

(12)

Patentschrift

(21) Anmeldenummer: A 50741/2020
(22) Anmeldetag: 01.09.2020
(45) Veröffentlicht am: 15.03.2022

(51) Int. Cl.: **H04L 12/24** (2006.01)
H04L 12/26 (2006.01)
H04L 29/06 (2006.01)
G06F 21/55 (2013.01)

(56) Entgegenhaltungen:
EP 3267625 A1
US 9843596 B1
WO 2016191231 A1
EP 3528162 A1
EP 3582443 A1
AT 520746 A4
AT 518805 B1
AT 514215 B1
M. Landauer et al. "System log clustering
approaches for cyber security applications: A
survey" 31.01.2020; Computers & Security
Volume 92, May 2020, 101739;
<https://doi.org/10.1016/j.cose.2020.101739>
EP 3719651 A1

(73) Patentinhaber:
AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
GMBH
1210 Wien (AT)

(72) Erfinder:
Höld Georg
3500 Krems (AT)
Landauer Max
1070 Wien (AT)
Wurzenberger Markus
1090 Wien (AT)
Skopik Florian DDr.
2000 Stockerau (AT)

(74) Vertreter:
Wildhack & Jellinek Patentanwälte OG
1030 Wien (AT)

(54) Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems

- (57) Die Erfindung betrifft ein Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems,
- wobei während des Betriebs des Computers oder von auf dem Computer ablaufenden Prozessen, Protokolle erstellt werden,
 - wobei den einzelnen Protokollzeilen basierend auf deren Syntax jeweils zumindest ein Ereignistyp zugewiesen wird,
 - wobei für jeden Ereignistyp ein Template umfassend eine Anzahl von Teilzeichenabfolgen der Protokollzeilen des jeweiligen Ereignistyp repräsentierenden Token (T), erstellt wird. Erfindungsgemäß ist vorgesehen,
 - dass in einer Initialisierungsphase
 - auf Grundlage einer Stichprobe umfassend eine Anzahl an Protokollzeilen Tokenpaare (TP) aus den, den diskreten Datentyp aufweisenden, Token (T) des Templates gebildet werden, und
 - für alle Tokenpaare (TP) und jeden angenommenen Wert des ersten Tokens (T_1) des Tokenpaares (TP) jeweils eine Korrelation ermittelt wird, und

- dass anschließend in einer Testphase
 - wiederholt, sobald eine Abfolge neu erstellter Protokollzeilen erstellt wurde, mittels eines Tests auf Homogenität getestet wird, ob die aufgefundenen Korrelationen erfüllt sind, und
 - ein anomaler Betriebszustand des Computersystems erkannt wird, wenn festgestellt wird, dass die Werte der Token (T) zumindest einer der ermittelten Korrelationen nicht den jeweils berechneten bedingten Verteilungen folgen.

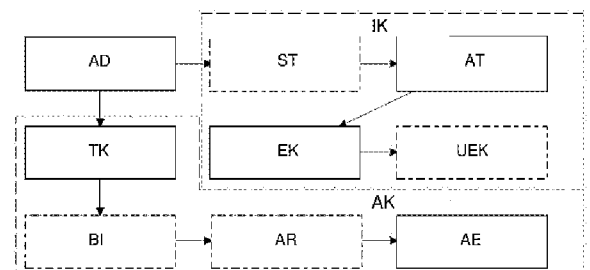


Fig. 1

Beschreibung

[0001] Die Erfindung betrifft ein Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems gemäß dem Oberbegriff von Patentanspruch 1.

[0002] Aus dem Stand der Technik ist bekannt, Protokolldateien, die von unterschiedlichen Prozessen erstellt werden, daraufhin zu untersuchen, ob die in den Protokolldateien beschriebenen Vorgänge einen anomalen Zustand der Prozesse bzw. des Computernetzwerks darstellen, in dem diese Prozesse ablaufen.

[0003] Bei den vorstehend genannten Verfahren bestehen wesentliche Probleme, die Werte der üblicherweise in für Menschen lesbarer Form geschriebenen Protokolldateien auf bestimmte Muster hin zu analysieren, um somit Betriebszustände zu erkennen, die ungewöhnlich oder einzigartig sind und auf anomale Betriebszustände hindeuten.

[0004] Solche Vorgehensweisen ermöglichen grundsätzlich das Auffinden von anomalen Zuständen in einem Computernetzwerk, sind jedoch relativ komplex und erfordern es, die Werte der Token auf Muster zu analysieren und anschließend neue Protokolldateien auf die gefundenen Muster hin zu analysieren, was insgesamt zu einem erhöhten Ressourcenbedarf der Protokoll-datenanalyse führt. Aus dem Stand der Technik sind jedoch keine konkreten Verfahren bekannt, die versuchen, die Werte aller Token der Protokollzeilen gleichzeitig auf Muster hin zu analysieren.

[0005] Aufgabe der Erfindung ist es daher, diese Probleme zu lösen und ein Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems zur Verfügung zu stellen, mit dem eine rasche und effiziente Untersuchung von, vom Computersystem erstellten, Protokollzeilen ermöglicht wird, sodass Anomalien im Verhalten des Computersystems zuverlässig und ressourcensparend detektiert werden können.

[0006] Die Erfindung löst diese Aufgabe bei einem Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems,

- wobei während des Betriebs des Computers oder von auf dem Computer ablaufenden Prozessen, Protokolle erstellt werden, indem bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils eine Protokollzeile in Form einer Zeichenabfolge erstellt wird, wobei der Aufbau der Protokollzeile einer vom jeweiligen Computer und/oder Prozess vorgegebenen Syntax entspricht,
- wobei die einzelnen Protokollzeilen an vorgegebenen Trennzeichen in Teilzeichenabfolgen unterteilt werden,
- wobei den einzelnen Protokollzeilen basierend auf deren Syntax, insbesondere aufgrund der Ähnlichkeit der einzelnen Protokollzeilen zueinander und aufgrund der die Protokollzeilen aufbauenden Teilzeichenabfolgen, jeweils zumindest ein Ereignistyp zugewiesen wird,
- wobei für jeden Ereignistyp ein Template umfassend eine Anzahl von Teilzeichenabfolgen der Protokollzeilen des jeweiligen Ereignistyps repräsentierenden Token, erstellt wird,
 - wobei das Template die Struktur der einem jeweiligen Ereignistyp zugeordneten Protokollzeilen beschreibt, und
 - wobei das Template eine Anzahl von Token mit statischem Inhalt, insbesondere statischen Werten, umfasst, die in ihrer Reihenfolge in ausgewählten, insbesondere allen, Protokollzeilen des jeweiligen Ereignistyps vorhanden sind, und wobei die Token mit statischem Inhalt des Templates durch Token mit variablen Inhalt, insbesondere variablen Werten, unterbrochen werden,
 - wobei das Template eine Anzahl von Token aufweist, deren Inhalt, insbesondere deren Werte, einen diskreten Datentyp aufweisen, und
- wobei die Protokollzeilen und deren Teilzeichenabfolgen charakteristisch für den Betriebszustand des Computersystems sind.

[0007] Erfindungsgemäß ist dabei vorgesehen,

- dass in einer Initialisierungsphase

- auf Grundlage einer Stichprobe umfassend eine Anzahl an Protokollzeilen des jeweiligen Ereignistyps Tokenpaare umfassend einen ersten Token und einen zweiten Token aus den, den diskreten Datentyp aufweisende, Token des Templates eines jeweiligen Ereignistyps gebildet werden,
 - wobei als Werte des jeweiligen Tokens jeweils diejenigen Teilzeichenabfolgen der Protokollzeilen der Stichprobe ermittelt werden, die den einzelnen Token der Tokenpaare zugehörig sind, und
- für alle Tokenpaare und jeden angenommenen Wert des ersten Tokens des Tokenpaares jeweils eine Korrelation ermittelt wird, indem die bedingten Verteilungen dafür berechnet werden, dass bei Auftreten des jeweiligen Werts im ersten Token des Tokenpaares, ein jeweiliger Wert im zweiten Token des Tokenpaares auftritt, und
- dass anschließend in einer Testphase
 - wiederholt, sobald eine Abfolge, insbesondere eine vorgegebene Anzahl, neu erstellter Protokollzeilen eines jeweiligen Ereignistyps erstellt wurde, mittels eines Tests auf Homogenität getestet wird, ob die aufgefundenen Korrelationen erfüllt sind, indem jeweils für alle Werte des ersten Tokens eines jeweiligen Tokenpaares überprüft wird, ob bei denjenigen neu erstellten Protokollzeilen, bei denen der jeweilige Wert im ersten Token auftritt, die Werte des zweiten Tokens den berechneten bedingten Verteilungen der jeweiligen ermittelten Korrelation folgen und
 - ein anomaler Betriebszustand des Computersystems erkannt wird, wenn, insbesondere wiederholt, festgestellt wird, dass die Werte der Token zumindest einer der ermittelten Korrelationen nicht den jeweils berechneten bedingten Verteilungen folgen.

[0008] Mit dieser Vorgehensweise ist es besonders zuverlässig möglich, die Muster der angenommenen Werte der Teilzeichenabfolgen der Protokollzeilen der üblicherweise, d.h. im Fall, dass keine Angriffe oder Veränderungen des Computersystems vorliegen, auf dem Computersystem erstellten Protokollzeilen, zu identifizieren, sodass besonders rasch und zuverlässig neu erstellte Protokollzeilen darauf überprüft werden können, ob diese dem Muster entsprechen und im Falle des nicht Entsprechens als Anomalien des Computersystems erkannt werden.

[0009] Dadurch, dass nur Token analysiert werden, welche diskrete Werte annehmen, d.h., deren Werte kategorisch sind und sich dadurch auszeichnen, dass die Auswahl der angenommenen Werte aus einer begrenzten Menge entstammt, können besonders einfache und rechenleistungssparende Verfahren für die Detektion von Mustern und darauf aufbauend von anomalen Betriebszuständen herangezogen werden.

[0010] Im Folgenden wird unter einer Protokolldatei eine Datei verstanden, welche eine sequenzielle zeilenweise Aufzeichnung von Ereignissen des Betriebssystems, oder eines Programmes beinhaltet. Diese Aufzeichnungen können Anmerkungen, Warnungen, System- oder Fehlermeldungen der Prozesse, die auf dem System ablaufen, beschreiben. Abhängig davon welche Prozesse in der Protokolldatei protokolliert werden, können sich die Protokollzeilen stark in ihrer Struktur, ihrem Inhalt und ihrem Detailgrad unterscheiden.

[0011] Protokollzeilen werden von Programmen erstellt, welche Protokollnachrichten, die einer gewissen Syntax entsprechend aufgebaut sind, ausgeben. Die Syntax ist hierbei im Hinblick auf den Aufbau der Protokollzeile, vergleichbar mit dem Satzbau in der Sprache, zu verstehen, wobei die Syntax das Zusammenfügen von Wörtern bzw. Wortgruppen zu ganzen Protokollzeilen, bzw. Sätzen, beschreibt. Es sind daher Teilzeichenabfolgen, die durch vorgegebene Trennzeichen (Separatoren), wie zum Beispiel Leerzeichen, Kommas und Strichpunkte, voneinander getrennt werden.

[0012] Basierend auf dieser Syntax können Protokollzeilen in verschiedene Klassen eingeteilt werden. Protokollzeilen einer bestimmten Klasse unterscheiden sich nur an den Stellen, an denen variable Zeichenabfolgen eingefügt wurden. Diese Klassen werden auch als Ereignistypen bezeichnet. Zu den Ereignistypen können Templates, bestehend aus Token, erstellt werden.

[0013] Token bezeichnen Teilzeichenabfolgen einer Protokollzeile, die durch Separieren der Zeichenkette an vorgegebenen Trennzeichen, wie zum Beispiel Leerzeichen oder Beistrichen ent-

stehen. Jeder Ereignistyp lässt sich daher durch ein Muster, bestehend aus einer Abfolge von Token, beschreiben. Diese Muster/Repräsentanten bestehen aus Token mit statischem Text, der in allen Protokollzeilen, die dem Ereignistyp entsprechen an derselben Stelle steht, und aus Token, die durch sogenannte Wildcards gekennzeichnet sind, die an Stellen eingefügt werden an denen die Zeilen variable Teilzeichenabfolgen beinhalten. Die Stellen mit variablen Zeichenketten werden auch als variable Token bezeichnet.

[0014] Im Folgenden werden die Begriffe „Token des Templates zu einem Ereignistyp“ und „Token eines Ereignistyps“ synonym verwendet.

[0015] Im Folgenden wird die betrachtete Abfolge von Protokollzeilen als Stichprobe bezeichnet. Die Werte eines Tokens bezeichnen die Teilzeichenabfolgen der Protokollzeilen der Stichprobe zu diesem Token, d.h., die Teilzeichenabfolgen, welche die Stichprobe im betreffenden Token annimmt. Im Folgenden wird unter einem „Wert eines Tokens“ also eine Teilzeichenabfolge einer Protokollzeile zu diesem Token verstanden.

[0016] Im Folgenden wird unter dem Datentyp eines Tokens eine Beschreibung der erwarteten Eigenschaften der Werte des Tokens verstanden. Dieser kann z.B. erkannt werden, indem getestet wird, ob die Werte des jeweiligen Tokens die Eigenschaften der Datentypen besitzen. Im Folgenden ist unter dem Datentyp „diskret“ zu verstehen, dass die Werte eines Tokens einer diskreten Verteilung folgen.

[0017] Im Folgenden wird unter der bedingten Verteilung $p_{T_2, T_1}(w_2|w_1)$, d.h. einer „durch einen Wert w_1 des ersten Tokens T_1 bedingten Verteilung der Werte w_2 des zweiten Tokens T_2 “, jene Verteilung der Werte w_2 des zweiten Token T_2 des jeweiligen Tokenpaares verstanden, welche durch die nachfolgende Verteilung der einzelnen Werte w_2 des zweiten Token T_2 definiert wird

$$p_{T_2, T_1}(w_2|w_1) = P(T_2 = w_2 | T_1 = w_1) = \frac{P(\{T_1 = w_1\} \cap \{T_2 = w_2\})}{P(T_1 = w_1)}$$

wobei w_1 ein Wert des jeweils ersten Tokens T_1 des jeweiligen Tokenpaares, w_2 ein Wert des zweiten Tokens T_2 des Tokenpaares und P die jeweilige Wahrscheinlichkeit ist.

[0018] Diese Definition ist analog zu der Definition der bedingten Verteilung in Park, K. I., & Park. (2018). Fundamentals of Probability and Stochastic Processes with Applications to Communications. Springer International Publishing.

[0019] Die Bezeichnung bzw. Indizierung eines Werts eines ersten Tokens mit $w_{„1“}$ bzw. $T_{„1“}$ und eines Werts eines zweiten Tokens mit $w_{„2“}$ bzw. $T_{„2“}$ ist jedoch exemplarisch zu verstehen. Die angegebene Definition einer „bedingten Verteilung“ wird im Folgenden analog auf Tokenpaare verschiedenster Token angewendet, wobei aus dem Kontext jeweils hervorgeht um welche Token und welchen Wert des jeweils ersten Tokens des Tokenpaares es sich handelt.

[0020] Im Folgenden wird unter dem normalen Betriebszustand eines Computersystems verstanden, dass das System bei gleichbleibenden Nutzerverhalten, äußeren Einflüssen und ohne Manipulation der intern ablaufenden Prozesse ausgeführt wird. Da Protokolldateien Informationen über Ereignisse und Abläufe von Prozessen eines Systems enthalten, können diese dazu genutzt werden, um Muster im Auftreten und in den Eigenschaften der Ereignisse zu finden. Muster können entdeckt werden, indem Protokolldateien, welche während eines normalen Betriebszustandes entstanden sind, über längere Zeit hinweg analysiert werden. Die Art von Mustern, welche in diesem Patent betrachtet werden, sind Muster in den Werten der Token der Ereignistypen. Dazu werden die Werte der Token verglichen und überprüft ob diese gewisse Eigenschaften besitzen. Unter normalem Betriebszustand treten diese Muster in den neu erzeugten Protokolldateien wieder auf, sodass man von einem Abweichen vom normalen Systemverhaltens spricht, sollte das Muster nicht mehr in neu erzeugten Protokollzeilen fortgeführt werden.

[0021] Im Folgenden wird unter Korrelationen zwischen Werten von zwei Token eines Ereignistyps eine spezielle Art von Mustern in Protokolldateien verstanden. Eine Korrelation zwischen zwei Token liegt vor, wenn während einer Beobachtung der angenommene Wert eines Tokens den Wert des anderen Tokens mitbestimmt, oder beeinflusst. Korrelationen können dabei Beziehun-

gen sein, die nur in eine Richtung oder auch in beide Richtungen wirken, d.h., dass, sobald eine Korrelation in eine Richtung vorhanden ist, auch in die andere Richtung ein Einfluss besteht. Um Unterscheidungen in der Verarbeitung von Korrelationen, welche in eine oder in beide Richtungen wirken, zu vermeiden, werden Korrelationen, die in beide Richtungen wirken, als zwei Korrelationen betrachtet, die jeweils in eine Richtung wirken. Im Folgenden werden daher nur noch Korrelationen in eine Richtung verarbeitet, jedoch ohne, dass Korrelationen verloren gehen. Im Folgenden wird derjenige Token, dessen Werte Einfluss auf die Werte eines anderen Tokens besitzen, als der erste Token bezeichnet und derjenige Token, dessen Wert beeinflusst wird, als der zweite Token. Im Folgenden wird die Verteilung der Werte des zweiten Tokens, wenn der erste Token einen gewissen Wert annimmt, die durch diesen Wert bedingte Verteilung genannt.

[0022] Im Folgenden wird unter einem anomalen Betriebszustand eines Computersystems ein Zustand des Computersystems verstanden, der nicht der normalen Benutzung des Systems entspricht. Dies kann aus einer Veränderung des Nutzerverhaltens, Änderungen äußerer Einflüsse und/oder einer Manipulation der intern ablaufenden Prozesse entstehen. Da Systeme unter anomalen Betriebszuständen i.d.R. eine Änderung in der Abfolge und Eigenschaften der Ereignisse verursachen, kann dies auch in den Protokolldateien erkannt werden. Eine mögliche Ausprägung dieser Veränderungen ist es, dass die im Abschnitt zum normalen Betriebszustand beschriebenen Muster, nicht mehr in neu erstellten Protokollzeilen vorhanden sind. Da Protokollzeilen, welche unter normalem Betriebszustand generiert wurden diesen Mustern folgen, kann im Umkehrschluss ein anomaler Betriebszustand erkannt werden, wenn das Muster in neu erzeugten Protokollzeilen nicht fortgeführt wird. Die Protokollzeilen, welche von den Mustern des normalen Betriebszustandes abweichen, werden auch Anomalien genannt.

[0023] Bei einer vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann vorgesehen sein, dass in der Testphase der Chiquadrattest auf Homogenität als Test auf Homogenität verwendet wird. Ein Beispiel eines derartigen Chiquadrattests auf Homogenität ist z.B. beschrieben in Bolboacă, S. D., Jäntschi, L., Sestraş, A. F., Sestraş, R. E., & Pamfil, D. C. (2011). Pearson-Fisher chi-square statistic revisited. Information, 2(3), 528-545.

[0024] Zur weiteren Einsparung von Rechenzeit und -leistung kann vorgesehen sein, dass bei der Bildung der Tokenpaare, die Verteilungen der einzelnen Werte aller Token ermittelt werden und jeweils nur Tokenpaare aus Token gebildet werden, deren Werte eine ähnliche Verteilung aufweisen. Die Überprüfung, ob die Verteilungen ähnlich sind, entspricht einem Test auf Homogenität und kann insbesondere mit einem Chiquadrat Test gelöst werden.

[0025] Um die Anzahl der zu überprüfenden Korrelationen zu reduzieren und derart weiter Zeit und Rechenleistung zu sparen, kann vorgesehen sein,

- dass bei der Bildung der Tokenpaare zu jedem Token eine Wertemenge ermittelt wird, wobei die Wertemenge als Werte diejenigen Teilzeichenabfolgen der Protokollzeilen der Stichprobe umfasst, die dem jeweiligen Token zugehörig sind, und
- dass jeweils nur Token gepaart werden, bei denen die Werte der jeweiligen Wertemengen einander, insbesondere zu einem vorgegebenen Prozentsatz, überschneiden.

[0026] Eine weitere Einsparung von Rechenzeit und -leistung kann erzielt werden, wenn bei der Bildung der Tokenpaare nur Token berücksichtigt werden, bei denen die Auftrittswahrscheinlichkeit des häufigsten Werts, den diejenigen Teilzeichenabfolgen der Protokollzeilen in der Stichprobe annehmen, die dem jeweiligen Token zugehörig sind, einen vorgegebenen Schwellwert überschreitet.

[0027] Um zufällige entstandene Korrelationen zwischen zwei Token, die auf keinem tatsächlich existierenden Zusammenhang zwischen den Werten der Token beruhen, besonders effektiv von einer weiteren Überprüfung auszuschließen, kann vorgesehen sein,

- dass in der Initialisierungsphase die einzelnen ermittelten Korrelationen zu den einzelnen Tokenpaaren in einem Überprüfungsschritt dahingehend überprüft werden, ob die jeweils ermittelten Korrelationen vorgegebene Bedingungen erfüllen, welche einem Test entsprechen, ob die einzelnen ermittelten Korrelationen nicht zufällig sind, und
- dass in der Testphase nur diejenigen ermittelten Korrelationen berücksichtigt werden, die diese

vorgegebenen Bedingungen erfüllen, indem sie in der jeweiligen Überprüfung als nicht zufällig erkannt werden.

[0028] Um zufällig entstandene Korrelationen zwischen Token besonders effektiv von einer weiteren Überprüfung auszuschließen, kann weiters vorgesehen sein, dass, im Fall, dass im Überprüfungsschritt mehrere verschiedene Überprüfungen der Korrelationen hintereinander vorgenommen werden, nur diejenigen Korrelationen in der Testphase berücksichtigt werden, die in zumindest einer Überprüfung, insbesondere allen Überprüfungen, als nicht zufällig erkannt wurden.

[0029] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann weiters vorgesehen sein,

- dass im Überprüfungsschritt die Korrelationen überprüft werden, indem mittels eines goodness of fit-Tests getestet wird, ob die angenommenen Werte des zweiten Tokens, des jeweiligen Tokenpaares in den Protokollzeilen der Stichprobe welche dem der Korrelation zugehörigen Wert im ersten Token besitzen, aus der Multinomialverteilung entstammen, in denen das Auftreten aller Werte gleich wahrscheinlich ist und
- dass die jeweilige Korrelation diese Überprüfung bestanden hat und als nicht zufällig erkannt wird, wenn der goodness of fit-Test negativ ist.

[0030] Ein Beispiel eines derartigen goodness of fit-Tests ist beispielsweise beschrieben in Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(2), 281-294. Die Multinomialverteilung ist beispielsweise beschrieben in: Epitropakis, M. G., Tasoulis, D. K., Pavlidis, N. G., Plagianakos, V. P., & Vrahatis, M. N. (2012, May). Tracking differential evolution algorithms: An adaptive approach through multinomial distribution tracking with exponential forgetting. In Hellenic Conference on Artificial Intelligence (pp. 214-222). Springer, Berlin, Heidelberg.

[0031] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann weiters vorgesehen sein,

- dass im Überprüfungsschritt zu den Korrelationen eines Tokenpaares jeweils die bedingten Verteilungen der Werte, die der zweite Token des jeweiligen Tokenpaares in der Stichprobe annimmt, zu den Korrelationen eines Tokenpaares, mittels eines Tests auf Homogenität, überprüft werden, ob diese derselben Verteilung angehören, und
- dass, sollte der Test auf Homogenität positiv ausfallen, alle Korrelationen dieses Tokenpaares die Überprüfung nicht bestanden haben und als zufällig erkannt werden.

[0032] Ein derartiger Test auf Homogenität ist z.B. beschrieben in: Kusolitsch, N. (2014). Maß- und Wahrscheinlichkeitstheorie: Eine Einführung. Springer-Verlag.

[0033] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann weiters vorgesehen sein,

- dass im Überprüfungsschritt die Korrelationen überprüft werden, indem
 - die Häufigkeiten der Werte des ersten Tokens des Tokenpaares ermittelt werden und
 - die Häufigkeiten derjenigen Werte des ersten Tokens aggregiert, insbesondere summiert, werden, für welche Korrelationen ermittelt wurden, und
- dass alle Korrelationen die Überprüfung nicht bestanden haben und als zufällig erkannt werden, wenn der derart berechnete aggregierte Wert, insbesondere die derart berechnete Summe, einen vorgegebenen Schwellenwert unterschreitet.

[0034] Für eine besonders effiziente Aktualisierung der Korrelationen basierend auf neu erstellten Protokollzeilen kann vorgesehen sein, dass in der Testphase zum Testen der Korrelationen die Korrelationen, insbesondere nur dann wenn das Testresultat positiv war, aktualisiert werden, indem für die Korrelationen die bedingten Verteilungen, die auf Basis der jeweiligen Stichprobe festgelegt wurden, auf Basis dieser Stichprobe zuzüglich der Stichprobe dieses Updateschrittes festgelegt werden.

[0035] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann in diesem Zusammenhang weiters vorgesehen sein, dass, insbesondere mittels eines Bi-

nomialtests, ermittelt wird, ob sich die negativen Testresultate einer jeweiligen Korrelation häufen und im Anschluss diese Korrelation verworfen wird, sofern dies der Fall ist.

Ein Beispiel eines derartigen Binomialtests ist z.B. beschrieben in Henze, N. (1997). Stochastik für Einsteiger (Vol. 8). Vieweg.

[0036] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann in diesem Zusammenhang weiters vorgesehen sein, dass neue Korrelationen gesucht werden, indem in der Testphase, insbesondere regelmäßig, einzelne Tokenpaare, für die zuvor keine Korrelationen gefunden wurden, oder deren Korrelationen verworfen wurden, einer neuerlichen Initialisierungsphase unterworfen werden, und dass die derart aufgefundenen neuen Korrelationen anschließend in der Testphase getestet werden.

[0037] Eine besonders zuverlässige Identifikation von anomalen Betriebszuständen kann erzielt werden, wenn zu den einzelnen Korrelationen, basierend auf einem Vergleich eines Abschnittes einer Liste der letzten Testresultate der Aktualisierungen der Korrelationen und einem, als Referenz dienenden, vergangenen Abschnitt der Liste der letzten Testresultate der Aktualisierungen der Korrelationen jeweils ein Korrelations-Indikator berechnet wird, der die Wahrscheinlichkeit dafür angibt, dass die jeweils zuletzt neu erstellten Protokollzeilen von der jeweiligen Korrelation abgewichen sind.

[0038] Eine weitere Verbesserung bei der Detektion von anomalen Betriebszuständen kann vorteilhafterweise erzielt werden, wenn für die Berechnung der Indikatoren zusätzlich zu den Testresultaten, die bei den Tests ermittelten p-Werte für die Berechnung des Indikators herangezogen werden.

[0039] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann in diesem Zusammenhang vorgesehen sein, dass alle Korrelations-Indikatoren eines Tokenpaares zu einem Tokenpaar-Indikator zusammengefasst werden und/oder alle Korrelations-Indikatoren eines Ereignistyps zu einem Ereignistyp-Indikator zusammengefasst werden, wobei der Tokenpaar-Indikator und der Ereignistyp-Indikator jeweils die Wahrscheinlichkeiten dafür angeben, dass die jeweils zugrundeliegenden Protokollzeilen von einer der zusammengefassten Korrelationen abgewichen sind.

[0040] Gemäß einer besonders vorteilhaften Variante eines erfindungsgemäßen Verfahrens kann weiters vorgesehen sein,

- dass in der Initialisierungsphase Ereignistypen ermittelt werden, deren Protokollzeilen in einem vorgegebenen zeitlichen Abstand zueinander, insbesondere unmittelbar nacheinander, vom Computersystem erstellt werden, und
- dass Tokenpaare erstellt werden, die jeweils einen, einen diskreten Datentyp aufweisenden, Token von verschiedenen der derart ermittelten Ereignistypen umfassen.

[0041] Ein Programm zur Durchführung eines Verfahrens kann vorteilhafterweise auf einem Datenträger abgespeichert und derart mobil sein.

[0042] Weitere Vorteile und Ausgestaltungen der Erfindung ergeben sich aus der Beschreibung und der beiliegenden Zeichnung.

[0043] Beispielhafte Ausführungsformen der Erfindung sind in der Zeichnung schematisch dargestellt und werden im Folgenden unter Bezugnahme auf die Zeichnung beispielhaft beschrieben.

[0044] Im Folgenden zeigt Fig. 1 eine schematische Darstellung eines erfindungsgemäßen Verfahrens mit den erforderlichen, sowie optionalen Verfahrensschritten.

[0045] Ein Beispiel von Protokollzeilen eines Ereignistyps mit dem zugehörigen Repräsentanten ist in Tabelle 1 (Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen) gegeben. Das Template des betrachteten Ereignistyps ist in der ersten Zeile der Tabelle gegeben. Die einzelnen Einträge der Spalten repräsentieren dabei den Wert eines statischen Tokens T bzw. eine Markierung (§) eines variablen Tokens T des Templates. Zeilen zwei bis sechs sind Protokollzeilen zu dem Ereignistyps, welche entlang der aller Leerzeichen getrennt wurden. Aus Über-

sichtlichkeitsgründen wurden aneinander angrenzende statische Token T zu jeweils einem zusammengefügt. Die Zeile sieben beschreibt die Datentypen der Token T. Dies ist vorteilhaft, da mit dem erfindungsgemäßen Verfahren diskrete Token T untersucht werden.

[0046] Tabelle 1: Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen

§		§	(§) sent request to	§
Jul-17-11:30:18		user1	(10.2.12.9) sent request to	10.4.0.5
Jul-17-11:32:27		user2	(10.2.12.6) sent request to	10.4.0.7
Jul-17-11:33:22		user2	(10.2.12.9) sent request to	10.4.0.7
Jul-17-11:35:58		user3	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:36:43		user1	(10.2.12.6) sent request to	10.4.0.5
chronologisch	statisch	diskret	statisch	diskret	statisch	diskret

[0047] Viele Systeme erlauben keinen direkten Einblick in die laufenden Prozesse, stellen jedoch eine Protokolldatei zur Verfügung. Auf Basis dieser Protokolldaten ist es möglich, das Normalverhalten bzw. ein Modell des erwarteten Verhaltens zu erstellen. Die Protokolldatei kann daher weiters dazu genutzt werden um zu schlussfolgern, welche Ereignisse in Folge des Ablaufs eines Prozesses ausgeführt werden/wurden. Da die Angriffe auf Systeme die Systemabläufe i.d.R. verändern, können diese Veränderungen auch in der Protokolldatei erkannt werden. Die Protokollzeilen einer Protokolldatei, die sich von den erwarteten Protokollzeilen unterscheiden, nennt man Anomalien. Anomalien können in Protokolldateien entdeckt werden, indem die Protokolldateien des Systems unter Normalbetrieb, über längere Zeit, auf Muster hin analysiert werden und anschließend im nachfolgenden Betrieb beobachtet wird, ob diese Muster weiter bestehen. Sollten Inhalte der Protokolldateien die Muster nicht fortführen, werden diese Inhalte als Anomalien erkannt. Aus diesem Grund können Protokolldateien zur Anomalieerkennung genutzt werden.

[0048] Eine spezielle Art von Mustern in Protokolldaten sind Korrelationen zwischen Werten von zwei Token T eines Ereignistyps. Eine Korrelation zwischen zwei Token T_1 , T_2 liegt vor, wenn während einer Beobachtung der angenommene Wert eines Tokens T_1 den Wert des anderen Tokens T_2 mitbestimmt, oder beeinflusst.

[0049] Korrelationen sind Beziehungen die in beide Richtungen wirken, da sobald in eine Richtung vorhanden auch in die andere Richtung ein Einfluss besteht. Im Allgemeinen ist man nicht an allen Korrelationen interessiert, sondern nur an jenen, die gewisse Anforderungen an die Protokolldaten und die Analysemethoden in einem spezifischen Anwendungsfall erfüllen. Dies können zum Beispiel eine maximale Anzahl an angenommenen Werten des beeinflussten Tokens T bei festgehaltenen Wert des jeweils anderen Tokens T sein, oder, dass nur Korrelationen gesucht werden, bei der die höchste Wahrscheinlichkeit, der angenommenen Werte des beeinflussten Tokens T_2 bei festgehaltenen Wert des jeweils anderen Tokens T_1 , einen gewissen Prozentwert überschreitet.

[0050] Hier sei noch erwähnt, dass, wenn eine Korrelation in eine Richtung besteht, die den Anforderungen genügt, nicht unbedingt auch eine Korrelation in die entgegengesetzte Richtung bestehen muss, die ebenfalls den Anforderungen genügt. Da für die Ausgabe des erfindungsgemäßen Verfahrens nur jene Korrelationen von Interesse sind, welche die Anforderungen erfüllen, werden die Richtungen in denen die Korrelationen wirken berücksichtigt.

[0051] Hier sei noch darauf hingewiesen, dass, sollte die Korrelation in beide Richtungen den Anforderungen genügen, beide Richtungen in den folgenden Ausführungsbeispielen getrennt behandelt werden. Im Folgenden wird nur noch von Korrelationen gesprochen, ohne explizit die Anforderungen oder die Richtungen zu erwähnen. Um die Notation kompakter zu halten, wird derjenige Token T dessen Werte Einfluss auf einen anderen Token T besitzen, als der erste Token T_1 bezeichnet und derjenige Token T, dessen Wert beeinflusst wird, als der zweite Token T_2 bezeichnet.

[0052] Der Ansatz eines erfindungsgemäßen Verfahrens ist für Token T , welche diskrete Werte annehmen, konzipiert. Diskrete Werte sind Werte, die kategorisch sind und sich dadurch auszeichnen, dass die Auswahl der angenommenen Werte aus einer begrenzten Menge entstammt. Eine Erweiterung der Methoden eines erfindungsgemäßen Verfahrens auf andere Datentypen, zB kontinuierliche Datentypen, ist nicht ohne weiteres möglich. Weiters ist es ebenfalls nicht ohne weitere Überlegungen möglich, Korrelationen zwischen Token T mit unterschiedlichen Datentypen zu generieren und zu testen.

[0053] Um Korrelationen in Protokolldaten zu finden, werden bei einem erfindungsgemäßen Verfahren die Verteilungen der Werte des zweiten Tokens T_2 , bei festgehaltenen Werten des ersten Tokens T_1 erzeugt und analysiert. Das heißt, dass für jeden Wert, den der erste Token T_1 annehmen kann, eine Verteilung generiert wird, welche der empirischen Verteilung der Werte des zweiten Tokens T_2 entspricht, die jeweils auf Basis all jener Protokollzeilen erstellt wird, die den Wert in dem ersten Token T_1 angenommen haben. Im Folgenden wird die Verteilung der Werte des zweiten Tokens T_2 , wenn der erste Token T_1 einen gewissen Wert annimmt, die durch diesen Wert bedingte Verteilung genannt. Sollte aus dem Kontext klar ersichtlich sein, welcher Wert gemeint ist, wird von der bedingten Verteilung gesprochen.

[0054] In der Analyse werden die bedingten Verteilungen überprüft, ob sie aus einer Korrelation hervorgegangen sein können, oder davon ausgegangen werden kann, dass keine Korrelation für diesen Wert des ersten Tokens T_1 besteht.

[0055] Ein Problem, welches mit dem Auffinden von Korrelationen in Protokolldaten verbunden ist, ist das Problem von zufälligen Korrelationen. Man spricht von zufälligen Korrelationen, wenn die eingelesenen Protokolldaten eine Korrelation zwischen zwei Token T suggerieren, zwischen denen jedoch keine Korrelation existiert. Um diesem Phänomen entgegenzuwirken kann die Anzahl an Protokollzeilen erhöht werden oder schärfere Anforderungen an die Korrelationen gestellt werden.

[0056] Eine spezielle Art von Korrelationen sind die $n:1$ - und $1:1$ -Korrelationen. $n:1$ -Korrelationen sind Korrelationen, in denen der Wert des ersten Tokens T_1 den Wert des zweiten Tokens T_2 mitbestimmt. Das heißt, dass es zu jedem angenommenen Wert des ersten Tokens T_1 einen eindeutigen Wert für den zweiten Token T_2 gibt, sodass für alle Protokollzeilen, die den gleichen Wert im ersten Token T_1 angenommen haben, der gleiche Wert im zweiten Token T_2 angenommen wird. $1:1$ -Korrelationen sind Korrelationen zwischen zwei Token T_1, T_2 , bei der in beide Richtungen eine $n:1$ -Korrelation besteht. Das heißt, dass der Wert jeder der beiden Token T_1, T_2 den Wert der jeweils anderen Token T_1, T_2 mitbestimmt.

[0057] Eine annähernde $n:1$ -Korrelation ist eine Korrelation, bei der zu jedem angenommenen Wert des ersten Tokens T_1 es einen eindeutigen Wert für den zweiten Token T_2 gibt, sodass für alle Protokollzeilen, die den ersten Wert im ersten Token T_1 angenommen haben, der prozentuelle Anteil der Protokollzeilen, welche den zweiten Wert im zweiten Token T_2 angenommen haben, höher ist, als ein vorab festgelegter Schwellwert. Ein für die meisten repräsentativen Anwendungen sinnvoll gewählter Schwellwert ist z.B. 95%. Analog zu der Definition einer $1:1$ -Korrelationen ist eine annähernde $1:1$ -Korrelation eine Korrelation zwischen zwei Token T_1, T_2 , bei der in beide Richtungen eine annähernde $n:1$ -Korrelationen besteht.

VORAUSSETZUNGEN

[0058] Voraussetzung für die Erkennung von Anomalien gemäß einem erfindungsgemäßen Verfahren ist, dass den einzelnen Protokollzeilen zunächst basierend auf deren Syntax, insbesondere aufgrund der Ähnlichkeit der einzelnen Protokollzeilen zueinander und der die Protokollzeilen aufbauenden Teilzeichenabfolgen, jeweils zumindest ein Ereignistyp zugewiesen wird.

[0059] Aus dem Stand der Technik sind verschiedenste Vorgehensweisen bekannt, wie jeder Protokollzeile basierend auf deren Syntax ein Ereignistyp oder ggf. mehrere Ereignistypen zugewiesen werden können. Beispiele hierfür sind:

1. Ein Parser der mittels eines Parsergenerators (siehe EP19169705.1, A50461/2018) erstellt wurde, kann dazu genutzt werden, um den einzelnen Protokollzeilen einen Ereignistyp zuzuordnen. Allerdings fehlt dem Parser die Möglichkeit die Werte einzelner Token T zu analysieren.
2. Protokollzeilen werden zunächst nach ihrer Ähnlichkeit geclustert (siehe EP3267625, A50601/2016) Daher entsprechen alle Protokollzeilen, die demselben Cluster zugeordnet wurden einem Ereignistyp. Allerdings fehlt dem Clustering die Möglichkeit die Werte einzelner Token T zu analysieren.

[0060] Weitere Voraussetzung für das erfindungsgemäße Verfahren ist, dass anschließend für jeden Ereignistyp ein Template umfassend eine Anzahl von Token T, erstellt wird, indem die einzelnen Protokollzeilen an vorgegebenen Trennzeichen in Teilzeichenabfolgen unterteilt werden.

[0061] Wie bereits zuvor erwähnt, bezeichnen Teilzeichenabfolgen die einzelnen Teile einer Protokollzeile, die durch Separieren der Zeichenabfolge an vorgegebenen Trennzeichen, wie zum Beispiel Leerzeichen oder Beistrichen etc. entstehen. Diese können durch Tokens T repräsentiert werden. Jeder Ereignistyp lässt sich daher durch ein Template, bestehend aus einer Abfolge von Token T, beschreiben.

[0062] Das Template beschreibt somit die Struktur, d.h. die Abfolge der Teilzeichenabfolgen, der zugeordneten Protokollzeilen, und weist eine Anzahl von statischen Token, die nur einen Wert in allen Protokollzeilen besitzen, auf, die in ihrer Reihenfolge in allen Protokollzeilen des jeweiligen Ereignistyps vorhanden sind, und durch variable Token, die mindestens zwei Werte in den Protokollzeilen annehmen können, unterbrochen werden.

[0063] D.h. die Templates bestehen aus statischen Token, z.B. statischem Text, der in allen Protokollzeilen, die dem Ereignistyp entsprechen an derselben Stelle steht, und aus variablen Token T, z.B. variablem Text, die durch sogenannte Wildcards gekennzeichnet sein können, die an Stellen eingefügt werden, an denen die Zeilen variable Werte beinhalten.

[0064] Die einzelnen ermittelten Templates und deren Token T sind charakteristisch für die Protokolldatei. Da die Werte der Token T Auskunft über die Prozesse des protokollieren Systems geben, können aus den Korrelationen der Token T Rückschlüsse auf das Systemverhalten gezogen werden.

[0065] Für eine derartige Unterteilung der einzelnen Protokollzeilen an vorgegebenen Trennzeichen in Teilzeichenabfolgen bzw. eine derartige Tokenisierung der einzelnen Protokollzeilen sind aus dem Stand der Technik ebenfalls verschiedenste Vorgehensweisen bekannt, wie z.B.

1. Es besteht die Möglichkeit aus einem durch einen mittels eines Parsergenerators (siehe EP19169705.1, A50461/2018) erstellten Parserbaums, Templates für die jeweiligen Ereignistypen zu generieren, indem man die Token T der Knoten entlang eines Pfades des Parserbaums aneinanderfügt. Da der Parserbaum sowohl statische als auch variable Token T enthält, erhält man ein Template, das aus statischen und variablen Token T besteht. Die vorliegende Erfindung kann nun dazu genutzt werden, um die Werte der als variabel erkannten Token T zu analysieren. Diese Funktionalität fehlt dem Parsergenerator und dem Parser.
2. Ein Templategenerator (siehe A50285/2019, EP20160854.4) kann genutzt werden um nachdem ein Clustering durchgeführt wurde, für jeden Cluster bzw. Ereignistyp ein Template zu erstellen, das aus statischen und variablen Token T besteht. Die vorliegende Erfindung kann nun dazu genutzt werden, um die Werte der als variabel erkannten Token T zu analysieren. Diese Funktionalität fehlt sowohl dem Clustering als auch dem Templategenerator.
3. Eine auf Datentypen basierende Anomalieerkennung (siehe A50642/2020) kann genutzt werden, um jene Token T zu identifizieren, welche einen diskreten Datentyp aufweisen. Die vorliegende Erfindung kann nun dazu genutzt werden, um die Korrelationen zwischen den Werten der Token T mit diskretem Datentyp zu analysieren. Diese Funktionalität fehlt den auf Datentypen basierenden Anomalieerkennungen, da diese Ansätze nur einzelne Token T, jedoch nicht Korrelationen zwischen Token T analysieren.

[0066] Im Folgenden wird das erfindungsgemäße Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems zunächst allgemein beschrieben:

ALLGEMEINE BESCHREIBUNG EINES ERFINDUNGSGEMÄßEN VERFAHRENS

[0067] Der Aufbau eines erfindungsgemäßen Verfahrens kann in acht Schritte unterteilt werden, wobei hier auch einige optionale Schritte umfasst sind. Der Aufbau und die Unterteilung in Schritte sind in Fig. 1 schematisch dargestellt.

[0068] Generell werden zunächst die Daten bzw. Protokollzeilen aufbereitet (bezeichnet mit Aufbereitung der Daten „AD“ in Fig. 1). In diesem Aufbereitungs-Schritt werden die Protokollzeilen in Teilzeichenabfolgen unterteilt und jeder Protokollzeile wird ein Ereignistyp zugeordnet, wie dies z.B. in Wurzenberger, M., Landauer, M., Skopik, F., & Kastner, W. (2019, April). AECID-PG: A Tree-Based Log Parser Generator To Enable Log Analysis. In 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM) (pp. 7-12). IEEE. beschrieben ist. Die einzelnen Teilzeichenabfolgen werden für die folgenden Analysen als Werte zu den einzelnen Token T gespeichert.

[0069] Anschließend werden folgende Abschnitte jeweils für die Protokollzeilen der einzelnen Ereignistypen separat durchgeführt, d.h. bei jeder Ausführung der im Folgenden beschriebenen Abschnitte werden nur Protokollzeilen eines Ereignistyps verarbeitet.

[0070] Ausgehend von (AD) ist Fig. 1 in zwei Abschnitte unterteilt. Der Abschnitt von (AD) nach (UEK) wird die Initialisierung der Korrelationen (IK) genannt und der Abschnitt von (AD) nach (AR) stellt die Testphase mit einem Aktualisieren der Korrelationen (AK) dar. Die einzelnen Schritte, aus denen die Abschnitte bestehen, werden im Anschluss beschrieben.

[0071] Für den Ablauf eines erfindungsgemäßen Verfahrens werden Protokollzeilen aufbereitet, indem ihnen, wie zuvor beschrieben, jeweils ein Ereignistyp zugeordnet wird und sie entsprechend des zugehörigen Templates in Teilzeichenabfolgen unterteilt werden. Sobald eine gewisse Anzahl aufbereitet wurde, werden die Korrelationen initialisiert. In der Initialisierungsphase werden alle bedingten Verteilungen erzeugt und anschließend einzeln überprüft und entfernt, falls es wahrscheinlich ist, dass einzelne bedingte Verteilungen zufälligen Korrelationen entsprechen.

[0072] Die Initialisierungsphase kann im Laufe des normalen Betriebs wiederholt werden, wenn regelmäßig neue Korrelationen initialisiert werden sollen. Das kann vorteilhaft sein, da im normalen Betrieb eines erfindungsgemäßen Verfahrens gefundene Korrelationen nicht mehr getestet werden, wenn eingelesene Protokolldaten diese Korrelationen über einen längeren Zeitraum nicht beinhalten. Um entgegenzuwirken, dass bei langen Beobachtungszeiträumen die überprüften Korrelationen immer weiter reduziert werden, können regelmäßig Korrelationen neu initialisiert werden.

[0073] Wenn die Initialisierung erfolgt ist, werden die erzeugten Korrelationen, mit einer gewissen Anzahl an aufbereiteten Protokollzeilen, regelmäßig getestet und deren Verteilungen aktualisiert. Bei der Anzahl der Protokollzeilen im Initialisierungs- und Updateschritt, kann der Benutzer vorteilhafterweise darauf achten, dass die Anzahl nicht zu gering ist, um eine adäquate Analyse der Protokollzeilen zu gewährleisten.

[0074] Generell umfasst ein erfindungsgemäßes Verfahren die im Folgenden beschriebenen zwingenden Schritte, die in Fig. 1 mit einer durchgehenden Umrandung gekennzeichnet sind, sowie eine Reihe von optionalen Schritten, die in Fig. 1 mit einer strichlierten Umrandung gekennzeichnet sind.

1. Aufbereiten der Daten (AD in Fig. 1): In diesem Schritt werden die Protokollzeilen in Token T unterteilt und jeder Protokollzeile wird ihr Ereignistyp zugeordnet, wie dies z.B. in Wurzenberger, M., Landauer, M., Skopik, F., & Kastner, W. (2019, April). AECID-PG: A Tree-Based Log Parser Generator To Enable Log Analysis. In 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM) (pp. 7-12). IEEE, beschrieben ist. Die Werte der einzelnen Token T werden für die folgenden Analysen gespeichert. Es ist zu beachten, dass bei jeder Ausführung der im Folgenden beschriebenen Schritte nur Protokollzeilen eines

Ereignistyps verarbeitet werden.

2. Selektieren (ST in Fig. 1) der überprüften Token T: Hier können optional diejenigen Token T ausgewählt werden, zwischen deren Werten nach Korrelationen gesucht wird. Es werden nur Token T ausgewählt, deren Werte diskret sind. Diese Aufgabe kann z.B. wie in der österreichischen Patentanmeldung A50642/2020, eingereicht am 28. Juli 2020, ausgeführt werden, da den gesuchten Token T der diskrete Datentyp zugeordnet wird. Dieser Schritt kann optional sein, wenn bekannterweise nur Token T mit diskreten Werten auftreten und somit keine Selektion notwendig ist.

Zusätzlich können, um den Prozess zu beschleunigen, Token T, die primär einen einzigen Wert annehmen, von der Auswahl ausgeschlossen werden. Die Korrelationen derartiger Token T beispielsweise ebenfalls durch das in der österreichischen Patentanmeldung A50642/2020 beschriebene Verfahren analysiert werden.

3. Auswahl (AT in Fig. 1) der Tokenpaare TP: In diesem Schritt werden nach Korrelationsrichtung geordnete Tokenpaare TP bestehend aus jeweils zwei Token gebildet. Die Tokenpaare TP werden in den nachfolgenden Schritten auf mögliche Korrelationen, die den Einfluss der Werte des ersten Tokens T_1 des Paares auf die Werte des zweiten Tokens T_2 angeben, überprüft.

Die Paarbildung, welche standardmäßig erfolgen kann, ist jeden Token T mit jedem anderen Token T zu paaren, wobei für jede Auswahl von zwei Token T zwei unterschiedlich geordnete Tokenpaare TP gebildet werden, die jeweils eine der beiden möglichen Richtungen der Korrelation repräsentieren. Diese Auswahl ermöglicht, dass in diesem Schritt keine Token T von gesuchten Korrelationen ausgeschlossen werden. Eine weitere Auswahlmöglichkeit für die Bildung von Tokenpaaren TP ist, die Verteilungen der angenommenen Werte der Token T zu vergleichen und nur Token T zu paaren, die eine ähnliche Verteilung ihrer Werte besitzen. Dieser Ansatz ermöglicht ein schnelleres Auffinden von annähernden 1:1-Korrelationen.

Eine weitere Möglichkeit ist es, die angenommenen Werte zu betrachten und Token T zu paaren die einen ähnlichen Wertebereich besitzen. Beispiele für Korrelationen die auf diese Weise auffindbar sind, sind Kommunikationen zwischen mehreren Geräten, Benutzern, etc.

4. Erstellen der Korrelationen (EK in Fig. 1): In diesem Schritt werden die bedingten Verteilungen der Korrelationen generiert, die später überprüft und aktualisiert werden. Dies erfolgt in der Form, dass für jeden Wert des ersten Tokens T_1 , die durch diesen Wert bedingte Verteilung berechnet und abgespeichert wird.
5. Überprüfen der Korrelationen (UEK in Fig. 1): Nachdem die Korrelationen generiert wurden, kann in diesem Schritt optional überprüft werden, ob sich diese Korrelationen wie echte Korrelationen zwischen den Werten verhalten. Dieser Schritt ist optional, da in gewissen Protokolldaten eine Korrelation zwischen allen Token T angenommen werden kann. Diese Protokolldaten besitzen meist eine geringe Anzahl an Ereignistypen und Anzahl von Token T pro Ereignistyp.

Sollte eine Korrelation einer Überprüfung nicht standhalten, wird die durch den betrachteten Wert bedingte Verteilung im Anschluss nicht mehr getestet oder aktualisiert. Für das Überprüfen der Korrelationen können verschiedene Ansätze, wie im Weiteren näher erläutert, gewählt werden, womit die resultierenden Korrelationen, durch das Anpassen der gestellten Anforderungen der Tests, angepasst werden können um die gesuchten Korrelationen zu erzeugen. Beispiele für diese Anforderungen können sein, dass nur nach annähernd n:1-Korrelationen gesucht wird, Korrelationen bei deren bedingten Verteilungen ein Wert zumindest 50% Auftrittswahrscheinlichkeit erreicht, etc.

Diese und weitere Ansätze können einzeln oder hintereinander gereiht ausgeführt werden, um die Korrelationen zu finden, die die gestellten Anforderungen erfüllen. Bei Verwendung von mehreren Ansätzen können auch jene Korrelationen ausgewählt werden, welche min-

destens eine der Anforderungen erfüllen, um alle Anforderungen mit den Korrelationen abzudecken.

Der erste Ansatz ist, die generierten bedingten Verteilungen auf Ähnlichkeit zu jener Verteilung zu überprüfen, bei der alle Werte gleich wahrscheinlich sind. Sind bedingte Verteilungen ähnlich zu dieser Verteilung, deutet dies darauf hin, dass diese bedingte Verteilungen zu keinen echten Korrelationen gehören.

Ein weiterer Ansatz ist es, alle bedingten Verteilungen von einem Tokenpaar TP untereinander zu vergleichen. Sollten sie eine große Ähnlichkeit aufweisen, deutet dies darauf hin, dass der angenommene Wert des ersten Tokens T_1 keinen Einfluss auf den Wert des zweiten Tokens T_2 besitzt und daher keine Korrelation zwischen den Werten besteht.

Ein dritter möglicher Ansatz benötigt mindestens eine andere vorangestellte Überprüfung, bei der Korrelationen entfernt werden, die der Überprüfung nicht standhalten, um zu funktionieren. Dieser Ansatz betrachtet nicht die einzelnen bedingten Verteilungen, sondern überprüft die Häufigkeiten der Werte des ersten Tokens T_1 jedes Tokenpaares TP. Wenn die Summe der Häufigkeiten der Werte, die für die Korrelation angenommen wurden, sehr gering ist, deutet dies darauf hin, dass keine Korrelation in der untersuchten Richtung zwischen den Token T besteht, sondern die Korrelationen nur zufällig entstanden sind. Sollte dieser Fall eintreten, werden alle Korrelationen dieses Tokenpaares TP gelöscht.

6. Testen der Korrelationen (TK in Fig. 1): In diesem Schritt werden die Korrelationen auf Basis von neu eingelesenen Daten bzw. Protokollzeilen getestet und optional nachfolgend aktualisiert. Dafür wird für jede gefundene Korrelation überprüft, ob die Werte der neu aufbereiteten Protokollzeilen den bedingten Verteilungen entsprechen, welche in der Korrelation definiert sind. Für die Überprüfung der Korrelationen können Tests auf Homogenität, wie in Kusolitsch, N. (2014). Maß- und Wahrscheinlichkeitstheorie: Eine Einführung. Springer-Verlag. beschrieben, insbesondere ein Chiquadratstest auf Homogenität, wie in Bolboacă, S. D., Jäntschi, L., Sestraş, A. F., Sestraş, R. E., & Pamfil, D. C. (2011). Pearson-Fisher chi-square statistic revisited. Information, 2(3), 528-545. beschrieben, eingesetzt werden. Das Resultat des Tests wird abgespeichert. Optional kann, insbesondere nur dann, wenn das Testresultat positiv war, die bedingte Verteilung mit Hilfe der neuen Werte aktualisiert werden, indem die Verteilung der Werte des Aktualisierungsschrittes proportional in die getestete bedingte Verteilung einfließt.

Sollten sich zu einer Korrelation die negativen Resultate häufen, kann optional die Korrelation verworfen werden. Für die Entscheidung, wann sich negative Resultate häufen, kann insbesondere ein Binomialtest, wie in Henze, N. (1997). Stochastik für Einsteiger (Vol. 8). Vieweg. beschrieben, eingesetzt werden.

Um das Auftreten neuer Korrelationen zu ermöglichen und um entgegenzuwirken, dass bei langen Beobachtungszeiträumen die Anzahl der überprüften Korrelationen immer weiter reduziert wird, können regelmäßig Tokenpaare TP ohne Korrelationen neu initialisiert werden und in den Testablauf eingebunden werden.

7. Berechnung (BI in Fig. 1) von Indikatoren: Dieser Schritt ist optional. Für manche Protokoll-daten sind die Korrelationen zwischen Token T zu schwach ausgeprägt, oder die Menge der zu analysierenden Protokollzeilen zu groß, als dass die Ausgabe aller negativ getesteten Korrelationen praktikabel wäre. In diesem Schritt werden die Testresultate der Aktualisierungen der Korrelationen analysiert und auf dessen Basis Korrelations-Indikatoren berechnet, die die Wahrscheinlichkeit des Abweichens der Protokollzeilen von dieser Korrelation angeben. Die Berechnung erfolgt auf Basis eines Abschnittes der Liste der letzten Testresultate M_{Akt} und einem vergangenen Abschnitt der Liste der letzten Testresultate M_{Ref} , die als Referenz dient, durchgeführt.

Alternativ können statt der Liste der Testresultate, die Liste der p-Werte der Tests benutzt werden, um mehr Information der Tests in die Berechnung des KorrelationsIndikators I_K einfließen zu lassen. Da man auf Basis des p-Werts das Testresultat ableiten kann, geht

dadurch keine Information verloren.

Das Resultat des Korrelations-Indikators I_K ist höher, je mehr sich die Resultate/Werte der Listen M_{Akt} und M_{Ref} unterscheiden.

Anschließend an die Berechnung der Korrelations-Indikatoren, können Tokenpaar-Indikatoren I_{TP} und/oder Ereignistyp-Indikatoren I_E berechnet werden, die die Information aller Korrelations-Indikatoren I_K aller Korrelationen eines Tokenpaares TP bzw. aller Korrelations-Indikatoren I_K aller Korrelationen aller Tokenpaare TP zu einem Ereignistyp, zusammenfassen.

8. Ausgabe der Resultate (AR in Fig. 1): In diesem optionalen Schritt kann der Verlauf der Häufungen von negativ getesteten Korrelationen bzw. berechneten Indikatoren ausgegeben werden, um für nachfolgende Untersuchungen oder Bewertungen zu dienen.
9. Anomalieerkennung (AE in Fig. 1): Ein anomaler Betriebszustand des Computersystems bzw. eine Anomalie wird erkannt, wenn beim Testen der Korrelationen (TK in Fig. 1) in der Testphase zumindest der Test von einer Korrelation ein negatives Resultat besitzt. Das bedeutet, dass die Werte der Token T der getesteten Protokollzeilen nicht der Regel der Korrelation entsprechen.

Optional kann, wie im Abschnitt Berechnung von Indikatoren „BI“ beschrieben, der Verlauf der Resultate der Tests zu den Korrelationen analysiert werden und ein Wert, der sogenannte Korrelations-Indikator I_K , berechnet, der die Wahrscheinlichkeit dafür angibt, dass die Regel zu einer Korrelation durch eine Anomalie gebrochen wurde. Zusätzlich können alle Ergebnisse der Korrelations-Indikatoren I_K eines Tokenpaares TP zu einem Tokenpaar-Indikator I_T oder allen Tokenpaaren TP eines Ereignistyps zu einem Ereignistyp-Indikator I_E zusammengefasst werden. In diesem Fall würde ein anomaler Betriebszustand des Computersystems bzw. eine Anomalie erkannt werden, wenn ein Wert der berechneten Korrelationen- und/oder Tokenpaar- und/oder Ereignistyp-Indikatoren einen gewissen Schwellwert überschreitet.

[0075] Man beachte, dass nur der Schritt 1 ein einziges Mal ausgeführt wird. Im Gegensatz zu den anderen Schritten, welche sich jeweils auf einen Ereignistyp beziehen und daher ausgeführt werden, sobald eine gewisse Anzahl an Protokollzeilen zu dem Ereignistyp aufbereitet wurden.

AUSFÜHRUNGSBEISPIELE EINES ERFINDUNGSGEMÄßEN VERFAHRENS

[0076] Bei den folgenden Ausführungsbeispielen werden zunächst die verwendeten Methoden und Tests ausformuliert. Danach wird das Verfahren schrittweise mit einem Beispieldatensatz ausgeführt. Die Tests sind ebenfalls beispielhaft und können für die Ausführung eines erfindungsgemäßen Verfahrens durch gleichwertige Tests ersetzt werden. Die Ausführungsbeispiele für die Schritte (ST) bis (UEK) werden dabei hintereinander ausgeführt und bilden gemeinsam die Initialisierung der getesteten Korrelationen. Die Ausführungsbeispiele für (UK) und (BI) bilden einen Aktualisierungsschritt und werden mit einem anderen Datensatz ausgeführt.

INITIALISIERUNGSPHASE BZW. INITIALISIEREN DER KORRELATIONEN (IK IN FIG. 1):

[0077] Die in diesem Abschnitt beschriebenen Ausführungsbeispiele behandeln jeweils einen Schritt der Initialisierungsphase der Korrelationen. Die Schritte werden immer auf Basis desselben Datensatzes ausgeführt und bauen aufeinander auf.

[0078] Der verwendete Beispiel-Datensatz ist in der Tabelle 2 (Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen) dargestellt:

[0079] Tabelle 2: Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen

§		§	(§) sent request to	§
Jul-17-11:30:18		user1	(10.2.12.9) sent request to	10.4.0.5
Jul-17-11:32:27		user2	(10.2.12.6) sent request to	10.4.0.7
Jul-17-11:33:22		user2	(10.2.12.9) sent request to	10.4.0.7
Jul-17-11:35:58		user3	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:36:43		user1	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:37:29		user1	(10.2.12.9) sent request to	10.4.0.5
Jul-17-11:39:38		user2	(10.2.12.6) sent request to	10.4.0.7
Jul-17-11:40:33		user2	(10.2.12.6) sent request to	10.4.0.7
Jul-17-11:43:09		user3	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:43:54		user1	(10.2.12.6) sent request to	10.4.0.5
chronologisch	statisch	diskret	statisch	diskret	statisch	diskret

[0080] Wie man aus dem Template ablesen kann (1. Zeile), besteht der Ereignistyp aus sieben Token T (4 Wildcards §, Leerzeichen „“, sowie Zeichenfolgen „(“ und „) sent request to“). Bei der Aufbereitung wurden die Werte der Token T in getrennten Mengen bzw. Listen gespeichert, sodass sie in den einzelnen Ausführungsschritten zur Verfügung stehen.

SELEKTIEREN DER ÜBERPRÜFTEN TOKEN (ST)

[0081] Dieses Ausführungsbeispiel beschreibt die Selektion der Token T, welche im Anschluss auf Korrelationen überprüft werden. Der Test in diesem Beispiel ob die Werte diskret sind und infolgedessen selektiert werden, beruht darauf, dass diskrete Werte nur eine begrenzte Anzahl an Werten annehmen. Der gewählte Test ist positiv, wenn mindestens zwei Werte unterschiedlich sind und maximal 35% der Werte unterschiedlich sind. Ist eine der Bedingungen nicht erfüllt ist der Test negativ. In diesem Beispiel besteht der Datensatz für die Berechnung aus 10 Zeilen, womit maximal drei verschiedene Werte erlaubt sind.

[0082] Der Test auf diskrete Werte für die Token T2, T4 und T6 mit den Werten „“, „(“ und „) sent request to“ sind negativ, da sie jeweils nur einen Wert beinhalten. Der Test für den ersten Token T1 ist ebenfalls negativ, da er zehn unterschiedliche Werte beinhaltet, und damit über dem Schwellwert von 3 liegt. Die anderen drei Token T3, T5 und T7 besitzen jeweils zwei oder drei unterschiedliche Werte, womit diese drei Token T selektiert werden.

AUSWAHL (AT IN FIG. 1) DER TOKENPAARE TP

[0083] Für die folgenden Beispiele könnte der Ansatz gewählt werden, dass jeder Token T mit jedem anderen gepaart wird. Damit wäre die Liste der Korrelationen [T3-T5; T3-T7; T5-T3; T5-T7; T7-T3; T7-T5]. Beispielhaft jedoch werden in den folgenden Schritten nur Korrelationen zwischen den Tokenpaaren [T3-T5; T3-T7; T5-T3] überprüft.

ERSTELLEN DER KORRELATIONEN (EK IN FIG. 1)

[0084] In diesem Schritt werden die Korrelationen erzeugt, indem die bedingten Verteilungen zu den einzelnen Tokenpaaren TP berechnet und abgespeichert werden. In diesem Beispiel werden nur die bedingten Verteilungen der Tokenpaare T3-T7 und T5-T3 aufgelistet, um die Übersicht zu gewährleisten.

[0085] Die Listen der bedingten Verteilungen zum Tokenpaar T3-T5 sind:

T3 = user1 → T7 = [10.2.12.6: 50%, 10.2.12.9: 50%]
T3 = user2 → T7 = [10.2.12.6: 75%, 10.2.12.9: 25%]
T3 = user3 → T7 = [10.2.12.6: 100%]

[0086] Die Listen der bedingten Verteilungen zum Tokenpaar T3-T7 sind:

T3 = user1 \rightarrow T7 = [10.4.0.5: 100%]

T3 = user2 \rightarrow T7 = [10.4.0.7: 100%]

T3 = user3 \rightarrow T7 = [10.4.0.5: 100%]

[0087] Die Listen der bedingten Verteilungen zum Tokenpaar T5-T3 sind:

T5 = 10.2.12.6 \rightarrow T3 = [user1: 29%, user2: 42%, user3: 29%]

T5 = 10.2.12.9 \rightarrow T3 = [user1: 67%, user2: 33%]

ÜBERPRÜFEN DER KORRELATIONEN (UEK IN FIG. 1)

[0088] Die Überprüfung in diesem Ausführungsbeispiel besteht aus zwei Ansätzen, die hintereinander ausgeführt werden. Der erste Ansatz ist, die generierten bedingten Verteilungen auf Ähnlichkeit zu jener Verteilung zu überprüfen, bei der alle Werte gleich wahrscheinlich sind. Der zugehörige Test ist

$$\sum_{i \in N, i \leq n} |p_i - \frac{1}{n}| \geq \tau$$

wobei n die Anzahl der Werte der bedingten Verteilung zu der Korrelation ist, p_i die Wahrscheinlichkeit des i -ten Wertes ist, und τ ein Schwellwert ist. Gilt die Ungleichung, ist der Test positiv, sollte sie nicht erfüllt sein, ist der Test negativ. Für dieses Beispiel sei der Schwellwert mit 0.5 angenommen.

[0089] In der zweiten Überprüfung werden alle Korrelationen zu den einzelnen Tokenpaaren TP jeweils zusammen überprüft. Wenn die Summe der Häufigkeiten der Werte, für die Korrelationen angenommen wurden, geringer ist als ein Schwellwert, werden sämtliche Korrelationen entfernt. Liegt die Summe über oder ist gleich dem Schwellwert, wird keine Korrelation entfernt. Der Schwellwert liegt in diesem Beispiel bei 0.6.

[0090] Zuerst werden die Korrelationen einzeln überprüft. Die Berechnungen zu den einzelnen Werten der Tokenpaare TP lauten:

[0091] Berechnungen zum Tokenpaar T3-T5:

user1: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| \frac{1}{2} - \frac{1}{2} \right| + \left| \frac{1}{2} - \frac{1}{2} \right| = 0 < 0.5$

user2: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| 0.75 - \frac{1}{2} \right| + \left| 0.25 - \frac{1}{2} \right| = 0.5 \geq 0.5$

user3: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| = 1 \geq 0.5$

[0092] Berechnungen zum Tokenpaar T3-T7:

user1: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| = 1 \geq 0.5$

user2: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| 0 - \frac{1}{2} \right| + \left| 1 - \frac{1}{2} \right| = 1 \geq 0.5$

user3: $\sum_{i \in N, i \leq 2} |p_i - \frac{1}{2}| = \left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| = 1 \geq 0.5$

[0093] Berechnungen zum Tokenpaar T5-T3:

10.2.12.6: $\sum_{i \in N, i \leq 3} |p_i - \frac{1}{3}| = \left| 0.29 - \frac{1}{3} \right| + \left| 0.42 - \frac{1}{3} \right| + \left| 0.29 - \frac{1}{3} \right| = 0.18 < 0.5$

10.2.12.9: $\sum_{i \in N, i \leq 3} |p_i - \frac{1}{3}| = \left| 0.67 - \frac{1}{3} \right| + \left| 0.33 - \frac{1}{3} \right| + \left| 0 - \frac{1}{3} \right| = 0.67 \geq 0.5$

[0094] Da die durch den Wert „10.2.12.6“ bedingte Verteilung des Tokenpaares T5-T3 die in der Beschreibung definierte Ungleichung nicht erfüllt, wird diese Korrelation gelöscht. Alle anderen Korrelationen erfüllen sie und haben damit die erste Überprüfung bestanden.

[0095] Als nächstes werden die zugehörigen Korrelationen zu den einzelnen Tokenpaaren TP als gesamtes getestet. Die Berechnungen für die Tests lauten:

T3-T5: $p_{\text{user2}} + p_{\text{user3}} = 0.4 + 0.2 = 0.6 \geq 0.6$

T3-T7: $p_{\text{user1}} + p_{\text{user2}} + p_{\text{user3}} = 0.4 + 0.4 + 0.2 = 1 \geq 0.6$

T5-T3: $p_{10.2.12.9} = 0.3 < 0.6$

[0096] Da das Tokenpaar T5-T3 die in der Beschreibung definierte Ungleichung nicht erfüllt, werden alle zu diesem Tokenpaar TP zugehörigen Korrelationen gelöscht.

[0097] Womit die erzeugten Korrelationen folgende sind:

[0098] T3-T5:

T3 = user2 \rightarrow T7 = [10.2.12.6: 75%, 10.2.12.9: 25%]

T3 = user3 \rightarrow T7 = [10.2.12.6: 100%]

[0099] T3-T7:

T3 = user1 \rightarrow T7 = [10.4.0.5: 100%]

T3 = user2 \rightarrow T7 = [10.4.0.7: 100%]

T3 = user3 \rightarrow T7 = [10.4.0.5: 100%]

TESTPHASE BZW. AKTUALISIEREN DER KORRELATIONEN (AK IN FIG. 1):

[00100] Dieser Abschnitt handelt von der Testung und Aktualisierung der Korrelationen. Hierfür werden die Korrelationen auf Basis der in Tabelle 3 (Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen) abgebildeten Protokollzeilen getestet und aktualisiert.

[00101] Tabelle 3: Beispiel eines Templates mit zugehörigen Protokollzeilen und Datentypen

§		§	(§) sent request to	§
Jul-17-11:30:18		user2	(10.2.12.6) sent request to	10.4.0.7
Jul-17-11:32:27		user2	(10.2.12.6) sent request to	10.4.0.6
Jul-17-11:33:22		user3	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:35:58		user2	(10.2.12.9) sent request to	10.4.0.6
Jul-17-11:36:43		user2	(10.2.12.9) sent request to	10.4.0.6
Jul-17-11:37:29		user2	(10.2.12.6) sent request to	10.4.0.6
Jul-17-11:39:38		user3	(10.2.12.9) sent request to	10.4.0.5
Jul-17-11:40:33		user3	(10.2.12.6) sent request to	10.4.0.5
Jul-17-11:43:09		user2	(10.2.12.9) sent request to	10.4.0.7
Jul-17-11:43:54		user2	(10.2.12.6) sent request to	10.4.0.6
chronologisch	statisch	diskret	statisch	diskret	statisch	diskret

[00102] Die abgebildeten Protokollzeilen besitzen denselben Ereignistyp, wie der Datensatz der in Tabelle 2 abgebildet ist. Bei der Aufbereitung wurden die Werte der Token T in getrennten Listen gespeichert, sodass sie in den folgenden Ausführungsschritten zur Verfügung stehen.

[00103] Die Schritte die in diesem Ausführungsbeispiel behandelt werden sind die Schritte Testen der Korrelationen (TK), Berechnen von Indikatoren (BI) und (im Zuge des Beispiels zur Berechnung von Indikatoren) Anomalieerkennung (AE).

TESTEN DER KORRELATIONEN (TK IN FIG. 1)

[00104] Dieses Beispiel beschreibt das Testen der Korrelationen. Sollte der Test zu einer Korrelation ein positives Resultat besitzen, wird die Korrelation mit Hilfe der Stichprobe des Aktualisierungsschrittes aktualisiert. In diesem Ausführungsbeispiel ist der Test, welcher benutzt wird, um zu überprüfen, ob die Protokollzeilen der Stichprobe den Korrelationen folgen, der Chiquadrattest auf Homogenität, welcher durch die folgende Teststatistik gegeben ist:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \approx \chi^2((r-1)(c-1))$$

$$E_{i,j} = (n_i * n_j) / \left(\sum_{j,j} o_{i,j} \right)$$

[00105] In dieser Berechnung bezeichnet r die Anzahl der getesteten Stichproben, welche im Anwendungsgebiet eines erfindungsgemäßen Verfahrens immer mit 2 angenommen wird und c die Anzahl der Kategorien. $O_{i,j}$ bezeichnet die beobachtete Anzahl an Werten der Kategorie j der Stichprobe i und $E_{i,j}$ die erwartete Anzahl an Werten der Kategorie j der Stichprobe i . Dabei ist n_i die Größe der Stichprobe i und n_j die Anzahl aller beobachteten Werte, die in die Kategorie j fallen.

[00106] Bei Korrelationen, bei denen sich keine Protokollzeilen in der Stichprobe befinden, so dass der Chiquadratstest nicht ausführbar ist, wird kein Testresultat ausgegeben. Bei Korrelationen mit nur einem Wert im zweiten Token T_2 , kann der Chiquadratstest nicht ausgeführt werden, wenn in allen Protokollzeilen, welche im ersten Token T_1 mit dem Wert des ersten Tokens der jeweiligen Korrelation übereinstimmen, ebenfalls nur dieser Wert im zweiten Token T_2 auftritt. In diesen Fällen wird ein positives Resultat gespeichert, da die Korrelation weiterhin besteht.

[00107] Das gewählte Signifikanzniveau dieses Tests ist $\alpha_1 = 0.05$. Damit ergibt sich der kritische Wert dieses Tests von $\chi^2_{(r-1)(c-1), 1-\alpha_1} = \chi^2_{(2-1)(2-1), 1-0.05} = \chi^2_{1, 0.95} = 3.8415$. Der Chiquadratstest ist positiv, wenn der Wert der Teststatistik kleiner oder gleich dem kritischen Wert ist. Man beachte, dass der kritische Wert auch abhängig von der Anzahl der angenommenen Werte ist und deshalb nicht immer ein einzelner Wert für mehrere Korrelationen oder mehrere Tokenpaare TP benutzt werden kann. Weiters kann vorteilhafterweise berücksichtigt werden, dass für Stichproben von geringem Umfang die kritischen Werte ungenau sein können.

[00108] Nachfolgend zu dem Test kann überprüft werden, ob die Hypothesen verworfen werden sollen. Dies ist jedoch wegen der Unschärfe des Tests bei kleinen Stichproben erst für spätere Aktualisierungsschritte sinnvoll. Dieser Test kann als Binomialtest ausgeführt werden, wobei die Wahrscheinlichkeit des Auftretens eines negativen Ergebnisses $p = \alpha_1$ als das Signifikanzniveau des Chiquadratstests und das Signifikanzniveau des Binomialtests als $\alpha_2 = 0.05$ gewählt werden könnte.

[00109] Da der Binomialtest in diesem Ausführungsbeispiel nicht berechnet wird, werden im Anschluss jene Korrelationen aktualisiert, welche ein positives Resultat im Chiquadratstest auf Homogenität besessen haben. Die Aktualisierung erfolgt, indem die bedingte Verteilung der Korrelation anteilig an die bedingte Verteilung der Korrelation der Stichprobe des Aktualisierungsschrittes angepasst wird. Gleichbedeutend dazu wäre, dass die bedingte Verteilung der Korrelation auf Basis der für deren Berechnung benutzte Stichprobe zuzüglich der Stichprobe des momentanen Aktualisierungsschrittes berechnet wird.

[00110] Die getesteten Korrelationen sind die im Ausführungsbeispiel des Initialisierens der Korrelationen erstellten Korrelationen, welche auf Basis von zehn Protokollzeilen erstellt wurden. Um eine bessere Übersicht für die Berechnung des Chiquadratstests zu gewährleisten wird in der folgenden Darstellung der Korrelationen nicht nur die Verteilung der Werte in absoluten Zahlen angegeben, sondern jeweils getrennt durch ein „ / “ die Verteilung der Werte der neuen Protokollzeilen in absoluten Zahlen angegeben.

T3-T5(user2): [10.2.12.6: 3, 10.2.12.9: 1] / [10.2.12.6: 4, 10.2.12.9: 3]
 T3-T5(user3): [10.2.12.6: 2] / [10.2.12.6: 2, 10.2.12.9: 1]
 T3-T7(user1): [10.4.0.5: 4] / [10.4.0.5: 0]
 T3-T7(user2): [10.4.0.7: 4] / [10.4.0.6: 5, 10.4.0.7: 2]
 T3-T7(user3): [10.4.0.5: 2] / [10.4.0.5: 3]

[00111] Die berechneten Teststatistiken der Chiquadratstests auf Homogenität mit den oben beschriebenen Verteilungen lauten:

T3-T5: [user2: 0.3508, user3: 0.8333]
 T3-T7: [user1: ---, user2: 5.2381, user3: ---]

[00112] Die Korrelationen, bei denen kein Wert steht, haben entweder keine eingelesenen Protokollzeilen, die getestet werden können, oder immer nur den gleichen Wert.

[00113] Da die kritische Distanz 3.8415 ist, sind von den drei durchgeführten Chiquadrattests, die zwei Tests des Tokenpaares T3-T5 positiv, da $0.3508 < 3.8415$ und $0.8333 < 3.8415$, und der Test von T3-T7(user2) negativ, da $5.2381 \geq 3.8415$.

[00114] Aus diesem Grund und weil der Test zu T3-T7(user3) auch positiv war, werden für die Korrelationen T3-T5(user2), T3-T5(user3) und T3-T7(user3) jeweils ein positives Testresultat gespeichert und die Korrelationen werden aktualisiert. Da sich keine passenden Protokollzeilen für Korrelation T3-T7(user1) in der Stichprobe befinden, wird für diese Korrelation weder ein positives noch ein negatives Testresultat abgespeichert, und für Korrelation T3-T7(user2) wird ein negatives Testresultat wegen dem negativen Chiquadrattest abgespeichert.

[00115] Für die Aktualisierung der Korrelationen werden die Werte der zwei getesteten Verteilungen zusammengefasst und diese Verteilung ergibt die neue Korrelation, die im nächsten Aktualisierungsschritt getestet wird. Für die drei Korrelationen sind die aktualisierten Korrelationen im Anschluss sowohl mit absoluten Zahlen als auch prozentueller Verteilung aufgeführt.

T3-T5(user2): [10.2.12.6: 7, 10.2.12.9: 4] / [10.2.12.6: 64%, 10.2.12.9: 36%]

T3-T5(user3): [10.2.12.6: 4, 10.2.12.9: 1] / [10.2.12.6: 80%, 10.2.12.9: 20%]

T3-T7(user3): [10.4.0.5: 5] / [10.4.0.5: 100%]

BERECHNUNG DER INDIKATOREN (BI IN FIG. 1)

[00116] In diesem Beispiel werden die Berechnungen von möglichen Indikatoren beschrieben. Um einen Indikator zu berechnen, können vorteilhafterweise zuvor eine Anzahl an Aktualisierungsschritten durchgeführt werden, um die benötigte Anzahl an Testresultaten zu generieren. Es werden jeweils die Korrelations-Indikatoren I_K , Tokenpaar-Indikatoren I_T und der Ereignistyp-Indikator I_E eines Ereignistyps berechnet.

[00117] Im folgenden Ausführungsbeispiel werden die Ergebnisse von 10 Aktualisierungsschritten betrachtet. Die Berechnung der Indikatoren basieren auf der folgenden, in Tabelle 4 zusammengefassten, Liste an p-Werten der Tests. Die p-Werte können aus den Teststatistiken berechnet werden, wie in Kusolitsch, N. (2014). Maß- und Wahrscheinlichkeitstheorie: Eine Einführung. Springer-Verlag. beschrieben.

[00118] Tabelle 4: Liste an p-Werten der Tests

Tokenpaar TP	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
T3-T5(user2)	0.554	0.836	0.495	0.463	0.738	0.537	0.438	0.783	0.845	0.432
T3-T5(user3)	0.361	0.945	0.834	---	0.784	0.574	0.784	0.643	0.854	0.957
T3-T7(user1)	---	1	1	1	---	1	1	1	---	1
T3-T7(user2)	0.022	0.035	0.031	0.073	0.094	0.064	0.054	0.025	0.043	0.063
T3-T7(user3)	1	1	1	1	1	1	0.072	0.034	0.053	0.062

[00119] In der Tabelle 4 steht „---“ für einen Test bei dem keine Protokollzeilen in der Stichprobe verfügbar sind, und numerische Werte stehen für den p-Wert der berechneten Teststatistik, bzw. wird für einen bestandenen Test, wobei nur ein möglicher Wert in der Korrelation zulässig war, der Wert 1 angenommen.

BESCHREIBUNG DES KORRELATIONS-INDIKATORS I_K

[00120] Die Berechnung des Korrelations-Indikators I_K erfolgt in diesem Ausführungsbeispiel anhand folgender Formel:

$$I_K = |\overline{M_{Ref}} - \overline{M_{Akt}}|$$

[00121] Die zugrundeliegende Idee dieser Berechnung ist es die p-Werte M_{Ref} einer Menge an vorangegangenen Tests, auch Referenz genannt, mit den p-Werten M_{Akt} einer Menge von aktuellen Tests zu vergleichen und einen Wert auszugeben, der angibt, um wie viel sich die Verteilung der Werte unterscheiden.

[00122] In diesem Ausführungsbeispiel ist die Menge der Referenz p-Werte die Menge der Teststatistiken der Tests eins bis fünf und die Menge der aktuellen p-Werte die Menge der Teststatistiken der Tests sechs bis zehn. Weiters ist die angewandte Teststatistik die Differenz der Erwartungswerte der Mengen.

[00123] Sollte einer der Mengen M_{Akt} , oder M_{Ref} leer sein, da keine Protokollzeilen für diese Korrelation in dem Zeitabschnitt eingelesen wurden, wird diese Korrelation nicht getestet.

[00124] Um eine Aussage über den berechneten Korrelations-Indikator I_K zu treffen, kann getestet werden, ob der Korrelations-Indikator I_K über einem Schwellwert liegt oder nicht. Der Schwellwert hängt sowohl von der Berechnung des Korrelations-Indikators I_K und der Art der eingelesenen Protokollzeilen ab, und kann daher vom Nutzer vorgeben werden oder durch ein anderes Verfahren geschätzt werden. Im dritten Ausführungsbeispiel sei der Schwellwert mit 0.5 angenommen.

BESCHREIBUNG DES KORRELATIONS-INDIKATORS I_K

[00125] Im Anschluss werden zu den einzelnen Korrelationen die Korrelations-Indikatoren berechnet.

$$\begin{aligned} I_{K,T3-T5(\text{user2})} &= |\overline{\{0.554, 0.836, 0.495, 0.463, 0.738\}} - \overline{\{0.537, 0.438, 0.783, 0.845, 0.432\}}| \\ &= |0.617 - 0.607| = 0.010 \end{aligned}$$

$$\begin{aligned} I_{K,T3-T5(\text{user3})} &= |\overline{\{0.361, 0.945, 0.834, 0.784\}} - \overline{\{0.574, 0.784, 0.643, 0.854, 0.957\}}| \\ &= |0.731 - 0.762| = 0.031 \end{aligned}$$

$$I_{K,T3-T7(\text{user1})} = |\overline{\{1, 1, 1\}} - \overline{\{1, 1, 1, 1\}}| = |1 - 1| = 0$$

$$\begin{aligned} I_{K,T3-T7(\text{user2})} &= |\overline{\{0.022, 0.035, 0.031, 0.073, 0.094\}} - \overline{\{0.064, 0.054, 0.025, 0.043, 0.063\}}| \\ &= |0.051 - 0.050| = 0.001 \end{aligned}$$

$$I_{K,T3-T7(\text{user3})} = |\overline{\{1, 1, 1, 1, 1\}} - \overline{\{1, 0.072, 0.34, 0.053, 0.062\}}| = |1 - 0.244| = 0.756$$

[00126] Da der berechnete Korrelations-Indikator $I_{K,T3-T7(\text{user3})}$ über dem Schwellwert von 0.5 liegt, deutet er auf eine Veränderung im Systemverhalten hin. Ein über dem Schwellenwert liegender Korrelations-Indikator I_K deutet auf eine schwerwiegende Veränderung im Systemverhalten hin und löst damit einen Alarm aus. Dadurch, dass der Alarm mit einer Korrelation assoziiert werden kann, bietet die vorliegende Erfindung außerdem einen Hinweis darauf wo im System die Anomalie aufgetreten ist.

[00127] Im Gegensatz zu gängigen Anomalieererkennungsansätzen, die meist punktuelle Ereignisse analysieren, lassen sich mit der vorliegenden Erfindung auch Anomalien erkennen, die eine schleichende Veränderung im Systemverhalten hervorrufen. Außerdem ist es möglich, Angriffe zu erkennen, die zwar Ereignisse, bzw. Protokollzeilen, erzeugen, die sich in ihrer Struktur und Werten nicht von denen des normalen Systemverhaltens unterscheiden, allerdings zu Veränderungen in den statistischen Merkmalen der Verteilungen der Werte führen.

BESCHREIBUNG DES TOKENPAAR-INDIKATORS I_{TP} UND EREIGNISTYP-INDIKATORS I_E

[00128] Die Berechnung des Tokenpaar-Indikators I_{TP} und des Ereignistyp-Indikators I_E werden gemeinsam beschrieben und berechnet, da beide denselben Ansatz zur Berechnung verfolgen. Im Allgemeinen müssen die Formeln nicht identisch sein, bzw. beide Arten von Indikatoren berechnet werden.

[00129] Die Berechnung in diesem Ausführungsbeispiel erfolgt anhand folgender Formeln:

$$I_{TP} = \max_{K \in TP} (I_K)$$

$$I_E = \max_{TP \in E} \max_{K \in TP} (I_K)$$

[00130] Um eine Aussage über die berechneten Indikatoren I_{TP} und I_E zu treffen, kann analog zum Korrelations-indikator I_K getestet werden, ob die Indikatoren über einem Schwellwert liegen oder nicht. Der Schwellwert hängt dabei sowohl von der Berechnung des Korrelations-Indikators I_K , der Berechnung des Tokenpaar-Indikators I_{TP} bzw. Ereignistyp-Indikators I_E und der Art der eingelesenen Protokollzeilen ab, und kann daher vom Nutzer vorgeben werden oder durch ein anderes Verfahren geschätzt werden. Im dritten Ausführungsbeispiel seien die Schwellwerte mit 0.6 für den Tokenpaar-Indikator I_{TP} und 0.65 für den Ereignistyp-Indikator I_E angenommen.

BESCHREIBUNG DES TOKENPAAR-INDIKATORS I_{TP} UND EREIGNISTYP-INDIKATORS I_E

[00131] Im Anschluss werden zuerst die Tokenpaar-Indikatoren I_{TP} und danach der Ereignistyp-Indikator I_E berechnet.

$$I_{TP,T3-T5} = \max \{0.010, 0.031\} = 0.031$$

$$I_{TP,T3-T7} = \max \{0, 0.001, 0.756\} = 0.756$$

$$I_E = \max \{0.031, 0.756\} = 0.756$$

[00132] Da der berechnete Tokenpaar-Indikator $I_{TP,T3-T5}$ über dem Schwellwert von 0.6 und der Ereignistyp-Indikator I_E über dem Schwellwert von 0.65 liegt, deutet dies auf eine Veränderung des Systemverhaltens hin. Wie bei den Korrelations-Indikatoren deutet ein Überschreiten der Tokenpaar- und Ereignistyp-Indikatoren der zugehörigen Schwellwerte auf schwerwiegende Veränderungen im Systemverhalten hin und lösen damit Alarme aus. Dadurch, dass ein Alarm mit einer Korrelation assoziiert werden kann, bietet die vorliegende Erfindung außerdem einen Hinweis darauf wo im System eine Anomalie aufgetreten ist.

[00133] Im Gegensatz zu gängigen Anomalieerkennungsansätzen, die meist punktuelle Ereignisse analysieren, lassen sich mit der vorliegenden Erfindung auch Anomalien erkennen, die eine schleichende Veränderung im Systemverhalten hervorrufen. Außerdem ist es möglich Angriffe zu erkennen, die zwar Ereignisse, bzw. Protokollzeilen, erzeugen, die sich in ihrer Struktur und Werten nicht von denen des normalen Systemverhaltens unterscheiden, allerdings zu Veränderungen in den statistischen Merkmalen der Verteilungen der Werte führen.

Token	T
Tokenpaar	TP
Erster/Zweiter Token eines Tokenpaares	T_1, T_2
Menge an aktuellen Testresultaten zu einer Korrelation	M_{Akt}
Menge an vergangenen Testresultaten zu einer Korrelation	M_{Ref}
Korrelations-Indikator	I_K
Tokenpaar-Indikator	I_{TP}
Ereignistyp-Indikator	I_E

Patentansprüche

1. Verfahren zur Detektion von anomalen Betriebszuständen eines Computersystems,
 - wobei während des Betriebs des Computers oder von auf dem Computer ablaufenden Prozessen, Protokolle erstellt werden, indem bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils eine Protokollzeile in Form einer Zeichenabfolge erstellt wird, wobei der Aufbau der Protokollzeile einer vom jeweiligen Computer und/oder Prozess vorgegebenen Syntax entspricht,
 - wobei die einzelnen Protokollzeilen an vorgegebenen Trennzeichen in Teilzeichenabfolgen unterteilt werden,
 - wobei den einzelnen Protokollzeilen basierend auf deren Syntax, insbesondere aufgrund der Ähnlichkeit der einzelnen Protokollzeilen zueinander und aufgrund der die Protokollzeilen aufbauenden Teilzeichenabfolgen, jeweils zumindest ein Ereignistyp zugewiesen wird,
 - wobei für jeden Ereignistyp ein Template umfassend eine Anzahl von Teilzeichenabfolgen der Protokollzeilen des jeweiligen Ereignistyps repräsentierenden Token (T), erstellt wird, wobei das Template die Struktur der einem jeweiligen Ereignistyp zugeordneten Protokollzeilen beschreibt, und
 - wobei das Template eine Anzahl von Token (T) mit statischem Inhalt, insbesondere statischen Werten, umfasst, die in ihrer Reihenfolge in ausgewählten, insbesondere allen, Protokollzeilen des jeweiligen Ereignistyps vorhanden sind, und wobei die Token (T) mit statischem Inhalt des Templates durch Token (T) mit variablen Inhalt, insbesondere variablen Werten, unterbrochen werden,
 - wobei das Template eine Anzahl von Token (T) aufweist, deren Inhalt, insbesondere deren Werte, einen diskreten Datentyp aufweisen, und
 - wobei die Protokollzeilen und deren Teilzeichenabfolgen charakteristisch für den Betriebszustand des Computersystems sind,
dadurch gekennzeichnet,
 - dass in einer Initialisierungsphase
 - auf Grundlage einer Stichprobe umfassend eine Anzahl an Protokollzeilen des jeweiligen Ereignistyps Tokenpaare (TP) umfassend einen ersten Token (T_1) und einen zweiten Token (T_2) aus den, den diskreten Datentyp aufweisenden, Token (T) des Templates eines jeweiligen Ereignistyps gebildet werden,
 - wobei als Werte des jeweiligen Tokens (T) jeweils diejenigen Teilzeichenabfolgen der Protokollzeilen der Stichprobe ermittelt werden, die dem jeweiligen Token (T) zugehörig sind, und
 - für alle Tokenpaare (TP) und jeden angenommenen Wert des ersten Tokens (T_1) des Tokenpaares (TP) jeweils eine Korrelation ermittelt wird, indem die bedingten Verteilungen dafür berechnet werden, dass bei Auftreten des jeweiligen Werts im ersten Token (T_1) des Tokenpaares (TP), ein jeweiliger Wert im zweiten Token (T_2) des Tokenpaares (TP) auftritt und
 - dass anschließend in einer Testphase
 - wiederholt, sobald eine Abfolge, insbesondere eine vorgegebene Anzahl, neu erstellter Protokollzeilen eines jeweiligen Ereignistyps erstellt wurde, mittels eines Tests auf Homogenität getestet wird, ob die aufgefundenen Korrelationen erfüllt sind, indem jeweils für alle Werte des ersten Tokens (T_1) eines jeweiligen Tokenpaares (TP) überprüft wird, ob bei denjenigen neu erstellten Protokollzeilen, bei denen der jeweilige Wert im ersten Token (T_1) auftritt, die Werte des zweiten Tokens (T_2) den berechneten bedingten Verteilungen der jeweiligen ermittelten Korrelation folgen und
 - ein anomaler Betriebszustand des Computersystems erkannt wird, wenn, insbesondere wiederholt, festgestellt wird, dass die Werte der Token (T) zumindest einer der ermittelten Korrelationen nicht den jeweils berechneten bedingten Verteilungen folgen.
2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, dass in der Testphase der Chi-Quadrat-Test auf Homogenität als Test auf Homogenität verwendet wird.

3. Verfahren nach Anspruch 1 oder 2, **dadurch gekennzeichnet**, dass bei der Bildung der Tokenpaare (TP), die Verteilungen der einzelnen Werte aller Token (T) ermittelt werden und jeweils nur Tokenpaare (TP) aus Token (T) gebildet werden, deren Werte in einem Test auf Homogenität eine ähnliche Verteilung aufweisen.
4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet**,
 - dass bei der Bildung der Tokenpaare (TP) zu jedem Token (T) eine Wertemenge ermittelt wird, wobei die Wertemenge als Werte diejenigen Teilzeichenabfolgen der Protokollzeilen der Stichprobe umfasst, die dem jeweiligen Token (T) zugehörig sind, und
 - dass jeweils nur Token (T) gepaart werden, bei denen die Werte der jeweiligen Wertemengen einander, insbesondere zu einem vorgegebenen Prozentsatz, überschneiden.
5. Verfahren nach einem der Ansprüche 1 bis 4, **dadurch gekennzeichnet**, dass bei der Bildung der Tokenpaare (TP) nur Token (T) berücksichtigt werden, bei denen die Auftrittswahrscheinlichkeit des häufigsten Werts, den diejenigen Teilzeichenabfolgen der Protokollzeilen in der Stichprobe annehmen, die dem jeweiligen Token (T) zugehörig sind, einen vorgegebenen Schwellwert überschreitet.
6. Verfahren nach einem der Ansprüche 1 bis 5, **dadurch gekennzeichnet**,
 - dass in der Initialisierungsphase die einzelnen ermittelten Korrelationen zu den einzelnen Tokenpaaren (TP) in einem Überprüfungsschritt dahingehend überprüft werden, ob die jeweils ermittelten Korrelationen vorgegebene Bedingungen erfüllen, welche einem Test entsprechen, ob die einzelnen ermittelten Korrelationen nicht zufällig sind, und
 - dass in der Testphase nur diejenigen ermittelten Korrelationen berücksichtigt werden, die diese vorgegebenen Bedingungen erfüllen, indem sie in der jeweiligen Überprüfung als nicht zufällig erkannt werden.
7. Verfahren nach Anspruch 6, **dadurch gekennzeichnet**, dass, im Fall, dass im Überprüfungsschritt mehrere verschiedene Überprüfungen der Korrelationen hintereinander vorgenommen werden, nur diejenigen Korrelationen in der Testphase berücksichtigt werden, die in zumindest einer Überprüfung, insbesondere allen Überprüfungen, als nicht zufällig erkannt wurden.
8. Verfahren nach Anspruch 6 oder 7, **dadurch gekennzeichnet**,
 - dass im Überprüfungsschritt die Korrelationen überprüft werden, indem mittels eines goodness of fit-Tests getestet wird, ob die angenommene Werte des zweiten Tokens (T_2), des jeweiligen Tokenpaares (TP) in den Protokollzeilen der Stichprobe welche dem der Korrelation zugehörigen Wert im ersten Token (T_1) besitzen, aus der Multinomialverteilung entstammen, in der das Auftreten aller Werte gleich wahrscheinlich ist und
 - dass die jeweilige Korrelation diese Überprüfung bestanden hat und als nicht zufällig erkannt wird, wenn der goodness of fit-Test negativ ist.
9. Verfahren nach einem der Ansprüche 6 bis 8, **dadurch gekennzeichnet**,
 - dass im Überprüfungsschritt zu den Korrelationen eines Tokenpaares (TP) jeweils die bedingten Verteilungen der Werte der zweiten Token (T_2) des jeweiligen Tokenpaares (TP) in der Stichprobe annimmt, mittels eines Tests auf Homogenität, überprüft werden, ob diese derselben Verteilung angehören, und
 - dass, sollte der Test auf Homogenität positiv ausfallen, alle Korrelationen dieses Tokenpaares (TP) die Überprüfung nicht bestanden haben und als zufällig erkannt werden.
10. Verfahren nach einem der Ansprüche 6 bis 9, **dadurch gekennzeichnet**,
 - dass im Überprüfungsschritt die Korrelationen überprüft werden, indem
 - die Häufigkeiten der Werte des ersten Tokens (T_1) des Tokenpaares (TP) ermittelt werden und
 - die Häufigkeiten derjenigen Werte des ersten Tokens (T_1) aggregiert, insbesondere summiert, werden, für welche Korrelationen ermittelt wurden, und
 - dass alle Korrelationen des Tokenpaares (TP) die Überprüfung nicht bestanden haben und als zufällig erkannt werden, wenn der derart berechnete aggregierte Wert, insbesondere die derart berechnete Summe, einen vorgegebenen Schwellenwert unterschreitet.

11. Verfahren nach einem der Ansprüche 1 bis 10, **dadurch gekennzeichnet**, dass in der Testphase zum Testen der Korrelationen die Korrelationen, insbesondere nur dann wenn das Testresultat positiv war, aktualisiert werden, indem für die Korrelationen die bedingten Verteilungen, die auf Basis von jeweiligen Stichproben festgelegt wurden, auf Basis dieser Stichprobe zuzüglich der Stichprobe dieses Updateschrittes festgelegt werden.
12. Verfahren nach einem der Ansprüche 1 bis 11, **dadurch gekennzeichnet**, dass, insbesondere mittels eines Binomialtests, ermittelt wird, ob sich die negativen Testresultate einer jeweiligen Korrelation häufen und dass die jeweilige Korrelation verworfen wird, sofern dies der Fall ist.
13. Verfahren nach einem der Ansprüche 1 bis 12, **dadurch gekennzeichnet**, dass neue Korrelationen gesucht werden, indem in der Testphase, insbesondere regelmäßig, einzelne Tokenpaare (TP), für die zuvor keine Korrelationen gefunden wurden, oder deren Korrelationen verworfen wurden, einer neuerlichen Initialisierungsphase unterworfen werden, und dass die derart aufgefundenen neuen Korrelationen anschließend in der Testphase getestet werden.
14. Verfahren nach einem der Ansprüche 1 bis 13, **dadurch gekennzeichnet**, dass zu den einzelnen Korrelationen, basierend auf einem Vergleich eines Abschnittes einer Liste der letzten Testresultate der Aktualisierungen der Korrelationen (M_{Akt}) und einem, als Referenz dienenden, vergangenen Abschnitt der Liste der letzten Testresultate der Aktualisierungen der Korrelationen (M_{Ref}), jeweils ein Korrelations-Indikator (I_K) berechnet wird, der die Wahrscheinlichkeit dafür angibt, dass die jeweils zuletzt neu erstellten Protokollzeilen von der jeweiligen Korrelation abgewichen sind.
15. Verfahren nach Anspruch 14, **dadurch gekennzeichnet**, dass für die Berechnung der Indikatoren zusätzlich zu den Testresultaten, die bei den Tests ermittelten p-Werte für die Berechnung des Indikators herangezogen werden.
16. Verfahren nach einem der Ansprüche 14 oder 15, **dadurch gekennzeichnet**, dass alle Korrelations-Indikatoren (I_K) eines Tokenpaares (TP) zu einem Tokenpaar-Indikator (I_T) zusammengefasst werden und/oder alle Korrelations-Indikatoren (I_K) eines Ereignistyps zu einem Ereignistyp-Indikator (I_E) zusammengefasst werden, wobei der Tokenpaar-Indikator (I_T) und der Ereignistyp-Indikator (I_E) jeweils die Wahrscheinlichkeiten dafür angeben, dass die jeweils zugrundeliegenden Protokollzeilen von einer der zusammengefassten Korrelationen abgewichen sind.
17. Verfahren nach einem der Ansprüche 1 bis 16, **dadurch gekennzeichnet**,
 - dass in der Initialisierungsphase Ereignistypen ermittelt werden, deren Protokollzeilen in einem vorgegebenen zeitlichen Abstand zueinander, insbesondere unmittelbar nacheinander, vom Computersystem erstellt werden, und
 - dass Tokenpaare (TP) erstellt werden, die jeweils einen, einen diskreten Datentyp aufweisenden, Token (T) von verschiedenen der derart ermittelten Ereignistypen umfassen.
18. Datenträger, auf dem ein Programm zur Durchführung eines Verfahrens nach einem der Ansprüche 1 bis 17 abgespeichert ist.

Hierzu 1 Blatt Zeichnungen

1/1

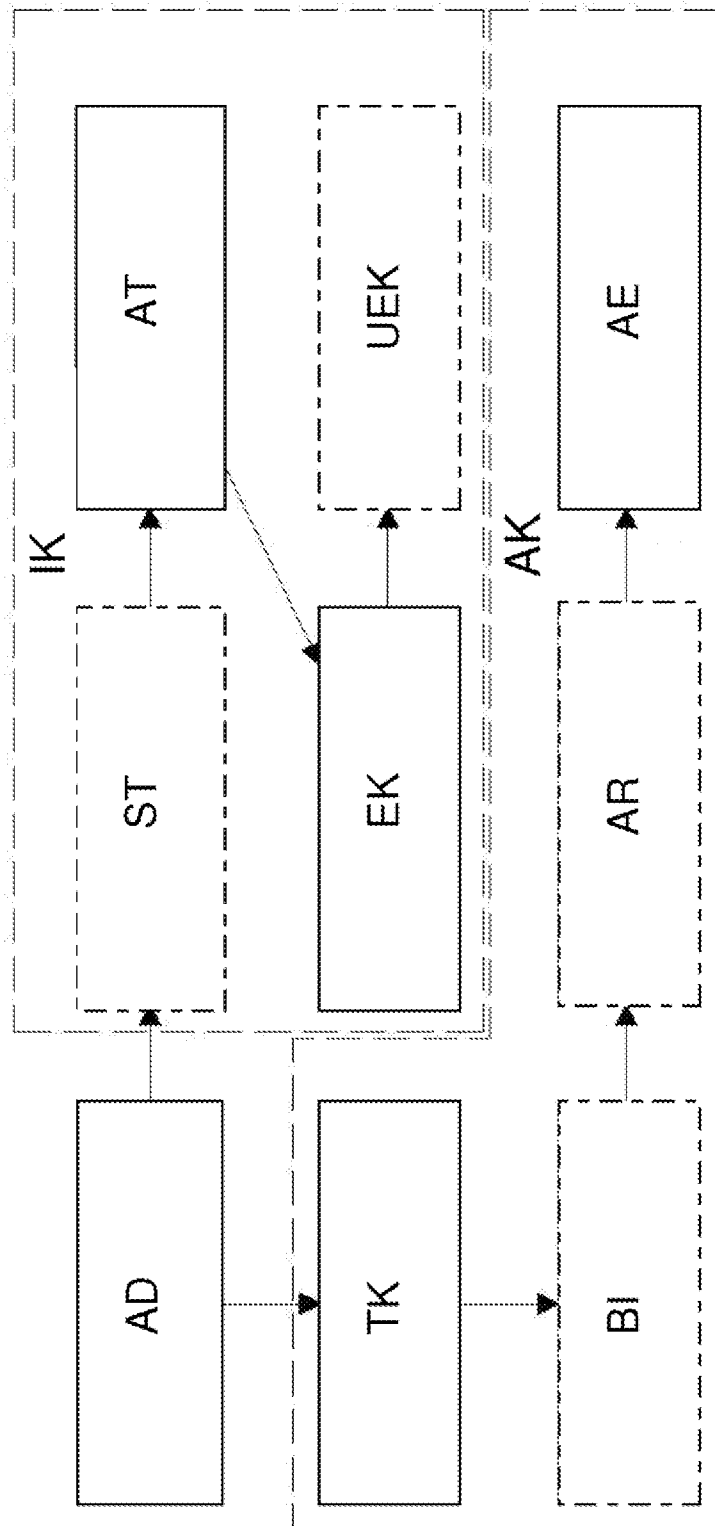


Fig. 1