Zur Identifikation von Strukturanalogien in Prozessmodellen

Jürgen Walter
Peter Fettke
Peter Loos

Veröffentlicht in:
Multikonferenz Wirtschaftsinformatik 2012
Tagungsband der MKWI 2012
Hrsg.: Dirk Christian Mattfeld; Susanne Robra-Bissantz

Braunschweig: Institut für Wirtschaftsinformatik, 2012

Zur Identifikation von Strukturanalogien in Prozessmodellen

Jürgen Walter, Peter Fettke, Peter Loos

Institut für Wirtschaftsinformatik (IWi) im Deutschen Forschungszentrum
für Künstliche Intelligenz (DFKI) und Universität des Saarlandes, 66123 Saarbrücken,
E-Mail: {juergen.walter, peter.fettke, peter.loos}@iwi.dfki.de

Abstract

Geschäftsprozessmodelle haben einen bedeutenden Stellenwert für die Modellierung betrieblicher Informationssysteme. Mit der immer größer werdenden Verbreitung nimmt auch die Komplexität solcher Modelle zu. Oft umfassen sie mehrere hunderte oder tausende Elemente, nicht selten mit erheblichen Redundanzen. Um eine Senkung dieses Umfangs zu erreichen, wurde in der Literatur auf die Verwendung von Analogien verwiesen. Da diese jedoch bisher nur für Datenmodelle definiert wurden, werden in diesem Beitrag einige Definitionen im Kontext von Prozessmodellen gegeben und entsprechende Maße zur Analogieberechnung vorgestellt sowie deren Anwendungspotentiale aufgezeigt.

1 Motivation

Im letzten Jahrzehnt hat das Geschäftsprozessmanagement (GPM) durch die Nutzung moderner Informations- und Kommunikationstechnik enorm an Bedeutung für Unternehmen gewonnen [12]. Ein wesentlicher Bestandteil von GPM sind Geschäftsprozessmodelle, die die notwendigen Prozessschritte umfassen, die für die Erstellung von Produkten oder Dienstleistungen notwendig sind [18]. Erfahrungsgemäß weisen diese im Allgemeinen eine sehr hohe Komplexität auf. Dies beruht im Wesentlichen auf folgenden drei Faktoren. Erstens umfassen Modelle mit praktischer Relevanz mehrere hundert Elemente, des Öfteren auch weitaus mehr, wie zum Beispiel das Handels-H-Modell von Becker und Schütte [2] mit mehr als 2000 Elementen. Zweitens wird die Modellkomplexität durch deren Anwendung in verschiedenen Wirtschaftsbereichen sowie durch verschiedene Akteure erhöht. Ebenso beeinflussen neue Technologien die Komplexität immens, wie z. B. die Möglichkeit der gemeinschaftlichen Entwicklung durch verschiedene Modellierer [11]. Weiterhin treten vielfältigste Änderungen während der langen Lebenszyklen der Prozessmodell auf [11].

Eine bekannte Methode zur Senkung der Komplexität basiert auf der Identifikation und Nutzung von Strukturanalogien [2;7;11]. Darunter können Ähnlichkeiten von Modellkonstrukten innerhalb eines oder mehrerer verschiedener Modelle verstanden werden. In der Literatur werden Strukturanalogien als relevant für die Geschäftsprozessmodellierung erachtet und meist intuitiv

anhand von leicht verständlichen Beispielen eingeführt. Sie sind vergleichbar mit industriellen Gleichteilen anderer Ingenieursdisziplinen [19]. Die Nutzung von Strukturanalogien weist ein hohes Potential für die Komplexitätssenkung von Prozessmodellen auf, wodurch letztendlich auch Entwicklungskosten reduziert werden können.

Strukturanalogien weisen neben dem praktischen Nutzen auch eine hohe theoretische Relevanz auf: Identifizierte Strukturanalogien können Hinweise auf bisher unbekannte Zusammenhänge innerhalb des Gegenstandsbereichs der Wissenschaft Wirtschaftsinformatik geben. Dadurch wird eine Beschreibung ähnlicher Sachverhalte auf einer höheren Abstraktionsstufe („Generalisierung“) ermöglicht [7]. Solch eine Abstraktion wird beispielsweise bei der Erstellung von Referenzmodellen genutzt [8], welche mittlerweile für nahezu jeden Bereich in der Wirtschaft aufgestellt wurden. Dies impliziert einige wichtige Fragen: *Existieren Konstrukte, die in einem Modell immer wieder oder gar gleichzeitig mit anderen auftreten*? *Wie häufig treten diese auf? Welche Gemeinsamkeiten weisen die verschiedene Modelle auf*? *Worin liegen die Unterschiede*?

Zur Beantwortung der genannten Fragenstellungen, wird folgender systematischer Vorgehensweise gefolgt (Bild 1):

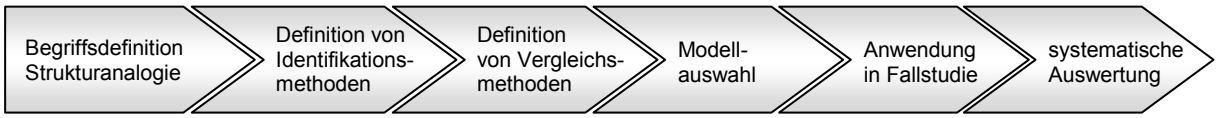

Bild 1: Vorgehensweise zur Identifikation von Strukturanalogien

Im zweiten Kapitel wird ein Literaturüberblick gegeben, welche Ansätze zur Ähnlichkeitsberechnung von Prozessmodellen im Allgemeinen bisher existieren. Kapitel drei führt neue Ansätze zur Berechnung der Strukturanalogie ein. Im Anschluss daran werden diese in Kapitel vier auf das Y-CIM-Referenzmodell von Scheer [18] sowie das Handels-H-Referenzmodell von Becker und Schütte [2] angewendet, von denen angenommen wird, dass diese Gemeinsamkeiten bzw. Analogien aufweisen. Kapitel fünf gibt abschließend eine Zusammenfassung des Beitrags sowie einen Ausblick auf zukünftige Arbeiten.

2 Literaturüberblick

Das Verständnis über den Begriff der Ähnlichkeit im Bereich der Prozessmodellierung geht weit auseinander. Es werden im Wesentlichen drei Sichtweisen unterschieden.

Zum ersten kann die Ähnlichkeit über die in einem Modell vorhandenen Prozesselemente, wie z. B. Ereignisse, Funktionen und Konnektoren etc., berechnet werden [4;6;13]. Hierfür werden Attribute der Elemente herangezogen - typischerweise die Bezeichner. Diese können hinsichtlich der Syntax und Semantik sowie ihrem Bezug auf ihre umgebenden Elemente untersucht werden [5]. Zur syntaktischen Analyse werden z. B. *String-Edit*-Distanzen berechnet, die angeben, in wie vielen Zeichen sich zwei Bezeichner (Wörter bzw. Wortgruppen) unterscheiden. Ein anderes Verfahren, das *Word-Stemming*, bezeichnet die Rückführung eines Wortes auf dessen Wortstamm, welcher für den Elementvergleich herangezogen wird. Ebenso wird das Stopp-Wort-Eliminations-Verfahren genutzt, bei dem sehr häufig auftretende Wörter ignoriert werden, da diese nur geringfügig zur Elementunterscheidung beitragen. Diese und weitere Verfahren werden meist in Kombination eingesetzt. Die semantische Analyse von Bezeichnen geht über die Berechnung der syntaktischen Ähnlichkeit hinaus, indem Synonyme, Homonyme,

Hyponyme und Hyperonyme sowie Antonyme mit Hilfe von Thesauri [9] oder Unternehmensontologien [20] identifiziert werden. Bei kontextbezogenen Analysen werden auch die umgebenden Elemente mitberücksichtigt, wodurch zusätzliche Informationen für eine präzisere Analyse verfügbar sind [4].

Die strukturelle Analyse von Prozessmodellen bildet den zweiten Ansatzpunkt zur Identifikation von Ähnlichkeiten. Die meisten dieser Verfahren beruhen auf der Analyse der einem Prozessmodell zu Grunde liegenden Graphstruktur [3;21]. Die einzelnen Prozesselemente bilden die Knoten des Graphen und der Kontrollfluss wird über die Kanten repräsentiert. Teilweise wird von der konkret eingesetzten Modellierungssprache wie EPK, BPMN oder Petrinetzen abstrahiert und ein allgemeineres graphbasiertes Modell eingeführt, auf dem die Analysen stattfinden. Vornehmlich wird die Ähnlichkeit zweier Modelle über die Berechnung von *Graph-Edit*-Distanzen vorgenommen [4;14]. Hier werden solange Knoten in einem Modell eingefügt, gelöscht oder substituiert, bis es dem anderen entspricht. Jeder Änderungsoperation werden Kosten zu Grunde gelegt, anhand derer der Ähnlichkeitsgrad bestimmt wird. Zwei Modelle gelten als ähnlicher, desto geringer die Gesamttransformationskosten sind.

Die dritte Sichtweise der Ähnlichkeitsanalyse basiert auf einer Verhaltensanalyse [4;24;1]. Bei vielen Ansätzen werden die verschiedenen „Ausführungssequenzen“, sogenannte *Process-Traces* [24], eines Modells mit denen eines anderen verglichen. Die notwendigen Daten werden während der Prozessausführung oder durch Simulation in sogenannten Log-Dateien erfasst. Die Analysemethode eignet sich hervorragend in Situationen, in denen keine Prozessmodelle vorhanden sind. Da bei diesem Verfahren jedoch die Äquivalenz von Elementen teilweise vorausgesetzt wird, welche entweder durch den Modellierer oder ein anderes Verfahren bestimmt werden muss, wie z. B. einem element-basierten Vergleich, ist der Anwendungshorizont etwas eingeschränkt. Dies bedeutet für die Analyse, dass sie sich vornehmlich auf Log-Dateien bezieht, die zum gleichen Prozess gehören, welche z. B. zu verschiedenen Ausführungen von Prozessinstanzen dieses Prozesses gehören. Dies wird z. B. zur Konformitätsprüfung (*conformance checking* etc.) genutzt [15]. Die Methode bietet zudem auch die Vergleichsmöglichkeit von Prozessmodellen an, denen unterschiedliche Modellierungssprachen zu Grunde liegen. Auch die Analyse der Häufigkeiten von Ausführungssequenzen ist möglich, wodurch relevante von irrelevanten Teilen eines Prozesses identifiziert werden können, was bei anderen Methoden nicht unmittelbar möglich ist, da alle Elemente als gleich wichtig angesehen werden.

Die in der Literatur vorgestellten Methoden beziehen sich weitestgehend auf den direkten Vergleich zweier vollständiger Prozesse und somit auf deren Ähnlichkeit. Der Unterschied zwischen Ähnlichkeit und Analogie besteht darin, dass sich ein Vergleich bei der Ähnlichkeitsanalyse mindestens auf zwei komplette Prozesse bezieht und nicht auf die Analyse eines einzelnen oder Teilen eines Prozesses ausgerichtet ist. Die Möglichkeit einer Einzel- bzw. Teilanalyse ist ein Ziel dieses Beitrags. Zwei Prozesse können analog zueinander sein, sie sind sich jedoch nicht ähnlich, weil sie z. B. für völlig unterschiedliche Aufgaben aus unterschiedlichen Anwendungsdomänen konzipiert wurden. Dennoch können sie eine identische Vorgehensweise bei der Bearbeitung der Aufgaben haben, was sich beispielsweise in den zu Grunde liegen Strukturen widerspiegelt.

Die folgenden Abschnitte befassen sich mit der strukturellen Analogie von Prozessmodellen.

3 Strukturelle Analogie von Prozessmodellen

Im vorangegangenen Abschnitt wurden einige verschiedene Ansätze zur Berechnung der Ähnlichkeit von Prozessmodellen angesprochen. Wie erwähnt, benötigt die Analyse von Ausführungssequenzen eine Implementierung und Ausführung der Prozessmodelle oder zumindest eine Simulation. Bei einer strukturellen Analyse ist dies nicht unbedingt notwendig. Im Folgenden werden zwei Verfahren zur strukturellen Analyse vorgestellt.

Der strukturelle Vergleich von Prozessmodellen erfordert die einheitliche Verwendung einer Modellierungssprache. Für den hier gewählten Ansatz sei dies die weit verbreitete Modellierungssprache EPK [23], in der auch die untersuchten Modelle vorliegen. Da nahezu allen Prozessmodellen eine Graphstruktur zu Grunde liegt, sind die vorgestellten Ansätze auch auf andere Sprachen übertragbar, weshalb die EPK für die vorliegende Arbeit einen geeigneten Ansatz darstellt. Eine EPK sei in Anlehnung an [23] folgendermaßen definiert:

Definition 1: Eine *EPK* ist ein gerichteter zusammenhängender Graph $G = (V_G, E_G)$ mit einer Knotenmenge V_G, bestehend aus den disjunkten Mengen von Ereignissen E, Funktionen F und Konnektoren C, sowie einer Kantenmenge E_G, wobei folgende Eigenschaften gelten:

- $V_G = E \cup F \cup C$, mit $E \cap F = E \cap C = F \cap C = \varnothing$
- $E_G \subseteq (V_G \times V_G) / ((E \times E) \cup (F \times F))$ ist eine nichtleere Menge von Kanten mit $\forall(v_1, v_2) \in E_G$: $v_1 \neq v_2$ und $\forall(v_1, v_2) \in E_G \Rightarrow (v_2, v_1) \notin E_G$
- t: $C \rightarrow \{\wedge, \vee, \times\}$ ist eine Funktion, die Konnektoren auf Konnektortypen abbildet
- $\bullet v = \{u \mid (u, v) \in E_G\}$; $v\bullet = \{u \mid (v, u) \in E_G\}$
- $C_{\wedge S} = \{c \in C \mid t(C) = '\wedge' \wedge | \bullet v | = 1 \wedge | v\bullet | > 1\}$
- $C_{\wedge J} = \{c \in C \mid t(C) = '\wedge' \wedge | \bullet v | > 1 \wedge | v\bullet | = 1\}$
- Für $\vee$ und $\times$ seien $C_{\vee S}$ und $C_{\vee J}$ sowie $C_{\times S}$ und $C_{\times J}$ äquivalent definiert

Der nachfolgende Ansatz bietet die Möglichkeit strukturelle Analogien zu identifizieren, die innerhalb eines oder mehrerer Modelle wiederholt auftreten. Die Berechnung bezieht die Elemente ein, jedoch nicht deren Beziehungen zu anderen.

In Moog [16] wird eine formale Definition für den Grad der strukturellen Analogie gegeben, welche sich jedoch auf Systeme im generellen bezieht. Das Maß, bezeichnet mit d, ist über zwei Mengen A und B definiert: $d = | A \cap B | / | A \cup B |$, wobei der Schnitt und die Vereinigung mit Hilfe eines element-spezifischen Vergleichsoperators berechnet werden. Das Ergebnis liegt im Intervall [0,1], wobei 0 besagt, dass keine Analogie vorliegt und je größer der Wert ist, desto größer auch die Analogie. Ist der d gleich 1, so sind die Mengen identisch.

Dieses Konzept wird nun auf Prozessmodelle übertragen, wobei verschiedene Maße für jeden Knotentyp (Ereignisse, Funktionen, Konnektoren und Kanten) definiert werden:

Definition 2: Der *Grad der Ereignisanalogie* zwischen einer EPK A und einer EPK B wird definiert als $d_E(A, B) = | E(A) \cap E(B) | / | E(A) \cup E(B) |$, wobei $E(A)$ and $E(B)$ die Mengen der Ereignisse der jeweiligen EPK sind.

Analog zu Definition 2 können Maße für die *Funktionsanalogie* (d_F), die *Konnektoranalogie* (d_C) sowie die *Kantenanalogie* (d_A) definiert werden, indem die entsprechenden Mengen herangezogen werden. Zwei Kanten werden als identisch angesehen, wenn ihre Start- und End-Knoten

gleich sind. Der für den Schnitt- und Vereinigungsoperator notwendige element-spezifische Vergleichsoperator kann beispielsweise auf einem string-basierten Vergleich der Bezeichner beruhen. Alternativ können auch Informationen genutzt werden, wie sie zum Beispiel das ARIS-Toolset der Software AG bietet. Hier können bestehende Elemente wiederverwendet werden, sogenannte „Ausprägungskopien“. Andere Modellierungswerkzeuge, die eine solche Unterstützung bieten, sind den Autoren nicht bekannt.

Um ein Gesamtmaß für die Analogie zweier Prozessmodelle zu bestimmen, kann z. B. der Mittelwert aus den oben definierten Maßen gebildet.

Definition 3: Der *Grad der Elementanalogie* zwischen einer EPK A und einer EPK B wird definiert als $d(A, B) = (d_E(A, B) + d_F(A, B) + d_C(A, B) + d_A(A, B)) / k$ mit $k = 4 \Leftrightarrow |C(A) \cup C(B)| > 0$ und $k = 3$ sonst.

Um genauere Werte für d zu erhalten, kann beispielsweise das Maße d_C weiter verfeinert werden, in dem beispielsweise nach den Konnektortypen $\wedge$, $\vee$, $\times$ bzw. Split und Join differenziert wird.

Wird Definition 2 auf die beiden EPK A und B aus Bild 2 angewendet, ist die Ereignisanalogie $d_E(A, B) = |E(A) \cap E(B)| / |E(A) \cup E(B)| = |\{e_1, e_2\} \cap \{e_1, e_2, e_3\}| / |\{e_1, e_2\} \cup \{e_1, e_2, e_3\}| = |\{e_1, e_2\}| / |\{e_1, e_2, e_3\}| = 2/3 \approx 0{,}67$. Die Funktionsanalogie $d_F(A, B) = |F(A) \cap F(B)| / |F(A) \cup F(B)| = 1$, weil die einzige Funktion in beiden EPK vorkommt. d_C ist 0, da der nur in einer der beiden EPK ein Konnektor vorkommt. Die Kantenanalogie $d_A(A, B)$ liegt bei 0.2. Die Elementanalogie nach Definition 3 liegt bei $d(A, B) = (d_E + d_F + d_C + d_A) / 4 = 28 / 60 \approx 0{,}47$.

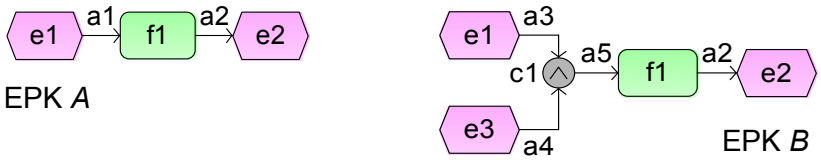

Bild 2: Ähnlichkeitsberechnung für teilweise strukturanaloge EPK

Dieses Beispiel zeigt, dass die Analogie-Maße nach Definition 2 und 3 sich sehr einfach berechnen lassen und entsprechende Resultate liefern können. Voraussetzung ist allerdings, dass Elemente als analog identifiziert werden müssen, was zwar bei der Nutzung von Ausprägungskopien trivial ist, ansonsten aber entsprechende Vergleichsverfahren erfordert. Liegt kein solches Verfahren vor und wird angenommen, dass alle Elemente verschieden sind, so sind diese Maße wenig hilfreich, da sie immer 0 liefern. Aus diesem Grund wird ein weiteres Verfahren zur Bestimmung der strukturellen Analogie eingeführt, welches auf der Analyse der zu Grunde liegenden Graphstruktur basiert.

Im Gegensatz zu anderen strukturellen Verfahren, werden hier nicht *Graph-Edit*-Distanzen berechnet, sondern (Sub-) Graphisomorphismen. Dies hat den Vorteil, dass bei diesem Verfahren auch alle Teilstrukturen identifiziert werden, und nicht nur eine einzige Abbildung.

Um die strukturelle Analogie zwischen zwei Prozessen mit Hilfe der Graphentheorie zu berechnen, ist für zwei gegebene Graphen G und H zu prüfen, ob ein Subgraph von H isomorph zu einem Subgraphen von G ist. Diese Problemstellung ist allgemein unter dem Namen *Subgraph Isomorphismus Problem* bekannt [10], das hier auf EPK übertragen wird.

Definition 4: Ein *Subgraph* $H = (V_H, E_H)$ eines Graphen $G = (V_G, E_G)$ ist ein Graph dessen Knotenmenge V_H eine Teilmenge der Knotenmenge V_G von G ist, wobei alle Kanten von H auch in G vorhanden sind: $H \sqsubseteq G \Leftrightarrow V_H \subseteq V_G$: $E_H \subseteq E_G$. Die Menge $S^k(G) = \{H \mid H \sqsubseteq G \wedge |V_H| = k\}$ sei die Menge aller zusammenhängenden Subgraphen von G mit k Knoten.

Definition 5: Zwei EPK $A = (V_A, E_A)$ und $B = (V_B, E_B)$ sind *isomorph* ($A \cong B$) genau dann, wenn eine bijektive Funktion f existiert, die alle adjazenten Knoten in A auf Knoten in B genau dann abbildet, wenn die Knoten in B adjazent sind: $A \cong B \Leftrightarrow \exists f$: $(E(A), F(A), C_{\wedge S}(A), C_{\vee S}(A), C_{\times S}(A), C_{\wedge J}(A), C_{\vee J}(A), C_{\times J}(A)) \rightarrow (E(B), F(B), C_{\wedge S}(B), C_{\vee S}(B), C_{\times S}(B), C_{\wedge J}(B), C_{\vee J}(B), C_{\times J}(B))$: $\forall u, v \in V_A$: $(u, v) \in E_A \Leftrightarrow (f(u), f(v)) \in E_B$

Definition 6: Zwei EPK A und B sind *strukturanalog* genau dann, wenn sie isomorph sind.

Definition 7: Eine EPK A ist *strukturell enthalten* in einer EPK B ($A \simeq B$) genau dann, wenn es einen Subgraphen in B gibt, der isomorph zu A ist: $A \simeq B \Leftrightarrow \exists B' \sqsubseteq B$: $B' \cong A$.

Definition 8: Eine EPK A ist *teilweise strukturanalog* zu einer EPK B ($A \sim B$) genau dann, wenn ein zusammenhängender Subgraphen in A' existiert, der isomorph zu einem zusammenhängenden Subgraphen in B' ist: $A \sim B \Leftrightarrow \exists A' \sqsubseteq A$: $\exists B' \sqsubseteq B$: $A' \cong B'$.

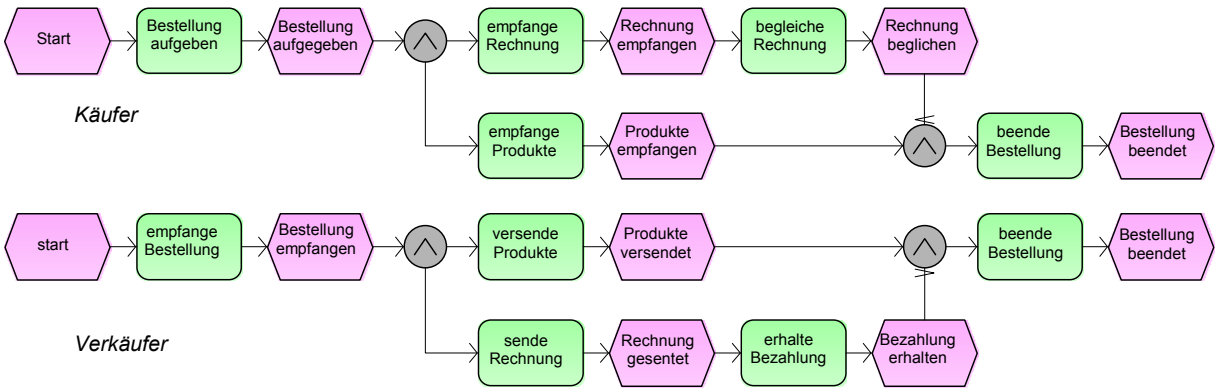

Bild 3: Zwei strukturanaloge EPK: Käufer - Verkäufer

Die in Bild 3 dargestellten EPK weisen unterschiedliche Bezeichner in den einzelnen Elementen auf. Da diese jedoch bei der Analogieberechnung nicht berücksichtigt werden, sind nach Definition 6 die beiden EPK strukturanalog.

Die Definitionen 6-8 determinieren zwar die Existenz einer Strukturanalogie, jedoch geben sie kein Aufschluss über den Grad der Analogie, weshalb in Definition 9 dieser bestimmt wird. Die Werte liegen zwischen 0 (keine Ähnlichkeit) und 1 (strukturanalog).

Definition 9: Der *Grad der Strukturanalogie* d_S zwischen EPK A und B wird definiert als: $d_S(A,B) = \frac{|S(A) \otimes S(B)|}{|S(A)| + |S(B)| - |S(A) \otimes S(B)|}$, mit $X \otimes Y = \{ x \mid x \in X, y \in Y: x \cong y\}$, wobei $S(X)$ die Menge aller in X vorhandenen Subgraphen sei: $S(X) = \{ S^k(X) \mid k = 1..|V_X| \}$.

Nachfolgend wird Definition 9 auf das Beispiel in Bild 2 angewendet, wobei die Elementindizierung nur zu deren Unterscheidung genutzt wird und nicht deren Gleichheit impliziert:

$|S(A)| = |\{(\{e\}, \varnothing), (\{f\}, \varnothing), (\{e, f\}, \{(e, f)\}), (\{e, f\}, \{(f, e\}), (\{e_1, e_2, f\}, \{(e_1, f), (f, e_2)\})\}| = 5$

$|S(B)| = |\{(\{e\}, \varnothing), (\{f\}, \varnothing), (\{c\}, \varnothing), (\{e, c\}, \{(e, c)\}), (\{c, f\}, \{(c, f)\}), (\{f, e\}, \{(f, e)\}),$
$(\{e_1, e_2, c\}, \{(e_1, c), (e_2, c)\}), (\{e, f, c\}, \{(e, c), (c, f)\}),$
$(\{e, f, c\}, \{(c, f), (f, e)\}), (\{e_1, e_2, f, c\}, \{(e_1, c), (e_2, c), (c, f)\}),$
$(\{e_1, e_2, f, c\}, \{(e_1, c), (c, f), (f, e_2)\}),$
$(\{e_1, e_2, e_3, f, c\}, \{(e_1, c), (e_2, c), (c, f), (f, e_2)\})\}| = 12$

$|S(A) \otimes S(B)| = |\{(\{e\}, \varnothing), (\{f\}, \varnothing), (\{e, f\}, \{(f, e\})\}| = 3$

$d_S(A, B) = |S(A) \otimes S(B)| / (|S(A)| + |S(B)| - |S(A) \otimes S(B)|) = 3 / 14 \approx 0{,}21$

Eine Analogie (hier $d_S \approx 0{,}21$) kann identifiziert werden, selbst wenn keine Informationen der Elemente genutzt werden, sondern nur deren Zusammenhang. Bei Anwendung von Definition 9 auf die EPK in Bild 3 ergibt sich für den Grad der Strukturanalogie ein Wert $d_S = 1$, wohingegen die Elementanalogie $d = (2/10 + 1/9 + 2/2 + 2/24) / 4 \approx 0{,}35$ ist.

4 Ähnlichkeit von Y-CIM und Handels-H

In diesem Abschnitt wird die Definition 9 „Grad der Strukturanalogie“ auf zwei Referenzmodelle angewendet, das Y-CIM-Modell von Scheer [18] und das Handels-H Modell von Becker und Schütte [2]. Die Elementanalogie nach Definition 2 wird nicht berücksichtigt, da davon ausgegangen wird, dass alle Elemente verschieden sind und somit $d_E = 0$ wäre.

Zur Analyse wurde ein Werkzeug entwickelt, welches alle vorhandenen Subgraphen zweier gegebener Modelle, deren Häufigkeiten sowie die Strukturanalogien zwischen den Modellen berechnet. Tabelle 1 enthält eine einfache Statistik der ausgewählten Referenzmodelle. Aus dem Verhältnis von Knoten (Summe der Ereignissen, Funktionen und Konnektoren) und Kanten geht hervor, dass die betrachteten Graphen dünn besetzt sind. Die EPK des Y-CIM-Modells umfassen durchschnittlich 17 Knoten und 17 Kanten, wohingegen die des Handels-H-Modells i m Durchschnitt 40 Knoten respektive 43 Kanten enthält. Die kleinsten EPK umfassen gerade einmal 3 Knoten und 2 Kanten; die größte 111 Knoten und 128 Kanten.

Bei der Analyse wurde auf die Unterscheidung der Konnektortypen ($\wedge$, $\vee$, $\times$, Split und Join) zu Gunsten der Übersichtlichkeit verzichtet und ebenso auf die Berechnung von Subgraphen mit mehr als acht Knoten. Tabelle 2 zeigt die Strukturanalogien der beiden Referenzmodelle. Es existieren drei Strukturen bestehend aus einem Knoten, welche den drei Knotentypen entsprechen. Gleiches gilt für die sieben Strukturen der Größe 2, welche gerade den Kantentypen entsprechen, die durch die Typen der Start- und Endknoten determiniert werden. Beide Male resultiert dies in einem Grad der Strukturanalogie von 1. In beiden Modellen treten auch die gleichen Strukturen mit drei Knoten auf. Überraschenderweise sind mehr als die Hälfte der Strukturen mit fünf Knoten in beiden Modellen vorhanden und immerhin noch 23% der Strukturen mit sieben und 14% bei acht Knoten. Insgesamt stimmt das Y-CIM- mit dem Handels-H-Modell in 20% der Strukturen mit bis zu acht Knoten überein. Nicht überraschend, sinkt der Grad der Strukturanalogie mit steigender Knotenanzahl.

Eines der in der Einführung genannten Ziele dieses Beitrags ist die Identifikation analoger Strukturen. In Tabellen 3 sind die Strukturen der Größe 2 und Tabelle 4 ausgewählte Strukturen der Größe 3 und deren relativen Häufigkeiten aufgeführt, wobei jeweils das Minimum bzw. Maximum der aufgetretenen Werte einer Spalte hervorgehoben wurde.

Die auftretenden Kantentypen, ersichtlich aus Tabelle 3, unterscheiden sich hinsichtlich ihrer relativen Häufigkeitsverteilung kaum. Die häufigsten Kanten gehen in beiden Modellen von einem Ereignis zu einem Konnektor. Der geringste Unterschied zwischen beiden Modellen tritt bei Kanten auf, die mit einer Funktion starten und einem Konnektor enden. Weiterhin geht hervor, dass im Handels-H-Modell wesentlich öfter Ereignisse auf Konnektoren folgen. Die häufigste Struktur der Größe 3 (siehe Tabelle 4) ist im Y-CIM-Modell mit rund 14% die Sequenz F→C→E. Im Handels-H-Modell ist dies mit 14% die Sequenz in umgekehrter Reihenfolge, E→C→F. Die Sequenz von F→E→C tritt in beiden Modellen mit ungefähr der gleichen relativen Häufigkeit auf.

Ein ermittelter Strukturanalogie-Messwert besagt, zu welchem Prozentsatz die verglichenen Modelle in ihren Strukturen übereinstimmen. Gleichzeitig gibt dies auch Aufschluss darüber, welche Strukturen im jeweils anderen Modell nicht auftreten. Je größer dieser Wert ist, desto höher ist der Grad der möglichen Wiederverwendung, z. B. in Form von „Referenz-Bausteinen“. Somit kann das Maß als mögliches Maß gesehen werden, was das Optimierungspotential widerspiegelt. Für eine konkrete Konstruktion solcher Bausteine sollten die Auftretenshäufigkeiten berücksichtigt werden, da sich nicht jede identifizierte Analogie hierfür eignet. Eine sehr markante Struktur der Größe acht, die in beiden Modellen sehr häufig auftritt, besteht aus zwei Ereignissen, gefolgt von einem Join-Konnektor und einer Funktion, auf die wiederum ein Split-Konnektor mit zwei Ereignissen folgt. An einem dieser Ereignisse ist eine Funktion angeknüpft.

Model	EPK-Anzahl	Ereignisse	Funktionen	Konnektoren	Knoten	Kanten	Graphdichte
Y-CIM	44	343 / 47%	240 / 33%	145 / 20%	728	752	0,001421
Handels-H	58	1031 / 45%	705 / 30%	573 / 25%	2309	2501	0,000469

Tabelle 1: Allgemeine Statistik Y-CIM und Handels-H

Subgraphengröße	Y-CIM	Handels-H	Vereinigung	Schnitt	Grad der Strukturanalogie
1	3	3	3	3	1,000
2	7	7	7	7	1,000
3	25	25	25	25	1,000
4	83	100	105	78	0,743
5	273	347	402	218	0,542
6	787	1018	1323	482	0,364
7	2077	2700	3876	901	0,232
8	5108	6491	10163	1436	0,141
Summe	8363	10691	15904	3150	0,198

Tabelle 2: Strukturähnlichkeit zwischen Y-CIM und Handels-H

Subgraph	Y-CIM	Handels-H	| Δ |
F → E	18,75	15,35	3,40
F → C	11,84	11,16	**0,68**
E → F	15,29	16,55	1,26
E → C	**22,87**	**19,71**	3,16
C → F	12,77	10,08	2,69
C → E	14,89	21,03	**6,14**
C → C	**3,59**	**6,12**	2,53

Tabelle 3: Identifizierte Subgraphen - Größe 2

Subgraph	Y-CIM	Handels-H	| Δ |
E → C → F	10,50	**13,75**	3.25
F → C → E	**13,95**	9,31	4.64
F → E → C	4,96	5,03	**0.07**
(E, E) → C	12.94	8.56	4.38
(F, F) → C	6.97	1.16	**5.81**
C → (F, C)	0,92	**0,08**	0,84
(F, C) → C	**0,08**	0,77	0,69

Tabelle 4: Ausgewählte Subgraphen - Größe 3

5 Diskussion

In der Literatur wurde angemerkt, dass eine Analyse von Analogien im Kontext des Geschäftsprozessmanagement für verschiedene Zwecke genutzt werden kann, zum Beispiel zur Konstruktion von Referenzmodellen. Allerdings wurde dieser Term nur intuitiv anhand von Beispielen eingeführt. In Anlehnung an die Methode der Analyse von Strukturanalogien in Datenmodellen [7], wurde im vorliegenden Beitrag eine formale Definition sowie eine Methode zur Berechnung von Strukturanalogien für EPK gegeben. Im nun nachfolgenden Abschnitt sollen deren Vor- und Nachteile diskutiert werden, auch in Bezug auf verschiedene Anwendungsszenarien.

Die Definitionen aus Kapitel 3 sind in Tabelle 5 noch einmal zusammengefasst. Sie beziehen sich auf die Analyse zweier Modelle, um daraus den Grad der Analogie abzuleiten. Die angegebenen Maße erlauben einen direkten Vergleich zweier EPK. Ebenso ermöglicht die gewählte Vorgehensweise die Analyse eines einzigen Modells, indem alle Subgraphen und deren Häufigkeiten identifiziert werden, was Auskunft über die dem Modelle inhärenten Strukturen gibt.

Analogietyp	Maß (Grad der Analogie zweier EPK A und B)
Ereignisanalogie	$d_E(A, B) = \| E(A) \cap E(B) \| / \| E(A) \cup E(B) \|$
Funktionsanalogie	$d_F(A, B) = \| F(A) \cap F(B) \| / \| F(A) \cup F(B) \|$
Konnektoranalogie	$d_C(A, B) = \| C(A) \cap C(B) \| / \| C(A) \cup C(B) \|$
Kantenanalogie	$d_A(A, B) = \| A(A) \cap A(B) \| / \| A(A) \cup A(B) \|$
Elementanalogie	$d(A, B) = (d_E(A, B) + d_F(A, B) + d_C(A, B) + d_A(A, B)) / 4$
Strukturanalogie	$d_S(A, B) = \| S(A) \otimes S(B) \| / (\| S(A) \| + \| S(B) \| - \| S(A) \otimes S(B) \|$

Tabelle 5: Typen von Analogien von EPK

Berechnungskomplexität. Die Berechnungskomplexität der vorgestellten Maße in Abschnitt 3 differiert stark. Die Berechnung der Elementanalogie ist trivial, geht man von dem reinen Vergleich auf Typen der Elemente und der bereits vorgenommenen Identifikation korrespondierender Elemente aus, da die zu Grunde liegenden Berechnungen auf einfachen Mengenoperationen beruhen. Um Elemente in Relation zu setzen, können auch linguistische Ansätze Verwendung finden, die im Literaturüberblick angesprochen wurden.

Die Berechnung der Strukturanalogie nach Definition 13 ist hingegen ein NP-vollständiges Problem, da alle möglichen Subgraphisomorphismen der Modelle berechnet werden müssen [10]. Die aktuellen Laufzeiten für die Berechnungen betragen für Subgraphen bis zur Größe 8 weniger als 4 Tage, bei der Verwendung einer Standardbiliothek für Graphanalysen. Da die Graphen sehr dünn besetzt und gerichtet sind und darüber hinaus verschiedene Knotentypen (Ereignisse, Funktionen und diverse Konnektortypen) existieren, ist hier ein Optimierungspotential zu sehen. Hier lassen sich *Pruning*-Verfahren [25] einsetzen, mit denen eine Vielzahl unnötiger Isomorphietests vermeidbar sind. Beispielsweise kann für jeden Subgraphen ein *Feature*-Vektor berechnet werden, der auf der Anzahl der jeweiligen Knoten- und Kantentypen basiert. Zwei Subgraphen können nur dann isomorph sein, wenn deren *Feature*-Vektoren identisch sind. Ebenfalls können sie auch nur dann isomorph zueinander sein, wenn mindestens eine Instanz in beiden Subgraphen vorkommt, was im Falle der Suche innerhalb eines Modells den Berechnungsaufwand erheblich reduziert. Wird diese Vorgehensweise auch für die zu vergleichenden Modelle herangezogen, muss letztendlich nur noch der Schnitt berechnet werden (siehe Definition 13). Auch eine inkrementelle Konstruktion eines Subgraphen-Graphen (DAG)

aus den zu untersuchenden EPK kann den Aufwand ebenfalls reduzieren (Konstruktion durch sukzessives Aufzählen der Erweiterungsmöglichkeiten eines Subgraphen um eine weitere Kante). Somit lässt sich der Berechnungsaufwand gegenüber einem naiven *Brute-Force*-Verfahren immens reduzieren. Für jedes *Pruning*-Verfahren muss gelten, dass der Gesamtberechnungsaufwand mit diesem Verfahren geringer sein muss als der des ursprünglichen Verfahrens. Für die Berechnung der Isomorphie können sowohl etablierte Verfahren [22] als auch Verfahren aus dem Bereich des *Graph-Mining* Anwendung finden. Ein guter Überblick hierzu ist in [25] gegeben.

Inexakte Strukturanalogien. Die im Abschnitt 3 vorgestellten Definitionen und Analogiemaße gehören zu den exakten *Matching*-Verfahren, welche ein Spezialfall der inexakten *Matching*-Verfahren darstellen, die im Bereich des *Data-Mining* hinlänglich bekannt sind. Wie in [17] gezeigt, können solche Verfahren auch auf Prozessmodelle angewendet werden. In dieser Arbeit wurde eine Abfragesprache für BPMN-Modelle geschaffen, die auf iexaktem *Matching* basiert, mit der ein konkretes BPMN-Fragment in einem *Repository* von Prozessmodellen gefunden werden kann. Solche auf *Graph-Edit*-Distanzen basierenden Verfahren können natürlich auch zur Definition von Strukturanalogien sowie den entsprechenden Maßen herangezogen werden. Ebenfalls ist es möglich, solche Verfahren zusätzlich zu dem in diesem Beitrag vorgestellten zu verwenden, um etwa die Frage zu beantworten, welche Subgraphen der Größe n mit Subgraphen der Größe m korrespondieren. Die Berechnungsergebnisse können in einem entsprechenden *Repository* hinterlegt und wiederverwendet werden, etwa zur weiteren semantischen Analyse korrespondierender Subgraphen bzw. Instanzen. Auch eine weiterführende Analyse von Single-Entry-Single-Exit-Blöcken ist möglich, worauf z. B. die Arbeit in [21] aufbaut. Allerdings ist bei inexakten Ansätzen schwierig, allgemeine Aussage zum Verwendungsgrad einer bestimmten Struktur abzuleiten.

Semantische Analysen. Die Analyse der Semantik von Elementen und deren Analogie ist im Allgemeinen sehr aufwendig, weil folgende Probleme auftreten können. Erstens können zwei identische Elemente mit Synonymen bezeichnet worden sein, oder zweitens, mit Homonymen, obwohl sie vollkommen verschieden sind. Diese Probleme resultieren z. B. aus, den bereits Eingangs erwähnten, mangelnden Modellierungskonventionen bzw. den unterschiedlichen Abstraktionsgraden etc. Synonyme Elementbezeichner können anhand ihrer Syntax oder Struktur identifiziert werden. Hierfür würden unterschiedliche Maße in der Literatur vorgeschlagen [6], die auf Datenbanken wie z. B. WordNet basieren, welche semantische und lexikalische Beziehungen zwischen Wörtern enthält. Im Gegensatz dazu ist die Identifikation von Homonymen viel schwieriger, weil der Unterschied der Bezeichner nicht offensichtlich ist. Durch die Analyse des Kontextes eines Elements, kann eine Identifikation von Homonymen vorgenommen werden. Hierfür kann beispielsweise die strukturelle Einbettung, also die Beziehung des Elements zu dessen umgebenden Elementen, herangezogen werden. Diese Elemente tragen Bezeichnungen, die zumindest in der Theorie, zu unterschiedlichen semantischen Kontexten gehören. Beispielsweise kann mit dem Wort „Bank“ zum einen ein Geldinstitut gemeint sein, zum anderen eine Sitzbank. Treten nun die Wörter „sitzen“ und „überweisen“ im jeweiligen Kontext der Elemente mit dem Wort „Bank“ auf, so kann davon ausgegangen werden, dass das Element zu zwei verschiedenen semantischen Kontexten gehört.

Übertragung auf andere Modellierungssprachen. Die in diesem Beitrag vorgestellten Definitionen und Methoden können verallgemeinert und auf andere Modellierungssprachen für die Geschäftsprozessmodellierung übertragen werden, sofern auf diese auch die Graphentheorie

anwendbar ist. Beispielsweise ist dies ohne größeren Aufwand für BPMN möglich. Die Definitionen müssen geringfügig angepasst werden, um die verschiedenen Knotentypen zu berücksichtigen. Je mehr verschiedene Knotentypen eine Modellierungssprache umfasst, desto besser lassen sich möglicherweise *Pruning*-Verfahren einsetzen. Auch die Transformation eines Modells in eine andere (abstrakte) Sprache ist möglich, sofern die Ausdruckskraft dadurch nicht beschränkt wird.

6 Resümee und Ausblick

In diesem Beitrag wurden verschiedene formale Definitionen und Maße für die Analogie von Prozessmodellen gegeben, da in der Literatur bisher nur intuitive Definitionen vorliegen. Um die Anwendbarkeit der präsentierten Konzepte zu belegen, wurden die Definitionen und Maße auf konkrete Beispiele sowie zwei Referenzmodelle angewendet. Als zu Grunde liegende Modellierungssprache wurden die Ereignisgesteuerten Prozessketten gewählt. Der strukturelle Teil dieser Modelle besteht aus einer Menge von Ereignissen, Funktionen, Konnektoren und Kanten. Das erste präsentierte Maß analysiert diese Mengen unabhängig voneinander, wobei die Anwendung vom Domänenwissen des Modellierers abhängig ist, da dieser Analogien von Elementen im Voraus definieren muss. Ist das Domänenwissen nicht verfügbar, können z. B. linguistische Verfahren eingesetzt werden. Ähnliche Schwierigkeiten treten auch bei verhaltensorientierten Ansätze zum Modellvergleich auf, wie z. B. Event-Log basierten Verfahren, die zwei unterschiedliche Prozessmodelle analysieren sollen. Weiterhin wurde ein graph-basiertes Verfahren zur Identifikation von Strukturanalogien vorgestellt, welches auf der Berechnung von (Subgraph-) Isomorphismen basiert, bei der keine Informationen über die Analogie von Elementen notwendig sind. Der größte Nachteil liegt jedoch in der relativ hohen Berechnungskomplexität, was aus dem Subgraphisomor-phismus-Problem resultiert. Um dem entgegenzuwirken, können *Pruning*-Verfahren erfolgreich genutzt werden. Die aus der Anwendung beider Verfahren gewonnen Informationen können z. B. für eine induktive Definition von Referenzmodellen genutzt werden, indem die Gemeinsamkeiten der jeweiligen Modelle identifiziert und von diesen dann abstrahiert wird.

Wie bereits im Diskussionsteil erwähnt wurde, existieren viele verschiedene Forschungsfelder im Bereich der Strukturanalogien. Zum Beispiel können weitere Maße entwickelt werden, die die Bezeichner von Elementen berücksichtigen, da diese bisher keinerlei Berücksichtigung finden. In diesem Zusammenhang sollte ebenfalls geprüft werden, inwiefern eine strukturelle Analyse bei der Identifikation von Homonymen unterstützend wirken kann. Die Einbindung von inexakten Verfahren sowie von *Graph-Edit*-Distanzen sollte im Kontext des Prozessmodellsvergleichs evaluiert werden. Ebenso ist eine Verallgemeinerung der präsentierten Ansätze zielführend, insbesondere dann, wenn Modelle verschiedener Modellierungssprachen analysiert werden sollen.

7 Literatur

[1] Becker, J; Bergner, P; Breuker, D; Räckers, M (2011): On Measures of Behavioural Distance between Business Processes. Zürich.

[2] Becker, J; Schütte, R (2004): Handelsinformationssysteme. Domänenorientierte Einführung in die Wirtschaftsinformatik, 2. Aufl. Redline Wirtschaft, Frankfurt am Main.

[3] Dijkman, R; Dumas, M; García-Bañuelos, L (2009): Graph Matching Algorithms for Business Process Model Similarity Search. In: Dayal U; Eder J; Koehler J; Reijers H (Hrsg.), Business Process Management, Bd. 5701. Springer Berlin / Heidelberg, 48-63.

[4] Dijkman, R; Dumas, M; van Dongen, B; Käärik, R; Mendling, J (2011): Similarity of business process models: Metrics and evaluation. Information Systems 36(2): 498-516.

[5] Dumas, M; Garcia-Banuelos, L; Dijkman, R (2009): Similarity Search of Business Process Models. IEEE Data Engineering Bulletin 32(1).

[6] Ehrig, M; Koschmider, A; Oberweis, A (2007): Measuring similarity between semantic business process models. Australian Computer Society, Inc. Ballarat, Australia.

[7] Fettke, P; Loos, P (2005): Zur Identifikation von Strukturanalogien in Datenmodellen – Ein Verfahren und seine Anwendung am Beispiel des Y-CIM-Referenzmodells von Scheer. Wirtschaftsinformatik 47(2): 89-100.

[8] Frank, U (2007): Evaluation of Reference Models. In: Fettke P; Loos P (Hrsg.), Reference Modeling for Business Systems Analysis. Idea Group Inc., Hershey, London, 118-140.

[9] Friedrich, F (2009): Measuring Semantic Label Quality Using WordNet. Berlin.

[10] Garey, MR; Johnson, DS (1979): Computer and Intractability: A Guide to the Theory of NP-Completeness. Freeman and Co., San Francisco.

[11] Houy, C; Fettke, P; Loos, P; van der Aalst, W; Krogstie, J (2010): BPM-in-the-Large – Towards a Higher Level of Abstraction in Business Process Management. In: Janssen M; Lamersdorf W; Pries-Heje J; Rosemann M (Hrsg.), E-Government, E-Services and Global Processes, Bd. 334. Springer, Boston, 233-244.

[12] Hung, RY-Y (2006): Business process management as competitive advantage: a review and empirical study. Total Quality Management & Business Excellence 17(1): 21-40.

[13] Koschmider, A; Oberweis, A (2007): How to detect semantic business process model variants? Proceedings of the 2007 ACM symposium on Applied computing. Seoul, Korea, 1263-1264.

[14] Li, C; Reichert, M; Wombacher, A (2008): On Measuring Process Model Similarity Based on High-Level Change Operations. In: Li Q; Spaccapietra S; Yu E; Olivé A (Hrsg), Conceptual Modeling - ER 2008, Bd. 5231. Springer Berlin / Heidelberg, 248-264.

[15] Medeiros, AKAd; Aalst, WMPvd; Weijters, AJMM (2008): Quantifying process equivalence based on observed behavior. Data Knowl. Eng. 64(1): 55-74.

[16] Moog, W (1985): Similarity and analogy theory. VDI-Verlag, Düsseldorf.

[17] Sakr, S; Awad, A (2010): A framework for querying graph-based business process models. ACM. Raleigh, NC, USA.

[18] Scheer, A-W (1998): Wirtschaftsinformatik - Referenzmodelle für industrielle Geschäftsprozesse, 2. Aufl. Springer, Berlin et al.

[19] Schütte, R (1998): Grundsätze ordnungsmäßiger Referenzmodellierung – Konstruktion konfigurations- und anpassungsorientierter Modelle. Gabler, Wiesbaden.

[20] Thomas, O; Fellmann, M (2009): Semantische Prozessmodellierung - Konzeption und informationstechnische Unterstützung einer ontologie-basierten Repräsentation von Geschäftsprozessen. Wirtschaftsinformatik (WI) 51(6): 1-13.

[21] Uba, R; Dumas, M; Garcia-Banuelos, L; Rosa, ML (2011): Clone detection in repositories of business process models. Springer-Verlag. Clermont-Ferrand, France.

[22] Ullmann, JR (1976): Algorithm for Subgraph Isomorphism. J Assoc Comput Mach 23(1): 31-42.

[23] van der Aalst, WMP (1999): Formalization and verification of event-driven process chains. Information and Software Technology 41: 639-650.

[24] van der Aalst, WMP; de Medeiros, A; Weijters, AJMM (2006): Process Equivalence: Comparing Two Process Models Based on Observed Behavior. In: Dustdar S; Fiadeiro J; Sheth A (Hrsg.), 4th International Conference on Business Process Management (BPM 2006). Bd. 4102, 129-144.

[25] Washio, T; Motoda, H (2003): State of the Art of Graph-based Data Mining. SIGKDD Explor. Newsl. 5(1): 59-68.