Begriffliche Grundlagen von Modelldifferenzen

Udo Kelter
Praktische Informatik,
Fachbereich Elektrotechnik und Informatik
Universität Siegen, D-57068 Siegen, Germany
kelter@informatik.uni-siegen.de

1 Motivation

Es gibt sehr unterschiedliche Konkretisierungen des Begriffs Differenz. Dies führt häufig zu Kommunikationsproblemen. Ziel dieses Papiers ist, die grundlegenden Begriffsvarianten und -Kontexte zu identifizieren und einen konsistenten Begriffsrahmen vorzuschlagen. Im folgenden verstehen wir unter Modellen die üblichen UML-Modelle und dazu ähnliche Modelle in anderen Modellierungssprachen.

Unter einer **Differenz** verstehen wir eine Darstellung der Unterschiede zwischen zwei Modellen. Folgende Aspekte sollte man strikt trennen:

1. den **konzeptuellen Inhalt** einer Differenz
2. die **physische Darstellung** einer Differenz in einem Speichermedium
3. die externe, oft **graphische Darstellung** einer Differenz, die i.d.R. auf der Standarddarstellung des jeweiligen Modelltyps basiert
4. den **Vergleich** zweier Modelle und die Berechnung einer möglichst guten Differenz. In der Mathematik ist die Differenz eine Funktion; dagegen ist die Differenz zwischen Modellen i.a. nicht eindeutig, die Unterschiede zwischen zwei Modellen kann man oft auf verschiedene Art darstellen. Der Vergleich von Dokumenten ist ferner nicht die einzige Methode zur Erzeugung von Differenzen; Differenzen können auch durch Protokollieren von Editieroperationen oder Verarbeiten vorhandener Differenzen erzeugt werden.
5. das **Mischen** von Modellen: dieses basiert auf einer Differenz, die i.d.R. durch Vergleich gewonnen wird. Über die reine Darstellung der Differenz hinaus geht es beim Mischen darum, ein drittes Dokument zu erzeugen, das die "Besonderheiten" der beiden Dokumente enthält.

Wir konzentrieren uns i.f. auf den ersten o.g. Punkt, da er die Grundlage für alle anderen Punkte bildet.

2 Asymmetrische Differenzen

Es gibt zwei grundlegende Varianten des Begriffs Differenz: asymmetrische und symmetrische.

Eine **asymmetrische Differenz** von einem Dokument D1 nach einem Dokument D2 ist eine Transformationsvorschrift, durch die D1 in D2 transformiert werden kann. Die Transformation besteht aus einer Sequenz von Operationsaufrufen, in denen auszuführende Operationen gemäß einem Editierdatentyp (s.u.) und passende Parameter angegeben sind.

Man kann mit einer asymmetrischen Differenz von D1 nach D2 nicht wieder das Dokument D2 nach D1 zurücktransformieren, hierzu wird eine andere Transformation benötigt.

Asymmetrische Differenzen sind primär dazu gedacht, ein Dokument D1 mit Hilfe der Differenz in ein Dokument D2 zu transformieren. Dies entspricht der Denkweise bei Patch-Werkzeugen und der internen Delta-Speicherung in Dokument-Repositorys. Daher müßte man eher von einem Summanden reden (oft benutzt wird die Bezeichnung **Delta**); der mathematischer Begriff Differenz unterstellt, daß das zweite Dokument vorher existiert bzw. daß die Differenz das Ergebnis einer Funktion, nicht ein Argument.

Für zwei gegebene Dokumente D1 und D2 kann es signifikant verschiedene Transformationsvorschriften geben, die D1 nach D2 transformieren. Dies gilt selbst dann, wenn man sinnlose Operationen in der Transformationsvorschrift (Beispiel: es wird etwas gelöscht und dann wieder eingefügt) ausschließt.

Editierdatentypen. Asymmetrische Differenzen beschreiben Transformationen zwischen Dokumenten. Wir unterstellen hierzu, daß Dokumente mittels bestimmter Operationen verändert ("editiert") werden können und daß diese Operationen durch einen abstrakten Datentyp definiert sind. Wir bezeichnen diesen Datentyp als den **Editierdatentyp**. Der Editierdatentyp muß Operationen zum Einfügen, Löschen, Ändern usw. von Dokumentelementen beinhalten, so daß letztlich alle Änderungen, die mit den Editoren dieses Dokumenttyps vorgenommen werden, nachvollziehbar sind.

Zu einem Dokumenttyp kann es unterschiedliche Editierdatentypen geben: beispielsweise kann man das Verschieben von Dokumentelementen als eigene Operation zulassen oder nicht. Die Wahl des Editierdatentyps ist nicht trivial.

3 Symmetrische Differenzen

3.1 Mengenbasierte Definition

Die Grundidee von symmetrischen Differenzen besteht darin, den aus der *Mengenlehre* gut bekannten Begriff der symmetrischen Differenz auf Dokumente zu übertragen. Hierzu muß man die Dokumente vereinfachend als Mengen von Dokumentkomponenten betrachten. Die formale Definition ist einigermaßen offensichtlich und kann in [1] nachgelesen werden.

Symmetrische Differenzen sind die Grundlage aller üblichen Differenzanzeigewerkzeuge, ferner Basis für Mischungen. Ferner basieren fast alle Algorithmen, die Modelle vergleichen, begrifflich auf symmetrischen Differenzen.

Die mengenbasierte Definition eignet sich vor allem für eine sehr informelle Betrachtung von Differenzen. Man kann sie allerdings fast nie praktisch anwenden, weil fast alle Dokumente Multimengen sind, also Dubletten enthalten können, oder sogar eine nichttriviale Struktur haben – selbst simple Textdokumente müssen als Folge (also geordnete Multimenge) von Textzeilen aufgefaßt werden. Die Betrachtung solcher Dokumente als Menge von Komponenten vereinfacht zu stark.

3.2 Multimengenbasierte Definition

Die an die Mengenlehre angelehnte Definition kann man auf Multimengen durch eine explizite Angabe erweitern, welche Dokumentkomponenten als korrespondierend angesehen werden sollen. Hierzu müssen wir den Begriff Dokumentkomponente ändern. Da eine Multimenge das gleiche Element mehrfach enthalten kann, reicht der Wert des Elements nicht aus, um eine der Dubletten zu identifizieren. Wir unterstellen daher irgendein Merkmal, z.B. eine Position im Dokument, das bei Bedarf einzelne Dokumentkomponenten identifiziert, aber nicht zu ihrem Inhalt zählt.

Definition: Eine **Differenz** zwischen zwei Dokumenten D1 und D2, deren Dokumentstruktur **eine Multimenge** ist, besteht aus folgenden Angaben:

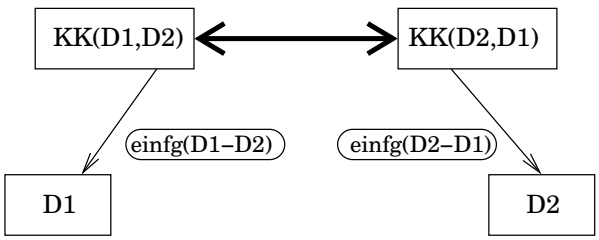
1. einer Menge von Korrespondenzen. Eine **Korrespondenz** ist ein Paar von je einer gleichen Komponente von D1 bzw. D2, die als einander entsprechend festgelegt werden. Jede Komponente kann in höchstens einer Korrespondenz auftreten. Eine Komponente, die in einer Korrespondenz auftritt, wird auch als **gemeinsame** Komponente bezeichnet. Die Menge aller gemeinsamen Komponenten bezeichnen wir auch als **Durchschnitt** beider Dokumente.
2. je einer **Einfüge-Transformation** pro Dokument, die ausgehend vom Durchschnitt das jeweilige Dokument D1 bzw. D2 rekonstruiert. Wir bezeichnen diese Transformationen als `einfg(D1-D2)` bzw. `einfg(D2-D1)`. Jede Einfüge-Transformation enthält ausschließlich einfügende Operationen, keine Änderungen oder Löschungen. Die durch diese Transformationsvorschriften eingefügten Komponenten werden als **spezielle Komponenten** von D1 bzw. D2 bezeichnet.

Die Operationen in den Einfüge-Transformationen müssen auch hier durch einen Editierdatentyp definiert sein. Im Gegensatz zu asymmetrischen Differenzen brauchen aber nur einfügende Operationen definiert zu sein.

Die symmetrische Differenz zweier Dokumente ist nicht eindeutig definiert und keine Funktion im mathematischen Sinne.

3.3 Definitionen für nichttriviale Dokumentstrukturen

Die Struktur von Mengen und Multimengen ist trivial und erlaubt es, den Mengendurchschnitt immer als neues korrektes Dokument anzusehen, das in D1 und D2 identisch eingebettet ist. Bei nichttrivialen Dokumentstrukturen kann man hiervon nicht mehr ausgehen. Folgende Probleme können auftreten; sei i.f. `KK(D1,D2)` die Menge der Komponenten von D1 mit einem Korrespondenzpartner in D2, zusammen mit der Struktur gemäß D1:



- `KK(D1,D2)` bildet kein gültiges Dokument.
- `KK(D1,D2)` und `KK(D2,D1)` haben inkonsistente Strukturen, z.B. infolge von Verschiebungen von Dokumentteilen.

Letztlich muß man bei nichttrivialen Dokumentstrukturen noch kompliziertere Definitionen den Begriffs Differenz einführen (s. [1]), die sich aber immer weiter vom intuitiv naheliegenden mengenbasierten Begriff entfernen und die teilweise nur noch für spezielle Dokumenttypen sinnvoll sind.

Literatur

[1] Kelter, U.: Lehrmodul Dokumentdifferenzen; 2007; `http://pi.informatik.uni-siegen.de/kelter/lehre/lm/dif`