ISSN: 1825-411X | DOI: 10.1473/stor299

TauRo – un sistema di ricerca e gestione avanzata di documenti XML

Paolo Ferragina, Alida Isolani, Dianella Lombardini, Tommaso Schiavinotto

Storicamente, 4 (2008).
ISSN: 1825-411X. Art. no. 35. DOI: 10.1473/stor299

Con l'avvento del Web 2.0 assistiamo a un cambiamento radicale nell'uso della Rete che non viene più vista solo come strumento da cui estrarre informazione prodotta da altri ma anche come mezzo per collaborare e condividere idee e contenuti (esempi sono Wikipedia, YouTube, Flickr, MySpace, LinkedIn, etc.). Nasce in quest'ottica TauRo, uno strumento di semplice utilizzo che consente di *creare, gestire, condividere e ricercare collezioni digitali di documenti XML* via Web. TauRo infatti è un sistema collaborativo attraverso il quale un utente, in possesso di un accesso a Internet e dotato di un browser, può pubblicare e condividere in maniera semplice ed efficace i propri documenti XML creando collezioni tematiche personali e/o condivise. TauRo inoltre fornisce meccanismi di ricerca estremamente avanzati grazie all'impiego di un motore di ricerca per documenti XML, denominato **TauRo-core**, progettato e realizzato presso il Centro di Ricerca Signum[1] della Scuola Normale Superiore di Pisa, sulla base di una pluriennale esperienza maturata nell'ambito dell'analisi testuale informatizzata.

TauRo-core: il motore di ricerca

La definitiva affermazione di XML come formato di interscambio dati su Internet da un lato, e dei motori di ricerca a la Google dall'altro, offrono una stimolante opportunità tecnologica per la fruizione di grandi quantità di dati

ISSN: 1825-411X | DOI: 10.1473/stor299

su normali PC o su dispositivi portatili quali smart-phone e palmari. Il motore di ricerca TauRo-core è uno strumento software innovativo, modulare e sofisticato che consente la memorizzazione compressa, e l'analisi/ricerca efficiente di pattern arbitrari in grandi collezioni di documenti XML disponibili sia su un unico PC che distribuite tra più PC possibilmente dislocati in varie parti della Rete. La flessibilità dell'architettura modulare di TauRo-core e l'utilizzo di tecniche di compressione avanzate per la conservazione dei documenti e per la memorizzazione degli indici lo rendono utilizzabile nei diversi scenari illustrati in Figura 1.

[[figure caption="Figura 1 – Alcuni scenari di utilizzo di TauRo-core: centralizzato, distribuito, P2P. Si noti la diversa distribuzione tra i nodi della rete dei moduli Interfaccia, TauRo-core e dei Documenti indicizzati dal motore di ricerca"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_01.jpg[[/figures]]

Rispetto ai motori di ricerca attualmente disponibili nel panorama internazionale, TauRo-core offre ulteriori e più sofisticate funzionalità di ricerca e analisi implementate per soddisfare le attuali esigenze della codifica di testi letterari. Questi ultimi infatti possono avere una marcatura che rende difficoltosa l'analisi ai motori di ricerca più comuni solitamente progettati per documenti non strutturati (es. motori di ricerca per il Web), o per documenti fortemente strutturati (es. database), o per documenti semi-strutturati (es. motori di ricerca per XML) ma nei quali non si fanno assunzioni sulla semantica della marcatura stessa. Ad esempio, si consideri la seguente porzione di un documento XML/TEI[2]:

…<c>W</c>illiam <c type='capital'>S</c>hakespeare was born some time in late April 1564…

Un motore di ricerca tradizionale non è in grado di reperire l'occorrenza della parola *Shakespeare* a seguito della presenza del tag <c> che identifica la

ISSN: 1825-411X | DOI: 10.1473/stor299

lettera maiuscola significativa. Questo problema diventa ancora più evidente se si considerano altre possibili "strutture" tipiche di un testo letterario, quali ad esempio note o glosse. TauRo-core invece offre la possibilità di indicizzare testi XML per i quali siano state definite delle opportune categorie di tag – denominate *smart-tag*[3] – alle quali siano state associate delle specifiche direttive di gestione/ricerca. Per cogliere la flessibilità del concetto di *smart-tag* ne illustriamo la classificazione:

- **jump-tag**: i tag di questo gruppo indicano un temporaneo cambio di contesto – come nel caso di un tag che indica una nota – e in questo modo il contenuto del tag è distinto dal testo vero e proprio e la ricerca avviene distinguendo i due piani semantici.
- **soft-tag**: questi tag non comportano un cambio di contesto, ma se l'elemento di apertura o di chiusura del tag è presente all'interno di una stringa di caratteri non separati da spazio questa stringa forma un'unica parola. Infatti nel nostro esempio la stringa <c>S</c>hakespeare viene considerata come Shakespeare.
- **split-tag**: rientrano in questa categoria i tag a cui viene attribuito un significato analogo al separatore di parola; non viene, dunque, cambiato il contesto e le parole vengono effettivamente considerate divise.

Nel caso dell'esempio, il tag <c> è classificato come uno *smart-tag* di tipo *soft*. TauRo-core offre inoltre un linguaggio di interrogazione proprietario, denominato TRQL, sufficientemente potente da consentire all'utente di effettuare delle analisi complesse sui testi che tengano conto della classificazione suddetta, e della relazione esistente tra contenuto e struttura dei documenti.

Questa flessibilità consente a TauRo-core di poter essere utilizzato anche in contesti diversi da quelli strettamente letterari, quali ad esempio, le collezioni documentali della pubblica amministrazione, gli archivi biologici, la manualistica, le collezioni legislative, le news, etc.. Il contesto letterario

ISSN: 1825-411X | DOI: 10.1473/stor299

rimane però il più complesso e quindi costituisce, per la sua peculiare mancanza di uniformità, il banco di provapiùstimolante e significativo.

TauRo: il sistema sul Web

TauRo è un sistema collaborativo che consente a un qualunque utente Web, previa regitrazione gratuita, di creare e condividere collezioni di documenti XML, e di sfruttare le potenzialità di TauRo-core per eseguire ricerche *full-text*, per espressioni regolari[4], per similitudine, e ricerche nella struttura dei documenti XML. Tali ricerche possono essere effettuate su un collezione per volta o su più collezioni in contemporanea, indipendentemente dalla loro natura. Illustriamo di seguito alcune caratteristiche del sistema aiutandoci con degli *screenshot*.

[[figure caption="Figura 2 - Home page di TauRo"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_02.jpg[[/figures]]

Le collezioni

Ogni utente registrato può caricare su **TauRo** le proprie collezioni di documenti XML. Una volta caricata, una collezione viene indicizzata da TauRo-core e resa così disponibile per successive operazioni di ricerca. In qualunque momento un utente può modificare le proprie collezioni aggiungendo/cancellando documenti, spostando documenti tra collezioni diverse, condividendo documenti tra più collezioni, o modificando lo stato di una collezione che può essere:

- **privata**: accessibile e ricercabile unicamente dal proprietario;
- **pubblica**: ricercabile da tutti gli utenti previa registrazione, modificabile solo dal proprietario;
- **su invito**: ricercabile e modificabile da tutti gli utenti su invito esplicito del proprietario. Lo scopo di quest'ultima tipologia di collezioni è la creazione di gruppi di lavoro, incentivando così l'utilizzo di TauRo anche come strumento collaborativo.

ISSN: 1825-411X | DOI: 10.1473/stor299

[[figure caption="Figura 3 - Form di inserimento/modifica di una collezione"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_03.jpg[[/figures]]
All'atto del caricamento o della modifica di una collezione, l'utente può impostare alcuni semplici parametri di configurazione, quali la specifica degli *smart-tag* e del tag di interruzione di pagina, al fine di sfruttare al massimo le potenzialità del motore di ricerca. Un ulteriore livello di personalizzazione offerto da TauRo consiste nell'associare a ciascuna collezione un proprio *foglio di stile XSL*[5] al fine di visualizzare in modo gradevole i risultati delle ricerche effettuate su di esse.

I documenti

Il sistema fornisce all'utente un insieme di funzioni che permettono di caricare, classificare e rimuovere i documenti XML dalle collezioni di sua proprietà. Durante il caricamento il sistema tenta di effettuare un riconoscimento automatico della DTD e dei *file di entity* utilizzati nel documento XML confrontandone il nome pubblico con quelli precedentemente salvati; nel caso di mancato riconoscimento viene data la possibilità di inserimento all'utente. Ogni documento può essere reso scaricabile liberamente a tutti oppure può essere selezionata una delle licenze *Creative Commons* che garantiscono il proprietario del documento da utilizzi impropri.

[[figure caption="Figura 4 - In primo piano form per la modifica di un documento, sullo sfondo elenco di documenti"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_04.jpg[[/figures]]

La ricerca

TauRo offre due differenti modalità di ricerca, la prima è pensata per cercare parole all'interno dei documenti (*ricerca semplice*), la seconda permette di costruire interrogazioni anche di tipo strutturale cioè sugli elementi – tag e attributi – della marcatura attraverso un'interfaccia grafica (*ricerca avanzata*

ISSN: 1825-411X | DOI: 10.1473/stor299

). In entrambi i casi le interrogazioni vengono tradotte in una sintassi interpretabile a TauRo-core e ad esso inviate. Il risultato della ricerca è l'elenco dei documenti della collezione che verificano l'interrogazione, affiancati alla distribuzione dei risultati all'interno dei documenti stessi. Selezionando un documento si accede all'elenco delle occorrenze contestualizzate, ossia inserite in un frammento di testo che le contiene, con la possibilità di accedere direttamente al testo nella sua interezza.

La ricerca semplice può essere esatta, per prefisso, suffisso, espressione regolare o per differenze. È possibile specificare più parole, e in questo caso esse risulteranno adiacenti nel documento. La ricerca semplice può essere eseguita su più collezioni contemporaneamente. Un interessante uso di questa modalità di ricerca consiste nella correzione dei refusi in un testo. Ad esempio, ricercando la parola *resto* con "1 differenza" troviamo le occorrenze delle parole che differiscono da *resto* per al più un carattere: si ottengono quindi parole come *resto*, *testo*, *presto*, *restò*, *sesto*, etc.. Poiché TauRo fornisce un risultato contestualizzato (Figura 6), è semplice per un utente individuare gli errori ortografici e procedere alla loro correzione.

[[figure caption="Figura 5 - Si richiedono le news identificate dal tag **<description>** che contengono l'espressione regolare se[tl].*[ae]"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_05.jpg[[/figures]]
Per illustrare le potenzialità della ricerca avanzata, forniremo un esempio. Supponiamo di voler effettuare una ricerca sulla collezione *Quotidiani*, l'archivio delle news che TauRo aggiorna in automatico settimanalmente prelevando i *feed RSS*[6] dai maggiori quotidiani italiani. Vogliamo selezionare i sommari delle news, identificati dal tag **<description>**, al cui interno siano presenti parole che iniziano con *set* o *sel* e terminano con *a* oppure *e*, come ad esempio “settembre”, “settimana” o “sella”. Questa ricerca è esprimibile attraverso la seguente espressione regolare **se[tl].*[ae]** dove '[]' indica l'alternativa tra più caratteri, '. ' indica un carattere qualsiasi e '.***** ' indica una serie di caratteri qualsiasi possibilmente vuota.

ISSN: 1825-411X | DOI: 10.1473/stor299

[[figure caption="Figura 6 – Occorrenze contestualizzate in un documento"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_06.jpg[[/figures]]
Per fare ciò l'interfaccia guidata di TauRo permette di selezionare il tag **<description>** dalla lista di tutti quelli presenti nella collezione. Una volta che questo è stato aggiunto nel *wizard* di costruzione dellaquery è possibile inserire l'espressione regolare semplicemente selezionando la voce "Aggiungi parole da ricercare" nel menu a tendina e digitandola nel campo di testo (Figura 5). Il risultato della ricerca è l'elenco dei documenti della collezione che verificano l'interrogazione. Selezionando un documento si accede all'elenco delle occorrenze (Figura 6) che costituisce il punto di partenza per l'accesso al testo: in questo caso è stato creato un foglio di stile ad hoc per la visualizzazione degli abstract con la possibilità di accedere al quotidiano relativo per la lettura del testo completo della news (Figura 7).

[[figure caption="Figura 7 - Visualizzazione del documento con un'occorrenza del risultato"]]figures/2009/ferragina-isolani-lombardini-schiavinotto/ferragina-isolani-lombardini-schiavinotto_209_07.jpg[[/figures]]
I risultati di interrogazioni così raffinate permettono di svolgere analisi che possono portare l'utente ad individuare nuove e originali chiavi di lettura dei testi. Gli esperimenti da noi condotti su numerose e ampie collezioni testuali, ci portano ad affermare che il sistema è particolarmente efficiente nella loro gestione e analisi: una ricerca richiede mediamente meno di un secondo per poter essere eseguita. Rimandiamo il lettore interessato al sito ufficiale per una descrizione più dettagliata delle caratteristiche del sistema o per provare direttamente le sue funzionalità sulle collezioni pubbliche attualmente disponibili.

Ci auguriamo che la semplicità d'uso affiancata alle numerose e sofisticate funzioni di ricerca offerte da TauRo, consentano a questo strumento di diventare un punto di riferimento per gli utenti del Web – umanisti e non –

ISSN: 1825-411X | DOI: 10.1473/stor299

che si trovano ogni giorno a gestire e analizzare collezioni di documenti XML senza doversi preoccupare di problemi di istallazione e aggiornamento di complessi pacchetti software.

Note

[1] Signum – Centro di ricerche informatiche per le discipline umanistiche. http://www.signum.sns.it

[2] http://www.tei-c.org/

[3] L. Lini, D. Lombardini, M. Paoli, D. Colazzo, C. Sartiani, *TReSy: A Text Retrieval System for XML Documents*. In Dino Buzzetti, Giuliano Pancaldi and Harold Short (eds), *Augmenting Comprehension: Digital Tools and the History of Ideas,* Atti del convegno *Informatica umanistica: filosofia e risorse digitali*, Bologna, *22–23 September 2002*. London: Office for Humanities Communication, King's College London 2004.

[4] L'espressione regolare è un insieme di caratteri, metacaratteri e operatori che definiscono una stringa o un gruppo di stringhe in un *pattern* di ricerca. Attraverso le espressioni regolari è possibile ricercare motivi ricorrenti all'interno di parole nel testo. Ne daremo un esempio di utilizzo successivamente.

[5] **XSL**, acronimo di e**X**tensible **S**tylesheet **L**anguage, permette di visualizzare uno stesso documento XML in formati diversi.

[6] RSS, acronimo di *Really Simple Syndication*, è uno dei più popolari formati per la distribuzione dei contenuti Web, basato su XML. RSS definisce una struttura adatta a contenere un insieme di notizie, ciascuna delle quali sarà composta da vari campi: nome autore, titolo, testo, riassunto, etc.