tesi di laurea

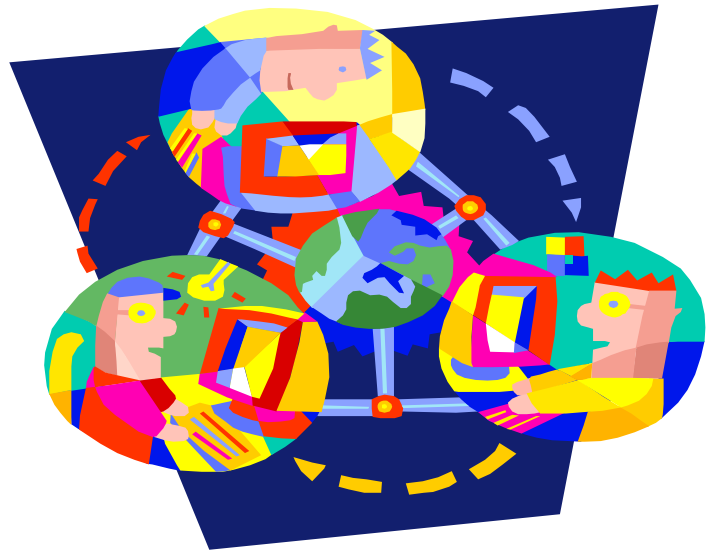
Sviluppo di un tool a supporto del recupero del modello dei contenuti di pagine Web

relatore
Ch.mo prof. Porfirio Tramontana

candidato
Claudio Quaranta
Matr. 534/1468

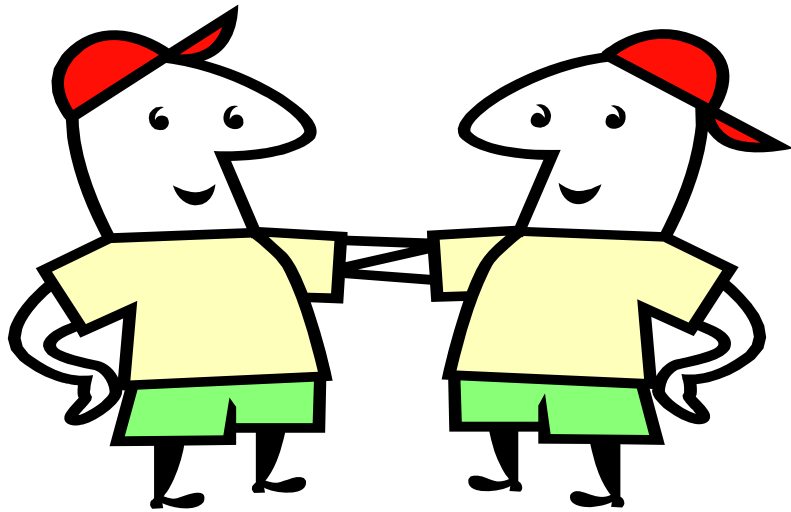
Ubiquitous Web Applications

- **Grazie all'enorme sviluppo di internet sempre un maggior numero di applicazioni vengono sviluppate per offrire servizi via Web**
- **Il progetto UWA (Ubiquitous Web Applications) ha come obiettivo quello di definire un'insieme di metodologie, notazioni, e strumenti per assistere la progettazione veloce delle future applicazioni Web**
- **Si desidera analizzare applicazioni già esistenti per ricavare con quali oggetti relativi alla metodologia UWA esse sono state realizzate**



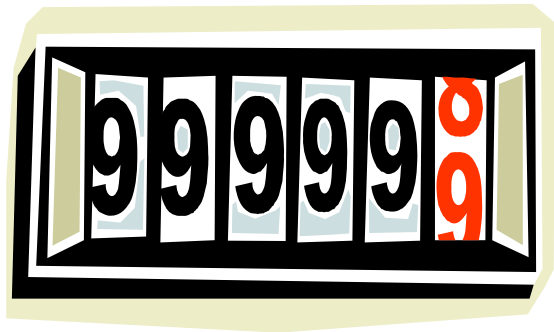
Candidati Label e Output

- **L'obiettivo è ricavare i candidati Label e Output (gergo UWA) di progettazione da applicazioni già esistenti**
- **Una delle strade possibili è quella di ricavare i pattern da applicazioni già esistenti, confrontando tra di loro pagine HTML che svolgono funzioni simili.**
- **Problemi :**
  - Quando due pagine HTML si dicono "simili"?
  - In che modo raggruppare le pagine "simili" tra di loro?



Distanza di Levenshtein

- **La distanza di Levenshtein tra due stringhe A e B è il numero minimo di modifiche elementari che consentono di trasformare A in B.**
- **Per modifica elementare si intende :**
  - inserimento di un carattere
  - eliminazione un carattere
  - sostituzione di un carattere con un altro

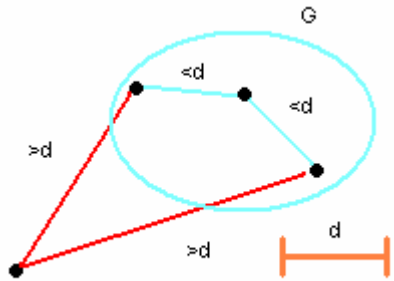


- **Sostituendo ad ogni Tag di una pagina html un carattere è possibile ricavare la distanza di Levenshtein tra pagine HTML**

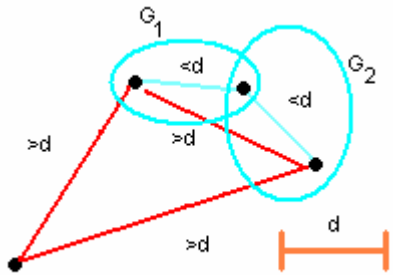
Classificazione pagine HTML

- **In che modo suddividere le pagine Html di un'applicazione?**
- **Tre i criteri proposti :**

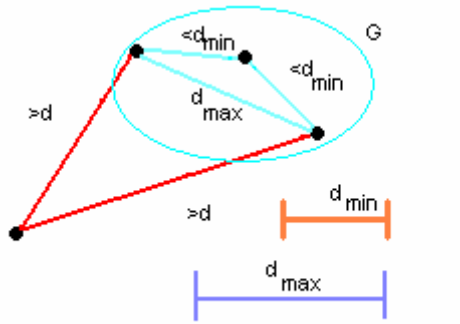
$$x_i \in G \Leftrightarrow \exists x_j \in G : d(x_i, x_j) < d_{\max}$$



$$x_i \in G \Leftrightarrow \forall x_j \in G : d(x_i, x_j) < d_{\max}$$



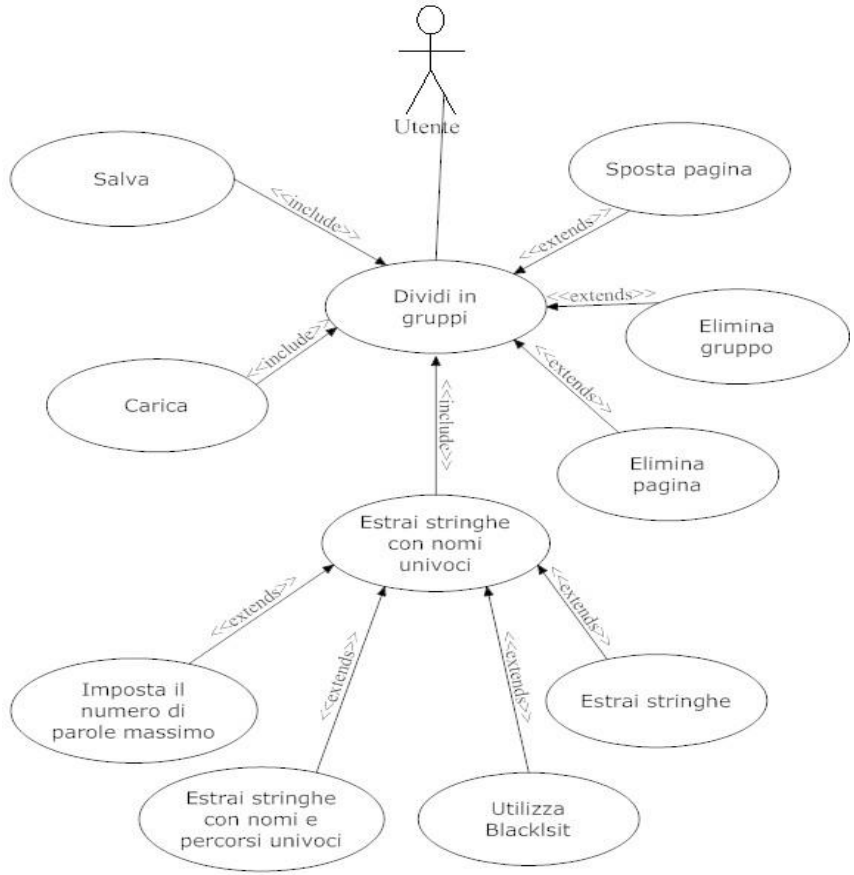
$$x_i \in G \Leftrightarrow \exists x_j \in G : d(x_i, x_j) < d_{\min} \wedge \forall x_j \in G : d(x_i, x_j) < d_{\max}$$



Use Case Diagram

Si è creato un tool scritto in Java in grado di supportare l'identificazione di tali oggetti.

In particolare il software implementa i tre tipi di raggruppamenti esposti e provvede all'estrazione automatica delle stringhe di testo dalle pagine html

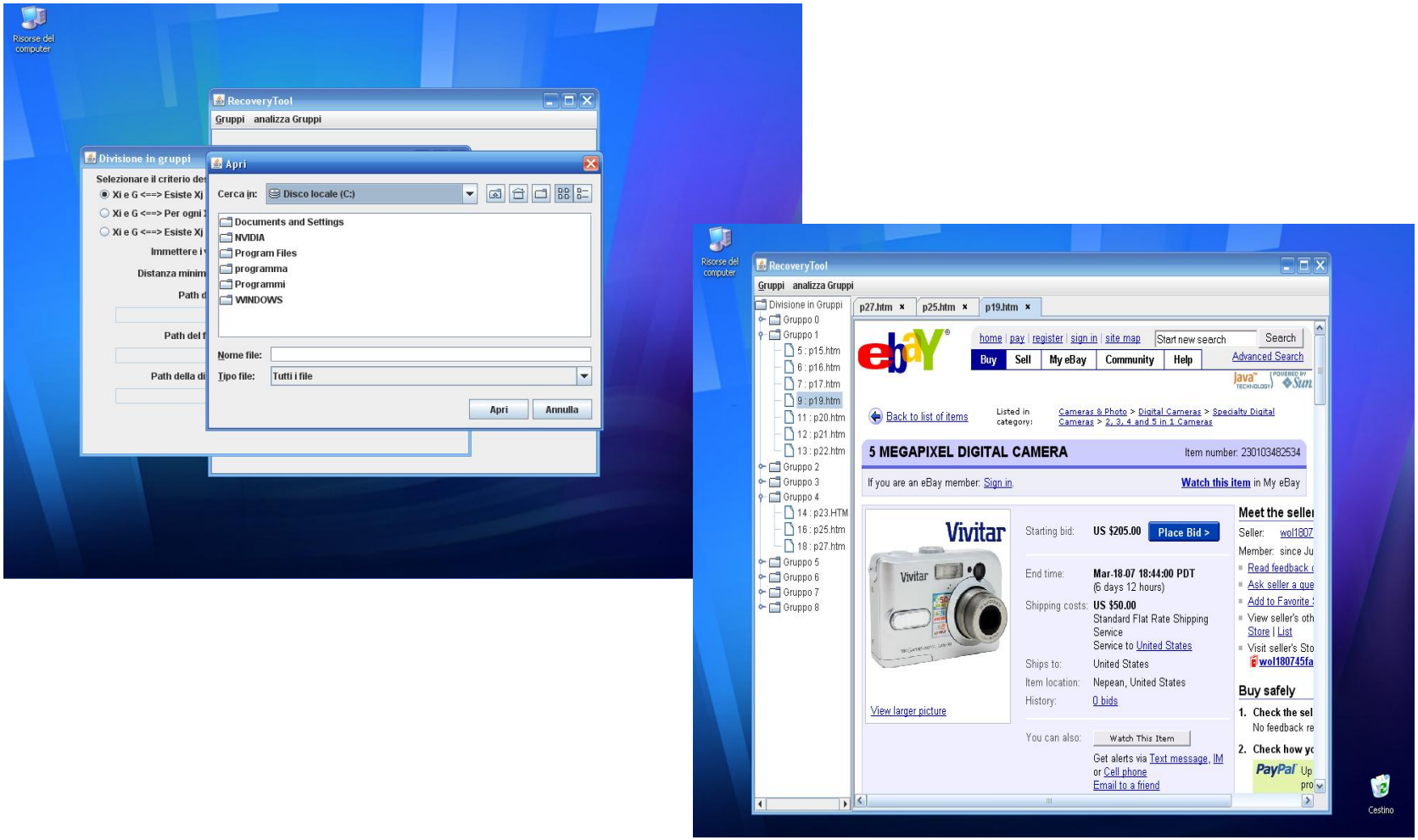


UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Facoltà di Ingegneria

Corso di Studi in Ingegneria Informatica

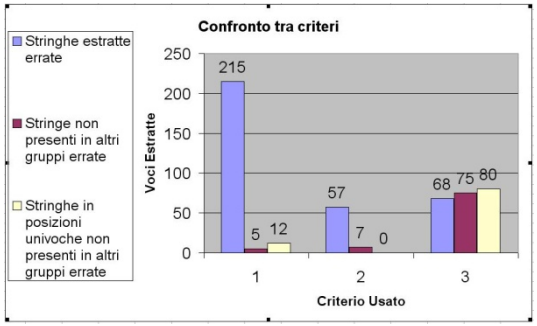
Sviluppo di uno strumento a supporto del recupero del modello dei contenuti di pagine Web



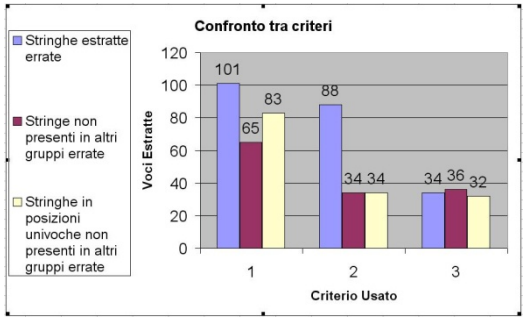
Casi di studio

- **Sono stati effettuati 3 casi di studio**
- **Si sono analizzati alcuni set di pagine provenienti da queste tre applicazioni Web, estraendone le stringhe e verificando la loro correttezza confrontando i risultati con un'insieme di dati corretti**

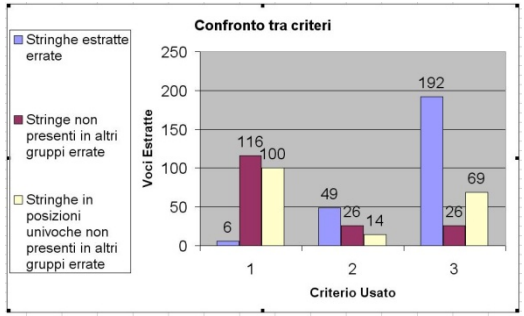
Ebay.it



Nga.gov



Chl.it



Conclusioni

- **Verificando i risultati ottenuti dai tre criteri di confronto si evince che:**
  - Il primo criterio è il meno adatto all'identificazione delle label e output in quanto molti candidati risultano essere errati
  - I restanti due criteri risultano equivalenti per quanto riguarda la correttezza dei candidati proposti, ma il secondo criterio permette di suggerire un valore di soglia "di default".
  - Per futuri esperimenti quindi si consiglia di utilizzare il secondo criterio proposto, impostando come soglia un valore compreso tra 30 e 50. Tale scelta deve essere fatta in base al set di pagine che si desidera analizzare e da quanto sono diverse tra di loro le pagine.