rimettere mano al codice sorgente del mapping
- Troppo tecnica:
  - gli script per il mapping sono stati scritti in Groovy/SQL
  - di fatto è codice usa e getta..

Integrazione: criticità della soluzione

- Costosa:
  - corrispondenza dei nomi dei fondi trovata in maniera completamente manuale
- Soggetta ad errori:
  - data la natura manuale del procedimento, si può sbagliare nel riportare qualche corrispondenza
- Non scalabile:
  - se fosse stato necessario trovare la corrispondenza per più del 1% dei fondi sarabbe stato infattibile manualmente

Riepilogo criticità e impatto

Fase	Sottofase	Criticità	Impatto
Navigazione	Inferenza algoritmo navigazionale di un sito	algoritmo di navigazione da inferire manualmente visionando il sito	medio
	Scaricamento delle pagine di un sito	scrittura manuale di un crawler specifico per ciascun sito	alto
Generazione dei wrapper	Sottoclassificazione delle pagine	sottoclassificazione svolta manualmente	alto
	Scelta del campione caratteristico	algoritmo “palla di neve” da raffinare	basso
Mapping	Mapping	trasformazioni strutturali degli schemi scritte manualmente in Groovy/SQL	alto
Integrazione	Record linkage	corrispondenza dei nomi trovata manualmente	alto

Manutenzione dei dati: criticità

- Non è stata provata la manutenzione periodica degli stessi dati estratti
  - se dopo 1-2 mesi fosse stato richiesto di estrarre gli stessi dati aggiornati?
- Possibili costi elevati dovuti a:
  - cambio della topologia del sito
  - cambio della struttura delle pagine:
    - rigenerare i wrapper
    - riscrivere le regole di mapping
  - cambiano i nomi dei fondi / bisogna estrarre dati da altri fondi:
    - raffinare / rieseguire l'integrazione

Studio di caso: diagnosi

- Verifica delle opportunità e limitazioni:
  - RoadRunner come nucleo del processo
  - mancanza di strumenti software opportuni per le altre fasi
- Soluzioni software per la situazione:
  - codificate manualmente e poco manutenibili
  - scarsamente generalizzabili ad altri contesti
- Risultano avere costi *elevati* e *comparabili*:
  - la prima applicazione del processo per l'estrazione dei dati richiesti
  - la manutenzione del processo per i successivi aggiornamenti periodici dei dati estratti

Necessità di un intervento di ingegnerizzazione

Necessità di un intervento di ingegnerizzazione

Intervento di ingegnerizzazione (1)

- Obiettivi dell'approccio metodologico e delle specifiche tecnologiche:
  - soluzione indipendente dal dominio
  - prima applicazione del processo di estrazione dati con costi ridotti rispetto al caso di studio
  - manutenzione periodica del processo con costi di ordini di grandezza inferiori rispetto alla prima applicazione
- Requisiti funzionali:
  - **flessibilità** rispetto alla *tipologia* e alla *struttura* delle informazioni
  - **scalabilità** rispetto al *numero* di siti, pagine, dati di interesse
  - **efficienza** nei *tempi* necessari a svolgere le attività
  - **efficacia** nella *qualità* dei risultati

Intervento di ingegnerizzazione (2)

- Vincoli temporali:
  - sistema realizzato e funzionante in un anno
  - costante rilascio di prototipi con sotto-insieme delle funzionalità
    - possibilità di testarne l'efficacia con altre demo durante l'anno
- Risorse umane a disposizione:
  - 3 sviluppatori a tempo pieno
  - un capo progetto
- Modalità realizzative:
  - team di sviluppo piccolo, requisiti che potrebbero evolvere:
    - metodologie di sviluppo agili
  - preferire tecnologie, librerie e strumenti software che abbiano:
    - costi di utilizzo nulli o molto bassi
    - licenza open source in modo da poter interagire con gli sviluppatori in caso di richiesta di nuove funzionalità o correzione di bug

Soluzione proposta

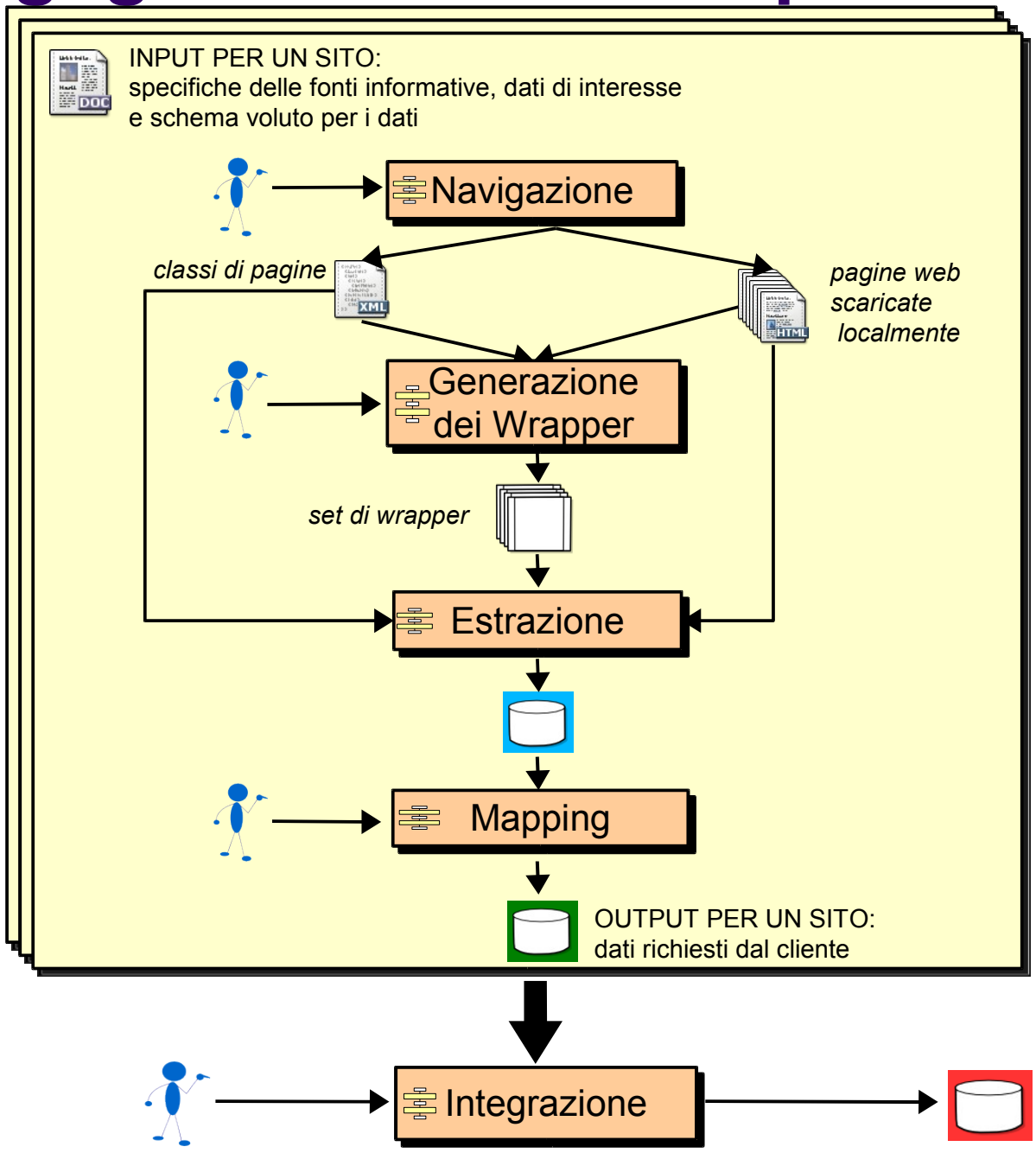
Scenario di riferimento

- Necessità di fornire un servizio competitivo di estrazione e fornitura periodica di informazioni presenti sul Web
  - (eventualmente) secondo gli schemi richiesti da un cliente
- Studio di fattibilità per individuare proposte metodologiche e tecnologiche che soddisfino i requisiti

Ingegnerizzazione del processo

ESTRAZIONE DATI DA VARI SITI WEB

INTEGRAZIONE



L'approccio adottato nel proporre una soluzione (1)

- Per ognuna delle fasi e sottofasi sono stati definiti i requisiti e le metodologie di funzionamento degli strumenti necessari all'assolvimento delle attività con maggiore semplicità e con costi ridotti

- Sono state fornite indicazioni sull'integrazione delle varie fasi fra di loro, principalmente attraverso l'uso di formati di scambio aperti e portabili

L'approccio adottato nel proporre una soluzione (2)

- Per la maggior parte degli interventi proposti è stata fornita una analisi dei rischi, costi e benefici delle possibili alternative tecnologiche e metodologiche:
  - valutando idee o prodotti presenti in letteratura o sul mercato ("*make vs buy*")
- In certi casi sono stati scritti prototipi o usati dei prodotti esistenti (anche commerciali) per valutare la bontà delle scelte:
  - dove possibile è stata fornita una preferenza

Riepilogo degli interventi proposti

Fase	Sottofase	Intervento realizzativo
Navigazione	Inferenza dell'algoritmo di navigazione di un sito	Estensione Firefox per inferire in maniera semi-automatica, col supporto dell'utente,l'algoritmo di navigazione di un sito
	Esecuzione dell'algoritmo di navigazione di un sito al fine di scaricare le pagine di interesse	Estensione Firefox (la stessa) per navigare un sito usando il suo algoritmo di navigazione e scaricare le pagine di interesse in modo automatico
Generazione dei wrapper	Sottoclassificazione delle pagine	Strumento visuale (prototipo attuale in Java con Jrex o nuova estensione per Firefox) per la selezione dei frammenti di interesse delle pagine, individuabili tramite espressioni `Xpath`
	Scelta del campione caratteristico	Raffinamento dell'algoritmo “palla di neve”
Mapping	Mapping dallo schema dei wrapper allo schema richiesto dal cliente	Formalizzazione e implementazione in `XQuery` degli operatori primitivi di mapping per le trasformazioni strutturali degli schemi
		Strumento grafico in Java Swing di supporto alla generazione delle regole di mapping
Integrazione	Record Linkage	*soluzione da valutare caso per caso*

Un esempio di analisi: la fase della Navigazione

Un esempio di analisi: la Navigazione

- Prima fase del processo, si vorrebbe che sia:
  - scalabile all'aumentare dei siti e pagine di interesse
  - flessibile al variare della struttura dei siti
  - efficiente nei tempi di scaricamento
  - efficace nel poter navigare anche siti che presentano difficoltà tecnologiche (link Javascript, frame HTML, form di login, cookie della sessione, HTTP Redirect e Continue, ..)
- Idea:
  1. inferire semi-automaticamente l'algoritmo di navigazione di un sito e renderlo persistente
  2. leggere tale algoritmo di navigazione per eseguire la navigazione del sito e scaricare automaticamente (e periodicamente) le pagine di interesse, preservandone la classificazione in classi di pagine

Navigazione: strumenti proposti

- Uno strumento visuale interattivo che col supporto dell'utente aiuti a generare l'algoritmo navigazionale per un sito Web
  - l'utente deve poter indicare i percorsi di navigazione (collezioni di link) e le tipologie di pagine eseguendo una navigazione d'esempio
  - l'algoritmo di navigazione deve essere reso persistene in formato XML
- Uno strumento che legga la descrizione dell'algoritmo di navigazione generato al punto precedente ed esegua in maniera automatica la navigazione
  - si ha un solo crawler
- I due strumenti possono così essere sviluppati separatamente o essere due funzionalità dello stesso strumento (“registrazione” della navigazione, “riesecuzione”)

Navigazione: inferenza dell'algoritmo di navigazione (1)

- Specifiche proposte:
  - *embedding* del motore di un browser web per il pieno supporto alla navigazione utente
  - inferire l'algoritmo di navigazione tramite la “registrazione” di una navigazione guidata dall'utente
  - salvataggio dell'algoritmo di navigazione in formato XML per successiva riproduzione dall'esecutore

Navigazione: inferenza dell'algoritmo di navigazione (2)

- Analisi rischi, costi, benefici alternative

	Jrex	Javascript	Firefox Extension	InternetMacros ™
Maturità e stabilità	bassa	bassa	alta	alta
Documentazione	bassa	alta	media	media/alta
Futuribilità	bassa	media	alta	media
Facilità di apprendimento	media	media	bassa/media	media
Facilità di installazione	bassa	alta	media	alta
Supporto e aiuto	bassa	alta	alta	media/alta
Portabilità	alta	media	alta	solo Windows + IE
Quantità di librerie a disposizione	alta	alta	alta	alta (molte commerciali)
Costi di utilizzo	nulli	nulli	nulli	licenza commerciale
Possibilità di scrittura dello strumento richiesto	media	bassa/media	media/alta	media
Conoscenze relative del team di sviluppo	bassa	bassa	bassa/media	bassa

Navigazione: esecuzione dell'algoritmo di navigazione (1)

- Specifiche proposte:
  - leggere la descrizione di un algoritmo di navigazione ed eseguire in maniera automatica la navigazione
    - necessità di emulare il comportamento di un vero browser web
  - salvataggio delle pagine in locale
  - scrittura di un file XML descrittore del risultato della navigazione:
    - es: per ogni classe di pagine, URL locale delle singole pagine scaricate

Navigazione: esecuzione dell'algoritmo di navigazione (2)

- Analisi rischi, costi e benefici delle alternative

	HttpClient	HtmlUnit	HttpUnit	Firefox Extension
Form HTML (anche metodo POST)	NO	SI	SI	SI
HTTP Redirect, HTTP Continue	SI	NO	SI?	SI
HTTPS	SI	SI	SI	SI
Javascript	NO	In parte	In parte	SI
Cookie	SI	SI	SI	SI
Frame HTML	NO	SI	SI	SI
Menu Flash	NO	NO	NO	SI
Necessità di server grafico attivo	NO	NO	NO	SI
Maturità e stabilità	alta	media/alta	media/alta	alta
Documentazione	media	media	media	media
Futuribilità	alta	media	media	alta

Navigazione: esecuzione dell'algoritmo di navigazione (3)

- (continua)

	HttpClient	HtmlUnit	HttpUnit	Firefox Extension
Facilità di apprendimento	media	alta	alta	bassa/media
Facilità di installazione	alta	alta	alta	media
Supporto e aiuto	medio	medio	medio	alto
Portabilità	alta	alta	alta	alta
Costi di utilizzo	nulli	nulli	nulli	nulli
Possibilità di scrittura dello strumento richiesto	bassa	bassa/media	bassa/media	media/alta
Riuso del codice dello strumento per la fase precedente	Solo se scelta è Jrex	Solo se scelta è Jrex	Solo se scelta è Jrex	Solo se scelta è Javascript o Firefox Extension
Conoscenze relative del team di sviluppo	alta	media/alta	media/alta	bassa/media

Indicazioni sulle criticità del procedimento di manuntezione periodica dei dati estratti

Manutenzione periodica del processo di estrazione dati da Web

- La prima estrazione e fornitura dei dati richiede costi che dipendono:
  - dagli strumenti a disposizione
  - dal *numero* di siti e classi di pagine web
  - dalla *struttura* delle informazioni desiderate
- Necessità di abbattimento dei costi per i successivi aggiornamenti:
  - tramite la definizione di *procedure automatiche*
  - possibilità di riuso, almeno parziale, dei risultati ottenuti in precedenza (algoritmi di navigazione dei siti, wrapper, regole di mapping, ..)

Criticità degli aggiornamenti periodici tramite procedure automatiche

- Navigazione
  - cambia il modello di un sito:
    - bisogna ridefinire le modalità di scaricamento delle pagine
- Generazione dei wrapper / Mapping
  - cambia la struttura delle pagine:
    - i wrapper risultano invalidati
    - le regole di mapping risultano invalidate
- Integrazione
  - vengono pubblicate nuove pagine:
    - bisogna trovare le correlazioni fra i nuovi documenti

Indicazioni per la gestione del progetto

Rischi del progetto (1)

- Complessità gestionale:
  - rilevanza strategica
  - interconnessione con altri progetti
  - eterogeneità degli sviluppatori
- Rischi per le dimensioni:
  - dimensione del sistema
  - numero complessivo mesi/uomo previsti

Rischi del progetto (2)

- Rischi per variabilità dei requisiti:
  - disponibilità, chiarezza, stabilità dei requisiti
  - comprensibilità degli strumenti esistenti
  - licello di formalizzazione del processo e delle fasi
- Rischi tecnologici
  - utilizzo di nuove tecnologie
  - mancanza di ambienti di sviluppo opportuni
  - integrazione di tecnologie eterogenee

Rischi del progetto: valutazione

Fattore di rischio	Alto	Medio	Basso
Rilevanza strategica del progetto	X		
Interconnessione con altri progetti		X	
Eterogeneità degli attori	X		
Numero complessi mesi/uomo previsti		X	
Dimensione del sistema		X	
Dimensione economica			X
Disponibilità e chiarezza requisiti		X	
Comprensibilità strumenti esistenti		X	
Livello di formalizzazione del processo			X
Utilizzo di nuove tecnologie	X		
Mancanza di ambienti di sviluppo adeguati		X	
Necessità di integrazione di tecnologie eterogenee	X		
Valutazione globale		X	

Metodologia di sviluppo agile: XP

- Per i fattori di rischio del progetto e la dimensione del team di sviluppatori è consigliabile una metodologia di sviluppo agile come XP
  - migliorare la comunicazione
  - cercare la semplicità
  - raccogliere ed ascoltare i feedback
  - avere coraggio nelle scelte
  - portare rispetto verso i componenti del team (sviluppatori, project manager, clienti)

Benefici

- Benefici:
  - maggiore automatizzazione del processo di estrazione e fornitura dati da Web (requisiti funzionali)
  - crescita formativa del personale (non solo sviluppatori)
    - tecnologie eterogenee e metodologia di sviluppo agile
- Metriche di riferimento:
  - contronfrare i tempi necessari alla fornitura dei dati quando il sistema non era presente, o in momenti in cui sono disponibili solo versioni parziali del sistema stesso

Conclusioni

Conclusioni

- L'impianto metodologico e le proposte:
  - offrono uno strumento di analisi per valutare le difficoltà di un intervento di estrazione con gli strumenti a disposizione
  - evidenziano i costi e le problematiche dell'applicazione del processo di estrazione dati e della sua manutenzione periodica
  - indicano quali siano i requisiti e le metodologie di funzionamento degli strumenti da realizzare intorno a RoadRunner al fine di ridurre i costi e semplificare le attività del processo
  - sono un punto di partenza per la CHI Technologies nella definizione e offerta di strumenti e servizi per l'estrazione di dati da Web
  - possono facilitare l'integrazione con altri servizi
    - es: indicizzazione dei dati estratti