WEB MINING

WEB MINING

“In un ambiente quale quello dell’*E-commerce,* altamente competitivo al giorno d’oggi, il successo di un sito spesso dipende dall’abilità del sito stesso nel trattenere i visitatori e nel trasformare casuali curiosi in potenziali clienti. La personalizzazione automatica basata sul Web Mining e le *recommender system technologies* stanno diventando strumenti critici in questa arena visto che aiutano ad adattare l’interazione tra sito e visitatore in base ai bisogni ed agli interessi di quest’ultimo.”

(Mobasher B. *et al.*, 2001)

WEB MINING

- Gli autori hanno voluto introdurre e sottolineare il motivo per cui sono nati i metodi di raccolta ed elaborazione di ingenti quantità di dati che si applicano al Web, ossia a dati ottenuti da Internet.
- Il motivo è che oggi un numero sempre crescente di aziende opera anche tramite Internet, presentando e vendendo i propri prodotti attraverso un sito. E’ chiara, quindi, l’importanza che assumono lo studio e la conoscenza del comportamento di un qualsiasi visitatore del sito in questione.

WEB MINING

- Quando si parla di grandi quantità di dati, riferendosi al Web, si intende dire che esistono dei database che ne possono contenere più di un terabyte… anzi ad oggi più di un yottabite.

- 1 terabyte = 1.000 gigabyte = 1.000.000.000.000 byte=10^12
- 1 yottabite=10^24

- Questi dati devono essere elaborati e trasformati dalle aziende interessate per ottenere le informazioni ad esse più utili. In particolare, ciò che vogliono sapere sono le caratteristiche dei consumatori, le loro richieste ed i loro bisogni per poter trasformare in potenziali clienti quelli che ancora non lo sono e per fidelizzare i clienti già esistenti, proponendo loro una sorta di offerta ad hoc. Tale fenomeno è chiamato *Mass Customization.*

WEB MINING

- *Mass Customization.* Fenomeno grazie al quale il venditore è spinto a personalizzare il suo prodotto e il suo messaggio per ogni tipologia di cliente, su larga scala con lo scopo di fidelizzarlo. L'interesse di molte aziende si è quindi spinto più in là della pura e semplice *Mass Customization*, ossia adattamento dell'offerta alle esigenze del consumatore, andando a focalizzarsi sulla
- *Web Personalization*, ossia quell'insieme di azioni che mirano a personalizzare il sito e, di conseguenza, l'offerta a seconda del visitatore-cliente. Gli strumenti principali tramite cui si attua la Web Personalization sono le tecniche di *Data Mining.* Di tali strumenti si avvale il *Web Mining*

Web Mining

- Nel 1996, Oren Etzioni per primo ha introdotto il termine *Web Mining* nelle sue opere e lo ha definito come “l’uso delle tecniche di Data Mining per scoprire ed estrarre automaticamente informazioni dai documenti raccolti nel World Wide Web (Web)” (Etzioni O., 1996).

La struttura del Web Mining

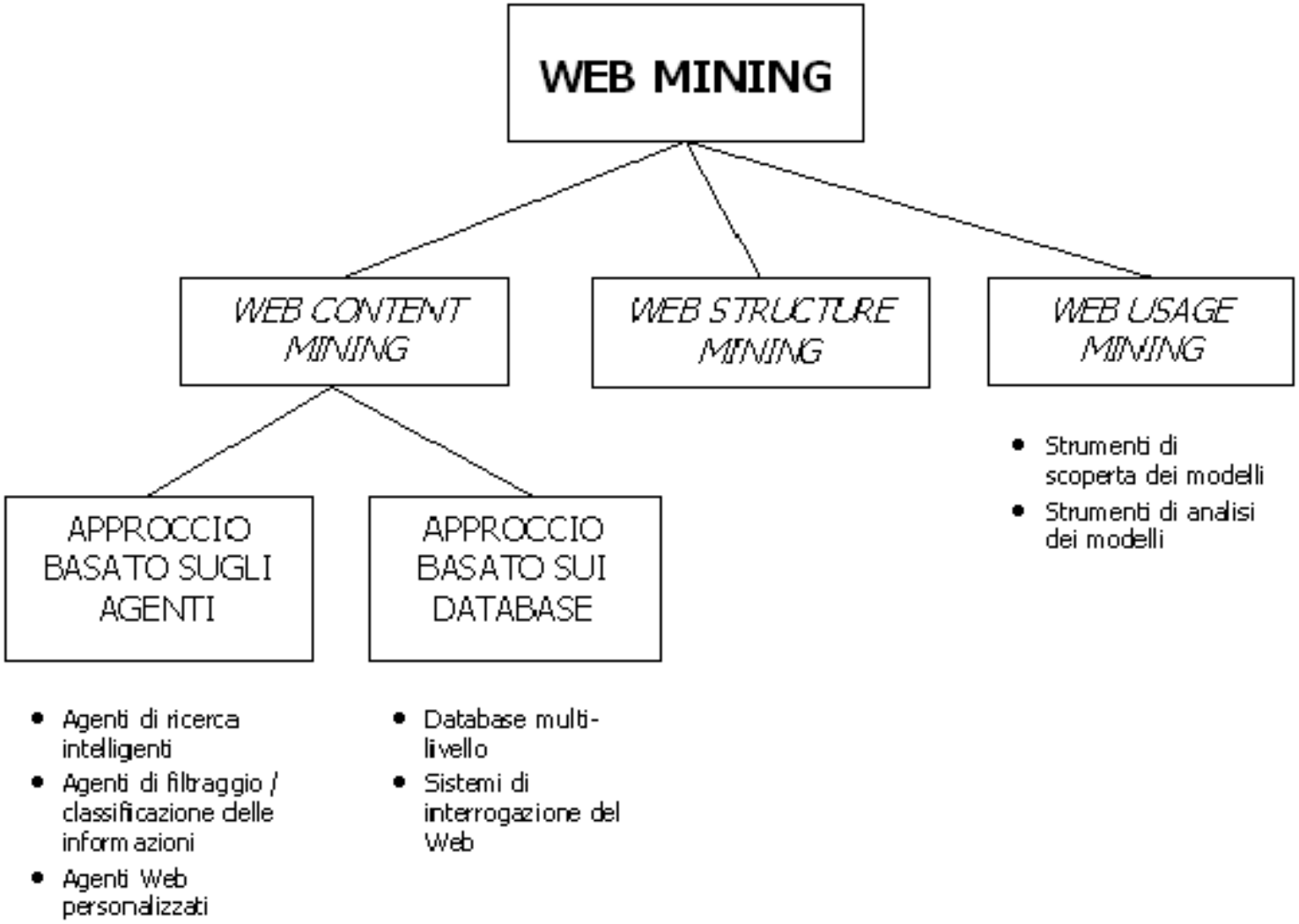
- In genere, si scompone il lavoro di Web Mining in quattro fasi principali:
- 1)Scoperta delle risorse (*Resource finding*).
- E' la fase di ricerca dei documenti Web voluti; indica il processo di raccolta dei dati sia on-line che off-line dalle fonti Web disponibili quali ad esempio *newletters* elettroniche, *newsgroup*, documenti HTML ecc.
- 2)Estrazione delle informazioni (*Information extraction and pre-processing*).
- In cui automaticamente si selezionano e preparano specifiche informazioni dai documenti Web scelti; include qualsiasi processo di trasformazione dei dati originali raccolti nella fase precedente.

- 3)Generalizzazione (*Generalization*).
- Si scoprono, sempre automaticamente, modelli generali e siti Web individuali; qui, in genere, si usano le tecniche di Data Mining e di *Machine Learning* e la componente umana gioca un ruolo importante nelle scelte che andranno a coinvolgere la quarta ed ultima fase.

- 4)Analisi (*Analysis*).
- Si convalidano e/o interpretano i modelli scoperti.

- A seconda dei dati considerati e del tipo di studio da effettuare, può essere classificato in:
  - Web Content Mining;
  - Web Structure Mining;
  - Web Usage Mining.

Web Mining: una classificazione



Web Content Mining

- Processo di ricerca automatica delle fonti di informazioni disponibili on-line e riguardante:
  - l'estrazione del contenuto dei dati del Web (file HTML, immagini, audio, video, ...);
  - la fornitura di strumenti maggiormente abili nel recupero delle informazioni (*Agenti Intelligenti del Web*) e che consentono di ampliare i *Database* e le tecniche di *Data Mining* al fine di offrire un livello di organizzazione più elevato per i dati semi-strutturati o non strutturati disponibili sul Web.
  - Il Text Mining è un esempio di Web Content Mining

Approccio basato sugli Agenti

- Sviluppo di sofisticati sistemi di intelligenza artificiale che possono agire in modo autonomo o semi-autonomo a nome di “utenti” particolari, per scoprire ed organizzare le informazioni derivanti dal Web.
- Tre categorie:
  - *Agenti di ricerca intelligenti.*
  - *Agenti di filtraggio / classificazione delle informazioni.*
  - *Agenti Web personalizzati.*

Cosa sono gli agenti?

SOFTWARE DI “INTELLIGENZA ARTIFICIALE” CHE:

- ***assistono gli utenti nel loro lavoro, o svolgono compiti per conto dell’utente sulla base dei suoi obiettivi, preferenze, criteri comportamentali o decisionali (Feldman and Yu, 1999)***
  - monitoraggio di eventi, situazioni, scenari
  - raccolta di informazione/conoscenza
- ***sono autonomi***
  - possono intraprendere azioni autonome “nell’interesse dell’utente” in relazione agli obiettivi loro assegnati
- ***sono in grado di adattarsi al contesto (Woolridge and Jennings, 1995)***
  - possono apprendere
  - possono reagire al cambiamento del contesto
  - possono “interagire socialmente”
- ***possono cooperare e “interagire socialmente”***
  - con operatori umani
  - con altri agenti software
- **possono aiutare utenti diversi a interagire/cooperare (intermediazione)**

Agenti di ricerca intelligenti

- *sistemi software in computer e/o reti che assistono gli utenti nell'esecuzione di attività legate all'uso delle tecnologie computer-based* (Maes, 2001)
- Hanno sviluppato la ricerca delle informazioni rilevanti utilizzando le caratteristiche di particolari domini ed i profili degli utenti per organizzare ed interpretare le informazioni trovate.

Agenti di filtraggio/classificazione delle informazioni

- *sistemi software in computer e/o reti che assistono gli utenti nell'esecuzione di attività legate all'uso delle tecnologie computer-based* (Maes, 2001)
- Per ritrovare le informazioni, filtrarle e classificarle, utilizzano diverse tecniche di recupero di tali informazioni e le caratteristiche di documenti Web ipertestuali.

Agenti Web personalizzati

- *sistemi software in computer e/o reti che assistono gli utenti nell'esecuzione di attività legate all'uso delle tecnologie computer-based* (Maes, 2001)
- Ottengono o apprendono le preferenze dell' "utente" ed in base a queste scoprono fonti di informazione Web corrispondenti a quelle preferenze e, presumibilmente, a quelle di altri individui con interessi simili.

Perché agenti intelligenti in Internet?

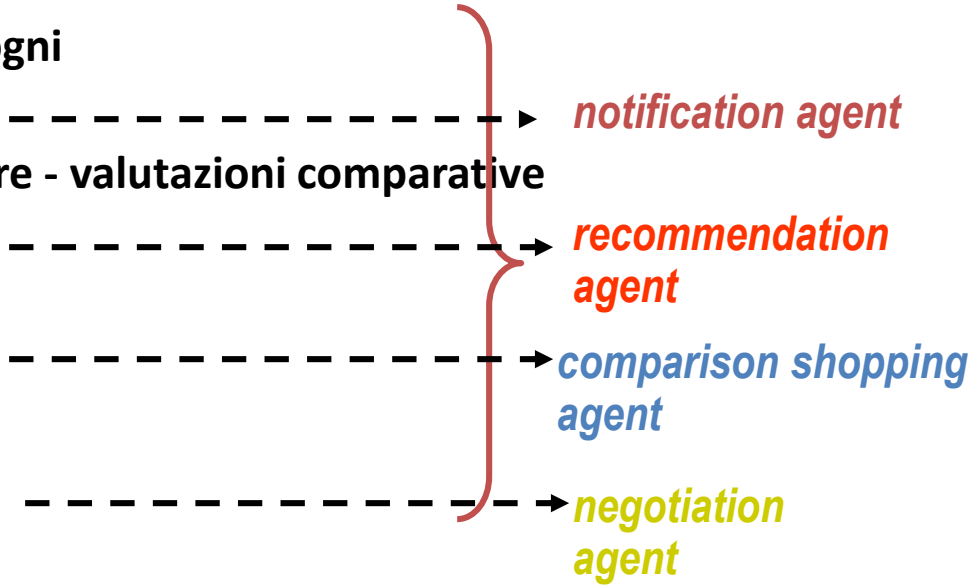
- **caratteristiche di Internet**
  - **dimensione (effetto “overload”)**
  - **complessità (contesto non strutturato)**
  - **dinamicità del contesto**
  - **elevata interazione**
  - **struttura aperta**
- **caratteristiche degli utenti**
  - **sempre più utenti; sempre meno esperti**
- **Internet e commercio elettronico come contesti “virtuali” (contrapposti al mondo fisico)**

Agenti nel Web: esempi di campi applicativi (potenzialmente utili)

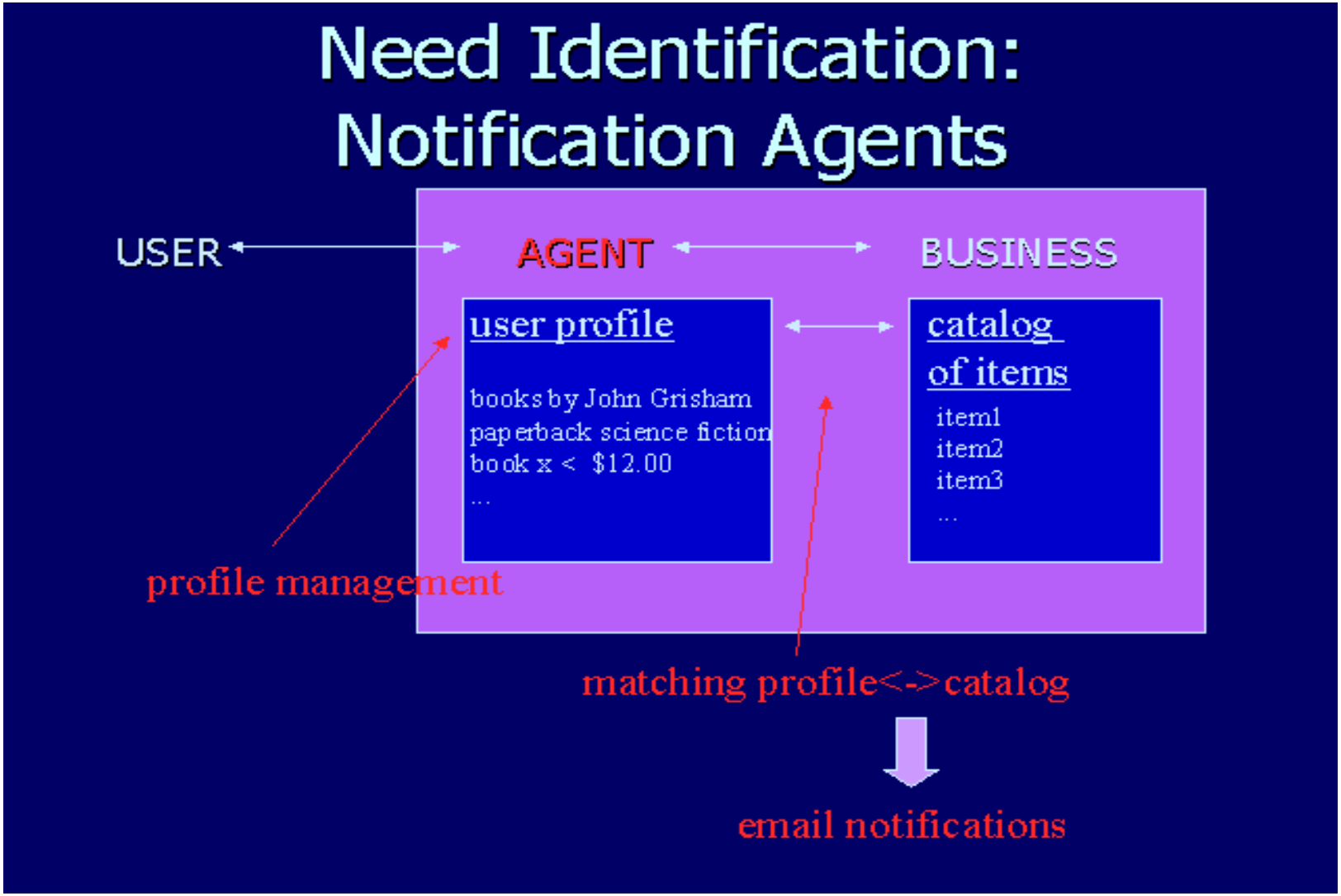
- Agenti di interfaccia (interface agent)
  - per task ripetitivi
  - customizzazione automatica (software, bookmark, indirizzi email, ecc.)
  - gestione sicurezza e accessi in rete
- Agenti di segreteria
  - Gestione comunicazione e contatti
  - schedulatori
  - profilatura del personale
- Intermediari informativi
  - Individuazione nuovi siti e nuove informazioni utili
  - Analisi, confronto, classificazione di documenti e risorse informative
  - Business intelligence (monitoraggio trend economici; analisi della concorrenza)
  - Data mining
- Formazione a distanza
- E-commerce

"mediazione" di agenti intelligenti (?)

- **identificazione ed esplicitazione dei bisogni**
  (need identification)
- **identificazione del prodotto da acquistare - valutazioni comparative**
  (product brokering)
- **identificazione del fornitore**
  (merchant brokering)
- **negoziazione**
  (negotiation)
- **acquisto; acquisizione della consegna**
  (purchase and delivery)
- **servizi post-vendita; valutazione del prodotto**
  (product service and evaluation)



Fasi nell'acquisto di un bene (fonte: Maes, 1999)



ESEMPIO: amazon

(fonte: Maes, 1999)

Approccio basato sui Database

- Concentra generalmente l'attenzione su:
  - tecniche per l'integrazione e l'organizzazione di dati eterogenei e semi-strutturati o non strutturati derivanti dal Web, in raccolte di risorse maggiormente strutturate e di più alto livello (come i *Database relazionali*);
  - tecniche per l'utilizzo di meccanismi standard di interrogazione di *Database* e tecniche di *Data Mining* per avere accesso e poter analizzare tali informazioni.
- Due categorie:
  - *Database multi-livello*.
  - *Sistemi di interrogazione del Web.*

Database multi-livello

- Sono ideati allo scopo di organizzare in maniera più efficiente le informazioni derivanti dal Web.
- Idea principale:
  - il *livello più basso* del Database contiene le informazioni semi-strutturate, memorizzate in diversi "magazzini" Web (come i documenti ipertestuali);
  - ai *livelli più alti*, invece, avviene l'estrazione dai livelli inferiori, dei dati che sono poi organizzati in raccolte strutturate (come i *Database relazionali* o *Object-Oriented*).

Sistemi di interrogazione del Web

- Esistono molti sistemi di interrogazione (*Query*) del Web e diversi linguaggi di recente sviluppo, che cercano di utilizzare:
  - linguaggi di interrogazione standard di Database (come *SQL*);
  - informazioni sulla struttura dei documenti Web;
  - il linguaggio naturale,

  allo scopo di generare una rappresentazione integrata delle informazioni estratte.

Data Mining Multimediale

- Ulteriore tecnica di *Web Content Mining.* Si occupa dell'estrazione di conoscenza ed informazioni di alto livello dalle vaste risorse multimediali on-line.

Web Structure Mining

- Processo di generazione di riassunti o schemi della struttura di un sito e di una pagina Web.
  - I Web Server registrano ed accumulano i dati riguardanti le interazioni degli utenti ogni volta che questi effettuano una richiesta di utilizzo di determinate risorse.
  - Analizzando i Log file di accesso a diversi siti, cerca di capire il comportamento degli utenti e la struttura del Web, allo scopo di migliorarne la progettazione.

Web Usage Mining

- Processo riguardante la scoperta automatica di modelli di accesso degli “utenti” alla rete da uno o più Web Server.
- Si concentra sullo studio di tecniche in grado di prevedere il comportamento degli “utenti” mentre stanno interagendo con il Web.
- Due categorie principali di strumenti e tecniche:
  - strumenti di *scoperta dei modelli* dalle transazioni Web;
  - strumenti di *analisi dei modelli* scoperti.

Scoperta dei modelli dalle transazioni Web

- L'analisi di come gli "utenti" accedono alla rete è un'operazione critica per poter determinare le strategie di Marketing più efficaci ed ottimizzare la struttura logica di un sito Web.
- Due fasi preliminari:
  - *pre-trattamento dei dati*, riguardante la *pulizia dei dati* e l'individuazione delle *transazioni Web*;
  - utilizzo di tecniche necessarie all'*individuazione dei modelli di accesso* derivanti dalle transazioni Web, identificati mediante la fase precedente.

Pagine Web

- La visualizzazione delle schermate, o *pagine*, del Web avviene attraverso l'utilizzo di un programma, il *Browser*, che serve a "sfogliare" le pagine ipertestuali.
- Una raccolta di pagine Web è chiamata *sito*.
- L'esplorazione di Internet ha inizio da un sito Web particolare, detto *home page*, il quale rappresenta il punto di partenza del sito Web stesso.
- Ogni pagina di un sito, compresa quella iniziale, ha un indirizzo univoco detto *Universal Resource Locator* (o semplicemente *URL*).
- Ogni utente che accede alla rete è identificato da un cosiddetto *indirizzo IP*, il quale è rappresentato da una serie di numeri del tipo: *xxx.yyy.zzz.ttt* ,in cui le singole terne possono assumere valori da *0* a *255*. La lunghezza di un indirizzo non è fissa ma può assumere valori che vanno da *1.1.1.1* a *255.255.255.255*

Log File

- I Log file sono semplici file di testo nei quali il Web Server scrive alcune informazioni relative agli accessi e agli errori riscontrati.
- Svolgono un fondamentale ruolo nella valutazione della quantità di accessi da parte degli utenti ad una pagina Web.
- Prima di effettuare una qualsiasi analisi su questo tipo di file, è necessario istruire il Web Server sulla locazione in cui i Log file devono essere conservati.

I dati da analizzare: i log file

130.93.25.19 - - [20/Dec/2000:10:19:44 +0100] "GET /mappa/01.jhtml HTTP/1.0" 200 2472 " -" "Mozilla/4.0"

146.58.31.12 - - [20/Dec/2000:10:19:42+0100] "GET /picts/index_27.gif HTTP/1.0" 200 312 " -" "Mozilla/4.0"

235.58.54.78 - - [20/Dec/2000:10:19:41 +0100] "GET /news/archivio.jhtml HTTP/1.0" 200 115 "-" "Mozilla/4.0"

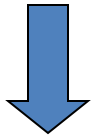
267.12.83.56 - - [20/Dec/2000:10:19:40+0100] "GET /news/01/01/01.jhtml HTTP/1.0" 200 793 "-" "Mozilla/4.0"

187.19.58.26 - - [20/Dec/2000:10:19:37 +0100] "GET /picts/index_21.gif HTTP/1.0" 200 949 "-" "Mozilla/4.0"

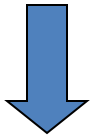
241.27.83.61 - - [20/Dec/2000:10:19:37+0100] "GET /favolando/01.jhtml HTTP/1.0" 200 88 "-" "Mozilla/4.0"

341.25.82.14 - - [20/Dec/2000:10:19:37 +0100] "GET /giochi/01.jhtml HTTP/1.0" 200 656 "-" "Mozilla/4.0"

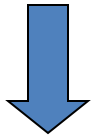
156.12.35.61 - - [20/Dec/2000:10:19:40+0100] "GET /picts/index_26.gif HTTP/1.0" 200 415 "-" "Mozilla/4.0"



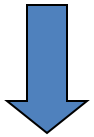
IP address



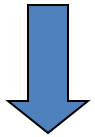
Date and Time



Requested Item



Return code and Bytes



Browser and OS

Pulizia dei dati

- Le tecniche per l'eliminazione degli articoli irrilevanti dai Log file del Server sono particolarmente importanti per qualunque tipo di analisi dei Web Log.
  - Una di queste prevede il controllo dell'estensione dell'*URL* (gif, jpeg, jpg,).
- Un problema collegato: determinare se ci sono stati degli accessi rilevanti che non sono stati registrati nei Log di accesso. I meccanismi come *Proxy Server* e *Local Cache* possono seriamente distorcere la panoramica dei passaggi dei diversi "utenti" attraverso un sito. Per esempio:
  - una pagina elencata una sola volta in un Log di accesso può essere stata in realtà in relazione con molti "utenti";
  - oppure, servirsi del nome di una macchina per identificare gli "utenti" in modo univoco, può condurre a considerare come uno unico diversi utenti erroneamente raggruppati insieme.

Identificazione delle transazioni

- Prima di effettuare una qualsiasi estrazione dai dati di utilizzo del Web, è necessario raggruppare le sequenze dei riferimenti di pagina in unità logiche: *sessioni utente* o *transazioni Web.*
  - Una *sessione utente* è composta da tutti i riferimenti di pagina realizzati da un utente durante una singola visita ad un sito della rete.
  - Una *transazione Web* differisce da una sessione utente sostanzialmente per la dimensione: a seconda dei criteri adottati per identificarla, la sua grandezza può variare da un singolo riferimento di pagina a tutti i riferimenti di pagina contenuti in una *sessione utente*.

Tecniche di scoperta dalle transazioni

- *Analisi dei percorsi*.
  - Nell'ambito del *Web Usage Mining*, si possono utilizzare i *grafi* per rappresentare le relazioni definite sulle pagine Web.
  - Esistono molti tipi di grafi che possono essere disegnati per eseguire un'analisi dei percorsi.
  - Il principale ed il più logico, rappresenta il *layout fisico* di un sito Web, dove: (social network analysis)
    - ⇨ i *nodi* indicano le pagine Web;
    - ⇨ le *linee* indicano i collegamenti ipertestuali tra le pagine stesse.

Tecniche di scoperta dalle transazioni (2)

- *Regole di associazione.*
  - Scoprire le regole di associazione esistenti tra i dati equivale a scoprire le correlazioni esistenti tra i riferimenti ai vari file sul Server a disposizione di un dato "utente".
  - In aggiunta, possono dare un'indicazione di come organizzare al meglio lo spazio Web a disposizione delle Società che operano sul Web.

Tecniche di scoperta dalle transazioni (3)

- *Modelli sequenziali.*
  - Si tratta di trovare i *modelli di inter-transazioni*: la presenza di un insieme di articoli è seguita da un altro articolo nell'insieme di transazioni ordinate cronologicamente.
  - Nei Log del Web Server riguardanti le transazioni, una visita effettuata da un "utente" viene registrata solo se supera un dato periodo di tempo.
  - La scoperta di modelli sequenziali nei Log di accesso del Web Server conduce le società che sviluppano la propria attività sul Web a predire i modelli delle visite degli "utenti" e le aiuta nello studio di obiettivi pubblicitari mirati a gruppi di "utenti", sulla base di tali modelli.

Tecniche di scoperta dalle transazioni

- *Clustering e classificazione.*
  - Le tecniche di *classificazione* consentono lo sviluppo di un profilo di tutti quegli "utenti" che accedono a particolari file del Server:
    - ⇨ sulla base delle informazioni demografiche disponibili su di essi,
    - ⇨ oppure tenendo conto dei loro modelli di accesso.
  - L'analisi dei *Cluster* consente di raggruppare insieme i dati che hanno caratteristiche simili. Il *Clustering* delle informazioni sugli "utenti" o dei dati sui Log di transazioni Web, può facilitare lo sviluppo e l'esecuzione di future strategie di Marketing sia on-line che off-line.

Analisi dei modelli scoperti

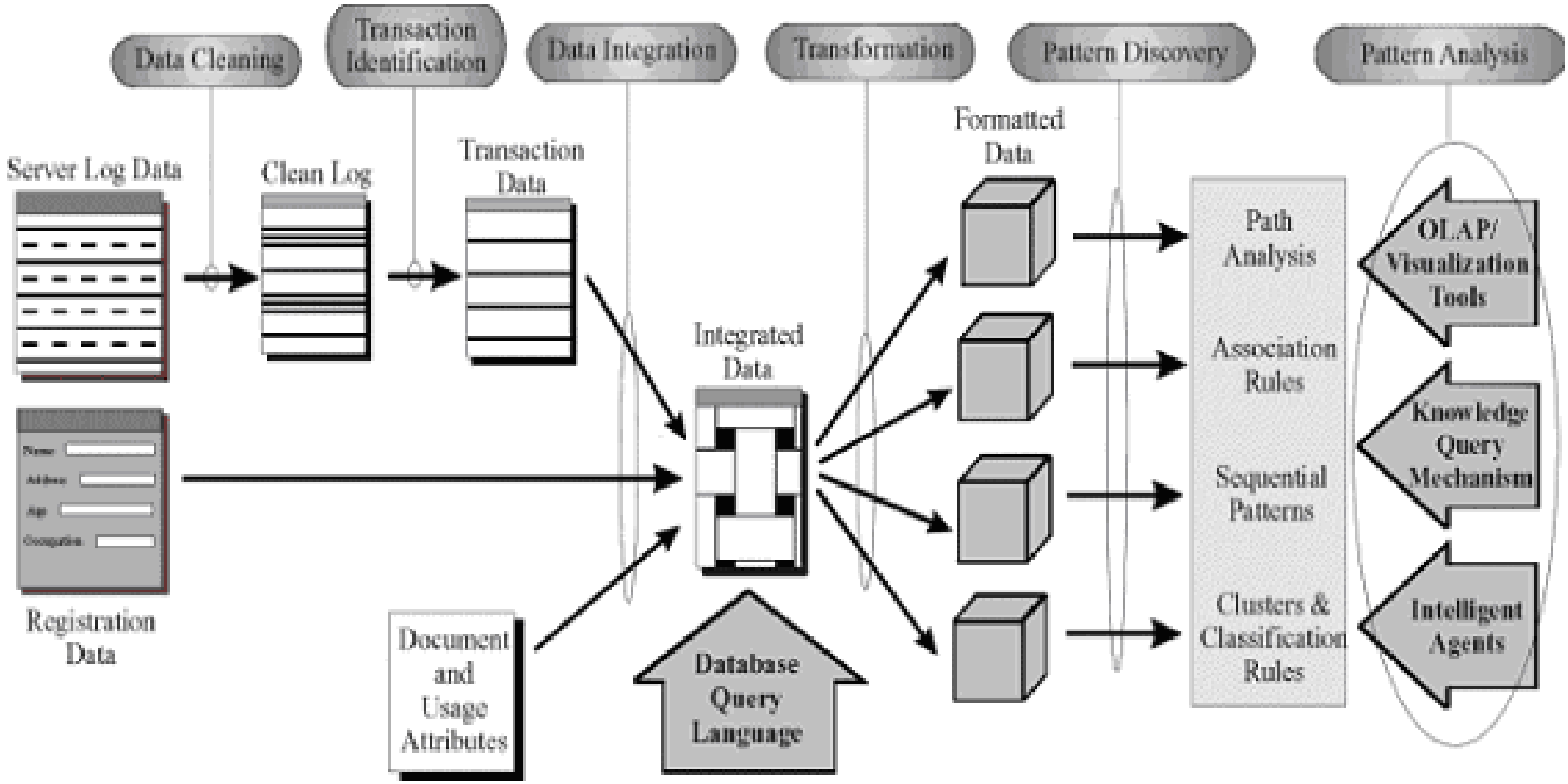
- Una volta scoperti i modelli di utilizzo del Web, è necessario analizzarli tramite meccanismi e strumenti utili per una loro migliore comprensione.
- Tali strumenti devono essere in grado di:
  - effettuare “estrazioni” da un certo numero di campi, concernenti statistiche;
  - disegnare grafici e visualizzazioni;
  - produrre analisi di utilizzabilità;
  - eseguire interrogazioni di Database.

Architettura di Web Usage Mining

- Il processo è diviso in due parti principali:
  - La prima parte comprende i processi che dipendono dal dominio di trasformazione dei dati del Web in moduli di transazioni adatti. Tali processi riguardano:
    - ⇨ il pre-trattamento dei dati;
    - ⇨ l'identificazione delle transazioni;
    - ⇨ le componenti di identificazione dei dati.
  - La seconda parte comprende la più ampia applicazione che non dipende dal dominio di Data Mining generico e le tecniche di corrispondenza dei modelli (come la scoperta di regole di associazione e modelli sequenziali).

Architettura di Web Usage Mining

Cooley et al (1999)



WEB 2 WEB 3 Web 4

- Web 2.0 is the term given to describe a second generation of the World Wide Web that is focused on the ability for people to collaborate and share information online.
- Web 2.0 basically refers to the transition from static HTML Web page to a more dynamic Web that is more organized
- and is based on serving Web applications to users.

- Other improved functionality of Web 2.0 includes open communication with an emphasis on Web-based communities of users, and more open sharing of information.

- Over time Web 2.0 has been used more as a marketing term than a computer-science-based term. Blogs, wikis, and Web services are all seen as components of Web 2.0.

- Blogs

  Short for *Web log*, a blog is a Web page that serves as a publicly accessible personal journal for an individual. Typically updated daily, blogs often reflect the personality of the author.

- wikis

  A **wiki** is a website which allows its users to add, modify, or delete its content via a web browser usually using a simplified markup language or a rich-text editor.Wikis are powered by wiki software. Most are created collaboratively.
- Wikis serve many different purposes, such as knowledge management and notetaking. Wikis can be community websites and intranets, for example. Some permit control over different functions (levels of access). For example, editing rights may permit changing, adding or removing material. Others may permit access without enforcing access control. Other rules may also be imposed to organize content.
- Ward Cunningham, the developer of the first wiki software, WikiWikiWeb, originally described it as "the simplest online database that could possibly work"

WEB 3

- Tim O'Reilly definisce il Web 3.0 ‘The term is used to describe the evolution of the Web as an extension of Web 2.0’.
- Nova Spivack definisce il Web 3.0 ‘as connective intelligence; connecting data, concepts, applications and ultimately people’.(Parola d’ordine connessione e connesso) (Web semantico, Intelligenza artificiale, web potenziato, fusione dei poli, web come database)

OSS: ‘While some call the The Semantic Web(An extension of the current Web that provides an easier way to find, share, reuse and combine information. It is based on machine-readable information ,'Web 3.0', Spivack's opinion is that The Semantic Web is just one of several converging technologies and trends that will define Web 3.0.’

- WEB 4
- Si parla del Web 4.0
- Ruolo fondamentale la realtà aumentata e i Big Data
- Si pensa che ogni individuo può avere un alter ego digitale
- Si pensa che ogni individuo dialoga con le nuove interfaccie come la domotica e le macchine intelligenti
- Ci sarà un sempre maggiore controllo delle informazioni