Abgleich von Film-Metadaten

Michel Piguet

Einleitung

Die Filmstelle VSETH ist der größte und älteste Filmclub der Schweiz. Er ist eine studentische Einrichtung der Hochschulen in Zürich. Die Filmstelle war schon in den 70er- und 80er-Jahren – noch lange vor den filmwissenschaftlichen Studiengängen – ein Ort, um interessierten Studenten die Möglichkeit zu geben, große filmgeschichtliche Kenntnisse zu erhalten. Als Videoaufzeichnung ab Fernsehen möglich wurde, bildete sich rasch ein umfangreiches Archiv von Filmaufzeichnungen, die sorgfältig ausgewählt worden waren und so zu einem großen und teilweise raren Bestand führten.

nanoo.tv wurde 2011 gegründet mit dem Zweck, für den Bildungsbereich online-Videoaufzeichnungen anzubieten. Die ZHdK hat das Projekt maßgeblich unterstützt und dadurch einen hochwertigen Filmbestand geschaffen, der für die Schweiz einmalig ist.

Da die Zukunft der Filmaufzeichnung zweifellos digital ist, bestand ein Interesse daran, zu wissen, ob Fernsehaufzeichnungen aus anderen Fremdbeständen leicht eingefügt werden können. Dazu wurde der Bestand der Filmstelle VSETH ausgewählt. Es war zu prüfen, welcher Anteil der VSETH-Fernsehaufzeichnungen im Bestand von nanoo nicht enthalten ist und ob sich eine Digitalisierung lohnt. Da die Bestände jedoch in verschiedenen Katalogen erfasst waren, blieb kein anderer Weg als einen externen Datenabgleich durchzuführen. Mit einem geschätzten Umfang von gegen 5000 Titeln wurde ein intellektueller Abgleich verworfen. Die Trialog AG hat einen automatisierten Abgleich vorgeschlagen, der zwar keine 100igen Ergebnisse liefert, jedoch klare Aussagen über die Schnittmenge machen kann. Die Ergebnisse wurden so aufbereitet, dass eine rasche intellektuelle Nachkontrolle unsicherer Treffer mit geringem Aufwand durchgeführt werden kann. Zudem erlaubt das Prozess-Design, Datenqualität und somit Datenabgleich bei Bedarf sukzessive zu verbessern, falls in einem weiteren Schritt Bedarf dazu besteht.

Schilderung des Verfahrens

Es wurden zwei Datensets gegeneinander abgeglichen.

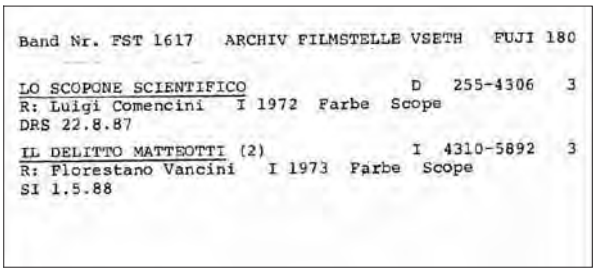
Datenset	Beschreibung
Filmstelle VSETH	Die Informationen zum Videobestand der Filmstelle VSETH lagen teilweise als Excel-Daten vor. Es zeigte sich, dass die Auswahlkriterien für die Datenerfassung unklar blieben. Deshalb musste auf die in den 90er-Jahren angelegten Karteikarten zurückgegriffen werden. Sie enthalten ca. 4700 Titel. Die Karteikarten wurden eingescannt und sind als TIFF-Dateien vorhanden.
nanoo	Die Daten von nanoo sind im Verbundsystem NEBIS vorhanden. Sie werden periodisch in den schweizerischen Bibliothekskatalog swissbib übertragen. Ein aus NEBIS mit den Einträgen „Filmmaterial" erstellter Datenauszug enthielt 43.440 Datensätze. Darin sind die Titel von nanoo enthalten.

Die Daten lagen in Form von HTML- bzw. XML-Dateien vor. Für den Erkennungs- und Datenextraktionsprozess werden Programmbibliotheken sowie die von der Trialog AG gepflegte Wissensbasis genutzt. Es ist eine Sammlung von Prozess- und Entscheidungsinformationen, die die Arbeitsschritte und die Verarbeitung der Daten steuern. Die einzelnen Prozessschritte waren:

Schritt	Aufgaben
Wissen ergänzen	Automatisierte Analyse der Datensets, um Information zu sammeln (Struktur, Informationselemente, vorkommende Werte, usw.)
„Vorreinigen“	Bereinigen der Daten, um die Struktur wahrnehmbar zu machen (Objekte „Karteikarte“ und „Film“)
„Reinigen“	Korrigieren typischer OCR-Fehler, die überall vorkommen. Zusammenfügen von getrennten Zeilen, u.a.
Erkennen der Elemente	Anwenden der Strukturmuster pro Filmtitel. Um Elemente in den Daten zu erkennen, werden sogenannte reguläre Ausdrücke verwendet.[1] Abgrenzen der Feldinhalte. Extraktion der verwendbaren Feldinhalte.
Visuelle Aufbereitung	Einfügen von Farben und Zeilenumbrüche pro Element Karteikarte, Film, Datenfeld für die visuelle Kontrolle des Prozesses.
Datenbankablage	Schreiben der Feldinhalte in die Datenbank. Statistische Kontrollen, weitere Reinigung und Vereinheitlichung der Feldinhalte, um den Abgleich zu verbessern.

Aufbereiten der Daten

Die Daten der Filmstelle VSETH (Karteikarten waren die Ausgangsdaten):


Band Nr. FST 1617 ARCHIV FILMSTELLE VSETH FUJI 180

LO SCOPONE SCIENTIFICO D 255-4306 3
R: Luigi Comencini I 1972 Farbe Scope
DRS 22.8.87
IL DELITTO MATTEOTTI (2) I 4310-5892 3
R: Florestano Vancini I 1973 Farbe Scope
SI 1.5.88

Die Karteikarten wurden gescannt. Der Text in den Bilddateien wurde mit OCR erkannt. Die Erkennung hat teilweise gravierende Mängel (Erkennungsfehler, aber vor allem auch Strukturmängel), die aus Effizienzgründen nur soweit korrigiert wurden, dass die Struktur von Karteikarte und die Filmaufzeichnung erkennbar wurden. Die Daten wurden mittels Mustererkennung automatisiert korrigiert. Ein großer Teil der Texterkennungsfehler wurde so korrigiert. Besonders die Strukturierungselemente (insbesondere Beginn und Ende von Karteikarte und von Filmen) mussten erkennbar gemacht werden. Mit Hilfe der in der Wissensbasis abgelegten Muster wurden die Informationselemente weiter bereinigt, separiert und normalisiert. Mit vertretbarem Aufwand ließen sich jedoch nicht alle Elemente sauber isolieren, da auf einigen Karteikarten wegen Verunstaltung durch Stempel und handschriftliche Einträge zu viele OCR-Fehler auftraten.

Band Nr. FST 1616 ARCHIV FILMSTELLE VSETH FUJI 180
TARZAN AND HIS-MATE——— D 177-3651 3
R: Cedric Gibbons USA 1934 s/w
0RF1 21.8.87
T-MEN D 3666-5871 3
R: Anthony Mann USA 1947 s/w
S3 3.9.87
Band Nr. FST 1617 ARCHIV FILMSTELLE VSETH FUJI 180
LO SCOPONE SCIENTIFICO D 255-4306 3
R: Lulgi Comenciñi I 1972 Farbe Scope
DRS 22.8.87
IL DELITTO MATTEOTTI (2) I 4310-5892 3
R: Florestano Vancini I 1973 Fßrbe Scope
SI 1.5.88
Band Nr. FST 1621 ARCHIV FILMSTELLE VSETH FUJI 180
SAME TIME, NEXT YEAR D 505-4522 :

Die Informationen auf den Karteikarten folgen einem klaren Schema. Die Typographie hat eine gute Qualität. Doch im Vor-Computer-Zeitalter wurde oft “O” (Buchstabe O) statt “0” (Null) geschrieben, eine „l“ (Buchstabe L) statt eine „1“ (eins), usw. Die Daten sind gut recherchiert. Filmtitel sind in Originalsprache, allerdings meist ohne original-sprachlichen Akzente und einheitliche Transkriptionsregeln. Für das Datenabgleichsverfahren wurden alle Informationen gespeichert, jedoch nur die nützlichen Informationselemente für den Titelabgleich aufbereitet. Aufgrund der statistischen Vorarbeiten waren das: Titel, Regisseur, eventuell Produktionsland und -jahr.

1 Zur Thema regulärer Ausdrücke siehe zum Beispiel: https://de.wikipedia.org/wiki/Regul%C3%A4rer_Ausdruck [28.04.2016]

Funktion der Wissensbasis

Die Wissensbasis stellt das eingespeicherte Wissen für die Datenverarbeitung zur Verfügung.

- Verwendung von Tags für die erkannten Inhaltselemente (<country> für Land; <lang> für Sprache; <station> für Sender)
- Gegebenenfalls Definition der vorkommende Inhalte (Beispiel Sender: ARD, ORF1, FR3) inklusive Zuordnung von häufig vorkommenden Alternativen wie „0RF1“, „ORF 1“, „ORFl“.
- Die hauptsächlich vorkommenden Kürzel (Beispiel Land: „I“ = Italien, „F“ für „Frankreich“). Da besonders Kürzel auch andere Funktion haben, ist die Angabe der Mehrdeutigkeit wichtig („I“ ist Land, aber auch die Sprache italienisch). Dies gibt dem Prozess die Möglichkeit, in mehrdeutigen Kontexten flexibel und richtig zu entscheiden.

Dies alles ist in der Wissensbasis abgelegt und wird von dort her benutzt.

Die Wissensbasis hat gelernt, welche Datenfelder eine beschränkte Anzahl von Ausprägungen haben. Zum Beispiel hat „Jahr“ nur Werte zwischen „1909“ und „1989“. Mit diesem Wissen ist es einfach, die Werte „19 34“ oder „13/3“ zu korrigieren.

Das Wissen in der Basis kommt auf mehrere Arten zusammen:

1. Schon früher bei anderen Aufträgen erworbenes Wissen,
2. Automatisierte Aufbereitung von Analyseergebnissen (z.B. Integration von feldspezifischen Vorschlägen für Fehlerbereinigungsmethoden),
3. Händische Anpassungen.

Visuelle Aufbereitung des Erkennungsprozesses

Der automatisierte Erkennungsprozess der Informationselemente wurde grafisch im Webbrowser darstellt, so dass erkannte Elemente in Farben dargestellt und nicht erkannte weiß bleiben. Dies half, von Auge rasch festzustellen, ob bei systematisch vorkommenden Fehlern allenfalls die Wissensbasis weiter ergänzt werden musste, um so die Erkennungsleistung zu verbessern.

Im obigen Fall sind nicht berücksichtigte „Schreibalternativen“ (Beispiele: „(2 l“ für Folgenummer „(2)“ und „F ßrbe Scope“ für „Farbe Scope“) schuld an nicht erkannten weißen Elementen. Im Vergleich: „Farbe+ Scope“ wurde als Muster erkannt und konnte so dem Informationsfeld „Material“ zugewiesen werden. „0RF1“ ist zwar mit einer Null geschrieben, wird aber als Sender erkannt, da die Wissensbasis schon gelernt hatte, dass dies „ORF1“bedeutet.

Dieses Beispiele zeigen, wie umfangreich das Wissen ist, das dem Erkennungsprozess mitgegeben werden muss: Reihenfolge der Informationselemente; vordefinierte Informationselemente und ihre alternativen Schreibweisen; Korrekturschemata für Zahlenwerte; erlaubte und mögliche Interpunktionen; Schwellenwerte für positive und negative Bewertung, usw.

Abgleichsprozess

Der von Karteikarten extrahierte Datenbestand hat insgesamt eine mittlere Qualität. Deswegen wurden spezielle Verfahren für den Abgleich angewandt, welche „weiche“ Vergleiche durchführen, so dass sich Fehler möglichst wenig auf die Erkennungsleistung auswirken. Die Informationselemente wurden in einer postgreSQL-Datenbank gespeichert. Jeder Eintrag der Filmstelle (Titel „0“) wurde mit allen Datensätzen von NEBIS verglichen, um möglichst viele *potentiell* gleich oder ähnlich bezeichnete Filme zu finden. Mit Hilfe der Wissensbasis wurden die Funde automatisch mit einer Punktzahl pro Informationsfeld bewertet. Das folgende Beispiel zeigt, wie die Titel als zusammengehörig erkannt und dies mit einem dunkelorangenen Balken entsprechend markiert wurde. In den Kolonnen ..._s stehen die Vergleichswerte („similarity“).

	di	ti	co	ye	id	di_s	ti_s	co_s	ye_s
0	Lee Daniels	The Butler	USA	2013	30263				
1	Lee Daniels	The Butler		2013		1	1	0	1
2	Lee Daniels	The butler		2014		1	1	0	0.428571
3	un Lee Daniels	The Butler	USA	2014		0.8	1	1	0.428571
4	un Lee Daniels	The Butler	USA	2014		0.8	1	1	0.428571

Abb. 1: Die Zuordnung der Titel ist korrekt

Im folgenden Beispiel wurde das Ergebnis als zweifelhaft erkannt und dies mit einem blauen-orangen Balken markiert. In diesem Fall ist der Abgleich richtig, aber nicht so sicher, weil zusätzliche Daten in den Feldern verglichen werden müssen, die als „Lärm“ wirken.

	di	ti	co	ye	id	di_s	ti_s	co_s	ye_s
0	Terence Young	DR NO	GB	1962	586				
1	Terence Young = James Bond jagt Dr. No	Dr. No		1962		0.411765	1	0	1
2	Terence Young ; Richard Maibaum, Johanna Harwood, Berkley Mather¬Monty Norman	James Bond jagt Dr. No	USA	2006		0.2	0.3	0	0

Abb. 2: Die Zusammengehörigkeit der Titel ist wahrscheinlich

Nachfolgend nochmals ein unsicherer Fall, doch diesmal hatte die Wissensbasis Recht. Dem Gesamtergebnis kann nicht getraut werden. Die Sortierung nach Ähnlichkeit macht jedoch deutlich, dass es ab Position 4 um einen anderen Film handelt. In diesem Fall hat der Lärm im Feld „di" dazu geführt, dass die Erkennungsrate so niedrig ist und entsprechend die Gesamtwertung nicht genügend differenzierte, um den anderen Film wegzuschneiden. Das Beispiel zeigt auch, dass Land („co") und Jahr („ye") keine verlässlichen Angaben liefern, um die richtig erkannten auszuzeichnen.

	di	ti	co	ye	id	di_s	ti_s	co_s	ye_s
0	Federico Fellini	Fellini's Satyricon	F	1969	90				
1	Federico Fellini ; Gaius Titus Petronius Arbiter¬Bernardino Zapponi ..	Satyricon	Italien¬Frankreich	2008		0.254237	0.526316	0.05	0
2	Federico Fellini	Satyricon		1968		1	0.526316	0	0.428571
3	Federico Fellini	Satyricon	Italien	1992		1	0.526316	0	0.25
4	Story und Drehbuch Federico Fellini und Bernardino Zapponi	Fellini's Roma		2003		0.288462	0.416667	0	0
5	story und Drehbuch Federico Fellini und Bernardino Zapponi	Fellini's Roma	Italien	2008		0.288462	0.416667	0	0
6	Federico Fellini	Fellini's Roma		1972		1	0.416667	0	0.25

Abb. 3: Die Zusammengehörigkeit aller Titel ist unsicher

Die Resultate

Die Datenbasis enthält zum Schluss alle Abgleiche der Filmtitel.[2]

	VSETH	NEBIS	nanoo	MIZ
Vorhandene Titel insgesamt:				
– Anzahl vorhandener Datensätze	Ca. 4800	43440	12423	19344
– Im Abgleich enthaltende Datensätze	4698[2]	43440	12423	19344
Titel der Filmstelle VSETH:				
– Wie viele davon sind in ...?	-	1095	415	573
– Welcher Anteil ist in ...? (%)		23%	9%	12%

Der Bestand der Filmstelle VSETH ist also zu 9% in nanoo vorhanden (inkl. Duplikate), jedoch zu 23% im NEBIS-Verbund insgesamt. Das hat damit zu tun, dass beim MIZ Video-Aufzeichnungen erst ab 2011 gemacht wurden.

Beurteilung des Abgleichsprozesses

Die Analyse Abgleichsgenauigkeit zeigt, wie oft Übereinstimmung bei den vier analysierten Feldern gefunden wurden (Wert: 0.0 = keine Übereinstimmung; Wert 1.0 = komplette Übereinstimmung). Wenn in beiden Vergleichsfeldern kein Wert vorhanden war, wurde der Wert 0.0 gesetzt.

Bei Regisseuren (di) erzielt der Abgleich oft keine 1.0. Das liegt daran, dass in einem Vergleichsfeld mehrere Namen aufgeführt sind und dass die Namen einander ähnlich sein können ohne dass entscheidbar wäre, ob es sich um denselben Regisseur handelt. Bei den Titeln (ti) ist der Vergleichsgrad ebenfalls breit gestreut, da sie nicht einheitlich angesetzt sind. Eine schlechte Überstimmung wird bei der Landesangabe (co) erzielt. Dies deshalb, weil viele Landesangaben fehlen (VSETH: 523; swissbib: 5392), weil nur teilweise ko-produzierende Länder genannt sind und weil verschiedene Editionen eines Films miteinander verglichen werden. Beim Jahr (ye) waren viele Informationen gar nicht vorhanden, so dass der Abgleich 0 ergab.

„Weiche" Abgleiche ergeben immer einen Zwischenbereich, bei dem zu einem gewissen Grad unsicher bleibt, ob es sich um einen Treffer handelt – oder nicht. Zur Erklärung der Abbildung 5: Die ..._sim-Werte bezeichnen die errechnete Ähnlichkeit der Filme. Diese

2 Ohne Datensätze, die ungenügende Information haben.

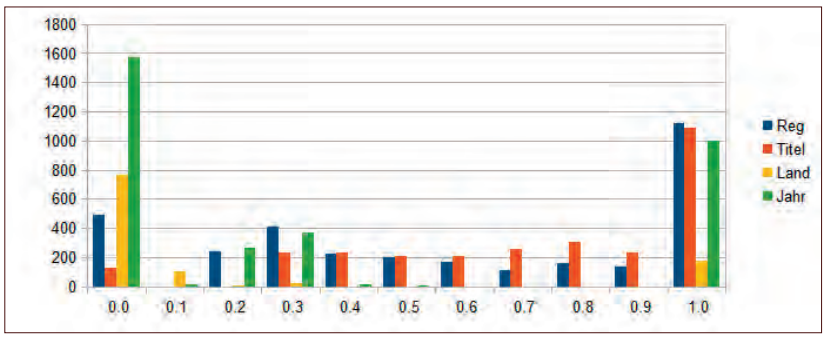


Abb. 4: Verteilung der Erkennungsrate nach Feld

Werte werden in den Datenfeldern mit Farben dargestellt (weiß = gering; rot = mittel; orange = gut; grün = sehr gut). Zum Kommentieren werden Abgleiche hier mit der „id“ zitiert.

783 Regisseur und Titel stimmen zu weniger als 50% überein, die zweite Landesangabe fehlt, die Jahre stimmen überein. => Treffer

818 Regisseur und Titel haben eine kleine Ähnlichkeit, die Jahre sind ganz verschieden. => kein Treffer

874 Regisseur hat nur geringe Ähnlichkeit, weil zusätzliche Namen genannt werden, der Titel ist in unterschiedlicher Sprache angegeben, die Landesangabe ist verschieden, die Jahre sind ganz verschieden. Dennoch => Treffer

Abb. 5: Die farbliche Darstellung erleichtert die intellektuelle Kontrolle

id	di_sim	ti_sim	co_sim	ye_sim	set_	director	title	country	year
783	0.50	0.45	0.00	1.00	vsethk	Sidney Lumet	TWELVE ANGRY MEN x	USA	1957
783	0.50	0.45	0.00	1.00	swissbib	Reginald Rose¬Sidney Lumet	12 angry men		1957
792	0.36	1.00	0.00	1.00	vsethk	Rouben Mamoulian	QUEEN CHRISTINA	USA	1933
792	0.36	1.00	0.00	1.00	swissbib	Rouben Mamoulian ; Salka Viertel ¬H.	Queen Christina		1933
818	0.32	0.32	0.00	0.25	vsethk	Alexander Mackendrick	THE MAN IN THE WHITE SUITE	GB	1951
818	0.32	0.32	0.00	0.25	swissbib	Alexander Rockwell	In the soup		1992
860	0.33	0.80	1.00	0.25	vsethk	Constantin Costa-Gavras	MISSING (2)	USA	1981
860	0.33	0.80	1.00	0.25	swissbib	Constantin Costa-Gavras. One flew ove	Missing	USA¬USA	1992
874	0.30	0.38	0.00	0.00	vsethk	Roberto Rossellini 1/	IL MESSIA	F	1975
874	0.30	0.38	0.00	0.00	swissbib	Roberto Rossellini ; Silvia d'Amico Ben	Der Messias	Italien	2006
883	0.41	0.92	0.00	0.00	vsethk	Terence Young	FROM RUSSIA WITH LOVE (2)	GB	1963
883	0.41	0.92	0.00	0.00	swissbib	Terence Young¬nach dem Ian Fleming	From Russia with love	USA	2000
3639	0.34	0.50	0.00	0.00	vsethk	Theo Angelopoulos	O THIASSOS	Griechenla	1975
3639	0.34	0.50	0.00	0.00	swissbib	Theo Angelopoulos ; Giorgos Arvanitis	O thiasos		2012

Eine intellektuelle Kontrolle einer Stichprobe von 150 Vergleichen hat ergeben, dass 140 Titel übereinstimmen (93%). Ein Abgleich kann mit immer größerem Aufwand immer weiter verfeinert werden, u.a. durch weitere Datenbereinigung, Datenergänzung, Einbezug von Serien-Information. Wie weit Bereinigungen sinnvoll sind, muss ökonomisch beantwortet werden: Wenn der Aufwand für das Verfahren ähnlich hoch ist wie das Neuerfassen der Daten (mit wahrscheinlich besserer Datenqualität), dann lohnt sich der erste Teil der Datenextraktion nicht. Der zweite Teil, die Daten zu vergleichen, wird sich dennoch lohnen. Es braucht jedoch beträchtliche Erfahrung, um im Voraus abzuschätzen, wie groß die zu leistende Arbeit ist.

Fazit

Die Zahl von 9% der Filmtitel der Filmstelle VSETH, die in nanoo vorhanden sind, bleibt eine ungefähre Angabe. Es kann jedoch davon ausgegangen werden, dass der Titelabgleich zu etwa 93% korrekt ist. Diese Aussage ist genügend, wenn man vorerst wissen will, wie groß die Überschneidung ist.

Das Verfahren ist so angelegt, dass in einem zusätzlichen Schritt die Abgleiche rasch beurteilt und akzeptiert oder verworfen werden können.

Der Abgleich kann weiter verfeinert werden. Dies lohnt sich dann, wenn die Daten später in einen Katalog integriert werden sollen. Die zusätzlichen Aufwände werden mit zunehmenden Ansprüchen rasch hoch, so dass bei eher kleinen Sets mit demselben Aufwand die Titel intellektuell durchgesehen werden können. Solche Abgleiche sind jedoch nötig, wenn die Daten nicht in einem einzigen Katalog vorhanden sind. Hier ein Ausblick auf spätere Artikel dieser Themenreihe: Man kann einwenden, dass bei bibliothekarischen Metadaten, die von ausgebildeten Bibliothekaren erstellt worden sind, solche differenzierten Abgleichsverfahren unnötig sind, weil die Datenqualität hoch ist. Wer dieser Meinung ist, darf gespannt auf die kommenden Artikel in dieser Zeitschrift sein. ❙

Abkürzungen

MIZ	Medien- und Informationszentrum (MIZ) der Zürcher Hochschule der Künste (ZHdK)
nanoo.tv	Online-Videoarchiv des MIZ
NEBIS	Bibliotheks-Verbundskatalog
swissbib	Elektronischer Bibliothekskatalog, der alle Daten der großen Bibliotheksverbünde der Schweiz enthält.
VSETH	Verband der Studierenden an der ETH Zürich
ZHdK	Zürcher Hochschule der Künste

lic. phil. Michel Piguet
Berater für Informations- und Knowledge Management und Mitinhaber der Trialog AG in Zürich
Ausgewiesener Spezialist für Datenbankmigrationen
Trialog AG
Josefstr. 178, CH-8005 Zürich
piguet@trialog.ch