Techniken zur Analyse von Logdaten

Yevgen Mexin

Institut für Informatik, Fakultät EIM
Prof. Dr. Kleine Büning, Dr. Anderka

- Logs
  - Protokolle
    - Arbeitsprozess
    - Erfolgreich erfüllte Aufgaben
    - Entstehende Fehler
    - Benutzeraktionen
  - Ziel
    - Analyse des System- und Benutzerverhaltens
    - Entdeckung von
      - Programmfehler
      - Angriffen
      - Effizienzschwachstellen etc.

Anwendungsbereich

- Applikation Server
  - Ein Programm, das die Dienstleistungen im Netz anbietet.
  - Verschiedene Typen
    - Mailserver
    - FTP-Server
    - Webserver
    - Datenbankserver

Anwendungsbereich

- Firewall
  - Sicherheitssysteme, die anhand festgelegter Regeln die ein- und ausgehende Pakete filtrieren.
  - Logdatei erfasst die Information über
    - Eingehende Pakete
    - Ausgehende Pakete
    - Verworfene Pakete

Anwendungsbereich

- IDS/IPS
  - Intrusion Detection Systems (IDS)
  - Intrusion Prevention Systems (IPS)
  - Sicherheitssysteme, die gegen Computersystem oder Netzwerk gerichtete Angriffe erkennen/abwehren.
  - In Logdateien wird die Information über verdächtigen Paketen erfasst.

Anwendungsbereich

- Betriebssysteme
  - Sicherheitslogs
    - Gültige/ungültige Anmeldeversuche
    - Ressourcennutzung
  - Systemlogs
    - Ereignissen in Systemkomponenten
  - Applikationlogs
    - Arbeitsprozess der Anwendungen

Verwendungszwecke

- Fehleranalyse (Debuggen)
  - Analyse der Ereignissen, die zur Fehler geführt haben.
- Sicherheit
  - Inspektion der Systemereignissen und Benutzerverhaltens nach einem Vorfall.
    - Firewalllogs, Protokolle der Login-Sessionen, Information über Ressourcennutzung und Paketaustausch.
    - Rekonstruktion der verdächtigen Session.

Verwendungszwecke

- Effizienzanalyse
  - Suche nach Performance-Problemen
    - Ressourcenverwendung, Interaktion der Komponenten, Sperroperationen von Ressourcen
- Prognostizieren
  - Vorhersage der zukünftigen Effizienzproblemen
  - Erstellen von analytischen Modellen/Simulatoren
    - Optimale Ressourcenplanung, Ablaufplanung, Systemkonfiguration etc. zu berechnen.

Verwendungszwecke

- Profilierung
  - Erstellen Profilen von
    - Benutzerverhalten,
    - Ressourcenverwendung,
    - Arbeitsbelastung, etc.
  - Sammeln von stochastischen Daten.
  - Web Usage Mining
    - Analyse des Benutzerverhaltens auf einer Webseite.

- Inhalt und Format von Logdatei
  - Abhängig von konkreten Anwendungsbereichen und von gestellten Zielen.
  - Kein universeller de-facto Standard.
  - Bekannt sind Formate für
    - Webserver Logdaten
      - Common Log file Format (CLF)
      - Extended Log file Format (ELF)
    - Übermittlung von Logdaten in *NIX Betriebssystemen und Netzwerken
      - Syslog

```
123.123.123.123 - - [17/dec/2013:18:23:48 +0100]
"GET/logo/UPB_LOGO_CMYK_12.zip HTTP/1.0" 200 1919610
```

- Common Log file Format (CLF)
  - National Center for Supercomputing Applications (NCSA)
    - IP-Adresse oder Hostname des Clients
    - Klientidentifikator und Benutzeridentifikator („-“, falls fehlt)
    - Zeitstempel
    - Anfragemethode und Anfragepfad
    - Protokoll
    - HTTP Anfragestatus
    - Anzahl der übergebenden Bits

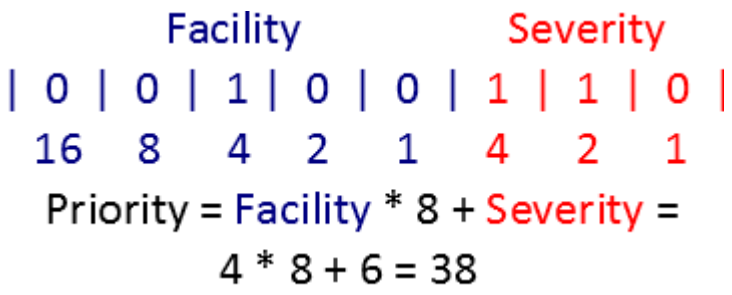
```
#Version: 1.0
#Date: 17-Dec-2013 24:00:00
#Fields: time c-ip cs-method cs-uri-stem cs-status cs-version
18:23:48 123.123.123.123 GET/logo/UPB_LOGO_CMYK_12.zip 200 HTTP/1.0
```

- Extended Log file Format (ELF)
  - World Wide Web Consortium (W3C)
  - Anpassbares Format

Inhalt und Formate

```
<38>Dec 15 18:03:24 charlie sshd[701]: Connection from 203.0.113.5 port 6010
 PR-----------TS----------HN--------TG----------------------CN--------------------------
<52>Dec 15 18:43:45 192.168.0.34 printer: paper out
 PR-----------TS----------------HN---------TG---------CN-----
```

- Syslog
  - Übermittlung von Logdaten im Netzwerk
    - PR - Priority
    - Header
      - TS - Timestamp
      - HN - Hostname
    - Message
      - TG - Tag
      - CN - Content

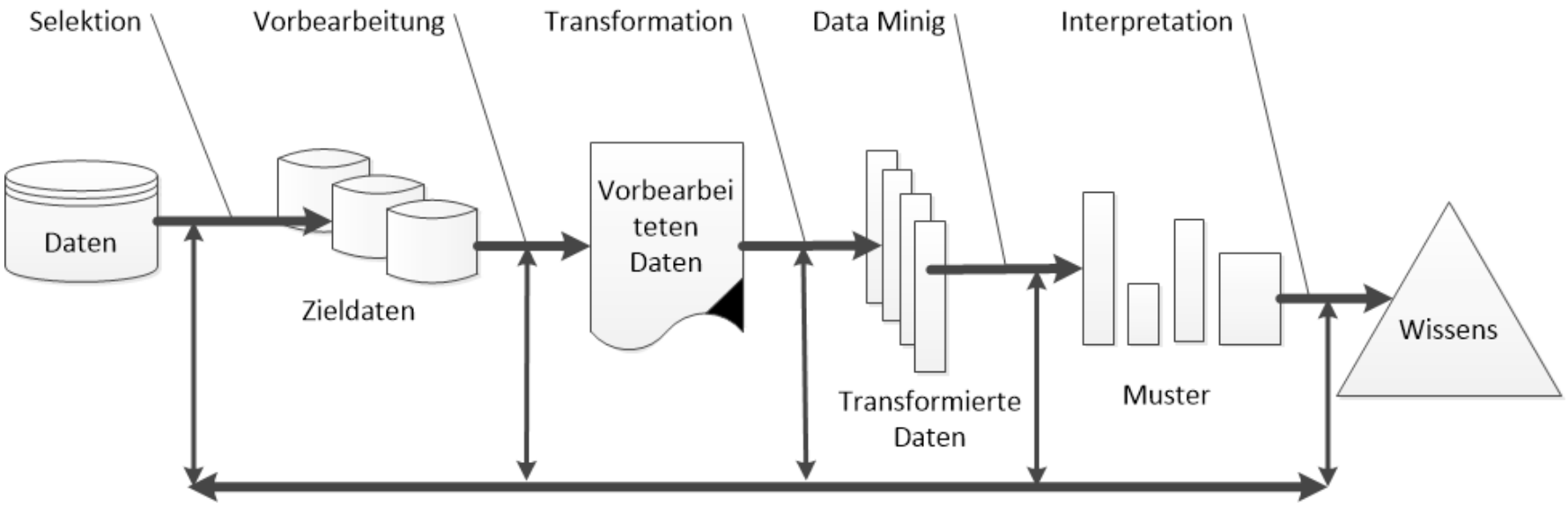


- Priorität
  - 1-Byte numerische Wert
    - Severity
      - Wichtigkeitsstufe der Nachricht.
      - von 0 (kritischste Stufe) bis 7 (Debugging-Nachricht)
    - Facility
      - Repräsentiert den Dienst, der die Nachricht auslöst.
      - von 0 (Kernelnachricht) bis 23 (nicht reserviert).
      - Von 16 bis 23 sind für lokale Verwendung erlaubt.

- Pattern Matching
  - Suche einer Bit-Sequenz, die einem der vorgegebenen Mustern entspricht.
- Stateful Pattern Matching
  - Pattern Matching im Kontext des ganzen Datenstroms
- Protocol Decode-Based Analysis
  - Erweiterung von Stateful Pattern Matching, bei der die Übertretungen der von Standard definierten Regeln gesucht werden.

- Heuristic-Based Analysis
  - Nimmt Entscheidungen anhand vorbereiteten Algorithmen. Oft werden statistische Methoden angewandt.
- Anomaly Detection
  - Suche abnormaler Ereignissen, die dem anhand von Mustern trainierten Modell nicht entsprechen.

- Knowledge Discovery in Databases (KDD)
  - Prozess der Erkennung von bislang unbekannten Zusammenhänge aus großen Datenmengen.
  - Oft verwendetes Synonym - Data Mining.



- Data Mining
  - Erkennung von Mustern in vorhandenen Daten.
  - Umfasst eine große Menge von Algorithmen und Verfahren.
    - Aus maschinelles Lernen, Statistik, Wahrscheinlichkeitslehre etc.
  - Einzelne Kategorien
    - Klassifikation und Vorhersage
    - Clusteranalyse

- Klassifikation/Vorhersage
  - Aufgabe: die Mitgliedschaft eines Objektes zu einem der endlichen Menge von Klassifikatoren bzw. den Ausgabewert vorauszusagen.
  - Überwachtes Lernen.
  - Verfahren
    - Entscheidungsbäume
    - Naive Bayes Klassifikator
    - Neuronale Netze
    - Genetische Algorithmen
    - Fuzzylogik Verfahren
    - Regression etc.

- Clusteranalyse
  - Aufgabe: die Objekte in einzelne Gruppen zu trennen.
    - Ähnlichkeit der Objekten
      - innerhalb einer Gruppe - maximieren
      - zwischen Gruppen - minimieren
  - Verfahren
    - Hierarchische
      - Single-Link, Group-Average-Link, Min-Cut etc.
    - Iterative
      - k-Means etc.
    - Dichtebasierte
      - DBSCAN etc.

- Distanz/Ähnlichkeit zwischen Objekten
  - Euklidische Distanz

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

  - Manhatten Distanz

$$d(x, y) = \sum_{i=1}^{p} (x_i - y_i)$$

  - Jaccard Distanz

$$J = \frac{f_{11}}{f_{11} + f_{10} + f_{01}}$$

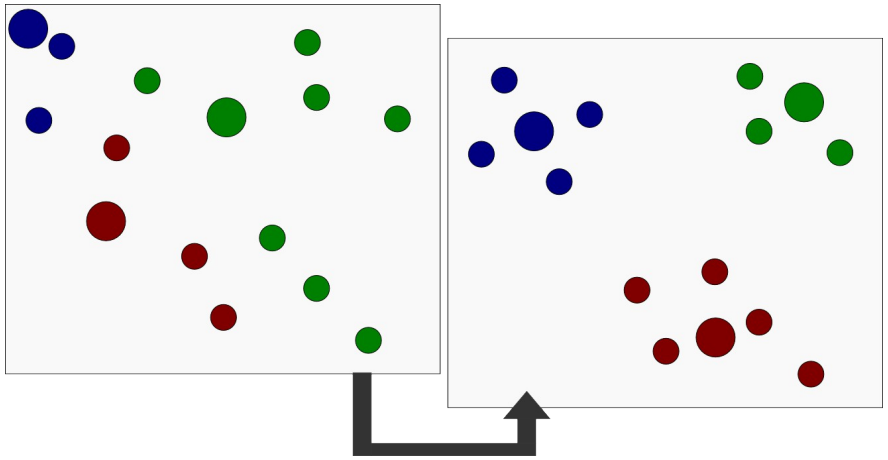
- Hierarchisch-Agglomerative Clustering
  1. Jedes Objekt ist ein Cluster.
  2. Verschmelze zwei am nächsten liegenden Cluster.
  3. Wenn die gewünschte Anzahl der Cluster erreicht ist, oder nur ein Cluster existiert – Abbruch.

     Sonst – wiederhole Schritt 2.
- Single-Link
  - Distanz zwischen zwei Cluster

$$d(C, C') = min(d(u, v)), u \in C, v \in C'$$

- Iterativ-exemplarbasiertes k-Means

1. Wähle zufällig k Clustermittelpunkte
2. Für alle Objekte
   - 2.1. Berechne Abstand zu dem Mittelpunkt
   - 2.2. Zuweise dem nächsten Mittelpunkt
3. Wenn keine Objektzuordnung geändert wird – Abbruch.
   Sonst berechne neue Mittelpunkte und wiederhole ab Schritt 2.



Clustering

- Execution anomaly detection in distributed systems through unstructured log analysis. [1]
  - Anomalien-Erkennungssystem anhand von Logdaten.
  - K-Means, endlicher Automat für Modell

- Social Intelligence In Completely Online Groups. Toward Social Prosthetics From Log Data Analysis and Transformation. [2]
  - Untersuchung von Benutzerverhaltens in sozialen Netzwerken.
  - hierarchisch-agglomerative Clustering

Clustering

- Mining Unstructured Log Files for Recurrent Fault Diagnosis. [3]
  - Fehler-Ursachen-Analys anhand von Logdaten.
  - Clustering Algorithmus und Entscheidungsbäume

- A Log Analysis Audit Model Based on Optimized Clustering Algorithm. [4]
  - Angriffserkennung anhand von Logdaten
  - Clustering Algorithmus

- Bayes-Klassifikation
  - Stochastisches Verfahren
    - vorhersagt Wahrscheinlichkeit von Zugehörigkeit eines Objektes zur bestimmte Gruppe.
    - Überwachtes Lernen – gegeben ist eine Trainingsmenge *D*
    - Betrachtet folgende Ereignisse
      - Attribut-Wert Paare $A_j : Attribut_j = Wert_j$
      - Klassifikation-Wert Paare $B_i : Klassifikation_i = Wert_i$
    - Satz von Bayes
      - a-Posteriori Wahrscheinlichkeit

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

- Naive-Bayes
  - Annahme: die Ereignisse $A_1, \ldots, A_p$ unter dem Ereignis B_i stochastisch unabhängig sind.
  - Unter dieser Bedingung kann die Klassifikation jedes Objektes berechnet werden.

$$B_{NB} = argmax_{B_i \in \{B_1, \ldots, B_k\}} P(B_i) \cdot \prod_{j=1,\ldots,p} P(A_j|B)$$

  - Bekannt sein müssen
    - $P(B_i)$ - relative Häufigkeit von Beispielen aus der Klasse B_i in der Trainingsmenge *D.*
    - $P(A_j|B_i)$ - relative Häufigkeit von Beispielen mit dem Attribut A_j aus der Klasse B_i in der Trainingsmenge *D*.

Bayes-Klassifikation

- Naive bayesian filters for log file analysis: Despam your logs. [5]
  - Kategorisierung von Logeinträge
  - Naive Bayes Filter
    - werden auch bei Email-Antispam Systemen verwendet.
- Mining log files for data-driven system management. [6]
  - Mining von Logdaten
  - Naive-Bayes-Klassifikation und Hidden Markov Model
- Bayesian analysis of online newspaper log data [7]
  - Analyse der bei einer online Zeitung gesammelten Logdaten.
  - Bayes Netz

Referenzen

[1] Fu, Q., Lou, J.G., Wang, Y., Li, J.: Execution anomaly detection in distributed systems through unstructured log analysis. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. (2009)

[2] Goggins, S., Gallagher, M., Laey, J., Amelung, C.: Social intelligence in completely online groups - toward social prosthetics from log data analysis and transformation. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on. (2010)

[3] Reidemeister, T., Jiang, M., Ward, P.: Mining unstructured log les for recurrent fault diagnosis. In: Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on. (2011)

[4] Yu, H., Shi, X.: A log analysis audit model based on optimized clustering algorithm. In: Network and Parallel Computing Workshops, 2007. (2007)

[5] Havens, R., Lunt, B., Teng, C.C.: Naive bayesian filters for log file analysis: Despam your logs. In: Network Operations and Management Symposium (NOMS). (2012)

Referenzen

[6] Peng, W.: Mining log files for data-driven system management. SIGKDD Explorations. (2005)

[7] Wettig, H., Lahtinen, J., Lepola, T., Myllymäki, P., Tirri, H.: Bayesian analysis of online newspaper log data. In: Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT'03 Workshops). (2003)

Vielen Dank!