

(12) **Patentschrift**

(21) Anmeldenummer:	A 50461/2018	(51) Int. Cl.:	H04L 12/26	(2006.01)
(22) Anmeldetag:	11.06.2018		H04L 29/06	(2006.01)
(45) Veröffentlicht am:	15.05.2021		G06F 11/00	(2006.01)
			G06F 11/07	(2006.01)
			H04L 12/24	(2006.01)
			G06F 21/55	(2013.01)

(56) Entgegenhaltungen:
 EP 3267625 A1
 EP 2800307 A1
 DE 202013011084 U1
 EP 2299650 A1
 OSTER, Norbert "Automatische Generierung optimaler datenflussorientierter Testdaten mittels Evolutionärer Verfahren" 11.06.2004; Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Software Engineering (Informatik 11); abgerufen im Internet am 08.01.2019 unter dem Link URL:<<http://www.gm.fh-koeln.de/~winter/tav/html/tav21/TAV21P4Oster.pdf>>

(73) Patentinhaber:
 AIT Austrian Institute of Technology GmbH
 1210 Wien (AT)

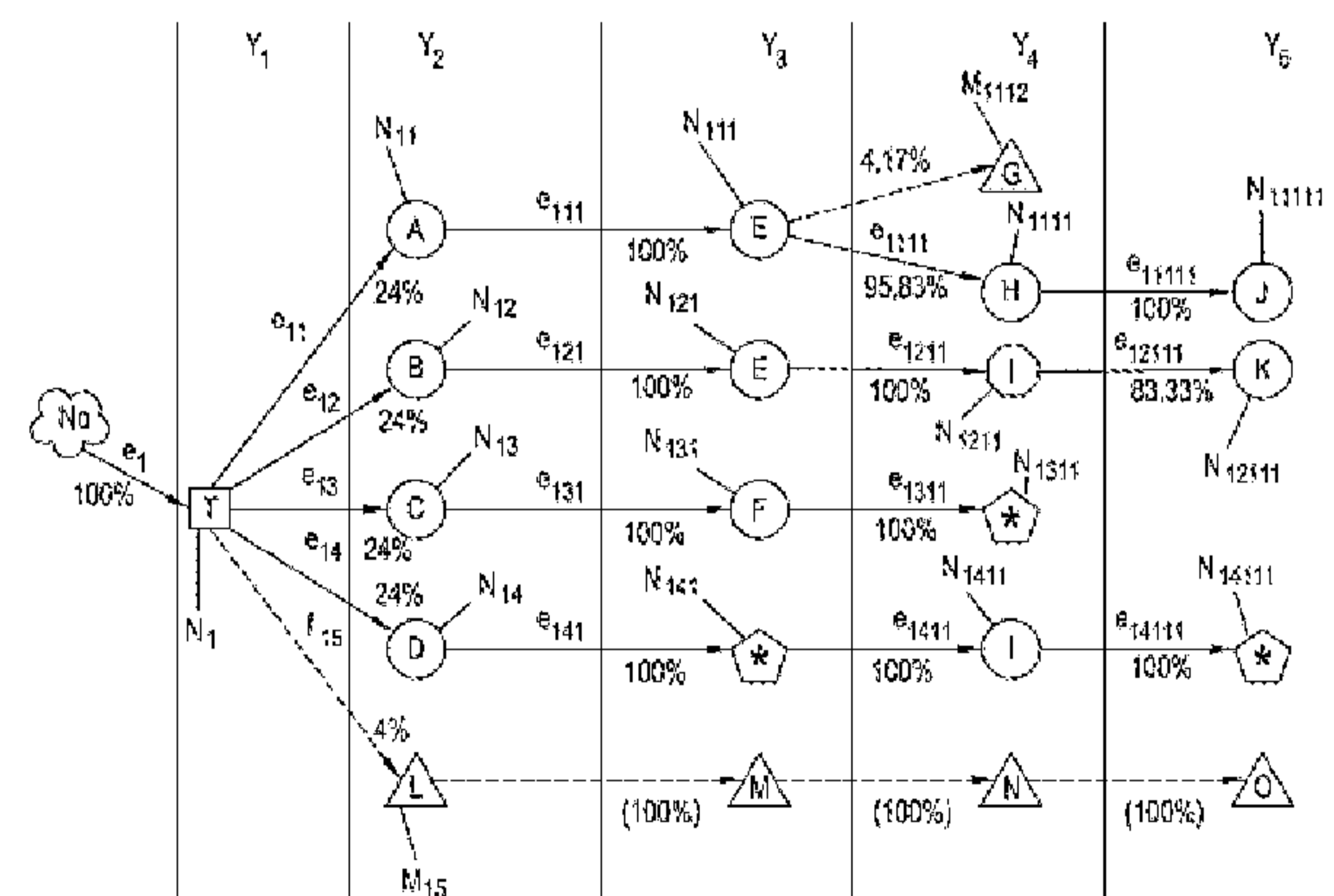
(72) Erfinder:
 Wurzenberger Markus
 1090 Wien (AT)
 Landauer Max
 1070 Wien (AT)
 Fiedler Roman
 8020 Graz (AT)
 Skopik Florian DDr.
 2000 Stockerau (AT)

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

(54) **Grammatikerkennung**

(57) Die Erfindung betrifft ein Verfahren zur Charakterisierung des Zustands eines Computersystems, wobei

- vom Computersystem oder von auf diesem ablaufenden Prozessen jeweils Protokolle erstellt werden, indem bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils eine Protokollzeile (L_1, \dots, L_{100}) erstellt wird und wobei die Protokollzeile (L_1, \dots, L_{100}) das jeweils protokollierte Ereignis beschreibt, und
- wobei jede derart erstellte Protokollzeile (L_1, \dots, L_{100}) in eine Anzahl von Teilzeichenketten unterteilt wird, wobei aufgrund der einzelnen Protokollzeilen (L_1, \dots, L_{100}) sowie der Abfolge der einzelnen, in den Protokollzeilen (L_1, \dots, L_{100}) enthaltenen Teilzeichenketten sowie aufgrund der Häufigkeit des Vorkommens der Protokollzeilen und der Teilzeichenketten in den Protokollzeilen ein die mögliche Abfolge von Teilzeichenketten beschreibender Syntaxbaum erstellt wird, und
- wobei dieser Syntaxbaum als charakteristisch für den Zustand des Computersystems angesehen wird.



Beschreibung

[0001] Die Erfindung betrifft ein Verfahren zur Charakterisierung des Zustands eines Computersystems. Als Ergebnis liefert die Erfindung eine strukturierte Beschreibung des Systemzustands, der insbesondere zur Detektion von anormalen Zuständen in Computersystemen genutzt werden kann.

[0002] Aus dem Stand der Technik sind unterschiedliche Verfahren bekannt, mit denen es möglich ist, den inneren Zustand eines Computersystems anhand bestimmter aussagekräftiger Parameter oder Datenstrukturen zu charakterisieren. Der grundsätzliche Zweck dieser Vorgehensweise liegt darin, durch den Erhalt von charakteristischen Parametern oder Datenstrukturen für ein Computersystem anormale Betriebszustände zu erkennen. Bei derartigen anormalen Betriebszuständen kann es sich beispielsweise um besondere Betriebszustände handeln, bei denen das System durch einen Hacker-Angriff modifiziert wird oder bei dem aufgrund unbeabsichtigter Änderungen eine Fehlfunktion des Systems besteht. Darüber hinaus werden jedoch auch andere abweichende Betriebsmodi detektiert, die beispielsweise mit Software-Updates oder ähnlichen Veränderungen des Systems des Computersystems einhergehen.

[0003] Mit der vorliegenden Erfindung wird eine weitere, einfach handhabbare und einfach auf weitere Protokollzeilen anwendbare Möglichkeit der Charakterisierung eines Computersystems geschaffen.

[0004] Der Erfindung liegt dabei die grundsätzliche Intention zugrunde, einzelne vom Computersystem oder den auf dem Computersystem ablaufenden Programmen erstellte Protokollzeilen zu analysieren und für diese Protokollzeilen eine Grammatik zu ermitteln, durch die die Gestalt der einzelnen Protokollzeilen charakterisiert werden kann.

[0005] Anders als bei Grammatiken von Computersprachen, die üblicherweise vorab vorgegeben werden, wird im Rahmen der Erfindung eine Grammatik bzw. ein die Grammatik der Protokollzeilen repräsentierender Baum aufgrund von bestehenden Protokollzeilen erstellt. Nachfolgende Protokollzeilen können anschließend daraufhin überprüft werden, ob sie im Sinne der so ermittelten Grammatik als grammatisch betrachtet werden und als dem normalen Systemzustand zugehörig betrachtet werden können.

[0006] Weicht die betreffende Protokollzeile jedoch von der Grammatik ab, so wurde eine Situation gefunden, die mit dem zuvor ermittelten normalen Betriebszustand bzw. mit diesem mit einer gewissen Wahrscheinlichkeit nicht in Übereinstimmung gebracht werden kann. In diesem Fall kann ein Systemzustand detektiert werden, der im vorliegenden Sinn als anormal angesehen werden kann. Ein solcher anormaler Betriebszustand kann insbesondere dann gefunden werden, wenn sich eine Mehrzahl oder eine einen Schwellenwert überschreitende Anzahl von neu ermittelten Protokollzeilen nachträglich als nicht grammatisch erweist.

[0007] Eine weitere vorteilhafte Möglichkeit, den Systemzustand zu charakterisieren und Änderungen des Systemzustands zu ermitteln, liegt darin, zu unterschiedlichen Zeitpunkten oder für unterschiedliche Zeitfenster jeweils separat einen die Grammatik charakterisierenden Graphen bzw. Syntaxbaum zu erstellen und diese so erstellten Syntaxbäume auf ihre Unterschiede hin zu untersuchen. Sofern diese Syntaxbäume signifikante Unterschiede aufweisen, kann eine Abweichung des Betriebszustands festgestellt werden.

[0008] Die Erfindung sieht ein Verfahren zur Charakterisierung des Zustands eines Computersystems vor, bei dem

- vom Computersystem oder von auf diesem ablaufenden Prozessen jeweils Protokolle erstellt werden, indem bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils eine Protokollzeile erstellt wird und wobei die Protokollzeile das jeweils protokollierte Ereignis beschreibt, und
- wobei jede derart erstellte Protokollzeile in eine Anzahl von Teilzeichenketten unterteilt wird.

[0009] Dabei ist weiters vorgesehen, dass aufgrund der einzelnen Protokollzeilen sowie der Ab-

folge der einzelnen, in den Protokollzeilen enthaltenen Teilzeichenketten sowie aufgrund der Häufigkeit des Vorkommens der Protokollzeilen und der Teilzeichenketten in den Protokollzeilen ein die mögliche Abfolge von Teilzeichenketten beschreibender Syntaxbaum erstellt wird, und

- dass dieser Syntaxbaum als charakteristisch für den Zustand des Computersystems angesehen wird.

[0010] Hierdurch wird erfindungsgemäß eine einfache und effiziente Möglichkeit geschaffen, den Systemzustand des Systems einfach zu charakterisieren und für weitere Untersuchungen zur Verfügung zu halten.

[0011] Zur Erstellung einer Repräsentation, die das Systemverhalten vorteilhaft beschreibt, und auch die Verteilung der Häufigkeit der einzelnen verfügbaren Protokollzeilen vorteilhaft abbildet, sieht vor, dass der Syntaxbaum anhand der einzelnen Protokollzeilen als azyklischer, gerichteter Graph erstellt wird,

- wobei der Syntaxbaum weitere Knoten (N_1, N_2) aufweist, denen jeweils ein Muster (P_1, P_2) zugeordnet wird, das bei Anwendung auf eine der Teilzeichenketten jeweils einen positiven oder negativen Übereinstimmungswert liefert,
- wobei der Syntaxbaum einzelne gerichtete Kanten aufweist, die ihren jeweiligen Quellknoten mit ihrem jeweiligen Zielknoten verbinden, wenn unter den einzelnen Protokollzeilen die bedingte Wahrscheinlichkeit dafür, dass
 - unter der Bedingung, dass Teilzeichenketten in der Protokollzeile (L_1, \dots, L_{100}) enthalten sind, die entsprechend ihrer Reihenfolge mit Mustern der einzelnen Knoten (N_1, N_{11}, N_{111}) auf einem vom Wurzelknoten (N) des Syntaxbaums zum Quellknoten (N_{11}) der Kante (e_{111}) führenden gerichteten Teilpfad entsprechend der Reihenfolge der Knoten (N_1, N_{11}, N_{111}) in diesem Teilpfad jeweils eine positive Übereinstimmung liefern,
 - als nächste Teilzeichenkette (s_{13}) in der betreffenden Protokollzeile (s_1) eine Teilzeichenkette vorliegt, die eine positive Übereinstimmung mit dem im Zielknoten (N_{111}) der betreffenden Kante (e_{111}) gespeicherten Muster aufweist,

einen vorgegebenen, allenfalls von der Position im Syntaxbaum abhängigen, Schwellenwert übersteigt, und

- insbesondere die bedingte Übergangswahrscheinlichkeit der betreffenden Kante zugeordnet wird.

[0012] Um eine Rasche und einfache Prüfung auf Veränderungen des Systemverhaltens zu erhalten, kann vorgesehen sein, dass a) anhand einer Anzahl von vorgegebenen oder innerhalb eines ersten Zeitraums erstellten Protokollzeilen (L_1, \dots, L_{100}) ein Syntaxbaum, insbesondere nach einem Verfahren nach einem der vorangehenden Ansprüche, erstellt wird,

b) die von den Computern oder den auf diesen Computern ablaufenden Prozessen bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils erstellten Protokollzeilen während eines zweiten Zeitraums ermittelt werden,

c) mittels des Parsers überprüft wird, ob und/oder in welchem Ausmaß die in Schritt b) ermittelten Protokollzeilen die vom Syntaxbaum vorgegebenen Regeln erfüllen, und

d) ein anormaler Zustand insbesondere dann festgestellt wird, wenn

- die Anzahl der während des ersten Zeitraums ermittelten und die vom Syntaxbaum vorgegebenen Regeln erfüllenden Protokollzeilen und
- die Anzahl der während des zweiten Zeitraums ermittelten und die vom Syntaxbaum vorgegebenen Regeln erfüllenden Protokollzeilen

voneinander um ein vorgegebenes Maß abweichen.

[0013] Eine alternative Möglichkeit, Änderungen des Systemverhaltens zu erkennen, sieht vor, dass für dasselbe Computersystem zu unterschiedlichen Zeitpunkten oder für unterschiedliche Systeme mit ähnlichem Aufbau und Verwendungszweck jeweils ein Syntaxbaum nach einem der vorangehenden Ansprüche ermittelt wird,

- dass zwischen den so erstellten Syntaxbäumen nach Abweichungen gesucht wird, und

- dass im Falle von Abweichungen, die einen vorgegebenen Schwellenwert überschreiten, ein abweichender, insbesondere kritischer oder anormaler, Zustand des Computersystems gemeldet wird.

[0014] Besonders vorteilhaft zur Speicherung einer Repräsentation des Systemverhaltens ist es, wenn es sich beim Syntaxbaum um einen gewurzelten Baum, vorzugsweise um einen gewurzelten Out-Tree, handelt.

[0015] Eine Methode zur effizienten Erstellung eines Syntaxbaums sieht vor, dass bei der Erstellung des Syntaxbaums die einzelnen Teilzeichenketten der Protokollzeilen in einem zweidimensionalen Speicher mit zwei Zugriffsindizes abgespeichert werden, wobei der erste Zugriffsindex als Zeilenindex die Protokollzeile und der zweite Zugriffsindex als Positionsindex die Position der Teilzeichenkette innerhalb der jeweiligen Protokollzeile angibt,

- dass für die Teilzeichenketten, denen der niedrigste Positionsindex zugewiesen ist,
 - nach einer Anzahl Mustern gesucht wird, die die Mehrzahl der Teilzeichenketten beschreiben,
 - für die einzelnen Muster jeweils die Wahrscheinlichkeiten dafür ermittelt werden, dass eine der Teilzeichenketten mit dem Muster übereinstimmt,
 - dass für die einzelnen Muster jeweils ein Knoten einer ersten Schicht in den Syntaxbaum eingefügt wird,
 - diesem Knoten das jeweilige Muster sowie diejenigen Protokollzeilen zugewiesen werden, deren herangezogene Teilzeichenketten mit dem Muster des Knoten übereinstimmen,
 - dieser Knoten als Zielknoten über eine gerichtete Kante mit dem Wurzelknoten des Syntaxbaums verbunden wird, und
 - dieser Kante die jeweilige zuvor ermittelte Wahrscheinlichkeit zugewiesen wird, und
- dass für inkrementell ansteigenden Positionsindex der Teilzeichenketten in den Protokollzeilen jeweils:
 - separat für einzelne Gruppen von Protokollzeilen, die jeweils einem Basisknoten der jeweils unmittelbar vorangehenden Schicht des Graphen zugeordnet sind, jeweils:
 - nach einer Anzahl Mustern gesucht wird, die die Mehrzahl der Teilzeichenketten an der durch den jeweiligen Positionsindex festgelegten Position beschreiben,
 - für die einzelnen Muster jeweils die Wahrscheinlichkeiten dafür ermittelt werden, dass die jeweilige Teilzeichenketten mit dem betreffenden Positionsindex mit dem Muster übereinstimmt,
 - dass für die einzelnen Muster jeweils ein Knoten einer dem Positionsindex entsprechenden Schicht in den Graphen eingefügt wird,
 - diesem Knoten das jeweilige Muster sowie diejenigen Protokollzeilen zugewiesen werden, deren herangezogene Teilzeichenketten mit dem Muster des Knoten übereinstimmen,
 - dieser Knoten als Zielknoten über eine gerichtete Kante mit dem Basisknoten verbunden wird, und
 - dieser Kante die jeweilige zuvor ermittelte Wahrscheinlichkeit zugewiesen wird.

[0016] Die folgenden Regeln dienen der vorteilhaften Erstellung von Syntaxbäumen. Dabei kann für die Erstellung von Knoten mit fixen Mustern vorgesehen sein, dass für die Teilzeichenketten kein Muster gefunden werden kann, das eine relative Häufigkeit aufweist, die einen Wahrscheinlichkeitsschwellenwert θ_1 übersteigt, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.

[0017] Alternativ oder zusätzlich kann vorgesehen sein, dass für den Fall, dass für die Teilzei-

chenketten ein Muster gefunden werden kann, sodass die relative Häufigkeit der dem Muster entsprechenden Teilzeichenketten einen ersten Wahrscheinlichkeitsschwellenwert θ_1 übersteigt,

- für den Fall, dass diese relative Häufigkeit auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 überschreitet, ein einziger Knoten mit dem betreffenden Muster eingefügt wird, wobei insbesondere der jeweiligen neu erstellten Kante der Anteil der dem Muster entsprechenden Teilzeichenketten zugewiesen wird und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen, die an der betreffenden Position eine dem Muster entsprechende Teilzeichenkette enthalten, dem neu eingefügten Zielknoten zugewiesen werden, und/oder
- für den Fall, dass diese relative Häufigkeit nicht auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 überschreitet, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird, alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.

[0018] Alternativ oder zusätzlich kann vorgesehen sein, dass für den Fall, dass für die Teilzeichenketten eine Mehrzahl der Muster gefunden werden kann, sodass die relative Häufigkeit der den einzelnen Muster entsprechenden Teilzeichenketten jeweils einzeln einen ersten Wahrscheinlichkeitswert θ_1 übersteigt,

- für den Fall, dass die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 überschreitet, eine Anzahl von Knoten mit jeweils einem der ermittelten Muster eingefügt wird, wobei insbesondere jeder Knoten über jeweils eine Kante mit dem ursprünglichen Knoten verbunden wird und/oder den so neu erstellten Kanten der Anteil der dem Muster entsprechenden Teilzeichenketten an der Gesamtzahl der dem ursprünglichen Knoten zugewiesenen Protokollzeilen zugewiesen wird und/oder die einzelnen Protokollzeilen auf die Knoten aufgeteilt werden, sodass jede Protokollzeile jeweils demjenigen Knoten zugeordnet wird, dessen Muster ihre jeweils betrachtete Teilzeichenkette entspricht,

und/oder

- für den Fall, dass die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 nicht überschreitet, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird, und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.

[0019] Alternativ oder zusätzlich kann vorgesehen sein, dass für den Fall,

dass die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position enden, einen vorgegebenen vierten Wahrscheinlichkeitsschwellenwert θ_4 überschreitet, und die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position nicht enden, einen vorgegebenen fünften Wahrscheinlichkeitsschwellenwert θ_5 überschreitet,

dem Muster des betreffenden Knoten die Option eines sofortigen Zeilenendes hinzugefügt wird, und bei der Prüfung einer Protokollzeile und einer Teilzeichenkette auf Übereinstimmung mit dem Muster eine Übereinstimmung dann als gegeben angesehen wird, wenn die jeweilige Teilzeichenkette mit dem Muster übereinstimmt und

- auch die nachfolgenden Teilzeichenketten der Protokollzeile den jeweils nachfolgenden Mustern des Syntaxbaums entsprechen, oder
- die Protokollzeile nach dieser Teilzeichenkette endet.

[0020] Eine verbesserte Maßnahme zur Erstellung von konkreten an den jeweiligen Systemzustand angepassten Bäumen sieht vor, dass die im Rahmen der Erstellung des Syntaxbaums verwendeten Schwellenwerte mit zunehmendem Abstand vom Wurzelknoten des Syntaxbaums oder mit zunehmender Pfadtiefe im Syntaxbaum ansteigen und/oder dem Abstand vom Wurzelknoten oder die Pfadtiefe angepasst werden, wobei insbesondere

- der erste bis vierte Wahrscheinlichkeitsschwellenwert mit zunehmendem Abstand vom Wurzelknoten oder fortschreitender Pfadtiefe im Syntaxbaum monoton ansteigt oder abfällt.

[0021] Um eine Anzahl von Grundbestandteilen von Protokollzeilen, wie beispielsweise IP-Adressen, als Musterbausteine in die Überprüfung integrieren zu können, kann vorgesehen sein, dass als Muster vorgegeben werden:

- vorab vorgegebene Grundmuster, insbesondere IP-Adressen oder andere strukturierte Daten, und/oder
- einzelne im Rahmen der Erstellung des Syntaxbaums festgelegte Zeichenketten.

[0022] Zur Anpassung des Syntaxbaums an geänderte Zustände des Systems kann vorgesehen sein, dass - für die einzelnen Pfade jeweils die bedingte Übergangswahrscheinlichkeit über eine vorgegebene Anzahl von Zeitfenstern, gebildet wird, und

- der weitere zeitliche Verlauf der so gebildeten bedingte Übergangswahrscheinlichkeiten der einzelnen Kanten, insbesondere nach dem jeweiligen Zeitfenster oder einer vorgegebenen Anzahl von Zeitfenstern, daraufhin untersucht wird, ob diese einen vorgegebenen Wahrscheinlichkeits-Schwellenwert unterschreiten und in diesem Fall
 - die ermittelten Pfade aus dem Graphen gestrichen werden und/oder
 - einzelnen Knoten des Graphen anstelle von unveränderlichen Teilen variable Teile der Protokollzeilen zugeordnet werden.

[0023] Um vorteilhaft erkennen zu können, wann der Syntaxbaum geändert werden soll, kann vorgesehen sein, dass zu einem späteren Zeitpunkt nach der Erstellung des Syntaxbaums eine signifikant hohe Anzahl von Protokollzeilen vorhanden ist, die keinem der durch Knoten und Kanten erstellten, gerichteten Pfade des Syntaxbaums zugeordnet sind, so kann für diese Protokollzeilen ein entsprechender, die einzelnen Protokollzeilen charakterisierender Pfad im Syntaxbaum durch Modifikation neu geschaffen werden und gegebenenfalls die den einzelnen Kanten zugeordneten Pfadwahrscheinlichkeiten werden in diesem Fall an die neu aufgetretenen Protokollzeilen angepasst werden.

[0024] Um einfach erkennen zu können, ob eine neu erstellte, weitere Protokollzeile den bisherigen Systemzustand repräsentiert oder ob diese neuartig ist und daher auf die Änderung des Systemzustands hindeutet, kann vorgesehen sein, dass vom Computersystem oder von auf diesem ablaufenden Prozessen zumindest eine weitere Protokollzeile erstellt wird, und dass mittels eines Parsers, der basierend auf dem Syntaxbaum erstellt wurde, untersucht wird, ob die weitere Protokollzeile mit dem Syntaxbaum übereinstimmt,

wobei gegebenenfalls eine fehlende Übereinstimmung als Hinweis für das Vorliegen eines abweichenden Systemzustands angesehen wird.

Bevorzugt kann ein Computerprogramm zur Durchführung eines erfindungsgemäßen Verfahrens auf einem Datenträger abgespeichert werden.

ERFASSUNG UND VORBEHANDLUNG VON PROTOKOLLZEILEN:

[0025] Konkret werden einige vorteilhafte Ausführungsformen der Erfindung anhand der vorliegenden Ausführungsbeispiele näher dargestellt:

[0026] In der folgenden Tabelle wird zum besseren Verständnis der Funktionsweise des im Folgenden gezeigten Parsergenerators ein Beispiel einer Anzahl von Protokollzeilen L_1, \dots, L_{100} dargestellt. Da die in der Praxis verwendeten Logfiles sehr unterschiedliche Strukturen aufweisen und üblicherweise über einen erheblichen Umfang verfügen, wird an dieser Stelle lediglich ein Logfile mit einer reduzierten Länge zu Demonstrationszwecken verwendet. Dabei sind die einzelnen, in den Logfiles verwendeten Wörter durch Großbuchstaben A, \dots, Z abstrahiert, wobei jeder der im Folgenden dargestellten Großbuchstaben A, \dots, Z in den einzelnen Zeilen jeweils einem bestimmten Wort entspricht. Alternativ besteht auch die Möglichkeit, dass einer der dargestellten Großbuchstaben A, \dots, Z in der folgenden Tabelle auch für die Verwendung eines bestimmten anderen Musters steht, beispielsweise kann einer der Buchstaben auch für das Auftreten einer

IP-Adresse oder einem Zeitstempel in einem vorgegebenen Format oder einer anderen strukturierten Zeichenkette stehen.

[0027] Die in der vorstehenden Tabelle dargestellten Protokollzeilen L_1, \dots, L_{100} wurden zu unterschiedlichen Zeitpunkten im betreffenden Computersystem aufgezeichnet, wobei die so erstellten Protokollzeilen L_1, \dots, L_{100} vorab mit einem Tokenisierungsalgorithmus in eine Vielzahl von Teilzeichenketten zerlegt wurden. Im Rahmen der Tokenisierung werden dabei unterschiedliche Sonder- oder Leerzeichen herangezogen, um eine Trennung der Protokollzeilen L_1, \dots, L_{100} in einzelne Teilzeichenketten zu bewirken. Im vorliegenden Fall wurden für die Trennung der einzelnen Protokollzeilen in Teilzeichenketten jeweils Leerzeichen herangezogen, um die Trennung in einzelne Teilzeichenketten zu gewährleisten. Es ist jedoch ohne Weiteres möglich, auch andere Sonderzeichen oder Teilzeichenketten heranzuziehen, mit denen typischerweise Protokollzeilen L_1, \dots, L_{100} in einzelne Wörter unterteilt sind. Dabei handelt es sich üblicherweise um Kommas, Tabulatoren, Klammern, Semikolons oder ähnliche Trenn- und Sonderzeichen.

Zeile	Position: 1	2	3	4	5
L ₁	Jul/18/00:00 :01	A	E	H	J
L ₂	Jul/18/00:00:04	B	E	I	K
L ₃	Jul/18/00:00:05	C	F	G	
L ₄	Jul/18/00:00:10	B	E	I	K
L ₅	Jul/18/00:00:18	A	E	H	J
L ₆	Jul/18/00:00:20	D	E	I	A
L ₇	Jul/18/00:00:21	B	E	I	K
L ₈	Jul/18/00:00:24	C	F	G	
L ₉	Jul/18/00:00:25	L	M	N	O
L ₁₀	Jul/18/00:00:26	A	E	H	J
L ₁₁	Jul/18/00:00:27	L	M	N	O
L ₁₂	Jul/18/00:00:28	A	E	H	J
L ₁₃	Jul/18/00:00:30	B	E	I	K
L ₁₄	Jul/18/00:00:31	A	E	H	J
L ₁₅	Jul/18/00:00:32	B	E	I	
L ₁₆	Jul/18/00:00:33	D	E	I	B
L ₁₇	Jul/18/00:00:35	A	E	H	J
L ₁₈	Jul/18/00:00:37	D	E	I	C
L ₁₉	Jul/18/00:00:38	D	E	I	D
L ₂₀	Jul/18/00:00:39	C	F	H	
L ₂₁	Jul/18/00:00:40	D	E	I	E
L ₂₂	Jul/18/00:00:41	A	E	H	J
L ₂₃	Jul/18/00:00:42	C	F	I	
L ₂₄	Jul/18/00:00:44	B	E	I	K
L ₂₅	Jul/18/00:00:45	D	X	I	F
L ₂₆	Jul/18/00:00:46	C	F	G	
L ₂₇	Jul/18/00:00:48	B	E	I	K
L ₂₈	Jul/18/00:00:49	D	E	I	G
L ₂₉	Jul/18/00:00:51	A	E	H	J
L ₃₀	Jul/18/00:00:55	C	F	H	
L ₃₁	Jul/18/00:00:58	A	E	H	J
L ₃₂	Jul/18/00:01:02	B	E	I	K
L ₃₃	Jul/18/00:01:03	D	E	I	H
L ₃₄	Jul/18/00:01:05	C	F	I	
L ₃₅	Jul/18/00:01:07	B	E	I	K
L ₃₆	Jul/18/00:01:09	C	F	G	
L ₃₇	Jul/18/00:01:10	D	E	I	I

L38	Jul/18/00:01:13	A	E	H	J
L39	Jul/18/00:01:14	B	E	I	
L40	Jul/18/00:01:15	D	E	I	J
L41	Jul/18/00:01:16	A	E	H	J
L42	Jul/18/00:01:17	D	E	I	K
L43	Jul/18/00:01:18	C	F	I	
L44	Jul/18/00:01:19	C	F	H	
L45	Jul/18/00:01:20	C	F	H	
L46	Jul/18/00:01:22	D	Y	I	L
L47	Jul/18/00:01:24	B	E	I	K
L48	Jul/18/00:01:25	D	E	I	M
L49	Jul/18/00:01:27	A	E	H	J
L50	Jul/18/00:01:28	B	E	I	K
L51	Jul/18/00:01:33	A	E	H	J
L52	Jul/18/00:01:34	C	F	G	
L53	Jul/18/00:01:36	L	M	N	O
L54	Jul/18/00:01:39	D	E	I	N
L55	Jul/18/00:01:41	C	F	G	
L56	Jul/18/00:01:42	D	E	I	O
L57	Jul/18/00:01:44	C	F	G	
L58	Jul/18/00:01:46	A	E	G	
L59	Jul/18/00:01:47	D	E	I	P
L60	Jul/18/00:01:49	C	F	W	
L61	Jul/18/00:01:50	B	E	I	K
L62	Jul/18/00:01:52	A	E	H	J
L63	Jul/18/00:01:54	C	F	I	
L64	Jul/18/00:01:55	A	E	H	J
L65	Jul/18/00:01:57	B	E	I	K
L66	Jul/18/00:01:59	D	E	I	Q
L67	Jul/18/00:02:00	B	E	I	K
L68	Jul/18/00:02:02	C	F	X	
L69	Jul/18/00:02:04	A	E	H	J
L70	Jul/18/00:02:06	A	E	H	J
L71	Jul/18/00:02:07	B	E	I	K
L72	Jul/18/00:02:08	D	E	I	R
L73	Jul/18/00:02:10	C	F	G	
L74	Jul/18/00:02:11	L	M	N	O
L75	Jul/18/00:02:13	D	E	I	S
L76	Jul/18/00:02:14	A	E	H	J
L77	Jul/18/00:02:16	D	E	I	T
L78	Jul/18/00:02:19	A	E	H	J
L79	Jul/18/00:02:22	C	F	Y	
L80	Jul/18/00:02:23	B	E	I	K
L81	Jul/18/00:02:25	A	E	H	J
L82	Jul/18/00:02:26	C	F	I	
L83	Jul/18/00:02:28	B	E	I	
L84	Jul/18/00:02:29	C	F	G	
L85	Jul/18/00:02:30	D	Z	I	U
L86	Jul/18/00:02:32	B	E	I	
L87	Jul/18/00:02:33	A	E	H	J
L88	Jul/18/00:02:36	D	E	I	V
L89	Jul/18/00:02:37	B	E	I	K
L90	Jul/18/00:02:39	C	F	G	

L ₉₁	Jul/18/00:02:41	C	F	Z	
L ₉₂	Jul/18/00:02:43	B	E	I	K
L ₉₃	Jul/18/00:02:46	A	E	H	J
L ₉₄	Jul/18/00:02:47	B	E	I	K
L ₉₅	Jul/18/00:02:48	C	F	I	
L ₉₆	Jul/18/00:02:51	B	E	I	K
L ₉₇	Jul/18/00:02:52	D	E	I	W
L ₉₈	Jul/18/00:02:53	A	E	H	J
L ₉₉	Jul/18/00:02:55	B	E	I	K
L ₁₀₀	Jul/18/00:02:59	D	E	I	X

[0028] Im vorliegenden Ausführungsbeispiel beginnt jede Protokollzeile L₁, ..., L₁₀₀ mit einem einzelnen Zeitstempel, der den Zeitpunkt der Erstellung der jeweiligen Protokollzeile L₁, ..., L₁₀₀ beschreibt. Dieser Aufbau von Protokollzeilen ist für L₁, ..., L₁₀₀ zwar grundsätzlich üblich, im Rahmen der Erfindung ist es jedoch nicht zwingend erforderlich, diesen Aufbau insgesamt zu wählen.

[0029] Für die einzelnen Knoten der Syntaxbäume können als Muster, denen die jeweiligen Teilzeichenketten zu entsprechen haben, fixe Teilzeichenketten vorgegeben werden. Daneben besteht auch die Möglichkeit, keine Vorgaben hinsichtlich einer Teilzeichenkette an einer bestimmten Position zu machen und dementsprechend einen Knoten mit variablem Inhalt zuzulassen. Darüber hinaus kann für die einzelnen im Rahmen der Erstellung von Syntaxbäumen verwendeten Knoten auch ein Muster, beispielsweise durch einen Regulären Ausdruck (engl. regular expression) vorgegeben werden. Dabei kann es sich beispielsweise um eine IP-Adresse oder einen Datumsstempel in einem bestimmten - mehr oder minder abstrakt gehaltenen - Datumsformat handeln, oder eingeschränkte Alphabete, wie zum Beispiel nur Ziffern und Punkte.

[0030] Die automatisierte Erstellung von Syntaxbäumen kann auf unterschiedliche Weise vorgenommen werden, wobei das im vorliegenden Fall dargestellte Ausführungsbeispiel der Erfindung eine besonders ressourcensparende, insbesondere zeit- und speichereffiziente, Vorgehensweise darstellt. Dabei wird nach Durchführung der Tokenisierung bzw. Aufteilung der einzelnen Protokollzeilen L₁, ..., L₁₀₀ in Teilzeichenketten eine spaltenweise Betrachtung der Protokollzeilen durchgeführt.

[0031] In diesem Zusammenhang werden jeder einzelnen Teilzeichenkette jeweils zwei Indizes zugeordnet, namentlich ein Zeilenindex, der die betreffende Protokollzeile L₁, ..., L₁₀₀, in der sich die Teilzeichenkette befindet, bezeichnet, und ein Positionsindex, der die Position der Teilzeichenkette innerhalb der jeweiligen Protokollzeile L₁, ..., L₁₀₀ angibt.

[0032] Nur beispielsweise wird die erste Protokollzeile, d.h. die Protokollzeile L₁ mit dem Zeilenindex 1, betrachtet, die wie folgt lautet:

JUL/18/00:00:01 A E H J

[0033] Die Protokollzeile L₁ enthält vier Leerzeichen, dementsprechend wird sie in fünf voneinander durch die Leerzeichen getrennte Teilzeichenketten unterteilt. Die einzelnen Teilzeichenketten sind nunmehr der Zeitstempel JUL/18/00:00:01 sowie die vier nachfolgenden zur Abstraktion von Wörtern verwendeten Buchstaben A, E, H und J. Die anderen Protokollzeilen werden auf dieselbe Weise in Teilzeichenketten unterteilt.

EFFIZIENTE VORGEHENSWEISE ZUR ERSTELLUNG EINES SYNTAXBAUMS:

[0034] Beim folgenden Vorgehen werden die einzelnen tabellarisch abgespeicherten Protokollzeilen spaltenweise betrachtet.

[0035] In einem ersten Schritt werden sämtliche Teilzeichenketten von Protokollzeilen betrachtet, deren Positionsindex den Wert 1 aufweist. Dabei handelt es sich im vorliegenden Ausführungs-

beispiel um die einzelnen Teilzeichenketten, in denen der Zeitstempel der jeweiligen Protokollzeilen enthalten ist.

[0036] Im Rahmen einer Analyse der einzelnen Teilzeichenketten mit Positionsindex 1 wird festgestellt, dass diese allesamt einem vorab bekannten und vorgegebenen Muster folgen, nämlich dem folgenden, vorab vorgegebenen, regulären Ausdruck entsprechen und dass die einzelnen Teilzeichenketten durch einen Knoten repräsentiert werden können, dem das folgende Muster zugewiesen ist:

$$(\text{Jan}|\text{Feb}|\text{Mar}|\text{Apr}|\text{May}|\text{Jun}|\text{Jul}|\text{Aug}|\text{Sep}|\text{Oct}|\text{Nov}|\text{Dec}) / (0[1-9] | [1-2][0-9] | 3[0-1]) / (2[0-3] | [01][0-9]) : [0-5][0-9] : [0-5][0-9]$$

[0037] Im Rahmen der Prüfung der einzelnen Teilzeichenketten der ersten Spalte werden für sämtliche der ermittelten Protokollzeilen Übereinstimmungen mit diesem Muster erkannt. Die Wahrscheinlichkeit dafür, dass eine Protokollzeile als erste Teilzeichenkette einen Zeitstempel enthält, der dem so festgelegten Muster entspricht, ist daher 1.

[0038] Der zu erstellende Syntaxbaum enthält einen Wurzelknoten, der den Einstieg in die Syntaxprüfung durch den Parser symbolisiert. Da im vorliegenden Fall die Wahrscheinlichkeit dafür, dass die Protokollzeilen an erster Stelle eine dem Muster entsprechende Teilzeichenkette aufweisen, einen ersten Wahrscheinlichkeitsschwellenwert $\theta_1 = 0,1$ sowie einen zweiten Wahrscheinlichkeitsschwellenwert $\theta_2 = 0,95$ überschreitet, wird in den Syntaxbaum ein einziger Knoten N_1 eingefügt. Dieser wird - als Zielknoten - über eine gerichtete Kante e_1 mit dem Wurzelknoten N_0 verbunden. Anhand des vorliegenden Beispiels wurde bereits eine Regel R2a zur rekursiven Erstellung von Syntaxbäumen näher dargestellt.

[0039] Wie an diesem Beispiel zu sehen war, wird dabei von einem Knoten, im vorliegenden Fall dem Wurzelknoten ausgegangen, dem sämtliche Protokollzeilen zugewiesen sind. Unter diesen Protokollzeilen wird mithilfe von statistischen Maßnahmen nach möglichst wenigen Mustern gesucht, die eine möglichst große Anzahl der ersten Teilzeichenkette gemeinsam haben. Im vorliegenden Fall konnte dies auf sehr einfache Weise gefunden werden, da die durch das Muster definierte Bedingung für sämtliche ersten Teilzeichenketten der Protokollzeilen zutrifft.

[0040] Zur Charakterisierung der relativen Häufigkeit, mit der die einzelnen Protokollzeilen bzw. deren erste Teilzeichenkette dem dem Knoten zugeordneten Muster entspricht, wird der Kante e_1 der Wert 1 bzw. 100% zugeordnet. Der im ersten Schritt erstellte Knoten N_1 wird einer ersten Schicht Y_1 des Syntaxbaums zugeordnet.

[0041] Im zweiten Schritt wird die zweite Spalte, die sämtliche Teilzeichenketten der Protokollzeilen enthält, deren Positionsindex 2 ist, betrachtet. In der vorstehend genannten Tabelle sind insgesamt fünf voneinander unterschiedliche Teilzeichenketten enthalten. Dabei ist zu beachten, dass die Teilzeichenketten A, B, C, D jeweils bei 24 der betrachteten Protokollzeilen an zweiter Stelle stehen, während die Teilzeichenkette L nur bei vier Protokollzeilen, nämlich bei den Protokollzeilen L_9 , L_{11} , L_{53} und L_{74} , an zweiter Stelle steht. Da die Protokolldatei insgesamt über 100 Zeilen verfügt, kann nun die Wahrscheinlichkeit bzw. relative Häufigkeit der Vorkommen der einzelnen Teilzeichenketten näher bestimmt werden. Im vorliegenden Ausführungsbeispiel betragen die relativen Häufigkeiten für das Auftreten von Teilzeichenketten A, B, C, D jeweils 0,24 bzw. 24 %, die Wahrscheinlichkeit für das Auftreten der Teilzeichenkette L an zweiter Stelle beträgt 0,04 oder 4 %.

[0042] Für den Aufbau eines Grammatikbaums ist im Folgenden zu untersuchen, welche der Teilzeichenketten mit ausreichender Häufigkeit in den Protokollzeilen vorkommen, dass sie in den Syntaxbaum übernommen werden können.

[0043] Dabei kann festgestellt werden, dass mehrere unterschiedliche Teilzeichenketten A, B, C, D bestehen, deren jeweilige relative Häufigkeit einen vorgegebenen zweiten Schwellenwert θ_1 überschreitet, während die relative Häufigkeit der Teilzeichenkette L diesen Schwellenwert θ_1 nicht überschreitet. Dieser Schwellenwert θ_1 wird im vorliegenden Ausführungsbeispiel mit $\theta_1=0,1$ festgelegt. Die Teilzeichenkette L, deren relative Häufigkeit den Schwellenwert θ_1 nicht

überschreitet, wird im Folgenden nicht beachtet und für die Erstellung des Syntaxbaums nicht weiter berücksichtigt.

[0044] In weiterer Folge wird die Summe derjenigen relativen Häufigkeiten für die Teilzeichenketten A, B, C, D ermittelt, die jeweils den ersten Schwellenwert überschreiten. Im vorliegenden Fall ergibt sich als Summe dieser relativen Häufigkeiten für die Teilzeichenketten A, B, C, D eine relative Häufigkeit von $0,24 + 0,24 + 0,24 + 0,24 = 0,96$. Dieser Wert überschreitet einen dritten Wahrscheinlichkeitsschwellenwert θ_3 , der im vorliegenden Ausführungsbeispiel einen Wert von $\theta_3 = 0,9$ aufweist. Da somit im vorliegenden Fall eine Mehrzahl von Teilzeichenketten aufgetreten ist, deren einzelne relative Häufigkeiten einen ersten Wahrscheinlichkeitsschwellenwert θ_1 überschreiten, und deren summierte relative Häufigkeit einen dritten Wahrscheinlichkeitsschwellenwert θ_3 überschreitet, werden bei der Erstellung des Syntaxbaums mehrere Knoten eingefügt, denen jeweils ein Muster zugeordnet ist, dem die Teilzeichenketten entsprechen. Die einzelnen Knoten N_{11} N_{12} N_{13} N_{14} werden über gerichtete Kanten als Zielknoten mit dem jeweils vorangehenden Knoten N_1 verbunden.

[0045] Anhand dieser Vorgehensweise wurde eine Regel R3a näher dargestellt: Für den Fall, dass eine einen zweiten Wahrscheinlichkeitsschwellenwert übersteigende Anzahl an Protokollzeilen, deren erste Teilzeichenketten mit den einzelnen Mustern des bis zu einem bestimmten Basisknoten zurückgelegten Teilpfads, d.h. bis zum Knoten N_1 des Syntaxbaums übereinstimmen, eine daran anschließende, hier zweite, Teilzeichenkette aufweisen, eine Anzahl von Mustern in Übereinstimmung bringbar ist,

wobei auch für jedes einzelne Muster A, B, C, D jeweils eine einen ersten Schwellenwert übersteigende Anzahl von Protokollzeilen in Übereinstimmung bringbar ist und die summierte Anzahl dieser Protokollzeilen einen dritten Schwellenwert übersteigt,

für jedes der Muster A, B, C, D jeweils ein separater, dieses Muster A, B, C, D enthaltende Knoten N_{11} , N_{12} , N_{13} , N_{14} erstellt wird, und diese Knoten mit dem Basisknoten über eine gerichtete Kante e_{11} , e_{12} , e_{13} , e_{14} verbunden wird.

[0046] Jedem Knoten N_{11} N_{12} N_{13} N_{14} wird jeweils ein Muster zugeordnet, dem die jeweiligen Teilzeichenketten entsprechen, d.h. der Syntaxbaum liefert bei der Überprüfung einer Protokollzeile bei der Verarbeitung der zweiten Teilzeichenkette dann einen positiven Übereinstimmungswert, wenn an der zweiten Stelle bzw. der Stelle mit dem Positionsindex 2 die Teilzeichenkette A, B, C oder D steht.

[0047] Im vorliegenden Ausführungsbeispiel werden wiederum den einzelnen Kanten noch diejenigen Wahrscheinlichkeitswerte bzw. relativen Häufigkeitswerte zugewiesen, die angeben, wie wahrscheinlich es für eine Protokollzeile, die dem jeweils vorangehenden Knoten N_1 zugewiesen ist, ist, dass sie an ihrer zweiten Stelle eine Teilzeichenkette enthält, die dem, dem Knoten zugeordneten, Muster entspricht.

[0048] Zur besseren Verständlichkeit der verbleibenden Wahrscheinlichkeit wurde in Fig. 1 eine zusätzliche, strichliert gekennzeichnete Kante f_{15} dargestellt, die zu einem nicht im Syntaxbaum enthaltenen Knoten M_{15} führt, der für die insgesamt vier Vorkommen der Teilzeichenkette \perp in den Protokollzeilen steht. Ebenso ist lediglich zu Illustrationszwecken der betreffenden, den Knoten mit dem Wurzelknoten verbindenden Kante der Häufigkeitswert 0,04 zugewiesen, mit dem eine Teilzeichenkette \perp an der betreffenden Position auftritt. Dieser Knoten ist zwar durch die Wahrscheinlichkeitsverteilung der einzelnen zweiten Teilzeichenketten begründet, wird aber nicht in den Syntaxbaum eingetragen und dient auch nicht der Charakterisierung des Computersystems.

Während aufgrund der ersten Teilzeichenketten der Protokollzeilen keine Unterscheidung möglich war und alle Protokollzeilen dem Knoten N_1 zugewiesen wurden, werden nunmehr die Protokollzeilen zur weiteren Verarbeitung bzw. zur weiteren Erstellung des Syntaxbaums auf die im zweiten Schritt erstellten Knoten N_{11} , N_{12} , N_{13} , N_{14} verteilt. Diejenigen Protokollzeilen L_9 , L_{11} , L_{53} und L_{74} die keinem der den Knoten N_{11} , N_{12} , N_{13} , N_{14} zugeordneten Mustern entsprechen, werden für die weitere Verarbeitung nicht herangezogen bzw. zu Illustrationszwecken dem nicht dem end-

gültigen Syntaxbaum zugehörigen Knoten M_{15} zugeordnet bzw. nicht mehr weiterverarbeitet. Dabei handelt es sich um diejenigen Protokollzeilen, die an zweiter Stelle die Teilzeichenkette \perp aufweisen.

[0049] Die im zweiten Schritt erstellten Knoten N_{11} , N_{12} , N_{13} , N_{14} werden einer zweiten Schicht Y_2 des Syntaxbaums zugeordnet.

[0050] In einem dritten Schritt werden nunmehr die Teilzeichenketten betrachtet, die an dritter Stelle in der jeweiligen Protokollzeile stehen bzw. deren Positionsindex gleich 3 ist. Bei der Verarbeitung der Teilzeichenketten, die sich in den einzelnen Protokollzeilen an dritter Stelle befinden, werden die einzelnen Teilzeichenketten nach der Zuordnung ihrer jeweiligen Protokollzeilen zu einem der Knoten N_{11} , N_{12} , N_{13} , N_{14} separat weiterverarbeitet. Die einzelnen Teilzeichenketten werden knotenweise getrennt verarbeitet, d.h. diejenigen Teilzeichenketten, die an zweiter Stelle die Teilzeichenkette A aufweisen, werden unabhängig von denjenigen Protokollzeilen behandelt, die an zweiter Stelle die Teilzeichenkette B aufweisen, etc.

[0051] Die zuvor beschriebenen Berechnungen der relativen Häufigkeiten, auch als Pfadfrequenzen bezeichnet, werden nun erneut durchgeführt. Da unter den Teilzeichenketten, die an erster Stelle einen Zeitstempel, die Teilzeichenkette A an zweiter Stelle aufweisen, alle Zeichenketten an dritter Stelle die Teilzeichenkette E aufweisen, ist die relative Häufigkeit unter diesen Protokollzeilen dafür, dass sie an dritter Stelle die Teilzeichenkette E aufweisen, gleich 1. Da diese relative Häufigkeit den ersten und zweiten Wahrscheinlichkeitsschwellenwert θ_1 , θ_2 überschreitet und es keine anderen Teilzeichenketten gibt, für die dies auch der Fall ist, wird entsprechend der Regel R2a in den Syntaxbaum lediglich ein einziger weiterer Knoten N_{111} über eine neu erstellte, gerichtete Kante e_{111} eingefügt. Dies bedeutet, dass der Syntaxbaum an der betreffenden Stelle lediglich eine Möglichkeit für Protokollzeilen umfasst. Der so erstellten gerichteten Kante e_{111} wird der Wahrscheinlichkeitswert 1 zugewiesen, sämtliche Protokollzeilen des Knotens N_{11} werden dem neu erstellten Knoten N_{111} zugewiesen.

[0052] Betrachtet man nun die Protokollzeilen, die einen Zeitstempel an erster Stelle und die Teilzeichenkette B an zweiter Stelle aufweisen, so kann festgestellt werden, dass unter diesen sämtliche Protokollzeilen ein \perp an dritter Stelle aufweisen.

[0053] Unter denjenigen Protokollzeilen, die an erster Stelle einen Zeitstempel und an zweiter Stelle die Teilzeichenkette C aufweisen, folgen ausschließlich Teilzeichenketten, die an dritter Stelle die Teilzeichenkette F aufweisen.

[0054] Entsprechend Regel R2a kann daher dem Knoten N_{12} über eine neu einzufügende Kante e_{121} ein einziger Knoten N_{121} mit einer Pfadfrequenz 1 nachgeordnet werden, dem sämtliche der Protokollzeilen des Knotens N_{12} zugeordnet werden.

[0055] Ebenso kann entsprechend Regel R2a dem Knoten N_{12} über eine neu einzufügende Kante e_{121} ein einziger Knoten N_{121} mit einer Pfadfrequenz 1 nachgeordnet werden, dem sämtliche der Protokollzeilen des Knotens N_{12} zugeordnet werden.

[0056] Betrachtet man in weiterer Folge diejenigen Protokollzeilen des Knotens N_{14} , an deren erster Stelle ein Zeitstempel und an deren zweiter Stelle eine Teilzeichenkette D enthalten ist, so kann festgestellt werden, dass von diesen 24 Protokollzeilen genau 21 Protokollzeilen ein E an dritter Stelle aufweisen, die restlichen drei Zeilen jedoch entweder ein X , ein Y oder ein Z . Die relative Häufigkeit der Teilzeichenkette E weist somit einen Wert von $21/24 = 0,876$ auf, die relative Häufigkeit der übrigen Teilzeichenketten X , Y , Z weist den Wert 0,04167 auf. Daher ist die relative Häufigkeit der Teilzeichenkette E die einzige relative Häufigkeit, die den ersten Wahrscheinlichkeitsschwellenwert θ_1 übersteigt.

Die relative Häufigkeit der Teilzeichenketten E liegt jedoch unter θ_2 , sodass im Folgenden aufgrund der Regel R2b kein eigener Knoten für die einzelnen Teilzeichenketten E , X , Y , Z gebildet wird, vielmehr wird ein Knoten N_{141} erstellt, der ein variables Muster enthält, das mit jeder Teilzeichenkette übereinstimmt. Der Knoten N_{141} wird über die Kante e_{141} als Zielknoten mit dem Knoten

N_{14} verbunden. Die Protokollzeilen des Knotens N_{14} werden dem Knoten N_{141} zugewiesen.

Knoten mit variablen Mustern werden eingesetzt, um Teile von Protokollzeilen, die häufigen Änderungen unterliegen, darzustellen. Das verhindert, dass der Syntaxbaum einem zu großen Wachstum unterliegt, was zu einer enorm großen Komplexität des Syntaxbaums führen könnte. Zusätzlich sind solche variablen Knoten ausschlaggebend, um nicht nur die verwendeten Protokollzeilen parsen zu können, sondern auch neue und unbekannte Protokollzeilen sinnvoll zu parsen. Das liegt daran, dass Teile der Protokollzeilen, die bei den Eingabedaten mit einer hohen Variabilität, d.h., ausschließlich oder fast ausschließlich unterschiedliche und sich nicht wiederholende Zeichenketten, beispielsweise einen sich stetig erhöhenden Zeilenindex, auftreten, vermutlich auch in den unbekanntem Daten mit einer ähnlichen Variabilität vorkommen, jedoch nicht genau die gleichen Zeichen umfassen. Im Syntaxbaum sind Knoten, die solche variablen Muster enthalten, als Fünfecke dargestellt.

[0057] Die im dritten Schritt erstellten Knoten N_{111} , N_{121} , N_{131} , N_{141} werden einer dritten Schicht Y_3 des Syntaxbaums zugeordnet.

[0058] Nachdem nun die dritte Spalte der vorstehenden Tabelle der Protokollzeilen abgearbeitet ist, werden in einem vierten Schritt die Teilzeichenketten mit Positionsindex 4 behandelt, d.h. die an vierter Stelle in der Tabelle der Protokollzeilen enthalten sind.

[0059] Sofern bei einigen der Protokollzeilen an erster Stelle ein Zeitstempel, an zweiter Stelle die Teilzeichenkette A und an dritter Stelle die Teilzeichenkette E vorhanden ist, wird nunmehr untersucht, mit welcher relativen Häufigkeit Teilzeichenketten an der vierten Stelle auftreten. Dabei sind grundsätzlich mehrere Teilzeichenketten G , H als potentielle Nachfolger erkennbar. Die relative Wahrscheinlichkeit für das Auftreten der Teilzeichenkette G beträgt $1/24 = 4,17\%$, die relative Wahrscheinlichkeit für das Auftreten der Teilzeichenkette H beträgt $23/24 = 95,83\%$. Die relative Wahrscheinlichkeit für das Auftreten der Teilzeichenkette H übersteigt damit den ersten und zweiten Wahrscheinlichkeitsschwellenwert θ_1 , θ_2 . In diesem Fall wird entsprechend Regel R2a lediglich ein einziger Knoten N_{1111} mit einem Muster eingefügt, das ausschließlich mit der Teilzeichenkette H , nicht jedoch mit der Teilzeichenkette G , eine Übereinstimmung liefert. Der Kante, die den Knoten N_{111} mit dem Knoten N_{1111} verbindet, wird eine relative Wahrscheinlichkeit von $95,83\%$ zugeordnet. Dem Knoten N_{1111} werden alle dem Knoten N_{111} zugeordneten Protokollzeilen zugeordnet, an deren vierter Position die Teilzeichenkette H enthalten ist. In Fig. 1 ist darüber hinaus auch noch der, nicht im Syntaxbaum enthaltene Knoten M_{1112} dargestellt, der die Übergangswahrscheinlichkeit für die Teilzeichenkette G symbolisiert. Die eine Teilzeichenkette L_{58} wird diesem Knoten zugeordnet und/oder nicht weiter für die Erstellung des Syntaxbaums herangezogen.

[0060] Da unter den Teilzeichenketten, die an erster Stelle einen Zeitstempel, den Teilzeichenkette B an zweiter Stelle und E an dritter Stelle aufweisen, alle Zeichenketten an vierter Stelle die Teilzeichenkette I aufweisen, ist die relative Häufigkeit unter diesen Protokollzeilen dafür, dass sie an vierter Stelle die Teilzeichenkette I aufweisen, gleich 1. Da diese relative Häufigkeit den ersten und zweiten Wahrscheinlichkeitsschwellenwert θ_1 , θ_2 überschreitet und es keine anderen Teilzeichenketten gibt, für die dies auch der Fall ist, wird entsprechend der Regel R2a in den Syntaxbaum lediglich ein einziger weiterer Knoten N_{1211} über eine neu erstellte, gerichtete Kante e_{1211} eingefügt. Dies bedeutet, dass der Syntaxbaum an der betreffenden Stelle lediglich eine Möglichkeit für Protokollzeilen umfasst. Der so erstellten gerichteten Kante e_{1211} wird der Wahrscheinlichkeitswert 1 zugewiesen, sämtliche Protokollzeilen des Knotens N_{121} werden dem neu erstellten Knoten N_{1211} zugewiesen.

[0061] Betrachtet man die Protokollzeilen, die einen Zeitstempel an der ersten Stelle, eine Teilzeichenkette C an der zweiten Stelle und eine Teilzeichenkette F an der dritten Stelle aufweisen, so sind mehrere Teilzeichenketten als potentielle Nachfolgerknoten an der vierten Position erkennbar. Die Pfadfrequenzen für das Auftreten der Teilzeichenketten an der vierten Stelle sind wie folgt: G : 0,4167; H : 0,167; I : 0,25; W , X , Y , Z : 0,04167.

Das bedeutet, dass die relativen Wahrscheinlichkeiten für das Auftreten der Teilzeichenketten G ,

H und I den ersten Wahrscheinlichkeitsschwellenwert θ_1 überschreiten, die relativen Wahrscheinlichkeiten für das Auftreten der Teilzeichenketten W, X, Y und Z jedoch nicht. Da gleich mehrere Wahrscheinlichkeitswerte für das Auftreten von Teilzeichenketten den ersten Wahrscheinlichkeitsschwellenwert θ_1 überschreiten, gelangt eine der Regeln R3a, R3b zur Anwendung.

Die Summe der Pfadfrequenzen der Teilzeichenketten, die deren relative Wahrscheinlichkeit den ersten Wahrscheinlichkeitsschwellenwert θ_1 überschreiten ist $0,4167 + 0,167 + 0,25 = 0,8337$ und überschreitet somit nicht den dritten Wahrscheinlichkeitsschwellenwert θ_3 . Aus diesem Grund werden hier keine eigenen Knoten für die Teilzeichenketten G, H und I erstellt, sondern ein variabler Knoten mit einem Muster gebildet, dem jede beliebige Teilzeichenkette entspricht.

[0062] In weiterer Folge werden diejenigen Protokollzeilen weiterverarbeitet, die dem Knoten N_{141} zugeordnet sind, dh es handelt sich dabei um Protokollzeilen, die an der ersten Stelle einen Zeitstempel aufweisen, an der zweiten Stelle die Teilzeichenkette D aufweisen und an der dritten Stelle eine beliebige Teilzeichenkette aufweisen. Es stellt sich heraus, dass alle diese Zeilen die Teilzeichenkette I an der vierten Stelle aufweisen. Ähnlich wie bereits für andere Knoten beschrieben wird entsprechend Regel R2a ein einziger Knoten N_{1411} mit einem der Teilzeichenkette I entsprechenden Muster erstellt.

Die im vierten Schritt erstellten Knoten N_{1111} , N_{1211} , N_{1311} , N_{1411} werden einer vierten Schicht Y_4 des Syntaxbaums zugeordnet.

[0063] In einem fünften Schritt werden die einzelnen Protokollzeilen hinsichtlich ihrer Teilzeichenketten an der jeweils fünften Stelle untersucht.

[0064] Unter den dem Knoten N_{1111} zugeordneten Protokollzeilen befinden sich ausschließlich Protokollzeilen, an deren fünfter Stelle ausschließlich Teilzeichenketten J enthalten sind. Entsprechend Regel R2a wird ein einziger Knoten mit einem Muster in den Syntaxbaum eingefügt, der lediglich auf die Teilzeichenkette J übereinstimmt.

[0065] In den Protokollzeilen ist allerdings auch ersichtlich, dass ein eine beliebige Teilzeichenkette auf I folgt. Jede der Teilzeichenketten ist eindeutig, somit ist die Pfadfrequenz einer jeden Teilzeichenkette 0,04167. Keine dieser Pfadfrequenzen überschreitet den ersten Wahrscheinlichkeitsschwellenwert θ_1 und somit tritt Fall 1 (Regel R1) ein. In diesem Fall wird aus den bereits genannten Gründen jedenfalls ein variabler Knoten erstellt.

Betrachtet man nun die Protokollzeilen des Knotens N_{1211} , die einen Zeitstempel an erster Stelle, ein B an zweiter Stelle, ein E an dritter Stelle aufweisen und ein I an vierter Stelle aufweisen, so erkennt man, dass nicht alle Protokollzeilen über weitere nachfolgende Teilzeichenketten verfügen, sondern an dieser Stelle enden (Protokollzeilen L_{15} , L_{39} , L_{83} , L_{86}). Gleichbedeutend haben nur 20 der 24 Protokollzeilen, die I an vierter Stelle aufweisen, die Teilzeichenkette K an fünfter Stelle stehen. Da der Anteil der endenden Logzeilen $4/24 = 0,167$ beträgt und den vierten Wahrscheinlichkeitsschwellenwert $\theta_4 = 0,01$ überschreitet, werden entsprechend Regel R4 alle nachfolgenden Knoten als optional betrachtet. Da der Anteil der Protokollzeilen, die an vierter Stelle die Teilzeichenkette I aufweisen, und an fünfter Stelle die Teilzeichenkette K stehen haben, $20/24 = 0,83$ beträgt, und den fünften Wahrscheinlichkeitsschwellenwert $\theta_5 = 0,01$ überschreitet, sind entsprechend Regel R4 nachfolgende optionale Knoten möglich. Um zu entscheiden, ob ein variabler Knoten oder ein bzw. mehrere fixe Knoten gebildet werden, werden nun die ursprünglichen drei Regeln überprüft, wobei als Grundwert 20, die Anzahl der nicht endenden Zeilen, herangezogen wird. Da die Zeilen, die an vierter Stelle die Teilzeichenkette I aufweisen, und an fünfter Stelle die Teilzeichenkette K stehen haben, demnach mit einer Häufigkeit von $20/20 = 1$ auftreten und damit sowohl θ_1 als auch θ_2 überschreiten kommt Regel R2a zum Einsatz und es wird ein fixer Knoten N_{12111} gebildet. In der Darstellung des Syntaxbaums ist dieser Umstand dadurch gekennzeichnet, dass der Knoten I als Achteck markiert ist. Das bedeutet, dass eine dem Syntaxbaum entsprechende Protokollzeile, entweder in diesem Knoten N_{1211} entspricht und sofort anschließend endet, oder aber alle nachfolgenden Knoten N_{12111} erfüllt.

Da die den Knoten N_{1311} , N_{1411} zugeordneten Protokollzeilen keine Teilzeichenketten an fünfter Stelle aufweisen, ist der Aufbau des Syntaxbaums bei diesen Knoten abgeschlossen und es wer-

den keine neuen Kanten oder Knoten eingefügt.

[0066] Da ebenso die den Knoten N_{11111} , N_{12111} zugeordneten Protokollzeilen keine Teilzeichenketten an sechsten Stelle aufweisen, ist der Aufbau des Syntaxbaums auch bei diesen Knoten abgeschlossen und es werden keine neuen Kanten oder Knoten eingefügt.

[0067] Die im fünften Schritt erstellten Knoten N_{11111} , N_{12111} werden einer fünften Schicht Y_5 des Syntaxbaums zugeordnet.

REGELN ZUR ERSTELLUNG DES SYNTAXBAUMS:

[0068] Bezogen auf die zu analysierenden Teilzeichenketten können die Regeln wie folgt zusammengefasst werden:

[0069] Regel R1: Kann für die Teilzeichenketten kein Muster gefunden werden, das eine relative Häufigkeit aufweist, die einen Wahrscheinlichkeitsschwellenwert θ_1 übersteigt, so wird an der betreffenden Stelle ein einziger Knoten eingefügt, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt. Der jeweiligen neu erstellten Kante wird eine Wahrscheinlichkeit von 1 zugewiesen. Alle dem Quellknoten der Kante zugewiesenen Protokollzeilen werden dem neu eingefügten Zielknoten zugewiesen.

[0070] Regel R2: Kann für die Teilzeichenketten ein einziges Muster gefunden werden, sodass die relative Häufigkeit der dem Muster entsprechenden Teilzeichenketten einen ersten Wahrscheinlichkeitsschwellenwert θ_1 übersteigt, so ist wie folgt zu unterscheiden:

[0071] Regel R2a: Überschreitet diese relative Häufigkeit auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 , so wird ein einziger Knoten mit dem betreffenden Muster eingefügt. Der jeweiligen neu erstellten Kante wird der Anteil der dem Muster entsprechenden Teilzeichenketten zugewiesen. Alle dem Quellknoten der Kante zugewiesenen Protokollzeilen, die an der betreffenden Position eine dem Muster entsprechende Teilzeichenkette enthalten, werden dem neu eingefügten Zielknoten zugewiesen.

[0072] Regel R2b: Überschreitet diese relative Häufigkeit nicht auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 , so wird an der betreffenden Stelle ein einziger Knoten eingefügt, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt. Der jeweiligen neu erstellten Kante wird eine Wahrscheinlichkeit von 1 zugewiesen. Alle dem Quellknoten der Kante zugewiesenen Protokollzeilen werden dem neu eingefügten Zielknoten zugewiesen.

[0073] Regel R3: Kann für die Teilzeichenketten eine Mehrzahl der Muster gefunden werden, sodass die relative Häufigkeit der den einzelnen Mustern entsprechenden Teilzeichenketten jeweils einzeln einen ersten Wahrscheinlichkeitswert θ_1 übersteigt, so ist wie folgt zu unterscheiden:

[0074] Regel R3a: Überschreitet die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 , so wird eine Anzahl von Knoten mit jeweils einem der ermittelten Muster eingefügt, wobei jeder Knoten über jeweils eine Kante mit dem ursprünglichen Knoten verbunden wird. Den so neu erstellten Kanten wird der Anteil der dem Muster entsprechenden Teilzeichenketten an der Gesamtzahl der dem ursprünglichen Knoten zugewiesenen Protokollzeilen zugewiesen. Die einzelnen Protokollzeilen werden auf die Knoten aufgeteilt, sodass jede Protokollzeile jeweils demjenigen Knoten zugeordnet wird, dessen Muster ihre jeweils betrachtete Teilzeichenkette entspricht.

[0075] Regel R3b: Überschreitet die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 nicht, so wird an der betreffenden Stelle ein einziger Knoten eingefügt, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt. Der jeweiligen neu erstellten Kante wird eine Wahrscheinlichkeit von 1 zugewiesen. Alle dem Quellknoten der Kante zugewiesenen Protokollzeilen werden dem neu eingefügten Zielknoten zugewiesen.

[0076] Regel R4: Wenn die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position enden, einen vorgegebenen vierten Wahrscheinlichkeitsschwellenwert θ_4 überschreitet und die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position nicht enden, einen vorgege-

benen fünften Wahrscheinlichkeitsschwellenwert θ_5 überschreitet, kann dem betreffenden Knoten ein Muster mit optionalem sofortigen Ende hinzugefügt werden. In diesem Fall wird bei der Prüfung einer Protokollzeile und einer Teilzeichenkette auf Übereinstimmung mit dem Muster eine Übereinstimmung dann als gegeben angesehen wird, wenn die jeweilige Teilzeichenkette mit dem Muster übereinstimmt und

- auch die nachfolgenden Teilzeichenketten der Protokollzeile den jeweils nachfolgenden Mustern des Syntaxbaums entsprechen, oder
- die Protokollzeile nach dieser Teilzeichenkette endet.

Als Grundwert für die Entscheidung, ob ein variabler Knoten, oder ein bzw. mehrere fixe Knoten entstehen wird die Anzahl der Protokollzeilen, die nicht endet herangezogen. Zur Entscheidung, welche Knoten erstellt werden sollen, werden die ersten drei Regeln herangezogen, mit dem eben beschriebenen Grundwert.

PRÜFUNG AUF ANORMALE ZUSTÄNDE:

[0077] Im Folgenden werden zwei Möglichkeiten der Prüfung, ob ein anormaler Zustand in dem die Protokollzeilen erstellenden Computersystem vorliegt, dargestellt. Bei der ersten Möglichkeit werden Änderungen im betreffenden Syntaxbaum untersucht. Im einfachsten Fall wird jeweils ein Syntaxbaum auf die vorstehend beschriebene Weise für zwei nicht idente Zeiträume separat erstellt. Dabei werden zwei Syntaxbäume erhalten, die bis identischer Funktionalität des Systems über die Zeit im wesentlichen dieselbe Form aufweisen sollten. Insbesondere dann, wenn hinreichend lange Zeiträume für die Untersuchung gewählt werden, in denen das Auftreten einzelner Typen von Protokollzeilen mit hinreichend großer Wahrscheinlichkeit bzw in hinreichend großer Anzahl vorhergesagt werden kann, sollten die beiden erstellten Syntaxbäume im wesentlichen dieselbe Struktur aufweisen.

[0078] In diesem Fall können die Bäume miteinander verglichen werden, und anhand des Vergleichs kann ein Übereinstimmungsmaß ermittelt werden, das angibt, wie stark die beiden Systemzustände voneinander abweichen. Bei der Ermittlung des Übereinstimmungsmaßes können auch die einzelnen bei der Erstellung des Syntaxbaums verwendeten Wahrscheinlichkeiten verwendet werden.

Zum Vergleich von zwei Syntaxbäumen können grundsätzlich unterschiedliche Verfahren herangezogen werden.

[0079] Eine vorteilhafte Methode besteht beispielsweise darin, die Protokollzeilen, die für die Erstellung des einen Syntaxbaums verwendet wurden, anschließend mithilfe des anderen Syntaxbaumes zu parsen. Wird dies auch umgekehrt durchgeführt, so können die Überschneidungen der Zuweisungen der Protokollzeilen als Ähnlichkeitsmaß herangezogen werden. Solche Ähnlichkeitsmaße sind beispielsweise der F-Score oder der Rand Index (Introduction to Information Retrieval, Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich, Cambridge University Press, 2008).

[0080] Eine weitere bevorzugte Variante der Prüfung, ob ein anormaler Systemzustand vorliegt, besteht darin, die einzelnen Protokollzeilen von einem Parser prüfen zu lassen, dessen Funktionalität durch den Syntaxbaum festgelegt wird, bzw der eine Übereinstimmung mit dem Syntaxbaum prüft. Da diejenigen Protokollzeilen, die für die Erstellung des Syntaxbaums herangezogen wurden, aufgrund des zuvor beschriebenen Vorgehens mit großer Wahrscheinlichkeit P_1 auch diesem Syntaxbaum entsprechen, ist dies für später erstellte Protokollzeilen nicht notwendigerweise der Fall. Sofern die Wahrscheinlichkeit P_2 , dass die während eines anderen, insbesondere nachfolgenden, Zeitraums erstellten Protokollzeilen dem Syntaxbaum entsprechen, gegenüber der Wahrscheinlichkeit P_1 signifikant bzw um einen vorgegebenen Schwellenwert verringert ist, kann ein anormaler Systemzustand detektiert werden.

[0081] Für die Prüfung, ob eine Protokollzeile dem Syntaxbaum entspricht, wird die Protokollzeile in ihre Teilzeichenketten untersucht. Dabei werden die Teilzeichenketten, die in der Protokollzeile L_1, \dots, L_{100} erhalten sind entsprechend ihrer Position in der Protokollzeile jeweils mit einzelnen

Mustern der Knoten N_1 , N_{11} , N_{111} , ... des Syntaxbaums in der durch diesen vorgegebenen Reihenfolge verglichen. Wird eine Übereinstimmung zwischen den Teilzeichenketten und den Knoten auf einem gerichteten Teilpfad des Syntaxbaums bzw deren Mustern gefunden, so entspricht die Protokollzeile insgesamt der durch den Syntaxbaum vorgegebenen Grammatik. Kann hingegen kein einziger gerichteter Teilpfad im Syntaxbaum gefunden werden, dem die Protokollzeile entspricht, so entspricht diese nicht der durch den Syntaxbaum vorgegebenen Grammatik und stellt daher einen anormalen Zustand dar.

VERÄNDERUNG DES SYNTAXBAUMS:

[0082] Der Syntaxbaum muss nicht notwendigerweise für einzelne Zeiträume oder Zeitabschnitte völlig neu ermittelt werden. Es besteht auch die Möglichkeit, den Syntaxbaum sukzessive anzupassen, um eine, beispielsweise tageweise, Neuermittlung des Syntaxbaums zu vermeiden. Dabei können die bei der Erstellung des Syntaxbaums vorgegebenen Pfadwahrscheinlichkeiten sukzessive anhand der neu erstellten Protokollzeilen angepasst werden. Für die einzelnen Kanten werden jeweils die bedingten Übergangswahrscheinlichkeiten über eine vorgegebene Anzahl von Zeitfenstern aktualisiert.

Dabei kann der Fall auftreten, dass die so gebildeten bedingte Übergangswahrscheinlichkeiten der einzelnen Kanten, insbesondere nach dem jeweiligen Zeitfenster oder einer vorgegebenen Anzahl von Zeitfenstern einen vorgegebenen Wahrscheinlichkeits-Schwellenwert unterschreiten. Dieser Umstand kann in vorgegebenen Zeitabständen untersucht werden. In diesem Fall können die folgenden Adaptierungen des Syntaxbaums vorgenommen werden.

- einzelne Kanten oder gerichtete Pfade, deren Wahrscheinlichkeit unter einen bestimmten Schwellenwert gesunken sind, werden aus dem Syntaxbaum gelöscht. Sofern ein Knoten aus dem Syntaxbaum gelöscht wird, werden auch alle ihm nachfolgenden Knoten und Kanten des Syntaxbaums gelöscht.
- das Muster von Knoten, denen ein vorgegebenes, nicht beliebiges, Muster zugewiesen ist, wird durch ein beliebiges Muster ersetzt.

[0083] Sofern zu einem späteren Zeitpunkt nach der Erstellung des Syntaxbaums eine signifikant hohe Anzahl von Protokollzeilen vorhanden ist, die keinem der durch Knoten und Kanten erstellten, gerichteten Pfade des Syntaxbaums zugeordnet sind, so kann für diese Protokollzeilen ein entsprechender, die einzelnen Protokollzeilen charakterisierender Pfad im Syntaxbaum durch Modifikation neu geschaffen werden. Die den einzelnen Kanten zugeordneten Pfadwahrscheinlichkeiten werden in diesem Fall an die neu aufgetretenen Protokollzeilen angepasst. Damit dies eintritt müssen auch wieder die zuvor genutzten Bedingungen erfüllt werden.

Patentansprüche

1. Verfahren zur Charakterisierung des Zustands eines Computersystems, wobei
 - vom Computersystem oder von auf diesem ablaufenden Prozessen jeweils Protokolle erstellt werden, indem bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils eine Protokollzeile (L_1, \dots, L_{100}) erstellt wird und wobei die Protokollzeile (L_1, \dots, L_{100}) das jeweils protokollierte Ereignis beschreibt, und
 - wobei jede derart erstellte Protokollzeile (L_1, \dots, L_{100}) in eine Anzahl von Teilzeichenketten unterteilt wird,

dadurch gekennzeichnet,

 - dass aufgrund der einzelnen Protokollzeilen (L_1, \dots, L_{100}) sowie der Abfolge der einzelnen, in den Protokollzeilen (L_1, \dots, L_{100}) enthaltenen Teilzeichenketten sowie aufgrund der Häufigkeit des Vorkommens der Protokollzeilen und der Teilzeichenketten in den Protokollzeilen ein die mögliche Abfolge von Teilzeichenketten beschreibender Syntaxbaum erstellt wird, und
 - dass dieser Syntaxbaum als charakteristisch für den Zustand des Computersystems angesehen wird und
 - dass der Syntaxbaum anhand der einzelnen Protokollzeilen als azyklischer, gerichteter Graph erstellt wird,
 - wobei der Syntaxbaum Knoten (N_1, N_2) aufweist, denen jeweils ein Muster (P_1, P_2) zugeordnet wird, das bei Anwendung auf eine der Teilzeichenketten jeweils einen positiven oder negativen Übereinstimmungswert liefert,
 - wobei der Syntaxbaum einzelne gerichtete Kanten aufweist, die ihren jeweiligen Quellknoten mit ihrem jeweiligen Zielknoten verbinden, wenn unter den einzelnen Protokollzeilen die bedingte Wahrscheinlichkeit dafür, dass
 - unter der Bedingung, dass Teilzeichenketten in der Protokollzeile (L_1, \dots, L_{100}) enthalten sind, die entsprechend ihrer Reihenfolge mit Mustern der einzelnen Knoten (N_1, N_{11}, N_{111}) auf einem vom Wurzelknoten (N) des Syntaxbaums zum Quellknoten (N_{11}) der Kante (e_{111}) führenden gerichteten Teilpfad entsprechend der Reihenfolge der Knoten (N_1, N_{11}, N_{111}) in diesem Teilpfad jeweils eine positive Übereinstimmung liefern, und
 - als nächste Teilzeichenkette (s_{13}) in der betreffenden Protokollzeile (s_1) eine Teilzeichenkette vorliegt, die eine positive Übereinstimmung mit dem im Zielknoten (N_{111}) der betreffenden Kante (e_{111}) gespeicherten Muster aufweist, einen vorgegebenen, allenfalls von der Position im Syntaxbaum abhängigen, Schwellenwert übersteigt, und
 - insbesondere die bedingte Übergangswahrscheinlichkeit der betreffenden Kante zugeordnet wird.
2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet,** dass
 - a) anhand einer Anzahl von vorgegebenen oder innerhalb eines ersten Zeitraums erstellten Protokollzeilen (L_1, \dots, L_{100}) ein Syntaxbaum, insbesondere nach einem Verfahren nach einem der vorangehenden Ansprüche, erstellt wird,
 - b) die von den Computern oder den auf diesen Computern ablaufenden Prozessen bei Auftreten vorgegebener Ereignisse für jedes dieser Ereignisse jeweils erstellten Protokollzeilen während eines zweiten Zeitraums ermittelt werden,
 - c) mittels des Parsers überprüft wird, ob und/oder in welchem Ausmaß die in Schritt b) ermittelten Protokollzeilen die vom Syntaxbaum vorgegebenen Regeln erfüllen, und
 - d) ein anormaler Zustand insbesondere dann festgestellt wird, wenn
 - die Anzahl der während des ersten Zeitraums ermittelten und die vom Syntaxbaum vorgegebenen Regeln erfüllenden Protokollzeilen und

- die Anzahl der während des zweiten Zeitraums ermittelten und die vom Syntaxbaum vorgegebenen Regeln erfüllenden Protokollzeilen voneinander um ein vorgegebenes Maß abweichen.
3. Verfahren nach einem der Ansprüche 1 oder 2, **dadurch gekennzeichnet**,
 - dass für dasselbe Computersystem zu unterschiedlichen Zeitpunkten oder für unterschiedliche Systeme mit ähnlichem Aufbau und Verwendungszweck jeweils ein Syntaxbaum nach einem der vorangehenden Ansprüche ermittelt wird,
 - dass zwischen den so erstellten Syntaxbäumen nach Abweichungen gesucht wird, und
 - dass im Falle von Abweichungen, die einen vorgegebenen Schwellenwert überschreiten, ein abweichender, insbesondere kritischer oder anormaler, Zustand des Computersystems gemeldet wird.
 4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet**, dass es sich beim Syntaxbaum um einen gewurzelten Baum, vorzugsweise um einen gewurzelten Out-Tree, handelt.
 5. Verfahren nach einem der Ansprüche 1 bis 4, **dadurch gekennzeichnet**,
 - dass bei der Erstellung des Syntaxbaums die einzelnen Teilzeichenketten der Protokollzeilen in einem zweidimensionalen Speicher mit zwei Zugriffsindizes abgespeichert werden, wobei der erste Zugriffsindex als Zeilenindex die Protokollzeile und der zweite Zugriffsindex als Positionsindex die Position der Teilzeichenkette innerhalb der jeweiligen Protokollzeile angibt,
 - dass für die Teilzeichenketten, denen der niedrigste Positionsindex zugewiesen ist,
 - nach einer Anzahl Mustern gesucht wird, die die Mehrzahl der Teilzeichenketten beschreiben,
 - für die einzelnen Muster jeweils die Wahrscheinlichkeiten dafür ermittelt werden, dass eine der Teilzeichenketten mit dem Muster übereinstimmt,
 - dass für die einzelnen Muster jeweils ein Knoten einer ersten Schicht in den Syntaxbaum eingefügt wird,
 - diesem Knoten das jeweilige Muster sowie diejenigen Protokollzeilen zugewiesen werden, deren herangezogene Teilzeichenketten mit dem Muster des Knoten übereinstimmen,
 - dieser Knoten als Zielknoten über eine gerichtete Kante mit dem Wurzelknoten des Syntaxbaums verbunden wird, und
 - dieser Kante die jeweilige zuvor ermittelte Wahrscheinlichkeit zugewiesen wird, und
 - dass für inkrementell ansteigenden Positionsindex der Teilzeichenketten in den Protokollzeilen jeweils:
 - separat für einzelne Gruppen von Protokollzeilen, die jeweils einem Basisknoten der jeweils unmittelbar vorangehenden Schicht des Graphen zugeordnet sind, jeweils:
 - nach einer Anzahl Mustern gesucht wird, die die Mehrzahl der Teilzeichenketten an der durch den jeweiligen Positionsindex festgelegten Position beschreiben,
 - für die einzelnen Muster jeweils die Wahrscheinlichkeiten dafür ermittelt werden, dass die jeweilige Teilzeichenketten mit dem betreffenden Positionsindex mit dem Muster übereinstimmt,
 - dass für die einzelnen Muster jeweils ein Knoten einer dem Positionsindex entsprechenden Schicht in den Graphen eingefügt wird,
 - diesem Knoten das jeweilige Muster sowie diejenigen Protokollzeilen zugewiesen werden, deren herangezogene Teilzeichenketten mit dem Muster des Knoten übereinstimmen,

- dieser Knoten als Zielknoten über eine gerichtete Kante mit dem Basisknoten verbunden wird, und
 - dieser Kante die jeweilige zuvor ermittelte Wahrscheinlichkeit zugewiesen wird.
6. Verfahren nach einem der Ansprüche 1 bis 5, dass für den Fall, dass für die Teilzeichenketten kein Muster gefunden werden kann, das eine relative Häufigkeit aufweist, die einen Wahrscheinlichkeitsschwellenwert θ_1 übersteigt, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.
7. Verfahren nach einem der Ansprüche 1 bis 6, **dadurch gekennzeichnet**, dass für den Fall, dass für die Teilzeichenketten ein Muster gefunden werden kann, sodass die relative Häufigkeit der dem Muster entsprechenden Teilzeichenketten einen ersten Wahrscheinlichkeitsschwellenwert θ_1 übersteigt,
- für den Fall, dass diese relative Häufigkeit auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 überschreitet, ein einziger Knoten mit dem betreffenden Muster eingefügt wird, wobei insbesondere der jeweiligen neu erstellten Kante der Anteil der dem Muster entsprechenden Teilzeichenketten zugewiesen wird und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen, die an der betreffenden Position eine dem Muster entsprechende Teilzeichenkette enthalten, dem neu eingefügten Zielknoten zugewiesen werden, **und/oder**
 - für den Fall, dass diese relative Häufigkeit nicht auch den zweiten Wahrscheinlichkeitsschwellenwert θ_2 überschreitet, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird, alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.
8. Verfahren nach einem der Ansprüche 1 bis 7, **dadurch gekennzeichnet**, dass für den Fall, dass für die Teilzeichenketten eine Mehrzahl der Muster gefunden werden kann, sodass die relative Häufigkeit der den einzelnen Muster entsprechenden Teilzeichenketten jeweils einzeln einen ersten Wahrscheinlichkeitswert θ_1 übersteigt,
- für den Fall, dass die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 überschreitet, eine Anzahl von Knoten mit jeweils einem der ermittelten Muster eingefügt wird, wobei insbesondere jeder Knoten über jeweils eine Kante mit dem ursprünglichen Knoten verbunden wird **und/oder** den so neu erstellten Kanten der Anteil der dem Muster entsprechenden Teilzeichenketten an der Gesamtzahl der dem ursprünglichen Knoten zugewiesenen Protokollzeilen zugewiesen wird **und/oder** die einzelnen Protokollzeilen auf die Knoten aufgeteilt werden, sodass jede Protokollzeile jeweils demjenigen Knoten zugeordnet wird, dessen Muster ihre jeweils betrachtete Teilzeichenkette entspricht,
- und/oder**
- für den Fall, dass die Summe der relativen Häufigkeiten der so erstellten Muster auch einen dritten Wahrscheinlichkeitsschwellenwert θ_3 nicht überschreitet, an der betreffenden Stelle ein einziger Knoten eingefügt wird, der ein Muster aufweist, das mit jeder Teilzeichenkette eine Übereinstimmung ergibt, wobei insbesondere der jeweiligen neu erstellten Kante eine Wahrscheinlichkeit von 1 zugewiesen wird, und/oder alle dem Quellknoten der Kante zugewiesenen Protokollzeilen dem neu eingefügten Zielknoten zugewiesen werden.

9. Verfahren nach einem der Ansprüche 1 bis 8, **dadurch gekennzeichnet**, dass für den Fall, dass die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position enden, einen vorgegebenen vierten Wahrscheinlichkeitsschwellenwert θ_4 überschreitet, und die Anzahl derjenigen Protokollzeilen, die bei der betreffenden Position nicht enden, einen vorgegebenen fünften Wahrscheinlichkeitsschwellenwert θ_5 überschreitet, dem Muster des betreffenden Knoten die Option eines sofortigen Zeilenendes hinzugefügt wird,
und bei der Prüfung einer Protokollzeile und einer Teilzeichenkette auf Übereinstimmung mit dem Muster eine Übereinstimmung dann als gegeben angesehen wird, wenn die jeweilige Teilzeichenkette mit dem Muster übereinstimmt und
 - auch die nachfolgenden Teilzeichenketten der Protokollzeile den jeweils nachfolgenden Mustern des Syntaxbaums entsprechen, oder
 - die Protokollzeile nach dieser Teilzeichenkette endet,
10. Verfahren nach einem der Ansprüche 1 bis 9, **dadurch gekennzeichnet**, dass die im Rahmen der Erstellung des Syntaxbaums verwendeten Schwellenwerte mit zunehmendem Abstand vom Wurzelknoten des Syntaxbaums oder mit zunehmender Pfadtiefe im Syntaxbaum ansteigen und/oder dem Abstand vom Wurzelknoten oder die Pfadtiefe angepasst werden, wobei insbesondere
 - der erste bis vierte Wahrscheinlichkeitsschwellenwert mit zunehmendem Abstand vom Wurzelknoten oder fortschreitender Pfadtiefe im Syntaxbaum monoton ansteigt oder abfällt.
11. Verfahren nach einem der Ansprüche 1 bis 10, **dadurch gekennzeichnet**, dass als Muster vorgegeben werden:
 - vorab vorgegebene Grundmuster, insbesondere IP-Adressen oder andere strukturierte Daten, und/oder
 - einzelne im Rahmen der Erstellung des Syntaxbaums festgelegte Zeichenketten.
12. Verfahren nach einem der Ansprüche 1 bis 11, **dadurch gekennzeichnet**, dass
 - für die einzelnen Pfade jeweils die bedingte Übergangswahrscheinlichkeit über eine vorgegebene Anzahl von Zeitfenstern, gebildet wird, und
 - der weitere zeitliche Verlauf der so gebildeten bedingte Übergangswahrscheinlichkeiten der einzelnen Kanten, insbesondere nach dem jeweiligen Zeitfenster oder einer vorgegebenen Anzahl von Zeitfenstern, daraufhin untersucht wird, ob diese einen vorgegebenen Wahrscheinlichkeits-Schwellenwert unterschreiten und in diesem Fall
 - die ermittelten Pfade aus dem Graphen gestrichen werden und/oder
 - einzelnen Knoten des Graphen anstelle von unveränderlichen Teilen variable Teile der Protokollzeilen zugeordnet werden.
13. Verfahren nach einem der Ansprüche 1 bis 12, **dadurch gekennzeichnet**, für den Fall, dass zu einem späteren Zeitpunkt nach der Erstellung des Syntaxbaums eine signifikant hohe Anzahl von Protokollzeilen vorhanden ist, die keinem der durch Knoten und Kanten erstellten, gerichteten Pfade des Syntaxbaums zugeordnet sind, so kann für diese Protokollzeilen ein entsprechender, die einzelnen Protokollzeilen charakterisierender Pfad im Syntaxbaum durch Modifikation neu geschaffen werden und gegebenenfalls die den einzelnen Kanten zugeordneten Pfadwahrscheinlichkeiten werden in diesem Fall an die neu aufgetretenen Protokollzeilen angepasst werden.
14. Verfahren nach einem der Ansprüche 1 bis 13, **dadurch gekennzeichnet**, dass vom Computersystem oder von auf diesem ablaufenden Prozessen zumindest eine weitere Protokollzeile erstellt wird, und dass mittels eines Parsers, der basierend auf dem Syntaxbaum erstellt wurde, untersucht wird, ob die weitere Protokollzeile mit dem Syntaxbaum übereinstimmt, wobei gegebenenfalls eine fehlende Übereinstimmung als Hinweis für das Vorliegen eines abweichenden Systemzustands angesehen wird.

15. Datenträger, auf dem ein Programm zur Durchführung eines Verfahrens nach einem der Ansprüche 1 bis 14 abgespeichert ist.

Hierzu 1 Blatt Zeichnungen

