インターネット計測とデータ解析 第10回

長 健二朗

2010年12月15日

前回のおさらい

インターネットの異常や問題を計る

- 異常検出
- スパム判定
- ベイズ理論

今日のテーマ

データの記録とログ解析

- データフォーマット
- ログ解析手法

いろいろなログ

- web server accesslog
- mail log
- syslog
- firewall log
- IDS log
- その他 あらゆる記録

なぜログ解析をするのか？

- 現状の把握
  - 新しい発見: 技術の進歩や利用形態の変化
  - そのうえで将来予測
- セキュリティ上の問題や機器故障、それらの兆候の把握
- 解析技術の向上
  - 自動化
- 障害のレポート、問題への対応
- 記録の必要性
  - 法的理由、その他

そもそも解析されないログには意味がない
(ログを取るだけで安心してはいけない)

ログ解析の問題

- 膨大なデータ量
- 必要な情報や精度の欠如、時刻情報や内容の信憑性
- (収集システムの障害などによる) 記録の欠落
- さまざまなフォーマットが存在
- 解析には時間と労力が必要
- 解析は難しいという思い込み

ログの管理

- ログ収集
  - プログラミング (syslog API の利用など)
  - 収集システム構築
- ログローテーション
  - 古いデータを一定期間保存した後削除
  - ログサイズ、時間、データの古さ
  - ローテーション時にデータを失わないよう工夫
- RRD (Round Robin Database)
  - 古いログを集約することで、データサイズを一定にする
  - 例: 5 分粒度で 1 週間、2 時間粒度で 1 カ月、1 日粒度で 1 年
- 可視化
  - グラフ化して web に貼ることで状況の把握を容易に

さまざまなフォーマット

- web server access log
- mail log
- DHCP server log
- syslog

web server access log

- Apache Combined Log Format
  - Common Log Format に referer と User-agent を追加
  - client_IP client_ID user_ID time request status_code size referer user-agent

例:

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] \
  "GET /apache_pb.gif HTTP/1.0" 200 2326 \
  "http://www.example.com/start.html" \
  "Mozilla/4.08 [en] (Win98; I ;Nav)"
```

mail log

受信、送信などのメール処理毎にログが取られる
例:

```
Oct 27 13:32:54 server3 sm-mta[24510]: m9R4WsBe024510:\
 from=<client@example.com>, size=2403, class=0, nrcpts=1 \
 msgid=<201012121547.oBCFlPX6032787@example.com>, \
 proto=ESMTP, daemon=MTA, relay=mail.example.co.jp [192.0.2.1] \
Oct 27 14:43:04 server3 sm-mta[24511]: m9R4WsBe024510: \
 to=<user@example.co.jp>, delay=01:10:10 xdelay=00:00:00, \
 mailer=local, pri=32599, dsn=2.0.0, stat=Sent
```

- 時刻
- ホスト名
- プロセスオーナー [プロセス番号]
- Queue ID: メールの内部 ID
- ...
- nrcpts: 受信者数
- relay: 次の送信先サーバ
- dsn: Delivery Status Notification, RFC3463
  - 2.X.X:Success, 4.X.X:Persistent Transient Failure, 5.X.X:Permanent Failure
- stat: Message Status
  - Sent, Deferred, Bounced, etc

DHCP server log

SYSLOG: メッセージの記録

```
Oct 28 15:04:32 server33 dhcpd: DHCPDISCOVER from 00:23:df:ff:a8:a7 via eth0
Oct 28 15:04:32 server33 dhcpd: DHCPOFFER on 192.168.2.101 \
  to 00:23:df:ff:a8:a7 via eth0
Oct 28 15:04:32 server33 dhcpd: DHCPREQUEST for 192.168.2.101 \
  from 00:23:df:ff:a8:a7 via eth0
Oct 28 15:04:32 server33 dhcpd: DHCPACK on 192.168.2.101 \
  to 00:23:df:ff:a8:a7 via eth0
Oct 28 15:09:32 server33 dhcpd: DHCPREQUEST for 192.168.2.101 \
  from 00:23:df:ff:a8:a7 via eth0
Oct 28 15:09:32 server33 dhcpd: DHCPACK on 192.168.2.101 \
  to 00:23:df:ff:a8:a7 via eth0
```

dhcpd.leases: 割り当てた IP アドレスの個別情報

```
lease 192.168.100.161 {
  starts 4 2010/12/09 23:13:39;
  ends 5 2010/12/10 00:13:39;
  tstp 5 2010/12/10 00:13:39;
  binding state free;
  hardware ethernet 5c:26:0a:17:06:00;
}
```

syslog

- UNIX 系 OS で任意のメッセージを通知したり保存する仕組み
  - もともとメールの処理記録保存用だったが広く使われるようになった
  - 他のサーバに転送も可能
  - ログのローテーション機能のサポート
- Windows Event Log

ログ解析手法

- 思いつくことを色々試す、グラフ化する
  - 手を動かしている内に分かる事、思いつく事が多い
- 処理スクリプトとコマンドラインツール (grep, sort, uniq, sed, awk, etc)
- 大量データを効率よく処理する工夫
- 繰り返し行う処理はできるだけ自動化する
  - いっぽうで自動化した処理を過信しないこと

大量データの扱い

- ナイーブにやると膨大なデータの読み込みや処理テーブルが必要
  - データ構造やアルゴリズムを勉強しておくと役立つ
- 大量データを扱う工夫
  - 集計に不要な情報の削除
  - 時間的、空間的に集約
  - 必要に応じて分割処理
  - 必要に応じて分散並列処理
- 中間ファイルに変換する、分割処理する
- 処理に必要なメモリ量の見積り
  - データ構造を工夫する
  - 一度に処理するサイズ、次元を押える工夫
- 全体の処理時間の見積り
  - 小さいデータセットで試行
  - スケールするアルゴリズム
- メモリサイズと処理時間のトレードオフに配慮

基本的手法

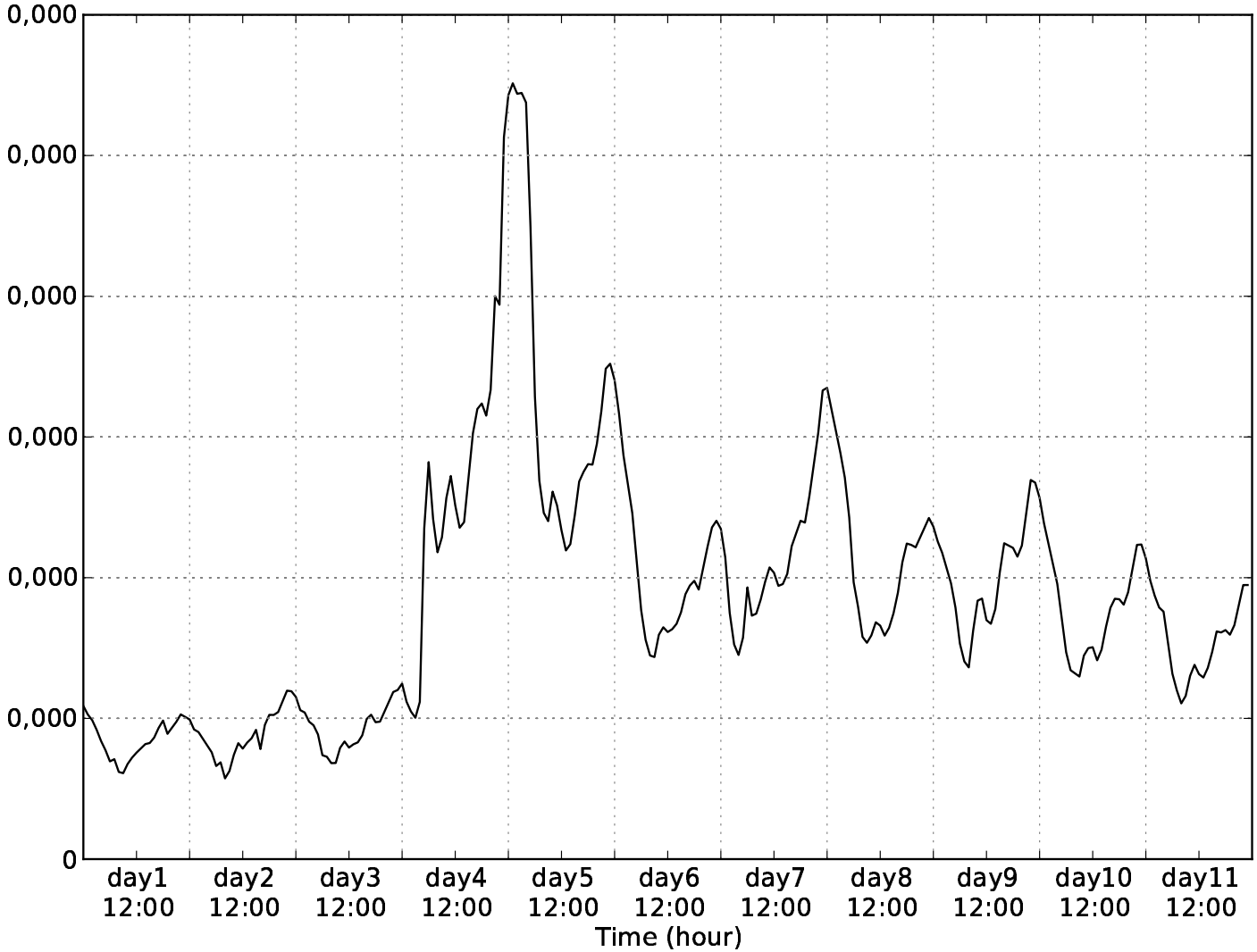
- 正規表現
- 集約
- 正規化
- 閾値
- 外れ値の扱い

ログ解析の具体例

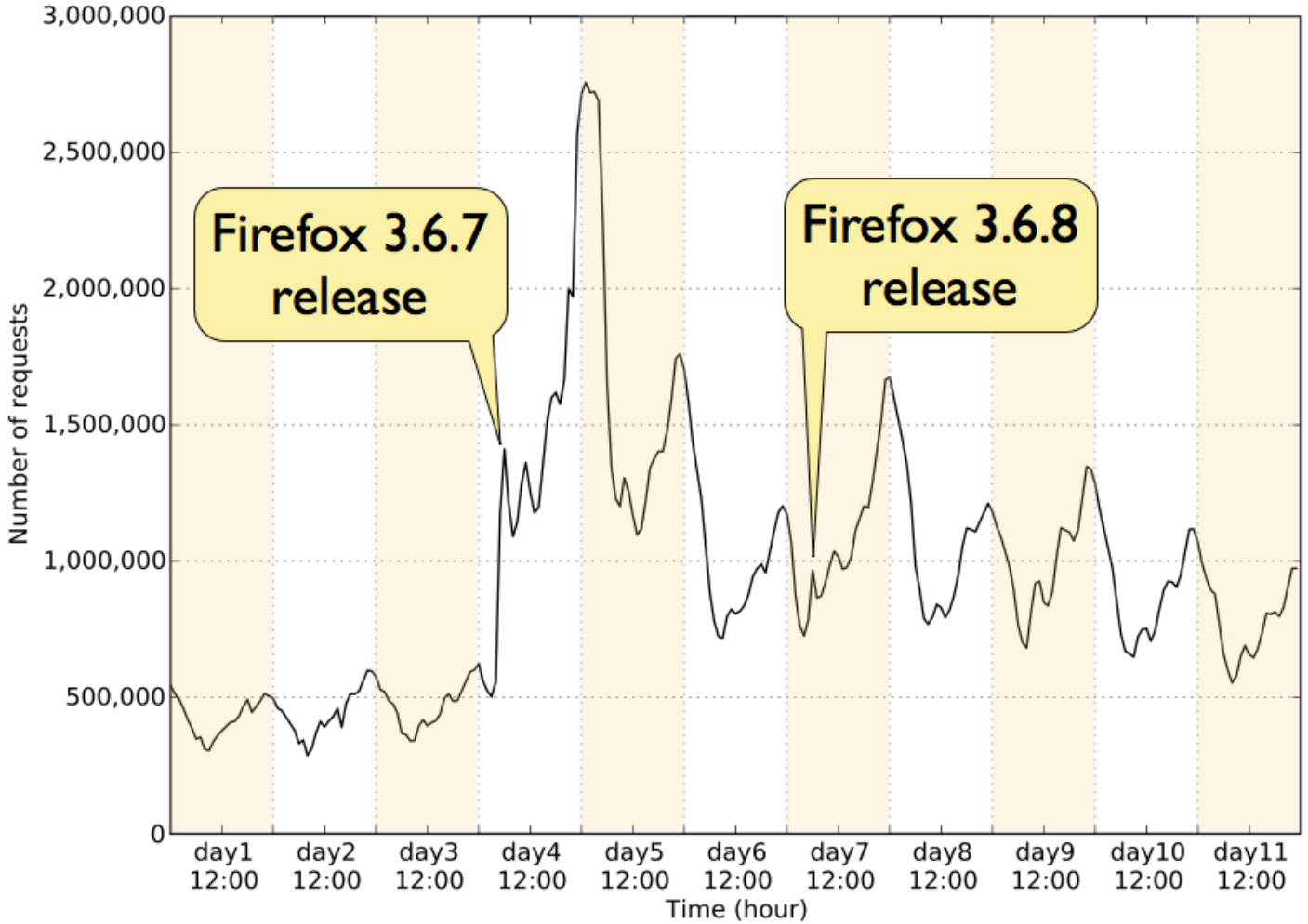
ftp.jaist.ac.jp の apache access log

- 自称日本最強のミラーサーバ (http://ftp-admin.blogspot.com/)
- ソフトウェア配布が主なので普通の web server ではない
- ftp という名前だが、http がメイン
- 期間: 2010/07/18-28
  - 2 回の firefox の更新
  - その間に mozilla.org のミラーサーバ選択ポリシー変更
- 簡単な解析
  - 1 時間ごとのリクエスト数の推移
  - コンテンツごとのリクエスト数の分布

1時間ごとのリクエスト数の推移

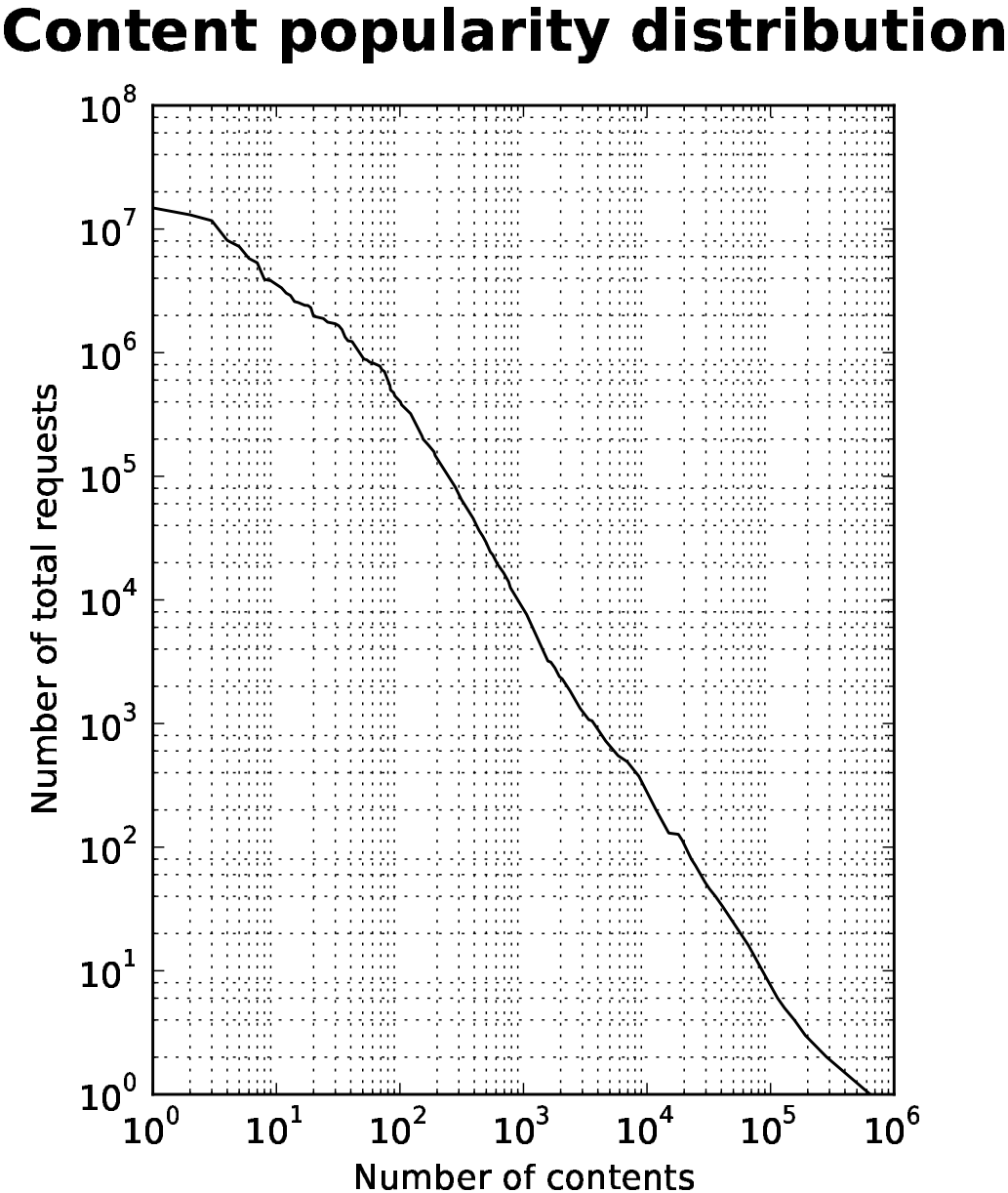


Firefox update

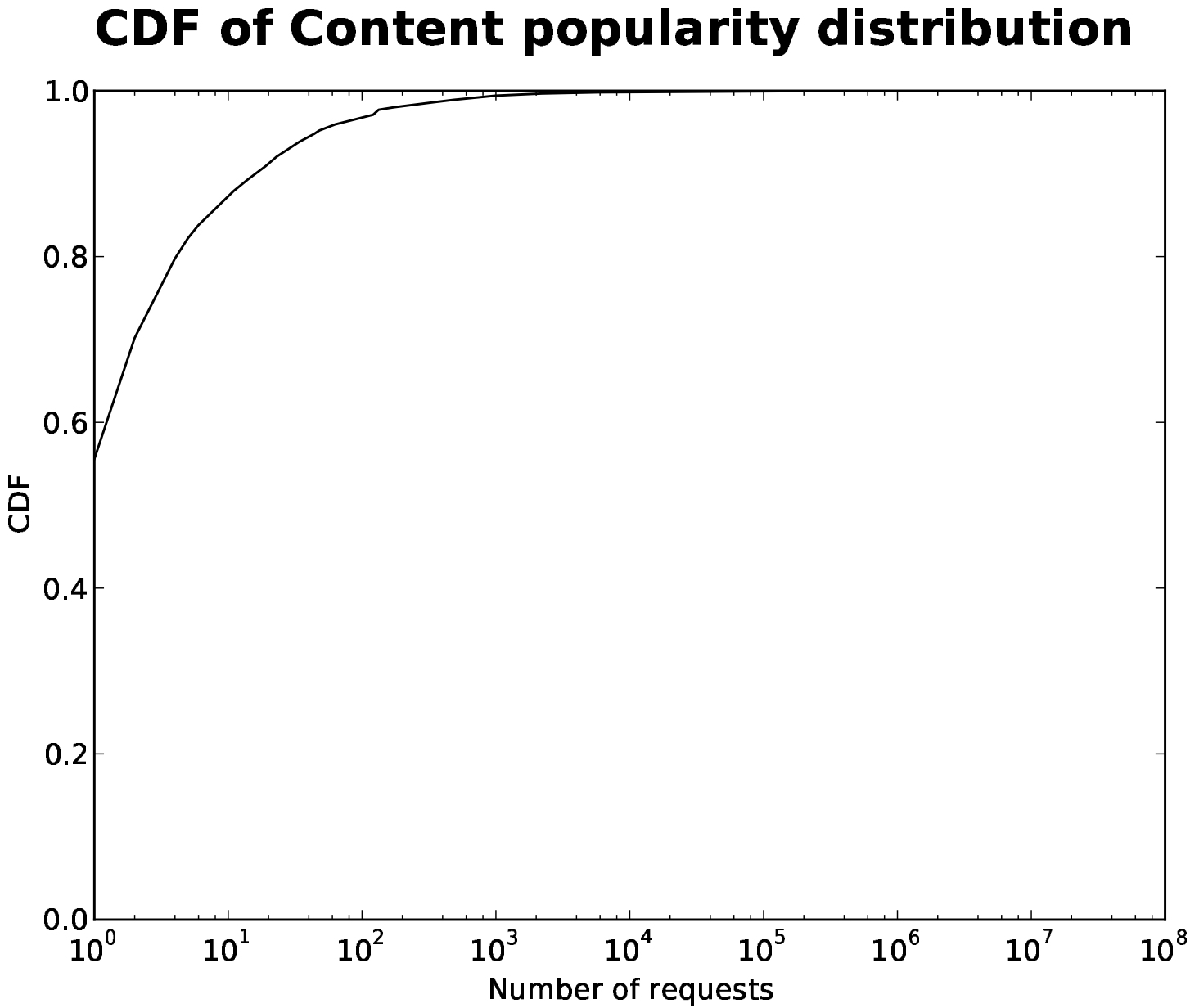


21

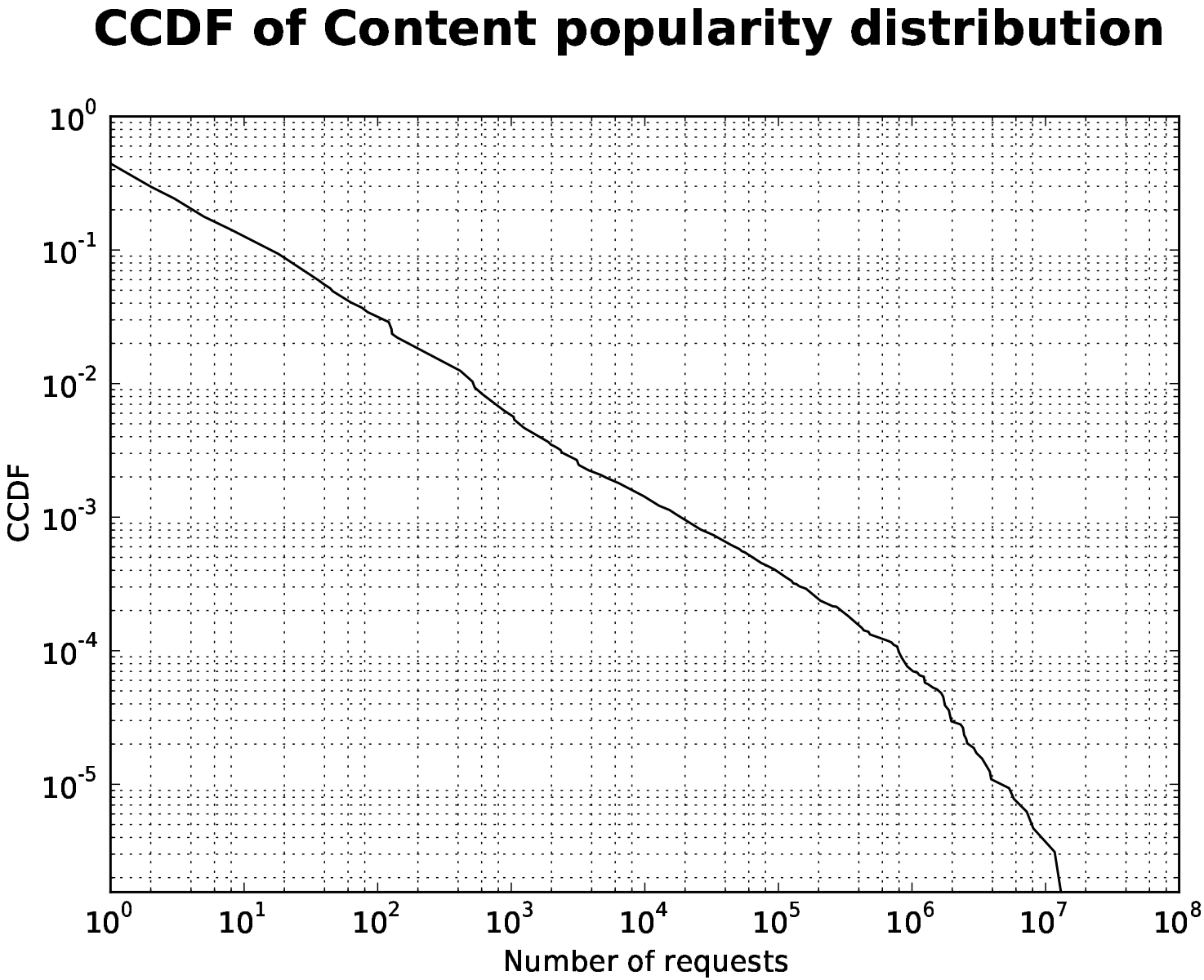
コンテンツごとのリクエスト数分布



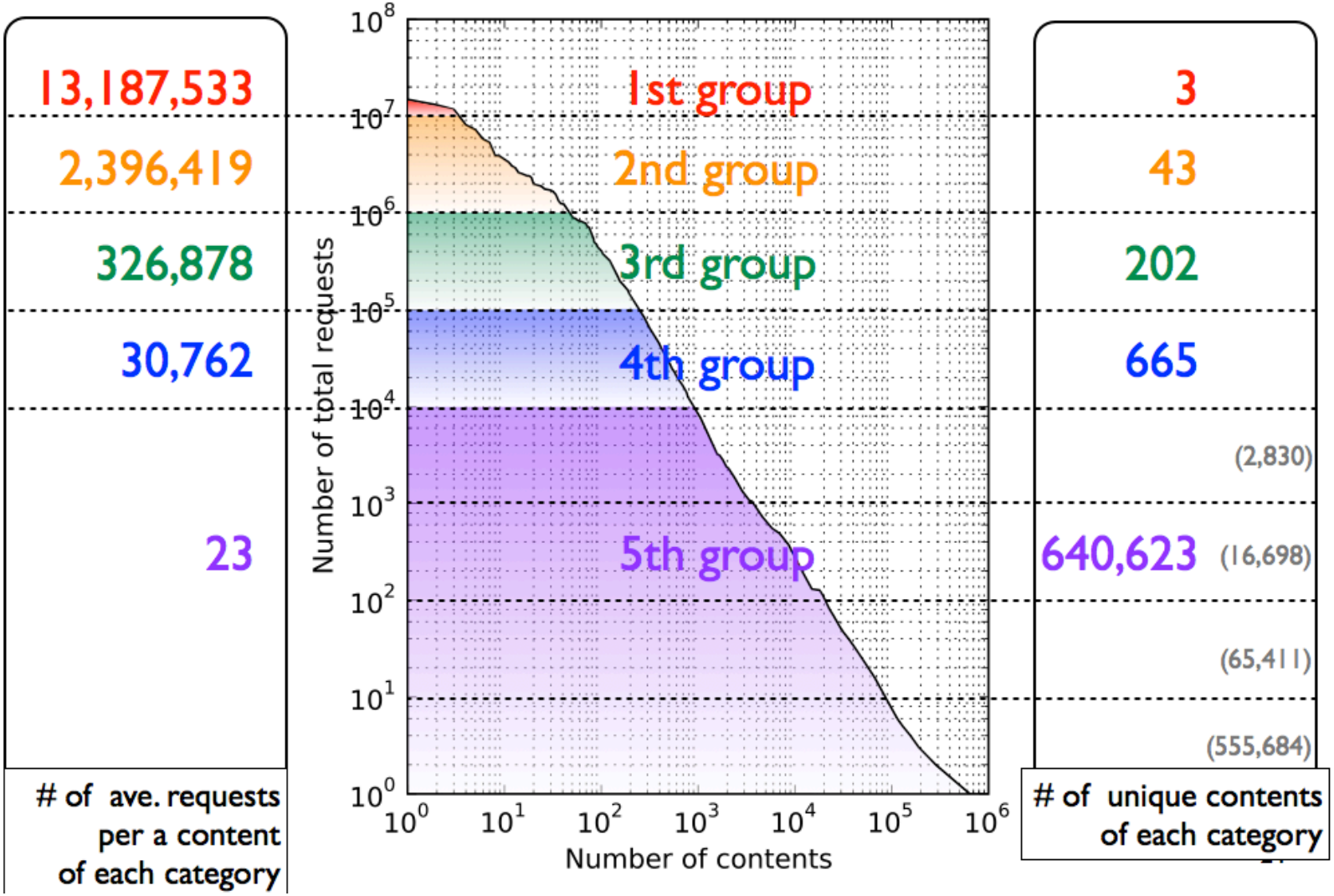
コンテンツごとのリクエスト数 CDF



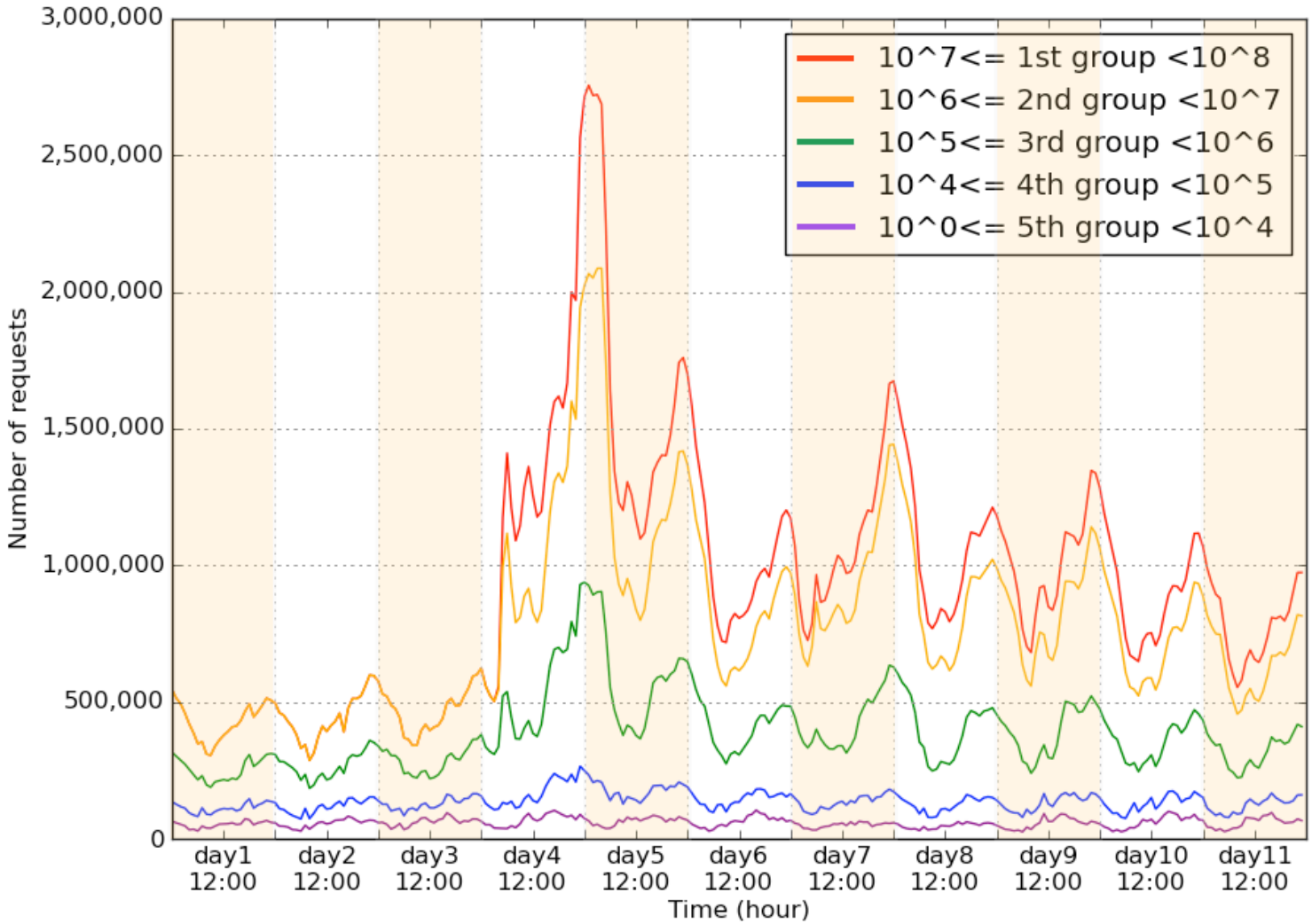
コンテンツごとのリクエスト数 CCDF



リクエスト数によるグループ分け



リクエスト数によるグループ分け



25

まとめ

データの記録とログ解析

- データフォーマット
- ログ解析手法

次回予定

第 11 回 データマイニング (12/22)

- パターン抽出
- クラス分類
- クラスタリング