tesi di laurea

Un processo a supporto della classificazione di pagine client

Anno Accademico 2006/2007

relatore

Ch.mo prof. Porfirio Tramontana

candidato

Marco Calandro

Matr. 885/73

Introduzione

- Il passaggio da un modello di comunicazione interattivo, ad un approccio di comunicazione di tipo batch rappresenta oggi una tendenza interessante oggi nel mondo del Web.
- In tale ottica, la classificazione di BCP prodotte da una server page ed il riconoscimento in maniera automatica della loro classe di appartenenza, nel rispetto di requisiti di accuratezza ed efficienza, può risultare utile.
- La corretta classificazione di tali pagine Web secondo questo approccio sarebbe utile per molteplici applicazioni:
  - Implementazione di wrapper.
  - Supporto all'analisi automatica dei risultati di test nelle diverse attività di validazione.
  - Ottenimento di un modello di UI per reingegnerizzare quest'ultima sulla base di una diversa architettura.
- Nel presente lavoro di tesi è stato studiato un processo per la classificazione di pagine Web prodotte da una server page, basato sulla comparazione della struttura HTML e del contenuto informativo delle pagine, ed è stato realizzato uno strumento software a supporto delle varie fasi di tale processo.

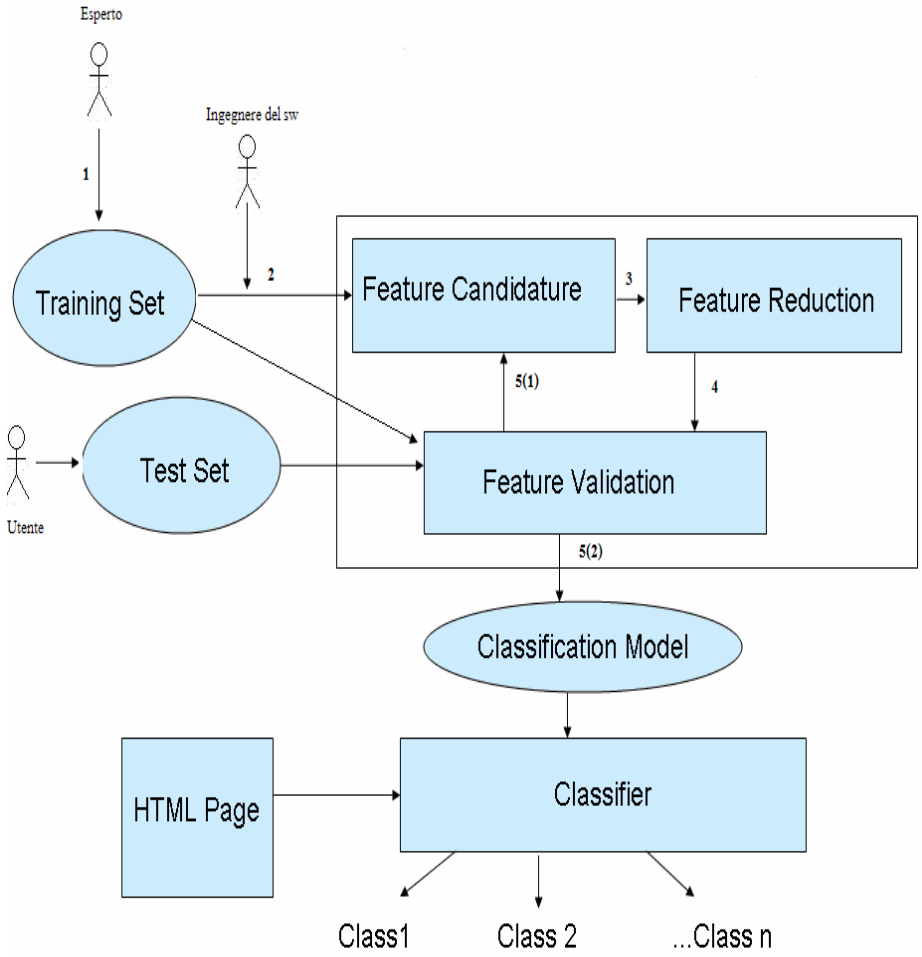
Descrizione del processo

1. Analisi sul Training Set

   - Definizione del Training Set
   - Candidatura delle feature
   - Valutazione delle feature candidate sul Training Set
   - Generazione delle espressioni discriminanti
   - Calcolo di Recall e Precision sul Training Set
   - Ripetere l'analisi finchè Recall e Precision non raggiungono la soglia fissata

2. Analisi sul Test Set

   - Definizione del Test Set
   - Valutazione delle espressioni discriminanti sul Test Set
   - Calcolo di Recall e Precision sul Test Set
   - Se Recall e Precision non raggiungono la soglia fissata ripetere l'analisi sul Training Set



La tipologia delle features

- Una feature può essere definita come una proprietà di cui sono dotate le istanze di una classe di pagine e che può essere espressa come un predicato che combini informazioni sugli elementi grafici e testuali delle pagine.
- Predicati di questo tipo possono essere pensati come delle query XPath applicate a pagine opportunamente convertite dal formato HTML al formato XHTML.
- E' possibile introdurre cinque categorie di feature relative ad una classe di pagine C:
  - Specific: una feature che è soddisfatta da tutte le pagine della classe C e solamente da queste.
  - Relevant: una feature che è soddisfatta da tutte le pagine di C, ma anche da altre pagine di altre classi.
  - CSPC (Conditionally Specific): una feature che è soddisfatta da un sottoinsieme delle pagine della classe C e da nessuna altra pagina delle altre classi.
  - Shared: una feature che è soddisfatta da un sottoinsieme delle pagine della classe C, ma anche da pagine di altre classi.
  - Irrelevant: una feature che non è soddisfatta da nessuna pagina della classe C.

Espressioni discriminanti e metriche di valutazione

- **Per ogni classe di equivalenza viene proposta una espressione booleana rispettando le seguenti regole.**
- **Primo candidato gli *SPECIFIC*:**
  - **se per una classe di equivalenza è presente almeno una feature che sia SPECIFIC essa da sola è sufficiente per la descrizione.**
- **Secondo candidato le *RELEVANT :***
  - **qualora non siano presenti featuresSPECIFIC, si ricorre alla combinazione secondo la *and* logica delle feature RELEVANT al fine di ottenere una maggiore precisione nella discriminazione.**
- **Terzo candidato le *CONDITIONALLY SPECIFIC (CSPC)*:**
  - **qualora non siano presenti neanche feature RELEVANT si procede combinandno con la *or* logica le feature CONDITIONALLY SPECIFIC, nel tentativo di riprodurre almeno parzialmente il potere discriminante delle feature SPECIFIC.**
- **In questo processo si valuterà l'efficacia di ciascuna espressione discriminante attraverso le metriche di Recall e Precision:**

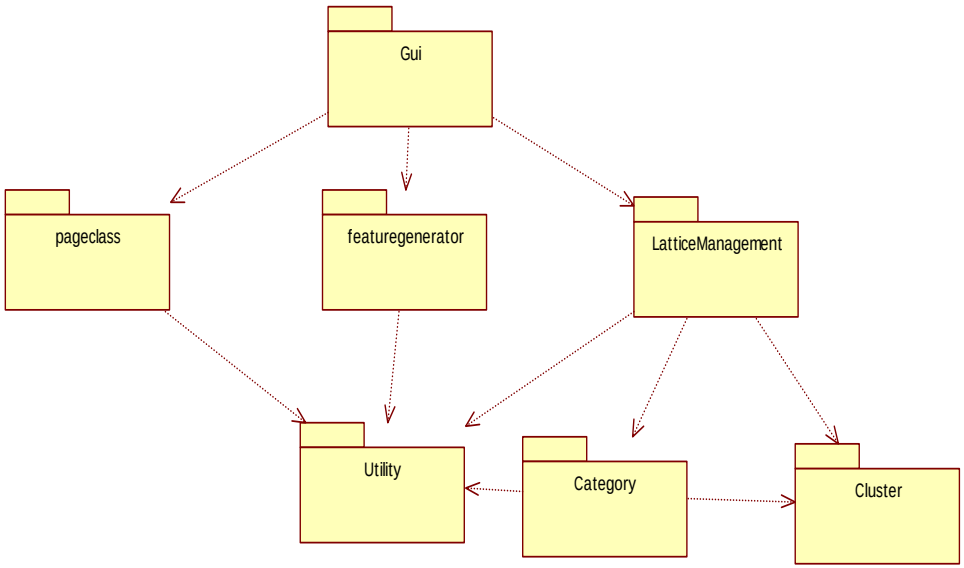
$$\mathrm{Pr}\,ecision = \frac{\#(pagesCorrectlyAttributedToClassC)}{\#(pagesAttributedToClassC)}$$

$$\mathrm{Re}\,call = \frac{\#(pagesAttributedToClassC)}{\#(AnalysedPpagesBelongingToClassC)}$$

La proposta automatica di feature

- E' stato elaborato ed implementato un algoritmo euristico per la proposta automatica di feature.
- La scelta di un criterio euristico nasce da motivazioni legate al tempo di esecuzione ed alla valenza non definitiva di tale fase nell'ambito del processo di classificazione, che lo fanno preferire ad un criterio esatto.
- Primo passo:
  - Estrazione pseudo-casuale di una pagina p1 dalla classe C e dei path dei suoi elementi (assoluti e relativi).
  - Generazione di un primo set di feature candidate.
- Secondo passo (per feature di tipo Relevant o Specific):
  - Estrazione pseudo-casuale di una seconda pagina p2 dalla classe C.
  - Eliminazione dal set iniziale, delle feature non soddisfatte da p2.
  - Se sono richieste feature di tipo Relevant il processo si arresta in questo punto.
- Terzo passo (per feature di tipo Specific o CSPC):
  - Estrazione pseudo-casuale di una pagina pj per ciascuna delle altre classi.
  - Eliminazione dalle feature candidate sopravvissute al passo precedente, di quelle che valgono anche per una sola delle pagine estratte dalle classi diverse da C.

L'architettura del tool



Esperimento 1

- **Server Page: http://www.wikipedia.org.**
- **Classi: *en*, *es*, *it*, *dan*, corrispondenti rispettivamente alle versioni nelle lingue inglese, spagnolo, italiano e danese.**
- **Feature Generation: ricerca automatizzata, impostata su features di tipo Specific e sui campi label (le feature proposte sono del tipo *path(assoluto o relativo)[contains(text(), 'testo')]*).**
- **Soglie richieste per Recall e Precision del 100%.**

Legenda feature

78-label	//a[contains(text(),'TilfÄ¦ldig artikel')]	dan
6-label	//a[contains(text(),'About Wikipedia')]	en
41-label	//a[contains(text(),'Cambios recientes')]	es
47-label	//a[contains(text(),'Puntano qui')]	it
66-label	//a[contains(text(),'Entra / Registrati')]	it

Espressioni discriminanti

Cluster	*Espressione disc.*
dan	*78-label*
en	*6-label*
es	*41-label*
it	*47-label*

Validazione Test Set 1

Cluster	Espressione discriminante	Recall	Precision
dan	*78-label*	*22/22*	*22/22*
en	*6-label*	*39/39*	*39/39*
es	*41-label*	*31/31*	*31/31*
it	*47-label*	*38/39*	*38/38*

Validazione Test Set 2

Cluster	Espressione discriminante	Recall	Precision
dan	*78-label*	*22/22*	*22/22*
en	*6-label*	*39/39*	*39/39*
es	*41-label*	*31/31*	*31/31*
it	*66-label*	*38/38*	*38/38*

Esperimento 2

- **Server Page: http://paginegialle.it.**
- ***Classi: empty, single*, *first*, *intermediate*, e *last* corrispondenti rispettivamente ad una pagina dei risultati di una ricerca fallita, ad una pagina dei risultati di una ricerca che contiene tutte le voci trovate, ad una pagina che riporta la prima parte delle voci trovate in una ricerca, ad una pagina che riporta una parte intermedia delle voci trovate in una ricerca, ad una pagina che riporta l'ultima parte delle voci trovate in una ricerca.**
- **Features Generation: ricerca automatizzata, impostata su features Specific su campi di i/o e su campi label per il cluster *empty*, e Relevant sugli stessi campi per gli altri cluster.**
- **Soglie richieste per Recall e Precision del 100%.**

Legenda feature

9-label	//tr/td[contains(text(),'Effettua una nuova ricerca:')]
4-i/o	//tr/@bgcolor
0-label	//td/span[contains(text(),'Nessun operatore trovato')]
18-label	//tr/td/a[contains(text(),'« prec')]
12-label	//tr/td/a[contains(text(),'succ »')]
32-i/o	//dd/select/option

Espressioni discriminanti

Cluster	*Espressione disc.*
empty	*0-label*
first	*! 18-label AND 12-label*
intermediate	*12–label AND 18-label*
last	*18-label AND !12-label*
single	*!18-label AND !32-i/o AND !0-label AND !12-label*

Validazione Test Set 1

Cluster	Espressione discriminante	Recall	Precision
empty	*4-i/o*	*50/50*	*50/51*
first	*!18-label AND 12-label*	*50/50*	*50/50*
intermediate	*12-label AND 18-label*	*50/50*	*50/50*
last	*18-label AND !12-label*	*50/50*	*50/50*
single	*!18-label AND !32-i/o AND 9-label AND!12-label*	46/50	*50/50*

Validazione Test Set 2

Cluster	Espressione discriminante	Recall	Precision
empty	*0-label*	*50/50*	*50/50*
first	*!18-label AND 12-label*	*50/50*	*50/50*
intermediate	*12-label AND 18-label*	*50/50*	*50/50*
last	*18-label AND !12-label*	*50/50*	*50/50*
single	*!18-label AND !32-i/o AND !0-label AND!12-label*	43/43	*43/43*

Tempi di esecuzione del Feature Generator

- **E' stato notato un andamento non lineare dei tempi di esecuzione rispetto al numero di feature inizialmente candidate a partire dalla prima pagina estratta, con tempi che crescono più velocemente del numero di feature.**
- **E' stato inoltre notato che generazioni di feature basate sui campi label delle pagine risultano caratterizzate da tempi di esecuzione maggiori di quelli ottenuti con generazioni basate sui campi di input o di output.**
- **Il FeatureGenerator è un tool che automatizza parzialmente la fase di Feature Candidature: è infatti comunque richiesta l'interazione con l'utente per impostare il processo di proposta automatica di feature, a valle di una preliminare osservazione del codice XHTML delle pagine raccolte. A tale scopo sarebbe utile una stima del tempo che impiega l'utente per l'analisi visuale.**
- **Restringendo il campo degli elementi da valutare nell'ambito delle pagine, si riduce il numero di feature candidate, e di conseguenza i tempi di esecuzione prodotti dal tool. Naturalmente, per un'analisi completa, bisognerebbe tener conto del tempo impiegato dall'utente per decidere su quali elementi incentrare l'analisi.**

Feature Specific su campi label

Cluster	Tempo di generazione (min,sec)	Numero di feature candidate inizialmente
empty	01:20	3261
first	41:25	74362
last	15:07	36417
intermediate	32:20	58446
single	04:05	12482

Feature Specific su campi i/o

Cluster	Tempo di generazione (min,sec)	Numero di feature candidate inizialmente
empty	00:51	1385
first	08:08	5801
last	03:55	4017
intermediate	06:43	5477
single	02:26	2905

Conclusioni

- Il framework realizzato nell'ambito del presente lavoro di tesi assiste l'ingegnere del software in tutte le fasi della processo di classificazione delle BCP prodotte da una applicazione Web.
- Limiti all'utilità del tool sono legati alle incompatibilità delle BCP con XHTML, ed agli strumenti software utilizzati per la conversione HTML – XHTML, che non sempre producono delle conversioni pulite ed ottimizzate.
- Il tool può essere ottimizzato con l'aggiunta di nuove euristiche per la proposta automatica di feature.