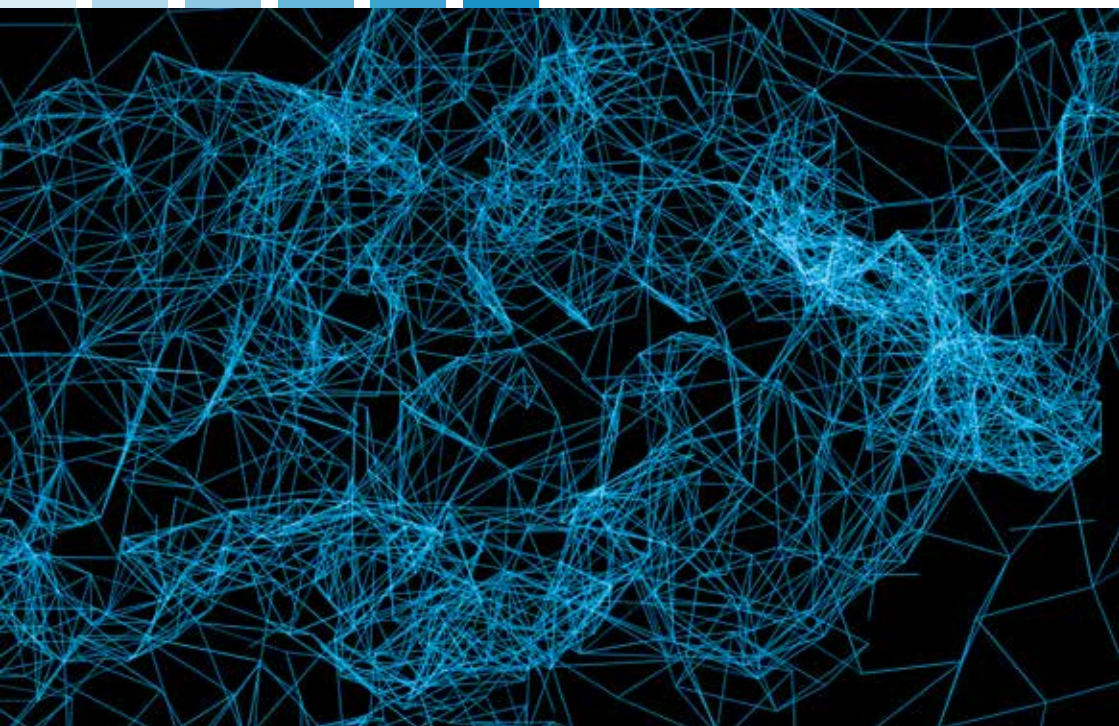
Modellierung thematischer Nähe in Organisationen durch Machine Learning

Thomas Thiele



Modellierung thematischer Nähe in Organisationen durch Machine Learning

Modeling Thematic Proximity
in Organizations using Machine Learning

Von der Fakultät für Maschinenwesen der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades
eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Thomas David Thiele

Berichter: Außerplanmäßige Professorin Dr. rer. nat. Sabina Jeschke

Universitätsprofessor Dr.-Ing. Dr. h. c. (UPT) Burkhard Corves

Tag der mündlichen Prüfung: 16. November 2018

Berichte aus der Informatik

Thomas Thiele

Modellierung thematischer Nähe in Organisationen durch Machine Learning

Shaker Verlag
Aachen 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Zugl.: D 82 (Diss. RWTH Aachen University, 2018)

Printed in Germany.

ISBN 978-3-8440-6542-8
ISSN 0945-0807

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen
Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9
Internet: www.shaker.de • E-Mail: info@shaker.de

Danksagung

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschlicher Mitarbeiter am Cybernetics Lab der RWTH Aachen University. Besonderer Dank gilt daher Prof. Dr. rer. nat. Sabina Jeschke für die Betreuung der Arbeit und die darüberhinausgehende vortreffliche Zusammenarbeit. Für die Übernahme des Kommisionsvorsitzes danke ich Prof. Dr.-Ing. Christian Brecher sowie Prof. Dr.-Ing. Burkhard Corves in seiner Rolle als zweitem Gutachter. Prof. Dr.-Ing. Tobias Meisen stand mir sowohl fachlich als auch persönlich stets zur Seite, auch ihm gebührt mein herzlichster Dank.

Als wesentlicher Beitrag für das Gelingen der Arbeit empfinde ich die besondere Atmosphäre des Intituts, die maßgeblich von meinen Kolleginnen und Kollegen gestaltet wird. Sie alle haben sich in den letzten Jahren als hervorragende Sparingspartner erwiesen. Die Forschungsgruppe Produktionstechnik hat sich hierbei als ebenso fachliches wie kameradschaftliches Umfeld erwiesen, den Kollegen der Gruppe gilt meine besondere Hochachtung. Ebenso haben der Austausch und die Diskussionen mit den Kollegen des Exzellenzclusters „Integrative Produktionstechnik für Hochlohnländer“ wesentliche Impulse zur Entwicklung der Arbeit gegeben. Mein Dank gilt hierbei im Besonderen Dr. phil. André Calero Valdez. Ebenso Dank gebührt Olivier Pfeiffer und Fabian Scheidt für ihre Unterstützung in der letzten Phase.

Neben dem universitäten und beruflichen Umfeld ist der private Rahmen ein ebenso zentraler Bestandteil für das Gelingen der Arbeit. Meine Eltern Renate Jeromin-Thiele und Harald Thiele nehmen hierbei eine wesentliche Rolle ein, sie haben mich von den ersten Ideen bis zur Fertigstellung der Arbeit stets in ganz besonderer Weise ermutigt, unterstützt und gefördert. Ebenso hervorheben möchte ich meine Partnerin Sarah Ginski, die mir besonders im finalen Endspurt trotz gleicher Situation mit viel Energie, Gelassenheit und Unterstützung in allen Lebenslagen beigestanden hat.

Inhaltsverzeichnis

I Zusammenfassung

Komplexe Problemstellungen nicht nur als disziplinäre Fragestellung zu begreifen, sondern über Fachgrenzen hinaus Lösungen zu entwickeln, erweist sich nicht nur als Trend, sondern auch als Notwendigkeit. Die Identifikation geeigneter Kooperationspartner und die Suche nach gemeinsamen Themen ist jedoch oftmals zeitaufwändig.

In der vorliegenden Arbeit wird daher ein System konzeptioniert und entwickelt, welches die Modellierung thematischer Nähe in Organisationen durch einen Machine Learning Ansatz erlaubt. Grundlage hierfür sind textuelle Daten, aus welchen zunächst mittels eines generativen Verfahrens inhärente Themen extrahiert werden. Danach werden diese Themen einem diskriminierenden Verfahren unterzogen, welches ein Matchmaking zwischen Themen unterschiedlicher organisationaler Entitäten ableitet. Die durch diese Verfahrenskette generierten Ergebnisse werden dann für den Nutzer in Form eines graphenbasierten Ansatzes visualisiert, es entsteht eine Landkarte verknüpfter Themen auf Basis eines automatisierten Prozesses.

I Summary

To understand complex problems not only as a disciplinary question, but to develop solutions beyond the boundaries of disciplines, proves to be not only a trend, but also a necessity. The identification of suitable cooperation partners and the search for common topics is often time-consuming.

In the present work, a system is therefore conceived and developed which allows the modelling of thematic proximity in organisations through a Machine Learning approach. This is based on textual data, from which inherent topics are extracted using a generative procedure. Thereafter, these topics are subjected to a discriminatory procedure which derives a matchmaking between topics of different organisational entities. The results generated by this process chain are then visualized for the user in the form of a graph-based approach, resulting in a map of linked topics on the basis of an automated process.

II Abkürzungsverzeichnis

ACM: Association for Computing Machinery
ASCII: American Standard Code for Information Interchange
BOW: Bag-of-words Modell
CBOW: Continous Bag of Words
CMS: Content Management Systeme
CoE: Exzellenzcluster
CRISP-DM: Cross Industry Standard Process for Data Mining
CSP: Cross-Sectional Processes
csv: Comma separated value
DFG: Deutsche Forschungsgemeinschaft
ECM: Entprise Content Management Systeme
h-Index: Hirsch-Index
HTTP: Hypertext Transfer Protocoll
ICD: Integrative Cluster Domain
IEEE: Institute of Electrical and Electronics Engineers
JSON: JavaScript Object Notation
kNN: k-nearest Neighbor
KNN: Künstliche Neuronale Netze
LDA: Latent Dirichlet Allocation
LSI: Latent Semantic Indexing
NLP: Natural Language Processing
OCR: Optical Character Recognition
pdf: Portable Document Format
POS-Tagging: Part-of-Speech-Tagging
relu: Rectified Linear Units
SVM: Support Vector Machine
tanh: Tangens Hyperbolicus
TDM: Term-Document Matrix
tf-idf: Term Frequency – Inverse Document Frequency
txt: Text Dateien
UCS: Universal Coded Character Set
XML: Extensible Markup Language

III Abbildungsverzeichnis

IV Tabellenverzeichnis

V Verwendete Formelzeichen

$1_A(x)$	Funktion	Indikatorfunktion nach Chaney und Blei (2012)
a_j	Maßzahl	Aktivierung eines KNN
α	Hyperparemeter	Initialer Hyperparemeter der LDA
ave	Funktion	Durschnittswert (z.B. Mittelwert oder Median)
β_{iv}	Verteilung	Topicverteilung des Themas i aus Topic Modell v
γ	Maßzahl	Gewichtungsfaktor für byte size basierte Normalisierung
$CharLength^{\gamma}$	Funktion	Funktion zur Ermittlung der Zeichenlänge eines Wortes oder Dokuments mit Gewichtungsfaktor γ
d_i	Mengenelement	Dokument in Menge D
df_t	Häufigkeit	Dokumenthäufigkeit aller Dokumente mit Begriff t
e_k	Menge	Menge der Begriffe einer Entität E_k
$\xi_{i,j}$	Maßzahl	Ähnlichkeit zwischen zwei Themen nach Chaney und Blei (2012)
$f(\cdot)$	Funktion	Allgemeine Klassifikationsfunktion
$GesamtPred_{E_j\|m}$	Maßzahl	Ergebnismatrix einer Prädiktion des m-ten Themas der Entität E_j
h	Maßzahl	Hirsch Index
$h(a_\mathrm{j})$	Funktion	Aktivierungsfunktion von a_j eines KNN
idf	Bezeichnung	Inverse Document Frequenzy
k	Hyperparemeter	Anzahl der Themen einer LDA
$\mathcal{L}$	Menge	Menge der Label

N	Menge	Anzahl der Dokumente in der Dokumentsammlung
n_{t_l}	Häufigkeit	Auftrittshäufigkeit des Wortes t_l im Korpus
$n_{t_l,d}$	Häufigkeit	Auftrittshäufigkeit des Wortes t_l im Dokument d
$\mathbb{N}$	Menge	Menge der natürlichen Zahlen
$P_{(t_l)}$	Wahrscheinlichkeit	Auftrittswahrscheinlichkeit eines Begriffes t_l
$Pred_{E_{j\|k}}$	Maßzahl	Ergebnis der Klassifikation des k-ten Begriffes der Entität E_j
p_j	Mengenelement	Person in Menge P
T_E	Menge	Menge aller Themen einer Entität E
t_l	Mengenelement	Wort in Wortmenge T
t_l^n	Menge	Wortsequenz bestehend aus n Worten
tf	Bezeichnung	Wortfrequenz
T_E	Menge	Topicraum einer Entität E
V	Menge	Menge aller *word2vec* Vektoren
w	Variable	Beliebiges Häufigkeitsmaß in einer Kosinusnormalisierung, z.B. tf oder $tf - idf$
$w_{ji}^{(1)}$	Hyperparameter	Gewicht zwischen dem j-ten Neurons und der i-ten Inputvariable im (1)-sten Layer eines KNN
$w_{j0}^{(1)}$.	Hyperparameter	Bias des j-ten Neurons im (1)-sten Layer eines KNN
x_i	Variable	Inputvariable eines KNN
$\overrightarrow{y_{E_j\|k}}$	Vektor	Ergebnisvektor zur Prädiktion des k-ten Begriffes der Entität E_j