

## Estrazione di conoscenza da testi letterari annotati

**Francesca Pietra**  
**Patrizia Michelassi**

**Pisa KDD Lab, CNUCE-CNR & Univ. Pisa**



Università di Pisa

8 Maggio 2003

## Introduzione

### OBIETTIVO

Descrivere il funzionamento di un sistema per estrarre conoscenza da un testo letterario annotato con XML in modo da:

- Mostrare come strumenti di recupero ed analisi della conoscenza operano su un testo letterario.
- Mostrare come può essere interpretato un testo letterario da un punto di vista semantico. I risultati ottenuti al termine del processo di elaborazione devono soddisfare il bisogno informativo dell'utente e devono essere interpretati da esperti del dominio.



## Introduzione

### Novità...

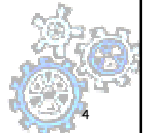
- 1) Uso dell'annotazione e dell'ontologia
- 2) *Uso di strumenti di Data Mining*: rappresenta la novità rispetto all'area dell'information extraction.
- 3) Sperimentazioni sulla Cantica "Inferno" della Divina Commedia



## Realizzazione

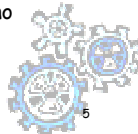
La realizzazione del sistema ha richiesto

- 1) Costruzione di una base di conoscenza
- 2) Utilizzo di strumenti per l'estrazione di conoscenza



## Caratteristiche del sistema

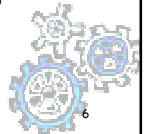
- ⌘ Una base di conoscenza strutturata in modo adeguato ai mezzi a disposizione per la manipolazione e l'estrazione dei dati (meta-rappresentazione del testo+descrizione del contesto semantico)
- ⌘ Capacità espressive avanzate per formulare le richieste dell'utente e per rispondere a domande complesse (queries)
- ⌘ Strumenti applicativi opportuni per eseguire le queries sfruttando le conoscenze descritte e ottenere come risultato le risposte al bisogno informativo espresso dall'utente.



## Problematiche da modellare

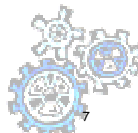
Il sistema deve essere capace di rispondere alle seguenti domande:

- ⌘ Quanti e quali sono i ghibellini e quanti e quali i guelfi presenti nella Cantica Inferno della Divina Commedia?
- ⌘ Quante e quali figure della mitologia classica vengono evocate nell'Inferno?
- ⌘ Classificare l'atteggiamento di Dante verso un personaggio in base alle caratteristiche del personaggio.



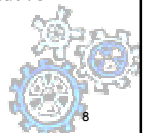
## Il punto di partenza...

- ⌘ Costruire la base di conoscenza, in particolare:
  - > Stabilire il linguaggio di rappresentazione della base di conoscenza.
  - > Stabilire il contenuto della base di conoscenza: annotazione del testo letterario e conoscenza concettuale del dominio.



## Breve cenno all'XML

- ⌘ XML: acronimo di eXtensible Markup Language;
- ⌘ È un linguaggio di rappresentazione
- ⌘ Crea documenti strutturati
- ⌘ Non ha tag predefiniti, ma opera come metalinguaggio
- ⌘ Tiene conto della struttura dinamica di un testo letterario
- ⌘ Si adatta alle diverse interpretazioni
- ⌘ Consente di incrementare dinamicamente la meta-rappresentazione testuale con l'aggiunta di nuove informazioni.



## Struttura della Base di Conoscenza

⌘ Tenendo presente la generale partizione tra conoscenza testuale e conoscenza contestuale, la base di conoscenza è stata suddivisa in due parti:

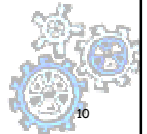
I) Metarappresentazione del testo

II) Ontologia



## Metarappresentazione del testo:

- 1) **NARRAZIONI**  
(parti di testo fuori dalle virgolette)
- 2) **DIALOGHI**  
(parti di testo tra virgolette)



## Narrazioni

- CANTO

- TESTO NARRAZIONE:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE

- COLLEGAMENTO DIALOGO:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE



## Dialoghi

- CANTO  
- DIALOGO:

INIZIO:  
    VERSO  
    CARATTERE  
FINE:

- PERSONAGGI DIALOGO  
- SEGMENTI DIALOGO:

TESTO:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE

CHI PARLA  
A CHI  
DI CHI

INTRODUZIONE:

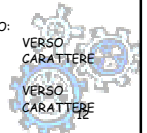
TESTO INTRODUZIONE:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE

INTERRUZIONE:

TESTO INTERRUZIONE:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE

CONCLUSIONE:

TESTO CONCLUSIONE:  
INIZIO:  
    VERSO  
    CARATTERE  
FINE:  
    VERSO  
    CARATTERE



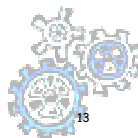
### Esempio di "Narrazione" in formato XML (Canto I, 1-64)

```

<Table>
<Narrazioni>
<Canto> I </Canto>
<TestoN>
<TestoNarrazione> Nel mezzo del
cammin...
mi ritrovai per una selva
oscura
.....
Quando vidi costui nel gran
diserto,
</TestoNarrazione>
<Inizio>
<Verso> 1 </Verso>
<Carattere> 1 </Carattere>
</Inizio>
<Fine>
<Verso> 64 </Verso>
<Carattere> 36 </Carattere>
</Fine>
    
```

```

<CollegamentoDialogo>
<Inizio>
<Verso> 65 </Verso>
<Carattere> 1 </Carattere>
</Inizio>
<Fine>
<Verso> 136 </Verso>
<Carattere> 37 </Carattere>
</Fine>
    
```



13

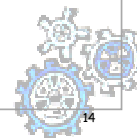
### Esempio di "Dialogo" in formato XML (Canto I, 65-136)

```

<Table>
<Dialoghi>
<Canto></Canto>
<Dialogo>
<PersonaggiDialogo>
<Nome>Dante</Nome>
<Nome>Virgilio</Nome>
</PersonaggiDialogo>
<Inizio>
<Verso>65</Verso>
<Carattere>1</Carattere>
</Inizio>
<Fine>
<Verso>136</Verso>
<Carattere>37</Carattere>
</Fine>
<SegmentiDialogo>
<Segmento>
<ChiParla>Dante</ChiParla>
<AChi>Virgilio</AChi>
    
```

```

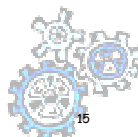
<Testo>
<TestoSegmento>Miserere di me qual che tu sii, od
ombra od omo certo!</TestoSegmento>
<Inizio> <Verso>65</Verso>
<Carattere>1</Carattere></Inizio>
<Fine>
<Verso>66</Verso>
<Carattere>43</Carattere>
</Fine>
<Interruzione>
<TestoInterruzione>gridai a lui, </TestoInterruzione>
<Inizio>
<Verso>65</Verso>
<Carattere>19</Carattere>
</Inizio>
<Fine>
<Verso>65</Verso>
<Carattere>31</Carattere>
</Fine>
<Interruzione>
<Testo>
</Segmento>
    
```



14

## Ontologia

- 1) TOPOGRAFIA
- 2) PERSONAGGI



15

## Topografia

```

- REGNO (Inferno / Purgatorio / Paradiso):
  LOCALIZZAZIONE:
  LUOGO
  NOME LUOGO
  PARTIZIONE LUOGO
  NOME PARTIZIONE
  INIZIO:
  CANTO
  VERSO
  FINE:
  CANTO
  VERSO
  PECCATO
  PENA
  CONTRAPPASSO:
  TIPO (per analogia / per contrasto)
  SPIEGAZIONE
    
```



16


### Esempio di "Topografia" in formato XML (XI Cerchio, 1° Zona)

```

<Table>
<Topografia>
<Regno> Inferno </Regno>
<Localizzazione>
<Luogo> IX Cerchio </Luogo>
<NomeLuogo> Cocito
</NomeLuogo>
<Partizione> I Zona </Partizione>
<NomePartizione> Caina
</NomePartizione>
<Inizio>
<Canto> XXXII </Canto>
<Verso> 1 </Verso>
</Inizio>
<Fine>
<Canto> XXXII </Canto>
<Verso> 72 </Verso>
</Fine>

```

<Peccato> Tradimento: traditori dei  
 parenti  
 </Peccato>  
 <Pena> I peccatori stanno immersi  
 nel lago ghiacciato fino al collo e  
 piangono tenendo il capo basso  
 </Pena>  
 <Contrappasso>  
 <Tipo> per analogia </Tipo>  
 <Spiegazione> come in vita ebbero  
 il cuore così duro e freddo da  
 tradire le persone più care, così  
 ora sono immersi nel duro e  
 freddo ghiaccio </Spiegazione>  
 </Contrappasso>  
 </Localizzazione>



17

### Personaggi


- NOME
- COGNOME
- NOTO COME
- TIPO (Storico / Letterario / Creatura):
  - x Storico: EPOCA (es. Medioevo, Antichità Greca, Antichità Latina, ecc.)  
 CATEGORIA (es. Politico, Ecclesiastico, ecc.)  
 SOTTOCATEGORIA (es. Capo Ghibellino, Papa, ecc.)  
 POSIZIONE POLITICA (es. Guelfo, Ghibellino, ecc.)  
 NOTE BIOGRAFICHE:
    - DATA
    - AVVENIMENTO
    - LUOGO
  - x Letterario: CARATTERISTICA  
 NOTE  
 AMBITO DI APPARTENENZA (es. Mitologia classica, Bibbia, Letteratura romanza, ecc.)  
 CATEGORIA  
 SOTTOCATEGORIA  
 CARATTERISTICA  
 NOTE
  - x Creatura: RUOLO  
 SIGNIFICATO ALLEGORICO



18

### Note

- IDENTIFICAZIONE:
  - IPOTESE: IDENTIFICATO CON PROPOSTO DA MOTIVO ACCETTATO DA SMENTITO: DA CHI PERCHÉ
  - SOPRANNOOME: APPELLATIVO MOTIVO FONTE
  - PROBLEMATICHE: SOGGETTO della PROBLEMATICA VERSIONI TESTIMONIANZE: TESTO FONTE
- INFORMAZIONI:
  - TESTO FONTE
  - COMMENTI:
    - TESTO COMMENTATORE MOTIVO CONDIVISO DA NON CONDIVISO
    - DA
  - EVENTI:
    - ACCADUTO MOTIVO LUOGO TESTO FONTE



19


### Esempio di "Personaggi" in formato XML (Farinata, Ulisse, lonza)

```

<Table>
<Personaggio>
<Personaggio>
<Nome> Manente </Nome>
<Cognome> degli Uberti </Cognome>
<NotaCome> Farinata </NotaCome>
<Tipo> Storico </Tipo>
<Epoca> Medioevo </Epoca>
<Categoria> Politico </Categoria>
<Sottocategoria> Capo ghibellino
</Sottocategoria>
<PosizionePolitica> Ghibellino
</PosizionePolitica>
<NoteBiografiche>
<Episodio>
<Data> 1239 </Data>
<Avvenimento> diventa capo del partito
ghibellino </Avvenimento>
<Luogo> Firenze </Luogo>
</Episodio>
<Episodio>
</Episodio>
</NoteBiografiche>

```

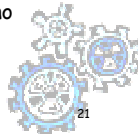
</Personaggio>  
 <Personaggio>  
 <Nome> Ulisse </Nome>  
 <NotaCome> Ulisse </NotaCome>  
 <Tipo> Letterario </Tipo>  
 <AmbitoAppartenenza> Mitologia  
 classica  
 </AmbitoAppartenenza>  
 <Categoria> Eroe greco </Categoria>  
 <Caratteristica> Re di Itaca  
 </Caratteristica>  
 </Personaggio>  
 <Personaggio>  
 <NotaCome> Lanza </NotaCome>  
 <Tipo> Creatura </Tipo>  
 <Ruolo> ostacolare l'ascesa di Dante al  
 colle della salvezza </Ruolo>  
 <SignificatoAllegorico> lussuria  
 </SignificatoAllegorico>  
 </Personaggio>



20

## Caratteristiche del sistema

- ⌘ Una base di conoscenza strutturata in modo adeguato ai mezzi a disposizione per la manipolazione e l'estrazione dei dati (meta-rappresentazione del testo+descrizione del contesto semantico)
- ⌘ Capacità espressive avanzate per formulare le richieste dell'utente e per rispondere a domande complesse (queries) ←
- ⌘ Strumenti applicativi opportuni per eseguire le queries sfruttando le conoscenze descritte e ottenere come risultato le risposte al bisogno informativo espresso dall'utente.



21

## Il linguaggio di rappresentazione delle queries

- ⌘ **Xquery**: è un XML Query Language definito dal W3C. Rappresenta il risultato di un processo di standardizzazione, di conseguenza eredita molte caratteristiche da altri XML Query Languages precedentemente studiati.
- ⌘ **MQL**: acronimo di Mining Query Language, definito e implementato al Dipartimento di Informatica dell'Università di Pisa, consente la modellazione di problemi complessi, linguaggio altamente espressivo.



22

## Strumenti utilizzati

Gli strumenti applicativi utilizzati per effettuare le estrazioni sono:

- ⌘ **XQuery Demo**
  - consente di eseguire query XML, necessarie per operare sulla base di conoscenza rappresentata in XML
- ⌘ **Strumenti di Data Mining**
  - consente di modellare e risolvere problemi di Data Mining (KDDML-MQL)

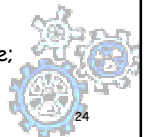


23

## Strumenti di Data Mining

KDDML-MQL

- ⌘ Ambiente di supporto per risolvere problemi di DM;
- ⌘ Il sistema è caratterizzato da:
  1. Un linguaggio per la formalizzazione di problemi di DM, noto come MQL;
  2. Un ambiente di supporto (KDDML), utilizzato come motore esecutivo delle queries.
- ⌘ Interamente basato su XML;
- ⌘ Architettura aperta e facilmente estendibile;



24

## Esempio 1

Supponiamo di voler modellare la seguente domanda:  
*"Quanti e quali sono i ghibellini e quanti e quali sono i guelfi presenti nella Cantica Inferno?"*

Come opera il meccanismo di ricerca:

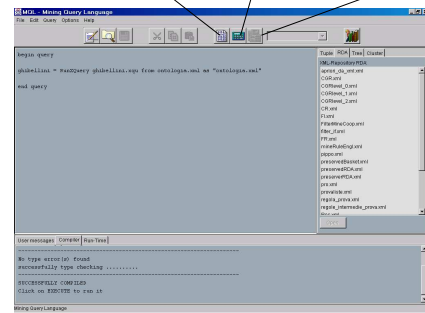
- ⌘ Modellazione della XQuery per estrarre le informazioni da uno o più documenti XML
- ⌘ Invocazione, tramite una query MQL, dell'operatore RunXQuery, allo scopo di eseguire la XQuery modellata al passo precedente
- ⌘ Visualizzazione dei risultati ottenuto tramite browser



25

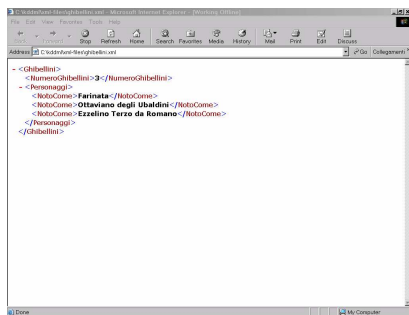
## Interfaccia grafica del sistema

Botone per la compilazione      Botone per l'esecuzione      Botone per la visualizzazione



26

## Risultato



27

## Esempio 2

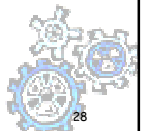
Il sistema deve modellare il seguente problema:

*"Classificare l'atteggiamento di Dante nei confronti di tutti i personaggi con cui parla in base a certi elementi selezionati in un dialogo"*

Come si procede?

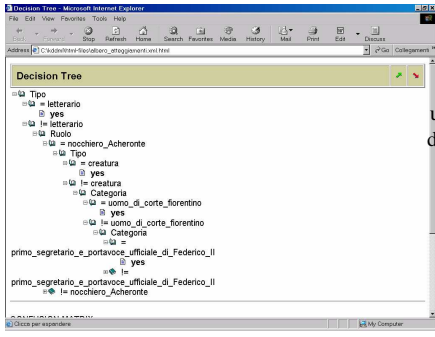
L'idea: costruire un albero di classificazione che, rispetto ai sentimenti che Dante prova in riferimento ad un personaggio, distingue l'atteggiamento di Dante in base all'attributo scelto per fare la classificazione

Per esempio classificare l'atteggiamento di Dante in Benevolo o Non Benevolo.




28

## Un esempio di classificazione




Il risultato consiste nell'aggiunta di un'informazione alla descrizione data dalla classificazione



29

## Conclusioni

- ⌘ E' stato mostrato come strumenti di recupero e di analisi della conoscenza operano su un testo letterario opportunamente annotato
  - ⌘ E' stato mostrato come può essere interpretato il testo letterario, in particolare come può essere estratta la conoscenza utilizzando strumenti di Data Mining
  - ⌘ In generale i risultati ottenuti soddisfano un bisogno informativo espresso dall'utente
  - ⌘ I risultati ottenuti, per ora, non devono essere considerati attendibili, in quanto ottenuti da una base di conoscenza in via di sviluppo.
- 
- 30