Automatic Ontology-Based Knowledge Extraction from Web Documents

5 gennaio 2009

1 Introduzione

Al fine di ottenere un effettivo WEB semantico bisogna essere in grado di costruire servizi che consentano l'estrazione di conoscenza dai documenti presenti in rete.

Una via che di solito si segue è quella dell'annotazione 'manuale' attraverso metadata; essa però risulta di difficile utilità sia per i costi, sia per i tempi impiegati. Conviene dunque sviluppare sistemi informatici che attuino l'estrazione automatica di una conoscenza e ciò può essere fatto se essi vengono guidati da una ontologia che specifica quali informazioni ricercare.

Un esempio interessante su questi servizi informatici è il progetto **Artequakt** sviluppato dall'Università di Southampton e presentato nel numero di Gennaio-Febbraio 2003 della rivista **Intelligent Systems** (IEEE).

Artequakt è stato sviluppato in due step:

step1 viene creata una ontologia che ha come dominio gli artisti e le loro opere; nello specifico Artequakt limita la sua ricerca ad informazioni legate al periodo impressionista (vedi Rembrandt). Sono in seguito utilizzati alcuni tools di "Information Extraction (IE)" e di linguistica computazionale (ad es. Wordnet) per popolare l'ontologia e per creare una Knowledge-Base (KB) da interrogare in seguito.

step2 viene sviluppato un applicativo che interroga la KB e produce documenti html "human-readable" in base ai testi ed alle informazioni salvate nel processo di estrazione delle informazioni.

2 Knowledge extraction

L'estrazione della conoscenza è una procedura piuttosto complessa: a tale fine Artequakt utilizza alcuni strumenti open-source quali ad esempio **Aple-Pie Parser**, **GATE** (General Architecture for Text Engineering) e **WordNet** (un database lessicale generico della lingua inglese).

Per meglio comprendere come avviene questo processo di estrazione conviene descrivere la sequenza di operazioni che il sistema compie.

Se ad esempio un utente richiede la biografia di un artista non inserito nel database, Artequakt inizia una ricerca di documenti in internet utilizzando i classici motori di ricerca. Le pagine individuate sono filtrate in base alla somiglianza (tramite vector-similarity) con pagine ritenute attendibili e prelevate da un sito affidabile.

A questo punto il sistema crea una lista di paragrafi e frasi e ne analizza la struttura sintattica ed il contenuto semantico al fine di individuare la conoscenza rilevante. Il tool Apple Pie Parser raggruppa le frasi che l'analisi sintatttica determina come essere grammaticalmente correlate. Grazie all'analisi semantica - tramite l'uso di Wordnet e GATE - il sistema individua gli elementi principali di una frase (soggetto, verbo, oggetto...) ed identifica le entità nominate (ad es. che Rembrandt è una persona o che Leiden è un luogo). Al fine di determinare se l'informazione processata risulta essere significativa, Artequakt utilizza l'ontologia sottostante e determina le relazioni fra le entità individuate (sia quelle necessarie, sia quelle aspettate).

Consideriamo ad esempio la frase che Artequakt ottiene dal processo di ricerca:

> "*Rembrandt was born on July 15 1606 in Leiden, the Netherlands*"

L'ontologia di fondo che ci serve in questo esempio è la semplice:

```
Person —> date_of_birth(date)
Person —> place_of_birth(place)
```

Per mezzo di GATE e WordNet si riescono ad individuare le entità:

```
Rembrandt —> Person
15 July 1606 —> Date
Netherlands —> Location
Leiden —> Location
```

Grazie all'utilizzo combinato di WordNet, GATE e dell'Ontologia del dominio di conoscenza, il sistema associa il verbo "*was born*" alle due proprietà date_of_birth e place_of_birth ed inserisce nella KB i due 'record' (triplette):

```
Rembrandt – date_of_birth – 15 July 1606
Rembrandt – place_of_birth – Leiden, Netherlands
```

3 Automatic ontology population

Con il termine di "ontology population" si intende la procedura che porta all'inserimento dell'informazione all'interno della KB per mezzo di una Ontologia. Di solito questa operazione avviene in modo semi-automatico in quanto

si accompagna all'utilizzo di tools per la creazione delle annotazioni di documenti il contemporaneo intervento esterno per la validazione delle informazioni individuate.

Artequakt tenta di superare questo processo semi-automatico sfruttando in modo completo l'ontologia: per ogni documento internet processato il sistema crea un file xml dove sono salvati i paragrafi e le frasi ritenute significative ed inoltre le strutture dell'ontologia utilizzate. Per fare un esempio pratico consideriamo il file:

```
<paragraph>
    <url>...</url>
    <text>
        Rembrandt was born on July 15 1606 in Leiden, the
        Netherlands. His father was a millar who wanted the
        boy to follow a learned profession... He was
        influenced by the work of Caravaggio and was
        fashinated by the work of many other italian artists
    </text>
    <sentence>
    <text>
        Rembrandt was born on July 15 1606 in Leiden,
        the Netherlands
    </text>
    <person>
        <name>Rembrandt</name>
        <place_of_birth>Leiden, the Netherlands</place_of_birth>
        <date_of_birth>
            <day>15</day>
            <month>7</month>
            <year>1606</year>
        </date_of_birth>
    </person>
    </sentence>
    ...
    <sentence>
    <text>
    He was influenced by the work of Caravaggio and was
    </text>
    <person>
        <name>Rembrandt</name>
        <inspired_by>the work of Cravaggio</inspired_by>
    </person>
    </sentence>
    ...
</paragraph>
```

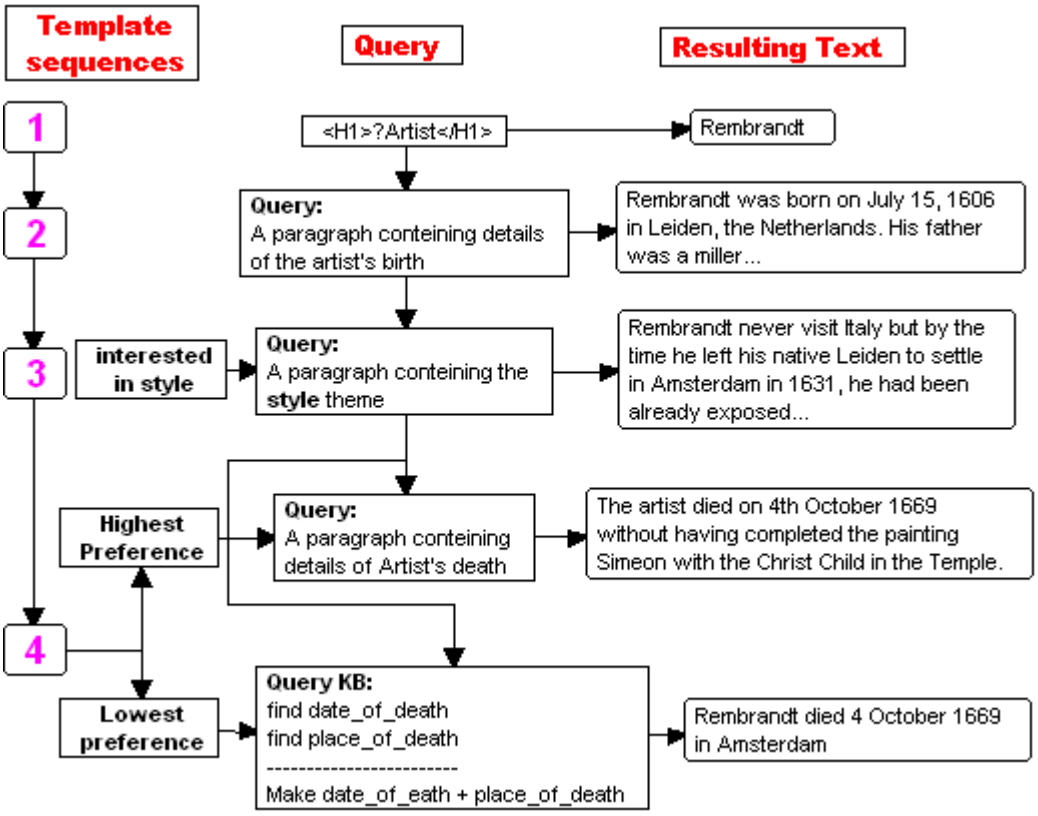
Questi file xml vengono inviati all'Ontology-Server. Il sistema inoltre è dotato di alcuni semplici tools di ricerca che permettono l'interrogazione del server al fine di estrarre le informazioni più comuni (ad esempio la ricerca di tutti i paragrafi che contengono la data di nascita dell'artista).

4 Narrative Generation and Biography Templates

Le macchine possono utilizzare ontologie strutturate per scambiarsi informazioni, le persone invece necessitano di un metodo più intuitivo: il racconto - tramite la descrizione delle informazioni presenti nella KB - rappresenta la via più semplice per sviluppare questa interfaccia.

La Narratologia suddivide la narrazione in *storie* - gli elementi base - e *discorsi*; questi ultimi sono una tecnica per presentare le prime. Nel caso di Artequakt la KB mantiene le storie (elementi base del discorso) in quanto la raccolta di informazioni sul WEB si riduce in ultima analisi al salvataggio di "estratti di testo" senza un esplicito ordine. Compito dunque del sistema è quello di organizzare le storie in discorsi - secondo il termine narratologico - per poi presentarli in documenti html.

Artequakt raggiunge quest'ultimo obiettivo con l'utilizzo di template supervisionati al cui interno sono presenti query da effettuare alla KB. I template sono in formato xml ed un esempio è rappresentato dal grafico seguente:



Il template presenta 4 step ad ognuno dei quali è associata una query nel database KB ed i corrispondenti risuldati espressi sotto forma di stringhe di

testo, la cui composizione genera il file html finale dove è presentata la biografia di Rembrandt. Analizziamo i 4 step in sequenza:

1. La query che viene fatta all'inizio è quella relativa all'artista di cui si vuole conoscere la biografia ed il risultato (Rembrandt) ha una formattazione <H1> in html.

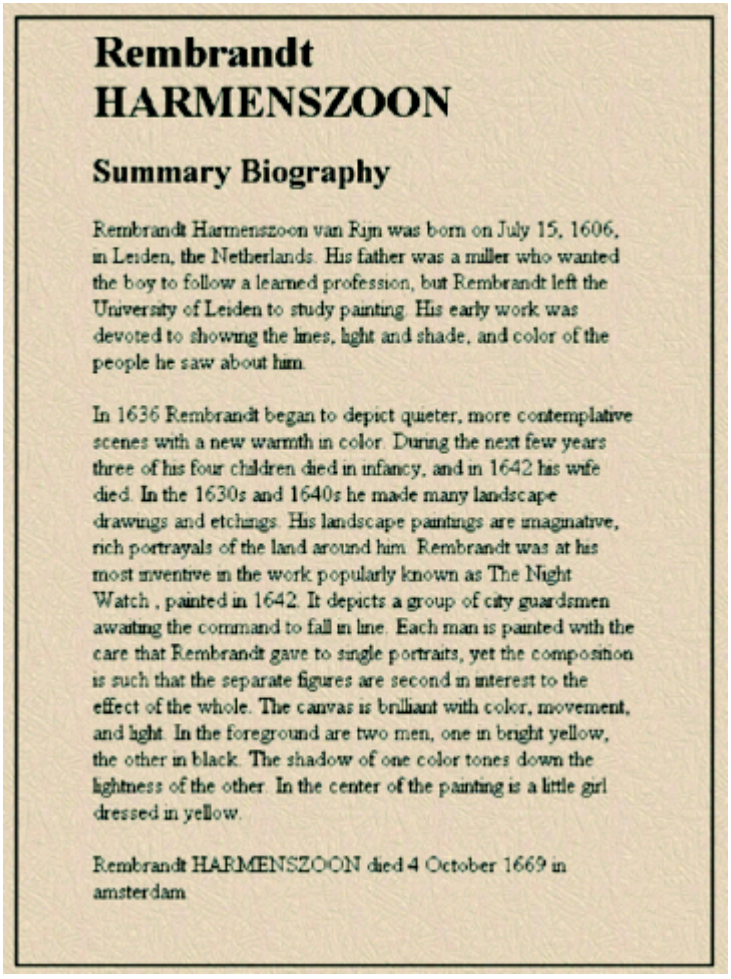
2. Il sistema richiede alla KB di estrarre un paragrafo contenente i dettagli della nascita dell'artista. Nel nostro caso la query da come risultato uno dei testi analizzati in precedenza: "*Rembrandt was born on July 15 1606 in Leiden, the Netherlands. His father was a millar...*"

3. Questo step è facoltativo. Se l'utente che ha richiesto la biografia è interessato allo stile dell'artista, Artequakt richiede un paragrafo che contenga le informazioni al riguardo. Nel nostro caso "*Rembrandt never visit Italy but by the time he left his native Leiden...*"

4. In quest'ultimo punto il sistema ha una scelta da compiere: vi sono due possibilità, una con preferenza più elevata dell'altra. Come primo obiettivo Artequakt richiede un paragrafo sulla morte dell'artista (Highest Preference), se la ricerca da risultato negativo il sistema compie una query alla KB (Lowest Preference) richiedendo la data ed il luogo di morte per poi creare una stringa intellegibile

5 Results

Il risultato finale che il sistema Artequakt propone all'utente è la pagina html della figura seguente:



Rembrandt
HARMENSZOON

Summary Biography

Rembrandt Harmenszoon van Rijn was born on July 15, 1606, in Leiden, the Netherlands. His father was a miller who wanted the boy to follow a learned profession, but Rembrandt left the University of Leiden to study painting. His early work was devoted to showing the lines, light and shade, and color of the people he saw about him.

In 1636 Rembrandt began to depict quieter, more contemplative scenes with a new warmth in color. During the next few years three of his four children died in infancy, and in 1642 his wife died. In the 1630s and 1640s he made many landscape drawings and etchings. His landscape paintings are imaginative, rich portrayals of the land around him. Rembrandt was at his most inventive in the work popularly known as The Night Watch , painted in 1642. It depicts a group of city guardsmen awaiting the command to fall in line. Each man is painted with the care that Rembrandt gave to single portraits, yet the composition is such that the separate figures are second in interest to the effect of the whole. The canvas is brilliant with color, movement, and light. In the foreground are two men, one in bright yellow, the other in black. The shadow of one color tones down the lightness of the other. In the center of the painting is a little girl dressed in yellow.

Rembrandt HARMENSZOON died 4 October 1669 in amsterdam