Tecnologie, strumenti e processi per l'accesso alle informazioni e l'estrazione della conoscenza

Maurizio Lancia – Alberto Salvati

CNR - Ufficio Sistemi Informativi

Sommario

- Scenario
- Obiettivi e linee guida
- Scelte tecnologiche
- Patrimonio Informativo
- Estrazione della conoscenza
- Esempi

Ufficio Sistemi Informativi

- Coordina lo sviluppo del sistema informativo dell'Ente in modo da assicurare la coerenza con standard tecnici e organizzativi comuni, ai sensi dell'art. 17 del Codice dell'Amministrazione Digitale

- Cura la progettazione, realizzazione e gestione di sistemi per la raccolta, l'elaborazione e la diffusione delle informazioni dell'Ente (applicativi, intranet, Data Warehouse, siti web, ecc.)

Scenario

- Sistemi Amministrativi e Gestionali
  - Personale
  - Contabilità
  - Attività Scientifiche (Previsione, Gestione e Consuntivazione attività di ricerca)
    - PdGP
    - Consuntivi
    - GECO
- Siti per l'accesso all'informazione e ai servizi
  - Sito CNR
  - Intranet
  - Data Warehouse
  - Albi elettronici (Diramazione interna, Comunicazione OO.SS.)
  - Aree Web (CDA, CSG, ...)

Obiettivi

- Integrazione dei sottosistemi (semplificazione gestione operativa)
- Rendere le informazioni coerenti e facilmente accessibili (tecnologia Internet)
- Soluzioni scalabili
- Creazione di competenze interne per progettazione, sviluppo e gestione del sistema

Linee guida

- Standard aperti
- Piattaforme e strumenti Open Source
- Riuso (in linea con direttive e normative vigenti)
- Pieno governo dell'intero processo di progettazione, realizzazione e manutenzione

Scelte tecnologiche

Tecnologie Internet

- Sistemi Applicativi: EAI (Enterprise Application Integration)
  - Standard comuni per cooperazione tra sottosistemi
  - Tecnologie abilitanti: Java 2 Enterprise Edition e altre piattaforme aperte
  - Architetture orientate ai servizi (SOA)
- Siti informativi: approccio dinamico Web-Database
- Strumenti a supporto
  - Estrazione conoscenza
  - Full-Text Indexing
  - Motori di ricerca

Volumi

- Software:
  - Source code: ~50,000 FPs (~ 1.500.000 linee di codice)
  - Data base tables: ~ 1.000
  - Users: ~10.000
- Produzione annua Dati-Documenti
  - > 10.000/anno documenti contabili (contratti, incarichi, ...)
  - > 25.000/anno documenti scientifici
  - > 200.000/anno documenti protocollati

Patrimonio informativo

- Dati strutturati
  - Contabilità
  - Gestione del Personale
  - Gestione attività di ricerca
- Dati non strutturati (testi liberi, news, pagine web, curricula, articoli, abstract, consuntivi, ...)
  - Sistema programmatico
  - Siti informativi (sito CNR, albi, Aree Web)
  - Intranet

Un grande e crescente patrimonio informativo

… e nel mondo?

- l'anno scorso si sono prodotti più transistor che chicchi di riso (IBM, 2007)
- l'universo digitale cresce del 60% ogni anno (IDC)
- nel 2011 un incremento di 10 volte in 5 anni (IDC)
- entro il 2011 solo metà delle informazioni potrà essere conservata, il resto non troverà spazio sui dispositivi di archiviazione (IDC)

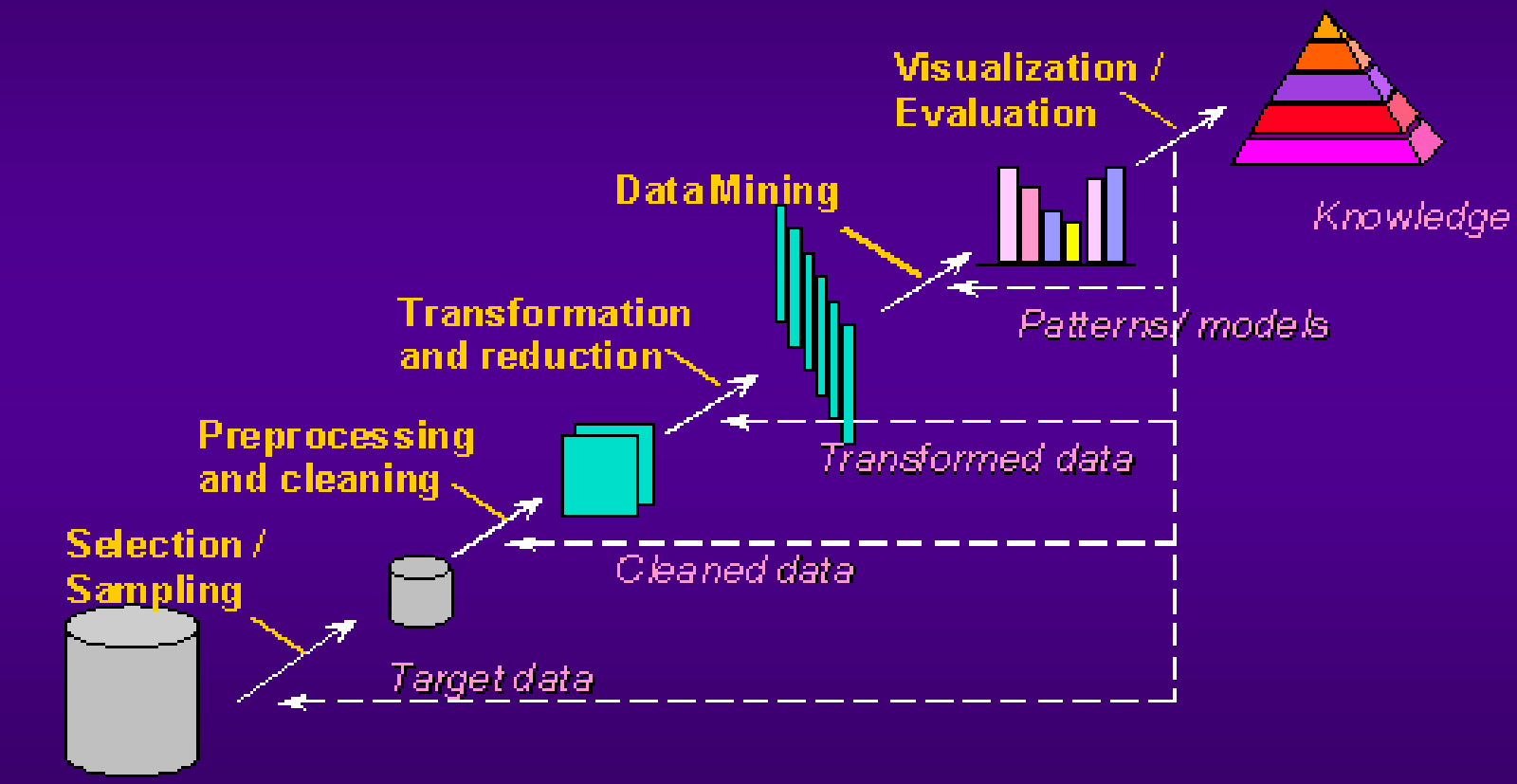
Informazioni ⇨ Conoscenza

- Disponibilità di enormi quantità di dati con una ricchezza di informazioni potenzialmente accessibili
- Indicazioni fondamentali per i processi decisionali
- Conoscenza preziosa per tutti

Estrazione della conoscenza

- KDD (Knowledge Discovery in Databases): l'intero processo strutturato di estrazione della conoscenza
- Data Mining: applicazione all'interno del processo KDD di specifici algoritmi per l'individuazione di correlazioni tra i dati. Generalmente si tratta di algoritmi di *clustering* (per il raggruppamento tematico) oppure di algoritmi di *machine learning* (per la classificazione automatica).
- Text Mining: applicazione di tecniche di Data Mining a testi non strutturati; utile per individuare gruppi tematici, classificare documenti, scoprire associazioni nascoste, addestrare motori di ricerca, estrarre concetti per creazione di ontologie
- Web Mining: applicazione di tecniche di Data Mining a dati (usage, content, structure) non strutturati presenti su siti web

Il processo di estrazione di conoscenza (KDD)



Usama Fayyad 13

Text Mining

Società dell'informazione

crescita smisurata del numero di documenti che possono contenere informazioni interessanti (agenzie stampa, pagine web, e-mail, libri e articoli in versione digitale, ...)

strumenti automatici per la loro catalogazione ed analisi

Text Mining

- Il Text Mining coinvolge l'applicazione di tecniche da aree come:
  - ***information retrieval:*** insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico
  - **elaborazione del linguaggio naturale (NLP): s**i occupa dell'analisi del linguaggio umano al fine di consentire la comprensione automatica del linguaggio naturale da parte del computer così come farebbe un essere umano.
  - **estrazione delle informazioni**: processo che consente di ottenere dati strutturati da un documento in linguaggio naturale non strutturato
  - **Data Mining**.

Esempi

- “metrica delle commesse”
  - Estrazione terminologica
  - Algoritmi di analisi (LSA)
  - Clustering
  - Correlazione
  - Generazione di nuovi metadati e parole chiave
  - Vicinanza/sovrapposizione/similitudine tra commesse
  - Rappresentazione spaziale
    - tecnologie-prodotti-metodologie
    - approccio-oggetto-finalità

Esempi

- “fisco e finanza”
  - Dichiarazioni fraudolente
  - Soci di persone che hanno partecipato ad altre società fallimentari
  - Commercio nascosto/evasione fiscale (eBay, Aste on line, Porta Portese, ...)

- “decisori e primo *screening*”
  - Sapere se una certa richiesta è probabilmente da accogliere o probabilmente da scartare

Esempi

- “letteratura biomedica”
  - Analizzare la letteratura biomedica, nel campo della genetica, allo scopo di individuare le eventuali interazioni tra geni
  - La conoscenza che si estrae analizzando le pubblicazioni specialistiche può essere considerata una fondamentale sorgente di informazioni che il ricercatore usa per interpretare e comprendere meglio i risultati sperimentali
  - Es. lista delle proteine presenti in un testo e nel tipo di relazione esistente tra loro
  - Es. scoperta di nuove interazioni che possono o no verificarsi, oppure la relazione tra tipi di interazioni e particolari malattie

Esempi

- “pubblicazioni”
  - Riconoscere coautori/collaboratori su dati non strutturati
  - Trovare testi “simili” in altra lingua
  - Riconoscere autori/località/fatti/norme citati all’interno di testi
  - Trovare autori che hanno trattato un certo argomento (ad esempio “inquinamento ambientale” (ricerca di esperti)
  - Costruire “*reti di collegamento*” tra autori o tra argomenti, rappresentabili anche graficamente e con archi pesati

Intranet
Istituti
Personale
Commesse
Proposte RSTL
Finanziamenti esterni
Partecipazioni societarie
Spin-off
Contratti
Doc. Programmatici
Motore di ricerca

Archivio Storico
Commesse

Motore di ricerca

Istruzioni per la ricerca

Parole da cercare:

fine

☐ tutte | ☐ una qualunque | ☐ la frase esatta | ☑ con operatori booleani [☐] **Includi sinonimi:** ☑

Caratteri speciali: * ? = % # ~ : w/ pre/ ~~ & | !

☑ **Suggerimenti** | ☑ Logs | ☑ Traduzioni | ☑ Termini correlati | ☑ Soggetti giuridici | ☑ Espressioni

Logs:

fine (v:33)
fine & progetto (v:3)
fine & progetto & mare (v:2)

Traduzioni:

fine: end fine finish sake
fine: bello contravvenzione elegante fine multa multare

Termini correlati: ?

fine: appendice, chiusa, chiusura, coda, concetto, conclusione, conclusione, cosa, decedere, decesso, desinenza, epilogo, esito, eutanasia, evento, evento, fiasco, finalità, geniale, idea, intendimento, intento, intenzione, martirio, meta, mira, morte, obbiettivo, obiettivo, oggetto, partizione, piccolo, preziosità, prezioso, proponimento, raffinato, record, ricercato, riuscita, scomparsa, scopo, snobismo, sofisticato, sonno, sorte, sottigliezza, sottile,

Soggetti giuridici:

fine:
fineschi silvia
vinci fine instruments di g.vinci

Espressioni:

fine: "Fine Arts Museum of San Francisco (USA)" "Montreal Museum of Fine Art (Montreal, Canada)" "Pushkin Museum of Fine Arts (Mosca, Russia)" "avere fine" "avere fine" "fine anno" "fine settimana" "fine stagione" "porre fine".

Cerca in:

☐ **Tutte le sezioni:** >5000 ☐ Istituti: 3202 ☑ Personale: >5000 ☐ Commesse: >5000 ☐ Fin. Esterni: 360
☐ Proposte RSTL: 668 ☐ Part. societarie: 101 ☐ Spin-off: 14 ☐ Contratti: >5000 ☐ Doc. Programmatici: 55

Grazie per l'attenzione

Alberto Salvati
alberto.salvati@cnr.it