

ht://Miner

Un sistema open-source di data mining e data warehousing per lo studio dei comportamenti degli utenti su Internet

Gabriele Bartolini
Comune di Prato
Piazza del Comune, 2
59100 Prato PO - Italia
+39 0574 1835214

g.bartolini@comune.prato.it

SOMMARIO

Conoscere gli utenti che navigano una rete civica è indispensabile per prendere decisioni e migliorare il servizio offerto all'utenza cittadina. Il progetto **ht://Miner**, totalmente sviluppato dal Comune di Prato all'interno della Rete Civica provinciale pratese *Po-Net* e **distribuito liberamente secondo la licenza GNU GPL**, si inquadra in questo contesto. L'ambizioso progetto prevede la creazione di un sistema completo di **web usage mining** e di **data warehousing** che, partendo dai semplici access log prodotti in modo automatico dai vari server web, permetta la scoperta di informazioni e la memorizzazione delle stesse in un archivio PostgreSQL [10] organizzato su due livelli: un database operativo ed un data warehouse. Il progetto ht://Miner è una suite di applicazioni scritte in C++, Perl e PHP che intervengono ai vari stadi del processo di scoperta di conoscenza, tipico dei sistemi di **KDD** (*knowledge discovery from data*) e di **data mining**. Il software utilizza in modo prevalente **tecnologia open-source**, in particolare sistemi operativi (GNU/Linux o FreeBSD), DBMS relazionali (PostgreSQL), linguaggi di programmazione (PHP, Perl), strumenti di sviluppo GNU (in particolare GCC, autotools e Gettext). Il progetto ht://Miner è l'ultimo tassello dell'esperienza pluriennale del Comune di Prato nello **sviluppo attivo** di software open-source. Sulla base delle direttive nazionali, regionali e comunali in materia di diffusione del software libero, uno degli obiettivi principali dell'ente è quello di promuovere il **riuso** delle tecnologie e degli applicativi con le altre realtà della Pubblica Amministrazione italiana, mettendo a disposizione l'esperienza maturata nel settore.

Categorie e descrittori dell'argomento

H.4.2 [Information Systems Applications]: Types of Systems – *Decision support*; H.2.8 [Database Management]: Database Applications – *Data mining*; H.2.7 [Database Management]: Database Applications – *Data warehouse and repository*; H.3 [Information

Systems]: Information storage and retrieval; J.4 [Social and behavioural sciences]: Sociology

Termini generali (ACM)

Algorithms, Management, Experimentation

Parole chiave

Web usage mining, data webhouse, web usage data warehouse, web log analysis, decision support system, open-source

1. INTRODUZIONE

Il software ht://Miner [1] rappresenta un esempio di applicazione verticale open-source interamente sviluppata dal Comune di Prato a partire da un progetto sperimentale intrapreso nel 2003. Il progetto si inquadra nel contesto della rete civica provinciale pratese *Po-Net* [2], una delle realtà più complesse, eterogenee e al contempo avanzate della pubblica amministrazione italiana su Internet: vi partecipano tutti i Comuni del territorio pratese, la Provincia, la Camera di Commercio, l'Azienda Sanitaria Locale, le aziende a partecipazione pubblica e poi biblioteche, musei, istituzioni culturali, scuole e associazioni. I gruppi di lavoro e le redazioni di Po-Net sono circa 140. Gli accessi alle pagine della rete civica registrati nel 2006 sono stati circa 37 milioni.

2. COMUNE DI PRATO E SVILUPPO DI SOFTWARE LIBERO

ht://Miner non rappresenta la prima esperienza del Comune di Prato con lo sviluppo attivo di software libero open-source (OSS). Già alla fine degli anni 90 il Comune di Prato decise di utilizzare il motore di ricerca ht://Dig [9] diventandone poi uno dei più attivi sviluppatori. Tuttavia, la rete civica basa le sue funzioni di ricerca su ht://Dig, soluzione che negli anni, a fronte di un investimento iniziale in termini di risorse umane, si è dimostrata vincente, e non solo dal punto di vista economico. Il coinvolgimento diretto del personale interno nel gruppo di sviluppo ha da un lato protetto

l'ente da modifiche del prodotto, e dall'altro ne ha notevolmente migliorato il *know-how* e le competenze tecniche. Soprattutto quest'ultimo aspetto ha permesso lo sviluppo di uno dei più utilizzati applicativi open-source per il controllo dei link: [ht://Check](http://Check) [3]. Il modello open-source adottato anche per questo applicativo ha portato notevoli vantaggi al Comune di Prato, soprattutto in termini di stabilità e robustezza del software dovuti al *testing diffuso* dell'applicazione con indicazioni volontarie di malfunzionamenti (*bug*) e suggerimenti da parte di utenti sparsi in tutto il mondo, con architetture hardware eterogenee ed in contesti diversi. Tali aspetti positivi non potevano essere ignorati al momento di *concepire e pianificare* lo sviluppo di una applicazione che permettesse di memorizzare gli accessi alle risorse Internet nel tempo e soprattutto di eseguire interrogazioni al volo. E' così che è nato il progetto per un sistema di data mining e data warehousing applicato a informazioni di utilizzo del web (*web usage*): [ht://Miner](http://Miner).

3. WEB USAGE MINING

Web usage mining (WUM) è un ramo specifico del Web mining, termine che fu definito per la prima volta nel 1996 da Etzioni [4] come "l'utilizzo di tecniche di data mining per scoprire automaticamente ed estrarre informazioni da documenti e servizi sul World Wide Web".

3.1 Data mining e KDD

Il data mining è l'applicazione, su grosse moli di dati, di tecniche e algoritmi per la scoperta di conoscenza. Il termine *mining* vuole enfatizzare il procedimento di estrazione di *conoscenze nascoste* all'interno dei dati. L'attività di data mining si inquadra in un contesto più ampio, quello dei **sistemi di supporto alle decisioni** (DSS) ed in particolare all'interno della disciplina nota come *Knowledge Discovery from Data* (KDD).

3.2 Tassonomia di Web mining

L'applicazione di tecniche di data mining al mondo del web viene pertanto definito *web mining*. Sulla base della tipologia dei dati presi in esame [5], si parla di:

- web structure data: struttura del web (relazioni fra pagine);
- web content data: contenuto del web (testo in particolare);
- web usage data: utilizzo del web da parte degli utenti.

Per web usage mining quindi si intende "un particolare processo di scoperta e analisi di modelli (*pattern*) che concentra l'attenzione sui dati relativi agli accessi effettuati dagli utenti (*Web usage data*)" (Figura 1). [ht://Miner](http://Miner) si inserisce in questo scenario.

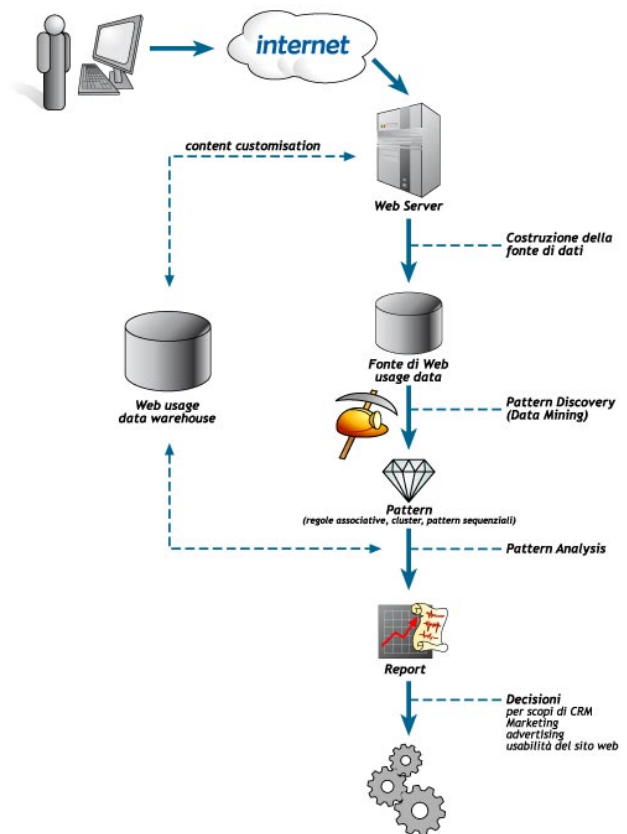


Figura 1. Panoramica sulle attività di WUM

4. STRUMENTI OS DI LOG ANALYSYS

La comunità open-source è in grado di fornire ottimi prodotti per il web usage mining e per la log analysis di dati provenienti da server HTTP. In particolare si distinguono: *AWStats*, *Webalizer*, *Analog*. Quest'ultimo ad esempio è stato utilizzato dalla rete civica Po-Net fino al 31 dicembre 2006 e per un arco di tempo quasi decennale. I prodotti succitati presentano caratteristiche simili: sono altamente configurabili e ricchi di report statistici prettamente descrittivi (*hit*, *visite*, *dettagli per browser*, *sistema operativo*, *ecc.*). Inoltre grazie all'esteso e diffuso utilizzo (e.g. ad oggi i sorgenti di AWStats sono stati scaricati da Sourceforge.net circa 1,4 milioni di volte), tali applicativi software si dimostrano molto robusti e **stabili** ed il supporto è notevolmente efficiente. Tuttavia ognuno di essi si limita a produrre report statici, ovvero istantanee ad un certo momento che non possono essere modificate, filtrate, scomposte a proprio piacimento se non attraverso una successiva rielaborazione dei dati. Nessuno di essi è inoltre in grado di identificare automaticamente gli *spider* dei motori di ricerca sulla base delle caratteristiche di navigazione; l'**integrazione** con fonti di dati ausiliarie a disposizione è limitata. Ad ogni modo, il limite maggiore è l'assenza di una **base di dati relazionale pluriennale modellata** sulle esigenze di *business* di un'organizzazione ed in grado di costituire il punto di partenza per applicazioni di data mining e di analisi. E' con questi presupposti funzionali che nasce l'idea di sviluppare [ht://Miner](http://Miner).

5. OBIETTIVI DI HT://MINER

Utilizzando l'esperienza maturata sia nel campo dello sviluppo di software libero che nell'utilizzo di software di log analysis, l'elevata specializzazione di natura accademica nel campo del Web usage mining [6,7,8] e constatando l'assenza di un prodotto open-source in grado di far fronte alle esigenze della realtà eterogenea della rete civica Po-Net, il Comune di Prato decide di investire risorse umane e tecnologiche nello sviluppo di ht://Miner. Nel 2003 investe in un progetto pilota dagli esiti incoraggianti e ne riprende lo sviluppo nel luglio 2006. Gli obiettivi funzionali principali sono:

- memorizzazione automatizzata dei dati in un database relazionale in grado di permettere:
 - disponibilità di rete via TCP/IP;
 - possibilità di interrogazione tramite SQL;
 - integrazione con altri sistemi, fonti di dati, applicazioni e interfacce (e.g. ODBC);
- flessibilità e scalabilità del sistema tramite un'architettura di dati su due livelli (stage):
 - transazionale*: dati al dettaglio, archivio normalizzato;
 - warehouse*: dati aggregati e storici, archivio denormalizzato e ridondante;
- creazione di un data warehouse per il supporto alle decisioni con garanzia e tutela della privacy dell'utente;
- predisposizione al data mining, in particolare per la scoperta di regole associative (*market basket analysis*) e lo studio dei percorsi di navigazione;
- predisposizione alla personalizzazione di massa (e.g. riuso delle informazioni di utilizzo per offrire suggerimenti agli utenti sulla base delle esperienze che i visitatori hanno fatto nel passato);
- personalizzazione del formato di input dei log (tramite espressioni regolari per Apache e automatica per IIS);
- individuazione automatica dei visitatori unici;
- individuazione automatica delle *sessioni* (visite): gruppi di richieste da parte di uno stesso visitatore con periodo di inattività inferiore a 30 minuti;
- individuazione automatica (tramite euristiche) delle transazioni e del tempo speso nella consultazione di ogni pagina da parte di un utente;
- classificazione delle richieste sulla base del tempo passato su una risorsa o del percorso seguito in:

- richieste di contenuto;
- richieste di navigazione;

- rilevazione sia supervisionata che automatica degli *spider* di indicizzazione dei motori di ricerca;
- supporto per la localizzazione degli indirizzi IP tramite libreria open-source GeoIP [6];
- classificazione degli *user agent* e dei principali sistemi operativi;
- classificazione delle URL in strutture gerarchiche organizzate a *categorie* (definite in XML);
- creazione di un framework di astrazione in PHP per la scrittura rapida di applicazioni di interrogazione online via web.

6. ARCHITETTURA DI HT://MINER

Il sistema ht://Miner è basato su una architettura modulare a stack basata su 5 livelli principali, ognuno dei quali si identifica con una particolare fase del processo di scoperta di conoscenza (KDD): pre-processing, processing, data warehouse, analisi (data mining, web interface e report) e personalizzazione di massa (Figura 2).

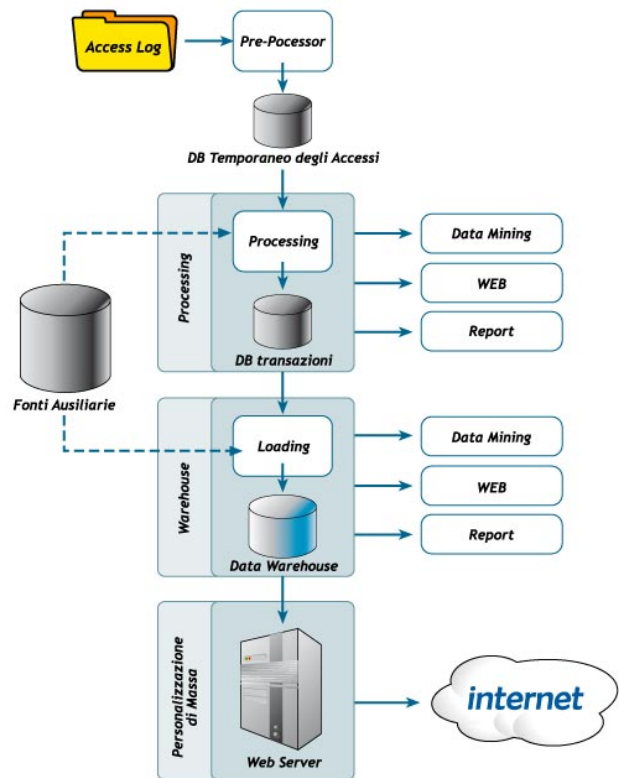


Figura 2. Architettura di ht://Miner

6.1 Pre-Processing

La fase di pre-processing è la prima fase del processo di KDD implementato in ht://Miner ed è quella che si occupa di:

1. creare l'archivio temporaneo;
2. caricare i file di access log effettuando operazioni di preparazione (controllo del formato di log) e di filtraggio dei dati;
3. risolvere gli indirizzi IP in nomi di host.

6.2 Processing

La fase di processing è la seconda fase del processo di KDD implementato in ht://Miner ed è quella che si occupa di:

1. organizzare i dati provenienti dal database temporaneo, individuando visitatori unici e sessioni;
2. trasformare i dati:
 - a. identificando le transazioni (particolari sottoinsiemi delle sessioni);
 - b. rilevando gli spider utilizzati dai motori di ricerca per l'indicizzazione;
 - c. classificando gli user agent (browser) secondo il modello e il sistema operativo;
 - d. localizzando le richieste sulla base dell'indirizzo IP (tramite libreria GeoIP) con precisione fino alla città;
3. memorizzare i dati nell'archivio delle transazioni.

6.3 Data Warehouse

Al fine di permettere un migliore e più efficiente recupero delle informazioni aggregate e di permettere la storicizzazione dei dati, ht://Miner prevede la gestione di un data warehouse. Il data warehouse di ht://Miner è organizzato su due livelli.

6.3.1 Data warehouse di primo livello:

Il data warehouse di primo livello presenta le caratteristiche seguenti:

1. è aggiornato tramite riorganizzazione dei dati presenti nel database delle transazioni, opportunamente aggregati (per giorno e per mese);
2. è strutturato secondo schema a stella tipico del data warehouse;
3. costituisce la base per successive elaborazioni;
4. attualmente memorizzai i *soggetti*: richieste di pagine e errori prodotti dai server web.

6.3.2 Data warehouse personalizzati (custom) o di secondo livello

Il data warehouse di secondo livello ha, seppur in tutta la sua semplicità, un ruolo fondamentale nell'architettura di ht://Miner, in quanto garantisce scalabilità e flessibilità necessarie all'attività di modellazione degli archivi personalizzati, costruiti sulla base:

1. del data warehouse di primo livello;

2. di fonti ausiliarie (e.g. struttura e contenuto).

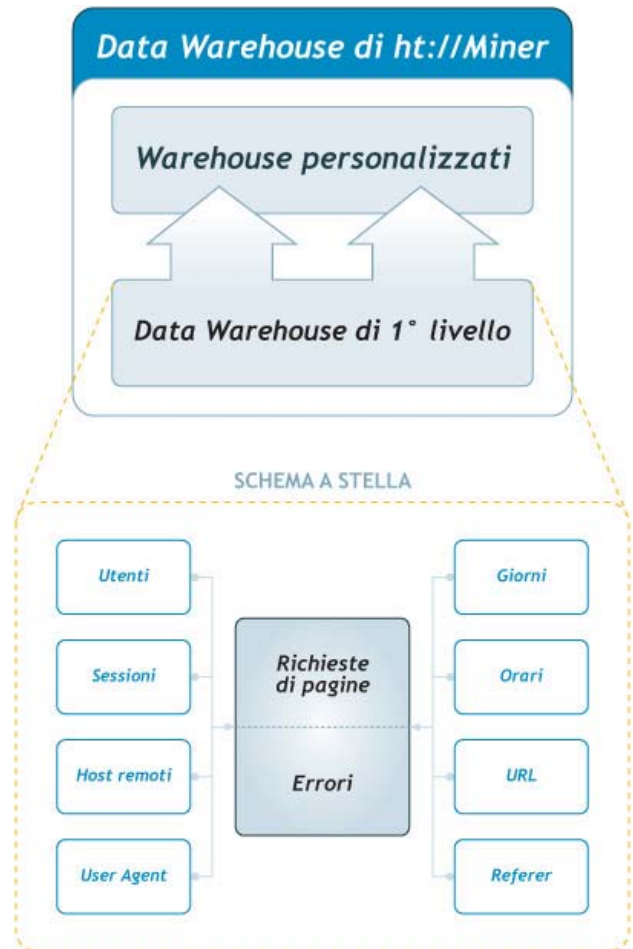


Figura 3. Panoramica del data warehouse

Al momento, ht://Miner è in grado di costituire un data warehouse personalizzato di secondo livello per le richieste anonime effettuate a pagine (esclusi gli accessi degli spider dei motori di ricerca). Per migliorare il grado di efficacia e di utilità delle analisi, ht://Miner mette a disposizione un **potente sistema per la classificazione delle URL in gerarchie di categorie a profondità variabile**. Tale fonte ausiliaria contiene informazioni circa la struttura del web di una certa organizzazione: ad esempio, la rete civica Po-Net attualmente utilizza una struttura gerarchica mista (organizzativa e tematica) di circa 750 categorie organizzate su 5 livelli di profondità a partire da quasi 4 milioni di differenti URL richieste. Dal punto di vista tecnico, la struttura è definita utilizzando un file XML e specificando regole di associazione fra URL e categorie tramite espressioni regolari.

Per sua natura, il data warehouse è una struttura di dati in continua evoluzione e modellazione; l'architettura a livelli di ht://Miner garantisce piena autonomia, indipendenza e quindi scalabilità del warehouse di secondo livello rispetto al resto del sistema.

6.4 Analisi

Il modulo di analisi è senza dubbio l'area in cui ht://Miner presenta maggiori possibilità di sviluppo futuro. L'obiettivo del modulo di analisi è quello di

utilizzare le basi di dati (sia transazionale che il data warehouse) di ht://Miner per scoprire nuove informazioni o per rappresentarle in modo da fornire supporto alle decisioni. Il modulo di analisi è composto da:

- modulo di data mining;
- modulo di reporting online (interfaccia web);
- modulo di reporting standard (procedure batch, strumenti OLAP, ODBC, ecc.).

La versione 1.0 di ht://Miner si prefigge di supportare il modulo di online reporting tramite la costituzione di una libreria middleware in PHP per l'interrogazione del data warehouse di secondo livello (Figura 4). Il framework in PHP è in grado di esportare i dati in XML e di creare trasformazioni al volo in documenti XHTML tramite fogli di stile XSL.

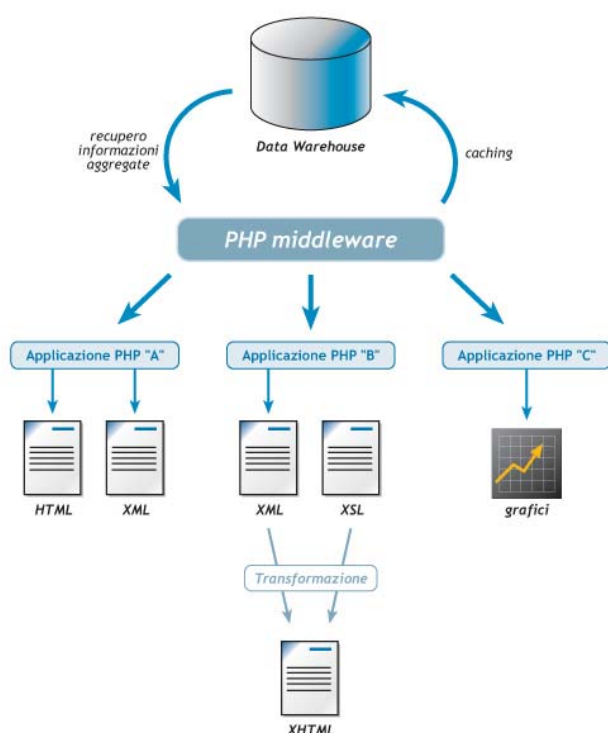


Figura 4. Libreria middleware in PHP

6.5 Personalizzazione di Massa

Il modulo di personalizzazione di massa rappresenta l'ultimo livello di completezza del sistema. L'obiettivo è quello di usufruire delle informazioni scoperte riguardanti l'utilizzo di Internet da parte degli utenti per fornire contenuti dinamici (e.g. sottoforma di suggerimenti stile *Amazon*).

7. SVILUPPO DI HT://MINER

Alla data di scrittura del presente documento, la release 1.0 di ht://Miner è ancora in fase di sviluppo. Tale *milestone* si prefigge di raggiungere **stabilità** in merito alle funzionalità attualmente già implementate e funzionanti di: pre-processing, processing, data warehouse di primo livello, parte del data warehouse personalizzato (richieste anonime di pagine e categorizzazione delle URL) e infine libreria *middleware*

in PHP. Lo sviluppo degli altri moduli, in particolare quello di data mining, unitamente ad un'attività di *refactoring* (soprattutto del processore di ht://Miner) è rimandato alla versione 2.x. Interessanti evoluzioni riguardano l'applicazione di algoritmi per la scoperta di regole associative [6] e per l'analisi dei percorsi di navigazione; la memorizzazione nel warehouse di secondo livello di altri *soggetti* descrittivi come: sessioni, categorie di ingresso e di uscita, ecc.; l'introduzione di librerie C++ stabili e portabili per la gestione della concorrenza (in particolare ACE framework); la generazione di grafici al volo utilizzando librerie per le immagini *raster* o vettoriali (e.g. SVG¹); l'inserimento del modulo di middleware all'interno della libreria globale PEAR² di PHP.

Il processo di *software development* è avvenuto grazie anche all'infrastruttura messa a disposizione gratuitamente da Sourceforge.net per lo sviluppo di applicazioni open-source. I sorgenti sono liberamente scaricabili dal sito ufficiale del progetto. La licenza GNU GPL pone vincoli (legali ma soprattutto morali) per coloro che intendono modificare ht://Miner e ridistribuirlo; nel rispetto dei principi etici del software libero è auspicabile che ogni modifica sia concordata con i responsabili del progetto, in modo da includere i contributi nel codice sorgente principale.

8. REQUISITI TECNICI

8.1 Hardware

Non esistono particolari necessità in termini di hardware. Trattandosi di un sistema per data warehouse, le esigenze in termini di spazio possono variare da situazione a situazione. Ad esempio, il sistema ht://Miner all'interno della Rete Civica Po-Net è installato su un server con le seguenti caratteristiche: Dell PowerEdge 1850 con 2 CPU Intel Xeon 3.00GHz e 4GB RAM; canale in fibra ottica SAN Emulex LP8000 connesso a una piattaforma CX500 (1Gbit); 200GB di spazio su disco allocato.

8.2 Software

Come detto in precedenza, ht://Miner è basato su tecnologia software open-source, in particolare:

- sistema operativo GNU/Linux con kernel 2.4 o 2.6 (preferibile per miglior supporto ai thread);
- compilatore C++ di GNU;
- PostgreSQL 8.x [10];
- librerie: GNU/Gettext, PCRE³, zlib, OpenSSL, eXpat, GeoIP di Maxmind [11];
- PHP 5;
- Perl.

¹ Scalable Vector Graphics

² PHP Extension and Application Repository

³ Perl Compatible Regular Expressions

9. PROPOSTE DI RIUSO E POSSIBILI COLLABORAZIONI

Uno degli obiettivi del Comune di Prato è quello di trovare *partner* per la costituzione di una comunità di sviluppatori di *ht://Miner*, che ne curino l'evoluzione rispettando i principi e le pratiche del modello open-source. Gli strumenti per il *collaborative development* sono già a disposizione (il software è ospitato su SourceForge) e il Comune di Prato si propone come coordinatore per lo sviluppo.

Altre forme percorribili di collaborazione fra Comune di Prato e pubblica amministrazione riguardano:

- il riuso, sotto l'assistenza del Comune di Prato, della tecnologia open-source *ht://Miner* da parte di altri enti della pubblica amministrazione italiana con servizi Internet per il cittadino;
- la fornitura di servizio (ASP) da parte del Comune di Prato per gli enti che non sono in grado di gestire direttamente e con le proprie risorse il sistema *ht://Miner*.

Tuttavia tali soluzioni devono essere analizzate singolarmente, descritte in modo dettagliato e ufficializzate tramite convenzioni che ne definiscano tempi e modalità.

10. CONCLUSIONI

Il sistema *ht://Miner* è stato ufficialmente adottato dalla rete civica Po-Net a partire dal 1 gennaio 2007 [12]. Dopo due mesi di utilizzo, i risultati sono molto soddisfacenti ed oltremodo incoraggianti. Con una media giornaliera nel primo bimestre 2007 di circa 1,8 milioni di richieste esaminate e circa 650 mila richieste accettate, il sistema è in grado di terminare le elaborazioni aggregate relative al giorno precedente in meno di tre ore. Il database transazionale contiene oltre 50 milioni di richieste complessive a partire dal 1 dicembre 2006 (circa 30GB), mentre il data warehouse, che per motivi di performance è organizzato su più livelli gerarchici di aggregazione, occupa approssimativamente 13GB di spazio su disco al mese.

Dettagli per le 750 categorie sono disponibili per giorno (con riepilogo orario), mese (con riepilogo per giorno della settimana) ed anno per quanto concerne: provenienza geografica (precisione fino alla città), browser e famiglia di browser utilizzati, sistema operativo. Le *metriche* al momento memorizzate sono le richieste di pagine, i byte trasferiti e i tempi di consultazione.

Si consideri che lo spazio richiesto (non indifferente per realtà piccole, nonostante il decrescente costo dei dispositivi di memoria di massa) è influenzato in modo determinante dalla complessità della struttura gerarchica a categorie ed in particolare dalla sua profondità. Utilizzando strutture gerarchiche più semplici si possono ridurre notevolmente i requisiti di spazio su disco (anche del 90% per gerarchie ad un solo livello).

In ogni caso, il punto forte del sistema è senza dubbio la **libertà del codice**. Il modello open-source offre infatti la possibilità a chiunque lo desideri di capirne il funzionamento interno e di adattarlo alle proprie esigenze di *business* e *tecnologiche*.

11. RIFERIMENTI

- [1] Po-Net. *ht://Miner Web Usage Mining system*, <http://www.htminer.it/>
- [2] Po-Net. *Rete Civica Provinciale Pratese*, <http://www.po-net.prato.it/>
- [3] Comune di Prato. *ht://Check, more than a link checker*, <http://htcheck.sf.net/>
- [4] Oren Etzioni, *The World-Wide Web: Quagmire or Gold Mine?*. Communications of the ACM, Volume 39, number 11, pages 65-68, 1996.
- [5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. *Grouping web page references into transactions for mining world wide web browsing patterns*. Technical Report TR 97-021, Dept. of Computer Science, Univ. of Minnesota, Minneapolis, USA, June 1997.
- [6] Gabriele Bartolini. *Web usage mining and discovery of association rules from HTTP server logs*. Data mining class: Research Paper. Department of Computer Sciences, Monash University, Melbourne, Australia. 2001. <http://www.prato.linux.it/~gbartolini/en/view-a/2/pdf/wum.pdf>
- [7] Gabriele Bartolini. *"Web usage mining": tecniche di data mining applicate al world wide web. Lo stato dell'arte*. Appendice a tesi Corso di Laurea in Statistica e Sistemi Informativi, AA 2001/02. Dipartimento di statistica "G. Parenti", Università degli Studi di Firenze, Facoltà di Economia.
- [8] Gabriele Bartolini. *"Web usage mining": costituzione di una fonte di dati per la valutazione del comportamento degli utenti di risorse di un server web*. Tesi Corso di Laurea in Statistica e Sistemi Informativi, AA 2001/02. Relatore Prof.ssa Cristina Martelli. Dipartimento di statistica "G. Parenti", Università degli Studi di Firenze, Facoltà di Economia.
- [9] ht://Dig Group. *ht://Dig Search Engine*, <http://www.htdig.org/>
- [10] PostgreSQL Global Development Group. *PostgreSQL, the world's most advanced open source database*. <http://www.postgresql.org/>
- [11] Maxmind. *GeoIP, IP Address Location Technology*. <http://www.maxmind.com/app/ip-location>
- [12] Po-net. *Statistiche degli accessi alla Rete Civica Po-Net (2007)*. <http://statistiche.po-net.prato.it/2007/>