Distanzberechnung zur Bestimmung von Ähnlichkeiten im Filmkontext

*Markus Deutschl*

Dept. of MultimediaTechnology

Fachhochschule Salzburg

Puch bei Hallein, Austria

mdeutschl.mmt-m2012@fh-salzburg.ac.at

Abstract — This electronic document is a “live” template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document. DO NOT USE SPECIAL CHARACTERS, SYMBOLS, OR MATH IN YOUR TITLE OR ABSTRACT. *(Abstract)*

Keywords — Distanzberechnung, Recommender, Dimensionsreduktion, Film, Ähnlichkeit.

# Einführung

Die Präsenz eines Recommender Systems (RS) gehört im modernen Web bereits zum Standard­repertoir von Verkaufs- und Informationswebsites. Dieses Werkzeug erlaubt es, den Benutzern jener Websites, Vorschläge aufgrund ihrer Präferenzen anzubieten. Durch diese Vorschläge erwächst den Benutzerinnen und Benutzern ein Mehrwert, da sie auf Produkte bzw. Informationen aufmerksam werden, die sie vorher noch nicht kannten, was sich in längerer Aufenthaltszeit auf Websites und in höheren Verkaufszahlen für die Websitebetreiber niederschlägt. Gerade im Bereich von Filmen und Serien ist das Interesse an RS sehr hoch, da das riesige Angebot von Benutzerinnen und Benutzern nicht überblickt werden kann und diese auch gerne auf Neuerscheinungen bzw. populäre Werke ihres Interesses hingewiesen werden wollen. Dieses Interesse der Industrie wurde durch den *Netflix prize*[[1]](#footnote-1) unterstrichen, welcher dem Team, das den *Netflix*-eigenen Vorschlagsalgorithmus um 10% verbessern konnte, eine Million Dollar Preisgeld bot.

Die Berechnung von Filmvorschlägen auf Basis von Benutzerinnen- und Benutzerinteressen ist ein sehr komplexes Fachgebiet, auf dem bereits einiges an Forschung betrieben wurde.

Neben den weithin bekannten Ansätzen *Content-based* RS [z.B. (Pazzani und Billsus 2007)], welche Vorschläge anhand des Benutzerprofils errechnen, sowie *Collaborative filtering* RS [z.B. (Herlocker, Konstan und Riedl 2000)], die die Vorschläge aufgrund von ähnlichen Benutzerinnen und Benutzern generieren, gibt es mittlerweile auch zahlreiche andere Ansätze zu diesem Problem.

In dieser Arbeit stelle ich eine Distanzfunktion vor, die Ähnlichkeiten zwischen Filmen aufgrund der Filmeigenschaften bestimmt. Aus diesen Distanzen kann in weiterer Folge ein Graph konstruiert werden, aus dem mit einer Breitensuche Vorschläge generiert werden können. Dieser Ansatz wurde gewählt, um Vorschläge für das RS von *MovLib*[[2]](#footnote-2) generieren zu können, dessen Filmdaten in einer graphenorientierten Datenbank abgelegt werden, um solche Breitensuchen möglichst effizient durchführen zu können. Durch die Benutzerunabhängigkeit der Distanzberechnung können auch Vorschläge ohne Nutzerdaten generiert werden, was wiederum das „Kaltstartproblem“ anderer RS und auch die Notwendigkeit von Trainingsdaten eliminiert.

In der Sektion V werden Überlegungen geschildert, wie Benutzerdaten iterativ zur Verbesserung der Vorschläge miteinbezogen, die Distanzfunktion verbessert und deren Ergebnisse verifiziert werden können.

# Verwandte Arbeiten

In der Vergangenheit haben sich bereits einige Forscherinnen und Forscher mit dem Themengebiet von Filmvorschlägen auseinandergesetzt. Es sollen hier nun einige wenige Ansätze aufgelistet werden, die in direktem oder indirektem Bezug auf diese Arbeit stehen, was aber keinesfalls wertend aufzufassen ist.

Die Arbeiten von Perny & Zucker (Perny und Zucker 2001), Li & Yamada (Li und Yamada 2004), sowie Das & ter Horst (Das und ter Horst 1998) gehen darauf ein, wie Vorschläge anhand von Nutzerpräferenzen generiert werden können. Dies umfasst Bewertungen, sowie latente Interessen von Benutzerinnen und Benutzern.

Symeonidis, Nanopoulos & Manolopoulos (Symeonidis, Nanopoulos und Manolopoulos 2009) und Adomavicius, Sankaranarayanan, Sen & Tuzhilin (Adomavicius, et al. 2005) wiederum beziehen Kontextinformationen über Userdaten und Items in ihre Vorschlagsberechnungen mit ein.

Des Weiteren sind noch die Arbeiten von Mak, Koprinska & Poon (Mak, Koprinska und Poon 2003), sowie Golbeck & Hendler (Golbeck und Hendler 2006) zu nennen, welche Textkategorisierung von Filmsynopsen bzw. vertrauensbasierte Daten von social networks in ihren RS verwenden.

Die Ähnlichkeitsberechnung dieser Arbeit unter­scheidet sich jedoch grundlegend von zuvor genannten Ansätzen, da rein nur die Eigenschaften und Domänenwissen von Filmen in sie mit­einbezogen werden.

# Ähnlichkeitsbestimmung

Für die Formulierung einer paarweisen Distanzfunktion mussten zu allererst die wichtigsten Eigenschaften von Filmen ermittelt werden, die einen Ähnlichkeitsvergleich ermöglichen. Dazu wurde auf domänenspezifisches Wissen zurück­gegriffen, um diese Faktoren zu identifizieren.

Als grundlegende Unterscheidungsbasis von Filmen wurden die Filmgenres definiert, da diese eine grobe Kategorisierung der Inhalte zulassen. Da die Definition der Genres z.T. stark in Art und Anzahl variiert, wurden die 20 Basisgenres des *MovieLens*-Datasets[[3]](#footnote-3) verwendet.

Einen weiteren einflussreichen Faktor stellt die Gesamtbewertung eines Films dar. Werden Filme ähnlich bewertet, kann davon ausgegangen werden, dass sie sich ähneln, da Benutzerinnen und Benutzer diese ähnlich gut einstufen. Diese Gesamtbewertung wurde vom *MovieLens*-Dataset extrahiert, welches diese wiederum von der Website *Rotten tomatoes[[4]](#footnote-4)* entnahm. Die Bewertungsskala liegt hier bei 1-5, wobei 5 die beste Bewertung darstellt.

Der Regisseur eines Films bestimmt zu einem großen Teil den Stil. Deshalb müssen Filme, die vom selben Regisseur stammen, auch als ähnlich angesehen werden. Natürlich existieren auch Filme die sehr unterschiedlich sind, obwohl sie vom gleichen Regisseur gedreht wurden. Diese Spezial­fälle sollen aber durch die Verwendung aller anderen Ähnlichkeitsfaktoren abgefangen werden. Weiteren Einfluss auf den Stil von Filmen üben Herkunftsland und Erscheinungsjahr aus. Diese beiden Faktoren haben aber nur begrenzte Wirkung im Vergleich zu den anderen und sollten die Ähnlichkeit deshalb nicht zu stark beeinflussen.

Das genaue Zusammenspiel und der Einfluss der einzelnen Faktoren werden gemeinsam mit einer Beschreibung der finalen Distanzfunktion im nächsten Abschnitt dargelegt.

# Distanzfunktion

Die Ausgangsbasis der Distanzberechnung stellen die Genres eines Films dar. Dafür wurden die Genres in einen 20-dimensionalen Vektor gebracht, welcher den Wert 100 beinhaltet, wenn ein Film einem Genre zugeordnet werden kann, bzw. 0 wenn das nicht der Fall ist. Die paarweise Genredistanz () wird danach mittels der euklidischen Distanz­funktion berechnet.

Der Einfluss der Gesamtbewertung wurde durch den Bewertungsfaktor () abgebildet, der die normali­sierte Bewertungsdifferenz abbildet. Die Variablen und stehen, wie bei allen Formel­darstellungen für die beiden Filme, die für die Distanzberechnung herangezogen werden.

Formel 1: Bewertungsfaktor

Der Faktor für das Erscheinungsjahr () bedient sich der Differenz der Erscheinungsjahre. Diese wurde allerdings abgeschwächt, um den Einfluss auf die Gesamtdistanz zu verkleinern, was vor allem für Remakes von Filmen wichtig ist.

Formel 2: Erscheinungsjahrfaktor

Des Weiteren wurden Faktoren bestimmt, die die Distanz bei übereinstimmenden Regisseuren () bzw. Produktionsländern () mindern. Der wurde bei Übereinstimmung mit und der mit definiert. Beide erhalten den Wert wenn es keine Übereinstimmung gibt.

Die finale Distanzfunktion mit all ihren Faktoren wurde zum Test auf die ersten 1.000 Filme des *MovieLens*-Datasets angewandt und dieser resul­tierende Graph anschließend mit dem Programm *Cytoscape* visualisiert. Bei einer maximalen Kantenlänge von 150 ergaben sich einige gut von­einander abgegrenzte Cluster, was auf eine vernünftige Partitionierung hinwies. Um die Ergebnisse noch genauer verifizieren zu können, müsste allerdings eine Dimensionsreduktion mit PCA oder MVE (Shaw 2011), um diese Daten im 2 oder 3-dimensionalen Raum unverzerrt betrachten zu können.

Formel 3: Finale Distanzfunktion

# Weiterführende Arbeiten

Benutzerdaten iterativ miteinbeziehen, Distanzfunktion ändern – leichter machen (Ginzi Tips), Dimensionsreduktion (Shaw), Clusteringalgorithmen zur Verifizierung vl. OPTICS,

# Referenzen

Adomavicius, Gediminas, Ramesh Sankaranarayanan, Shahana Sen, und Alexander Tuzhilin. *Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach.* Herausgeber: ACM. New York, 1. Januar 2005.

Das, Duco, und Herman ter Horst. *Recommender Systems for TV.* Herausgeber: AAAI. Eindhoven, 1998.

Golbeck, Jennifer, und James Hendler. *FilmTrust: Movie Recommendations using Trust in Web-based Social Networks.* Maryland, 2006.

Herlocker, Jonathan J., Joseph A. Konstan, und John Riedl. *Explaining Collaborative Filtering Recommendations.* Herausgeber: ACM. Minneapolis, Minnesota, Dezember 2000.

Li, Peng, und Seiji Yamada. *A Movie Recommender System Based on Inductive Learning.* Herausgeber: IEEE. Tokyo, 2004.

Mak, Harry, Irena Koprinska, und Josiah Poon. *INTIMATE: A Web-Based Movie Recommender Using Text Categorization.* Sydney, 2003.

Pazzani, Michael J., und Daniel Billsus. *Content-Based Recommendation Systems.* Herausgeber: Springer Verlag Berlin Heidelberg. 2007.

Perny, Patrice, und Jean-Daniel Zucker. *Preference-base Search and Machine Learning for Collaborative Filtering: the "Film-Conseil" Movie Recommender System.* Herausgeber: CEPAD. Paris, 2001.

Symeonidis, Panagiotis, Alexandros Nanopoulos, und Yannis Manolopoulos. *MoviExplain: A Recommender System with Explanations.* Herausgeber: ACM. New York, Oktober 2009.

# Copyright Forms

You must submit the IEEE Electronic Copyright Form (ECF) as described in your author-kit message. THIS FORM MUST BE SUBMITTED IN ORDER TO PUBLISH YOUR PAPER.

# Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try   
“R. B. G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

# References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example: [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in “[3]”—do not use “Ref. [3]” or “reference [3]”. Do not use reference citations as nouns of a sentence (e.g., not: “as the writer explains in [1]”).

Unless there are six authors or more give all authors’ names and do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

1. <http://www.netflixprize.com/> [↑](#footnote-ref-1)
2. *MovLib* ist eine freie und quelloffene Filminformationsseite. <http://movlib.org> [↑](#footnote-ref-2)
3. <http://www.grouplens.org/node/462> [↑](#footnote-ref-3)
4. <http://www.rottentomatoes.com/> [↑](#footnote-ref-4)