(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
25. Februar 2021 (25.02.2021)

WIPO | PCT

(10) Internationale Veröffentlichungsnummer
WO 2021/032824 A1

(51) Internationale Patentklassifikation:
G06F 16/33 (2019.01)

(21) Internationales Aktenzeichen: PCT/EP2020/073304

(22) Internationales Anmeldedatum:
20. August 2020 (20.08.2020)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
10 2019 212 421.6
20. August 2019 (20.08.2019) DE

(71) Anmelder: **FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E. V.** [DE/DE]; Hansastr. 27 c, 80686 München (DE).

(72) Erfinder: **HOPPE, Thomas**; Baerwaldstr. 39, 10961 Berlin (DE).

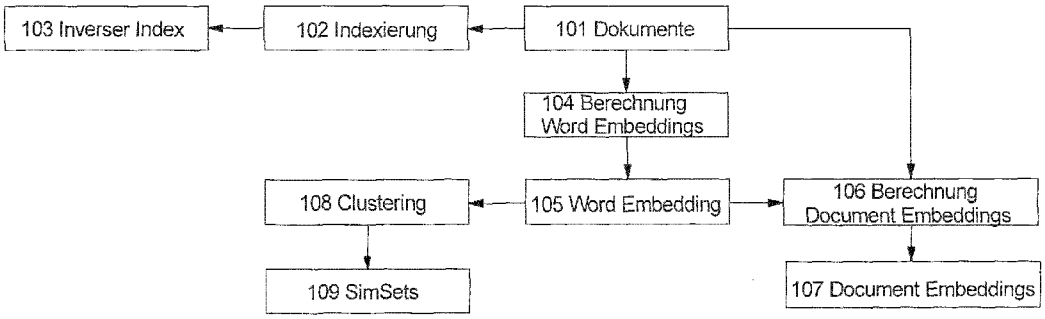
(74) Anwalt: **MAIKOWSKI UND NINNEMANN PATENTANWÄLTE PARTNERSCHAFT MBB**; Postfach 15 09 20, 10671 Berlin (DE).

(81) Bestimmungsstaaten *(soweit nicht anders angegeben, für jede verfügbare nationale Schutzrechtsart)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM,

(54) **Title:** METHOD AND DEVICE FOR PRE-SELECTING AND DETERMINING SIMILAR DOCUMENTS

(54) **Bezeichnung:** VERFAHREN UND VORRICHTUNG ZUR VORAUSWAHL UND ERMITTLUNG ÄHNLICHER DOKUMENTE

Fig. 2



101 Documents
102 Indexing
103 Inverse index
104 Calculation of word embeddings
105 Word embedding
106 Calculation of document embeddings
107 Document embeddings
108 Clustering
109 SimSets

(57) **Abstract:** The invention relates to a method for pre-selecting and determining similar documents from a quantity of documents (101), wherein the documents (101) comprise tokenised character strings, characterised in that a) an inverted index for at least one sub-quantity of the documents (101) is calculated by means of an indexing method (102), b) word embeddings (105) are calculated for the at least one sub-quantity of the documents (101), c) for the at least one sub-quantity of the documents (101) a document embedding (107) is calculated for each of these documents (101) in that, for each document (101), the word embeddings (105) of all character strings, in particular words of the document (101), are added and standardised (106) with the number of character strings, in particular words, wherein, beforehand, subsequently or in parallel, d) SimSet groups (109) of similar character strings are calculated with the

WO 2021/032824 A1

[Fortsetzung auf der nächsten Seite]

TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Bestimmungsstaaten** *(soweit nicht anders angegeben, für jede verfügbare regionale Schutzrechtsart)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasisches (AM, AZ, BY, KG, KZ, RU, TJ, TM), europäisches (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *mit internationalem Recherchenbericht (Artikel 21 Absatz 3)*

calculated word embeddings (105) with the aid of a clustering method, and then e) in a query phase (200) a query expansion (202) is performed, in which i) query terms which occur in SimSet groups (109), or ii) query terms which do not occur in the SimSet groups (109) but in the documents (101), or iii) query terms which do not occur in the documents (101), in particular also incorrectly written query terms, are used for a pre-selection (203) of the documents in order to limit the number of hits, and then a query embedding (205) is determined; and then f) the query embedding (205) is compared with the document embeddings (107) of the documents pre-selected with use of the SimSet groups (109) formed in step d) with the clustering method in order to limit the number of document embeddings (109) to be compared, so as to automatically determine a ranking of the similarity of the documents (101) and display and/or store these. The invention also relates to a device.

(57) Zusammenfassung: Die Erfindung betrifft ein Verfahren zur Vorauswahl und Ermittlung ähnlicher Dokumente aus einer Menge an Dokumenten (101), wobei die Dokumente (101) tokenisierte Zeichenketten aufweisen, dadurch gekennzeichnet, dass a) mit einem Indexierungsverfahren (102) ein invertierter Index für mindestens eine Teilmenge der Dokumente (101) berechnet wird, b) für die mindestens eine Teilmenge der Dokumente (101) werden Word Embeddings (105) berechnet, c) für die mindestens eine Teilmenge der Dokumente (101) wird für jedes dieser Dokumente (101) jeweils ein Document Embedding (107) berechnet, indem für jedes Dokument (101) die Word Embeddings (105) aller Zeichenketten, insbesondere Worte des Dokuments (101) addiert und mit der Anzahl der Zeichenketten, insbesondere Worten, normiert werden (106), wobei vorher, nachfolgend oder parallel d) mit den berechneten Word Embeddings (105) mithilfe eines Clusteringverfahrens SimSet-Gruppen (109) von ähnlichen Zeichen ketten berechnet werden, und dann anschließend e) in einer Anfragephase (200) eine Anfrageexpansion (202) vorgenommen wird, bei der i) Anfrageterme, die in SimSet-Gruppen (109) verkommen, oder ii) Anfrageterme, die nicht in den SimSet-Gruppen (109) aber in den Dokumenten (101) Vorkommen, oder iii) Anfrageterme, die nicht in den Dokumenten (101) verkommen, insbesondere auch falsch geschriebene Anfrageterme für eine Vorselektion (203) der Dokumente verwendet werden, um eine Einschränkung der Treffermenge zu erreichen und dann ein Query Embedding (205) bestimmt wird; und dann f) ein Vergleich des Query Embeddings (205) mit den Document Embeddings (107) der, unter Verwendung der im Schritt d) mit dem Clusteringverfahren gebildeten SimSet-Gruppen (109) zur mengenmäßigen Einschränkung der Anzahl der zu vergleichenden Document Embeddings (107), vorselektierten Dokumewnte durchgeführt wird, um automatisch ein Ranking der Ähnlichkeit der Dokumente (101) zu ermitteln und diese anzuzeigen und / oder abzuspeichern. Die Erfindung betrifft auch eine Vorrichtung.

Verfahren und Vorrichtung zur Vorauswahl und Ermittlung ähnlicher Dokumente

Die Erfindung betrifft ein Verfahren zur Ermittlung ähnlicher Dokumente mit den Merkmalen des Anspruchs 1 und eine entsprechende Vorrichtung mit den Merkmalen des Anspruchs 8.

Suchfunktionen und -verfahren stellen grundlegende Funktionalitäten von Betriebssystemen, Datenbanksystemen und Informationssystemen dar, die insbesondere in Content- und Document-Management-Systemen, Information-Retrieval-Systemen von Bibliotheken und Archiven, Suchfunktionen von Webauftritten in Intranets und Extranets eingesetzt werden. Diese Suchfunktionen und –verfahren beziehen sich auf elektronische Dokumente (im Folgenden nur Dokumente genannt), die mindestens teilweise einen Text aufweisen und die durch Digitalisieren (Umwandlung in einen Binärcode) in Dateiform angelegt oder überführt wurden.

Ohne Suchfunktionen wäre die Recherche in umfangreichen Dokumentbeständen, wie z.B. Patentschriften, kaum noch zu bewältigen.

Suchfunktionen, -verfahren und -maschinen beruhen auf informationstechnologischen Prinzipien des Information und Document Retrievals (Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press. 2008), wie z. B. Algorithmen zur Konvertierung und syntaktischen Analyse von Dokumenten, effizienten Datenstrukturen zur Indexierung der Dokumentinhalte, Zugriffsalgorithmen, die für diese Indexstrukturen optimiert sind, die Vermeidung wiederholter Berechnung durch Zwischenspeicherung von Ergebnissen (sogenanntes caching) (siehe DE10029644) und Messverfahren, mit denen der Grad der Übereinstimmung (als „Relevanz“ bezeichnet) von Dokumenten bzgl. einer Suchanfrage gemessen werden kann.

Konventionelle Methoden des Information Retrievals von unstrukturierten, textuellen Informationen bewerten die „Relevanz“ von Dokumenten anhand des

Vorkommens von Suchbegriffen durch statistische, probabilistische und informationstheoretische Bewertungen.

Eine wesentliche Charakteristik von Suchmaschinen ist die Interpretation der Verknüpfungsart von eingegebenen Stichworten. In der Praxis haben sich zwei Verknüpfungsarten durchgesetzt: AND und ANDOR.

Bei AND werden nur die Dokumente gesucht, die alle Suchbegriffe enthalten. Bei ANDOR hingegen wird die Suchanfrage als disjunktiv verknüpft interpretiert, die Ergebnisdokumente jedoch werden anhand der Anzahl der gefundenen Suchbegriffe pro Dokument gewichtet, um auch ähnliche Dokumente noch finden zu können.

Diese konventionellen Verfahren basieren in der Regel auf Termvektoren, die Dokumente als Vektoren in einem hochdimensionalen Raum (z. B. mit tausenden bis zu hunderttausenden Dimensionen) symbolisch repräsentieren. Jede Dimension eines solchen Vektorraums repräsentiert hierbei ein Wort. Alle Dimensionen zusammen genommen bilden die Orthonormalbasis des Vektorraums.

Dateienvektoren, oder Dokumentvektoren werden hierbei als Linearkombination der Worthäufigkeiten oder normierten Worthäufigkeiten über der Orthonormalbasis gebildet.

Da Dokumente in der Regel nur aus einem Bruchteil aller möglichen Worte bestehen, sind Dokumentenvektoren

a) in der Regel „sparse“ (nur dünn besetzt, viele der Vektorkomponenten sind Null),

b) diskret (jede Dimension erfasst nur die Bedeutung eines Wortes) und

c) tendiert diese Repräsentation allein durch die Struktur hochdimensionaler Räume dazu, „hartnäckige“ (obstinate) Dokumente (Dokumente, die bei unterschiedlichsten Anfragen als Ergebnisse gefunden werden) zu produzieren.

(On the existence of obstinate results in vector space models, Milos Radovanovic, Alexandros Nanopoulos, Mirjana Ivanovic, Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, DOI: 10.1145/1835449.1835482).

Insbesondere der diskrete Charakter dieser symbolischen Repräsentation führt dazu, dass Worte mit ähnlichen Bedeutungen auf voneinander unabhängige Dimensionen der Orthonormalbasis und damit auf unabhängige Komponenten der Dokumentvektoren abgebildet werden.

Um Begriffsabhängigkeiten zu berücksichtigen, ist es bei dieser Form der Repräsentation daher notwendig, zusätzliches Wissen zu nutzen, um die Dokumentvektoren mit Informationen über ähnliche Begriffe und den Grad dieser Begriffsähnlichkeiten anzureichern.

Sogenannte semantische Suchverfahren ermitteln die den Dokumenten zugrundeliegenden Themen wahrscheinlichkeitsbasiert (US4839853, Latent dirichlet allocation. David M. Blei, Andrew Y. Ng, Michael I. Jordan In: *Journal of Machine Learning Research*, Jg. 3 (2003), S. 993–1022, http://jmlr.csail.mit.edu/papers/v3/blei03a.html, (letzter Zugriff 6.2.2019) und seine Varianten) oder ermitteln Ähnlichkeiten zwischen Dokumenten auf der Basis explizit vorgegebener Wissensmodelle, in Form von Begriffsmodellen (linguistischen Modellen, semantischen Netzen, Wortnetzen, Taxonomien, Thesauri, Topic Maps, Ontologien, Wissengraphen).

Die durch die erste Gruppe semantischer Suchverfahren, auch als Topic Modelling Verfahren bezeichnet, ermittelten Themen wirken in der Regel künstlich, sind selten von Menschen interpretierbar und erzeugen oft kaum zuordenbare Suchergebnisse.

Die zweite Form von semantischen Suchverfahren nutzt vorgegebene Wissensmodelle, um die Dokumente und Anfragen auf ein gemeinsames kontrolliertes Vokabular, das durch das Wissensmodell definiert wird, abzubilden

[EP 2199926 A3 / US 000008156142 B2] und damit die Suche zu vereinfachen. Die Abbilder von Dokumenten auf das Wissensmodell werden als Annotationen bezeichnet, welche ggf. über Begriffsähnlichkeiten mit zusätzlichen Begriffen des Wissensmodells angereichert werden.

Um zu bestimmen, mit welchen Begriffen eine Annotation zusätzlich anzureichern ist, werden Wissensmodelle genutzt, um zu ermitteln, dass sich synonyme Begriffe gegenseitig implizieren, dass Unterbegriffe deren Oberbegriffe oder Begriffe, die in anderen Beziehungen zueinander stehen, implizieren. Der Grad der Begriffsähnlichkeit kann anhand der semantischen Distanz (Conceptual Graph Matching for Semantic Search. Zhong J., Zhu H., Li J., Yu Y. In: Priss U., Corbett D., Angelova G. (eds) Conceptual Structures: Integration and Interfaces. ICCS 2002. Lecture Notes in Computer Science, vol. 2393. Springer, Berlin, Heidelberg) bzw. der Länge dieser Implikationsketten aus den Wissensmodellen ermittelt werden.

Die Menge der Annotationen, erweitert um solche zusätzlichen Begriffe, entspricht einer Anreicherung des aus den Annotationen bestehenden Dokumentvektors um weitere aus den Begriffsähnlichkeiten ermittelte Vektorkomponenten.

Auf Begriffsmodellen basierende Suchverfahren stellen durch die hohe Qualität der Sucherergebnisse und die potentielle Erklärbarkeit der Ergebnisse anhand der Netzstruktur die derzeit am weitesten verbreitete Form semantischer Suche dar [EP2562695A3, EP2045728A1, EP2400400A1, EP2199926A2, US20060271584A1, US20070208726A1, US20090076839A1, WO2008027503A9, WO2008131607A1, WO2017173104].

Verbunden mit dieser letzten Art von semantischen Suchverfahren sind jedoch mehrere Nachteile

1) Die Verfahren sind abhängig von explizit vorgegebenen Begriffsmodellen.
2) Falls diese Modelle für ein Anwendungsgebiet nicht existieren, müssen sie erst modelliert werden.

3) Die Qualität der Suchergebnisse ist darüber hinaus abhängig von der Qualität dieser Modelle.
4) Durch diese Modellabhängigkeit sind diese semantischen Suchverfahren nicht auf andere Anwendungsgebiete übertragbar.
5) Diese Verfahren versagen in der Regel bei Schreibfehlern und Begriffen, die nicht in den Begriffsmodellen enthalten sind.
6) Da fehlerhaft geschriebene Begriffe in der Regel nicht Bestandteil der Begriffsmodelle sind und unbekannte Begriffe nicht Bestandteil der Begriffsmodelle sein können, müssen diese Verfahren durch zusätzliche Verfahren zur Rechtschreibfehlererkennung oder –korrektur und durch konventionelle Volltextsuche ergänzt werden.

Das zu lösende Problem des „Semantischen Information Retrievals auf der Basis von Word Embeddings“ (SIR) besteht daher darin, eine Suchfunktion zu realisieren, die ohne explizit vorgegebenes Hintergrundwissen arbeitet. Die Suche soll über eine beliebige Dokumentenmenge so effizient wie konventionelle Information Retrieval Verfahren durchgeführt werden. Sie soll passende Dokumente unter Berücksichtigung der Ähnlichkeit der in Ihnen verwendeten Begriffe nach ihrer Ähnlichkeit sortiert ausgeben. Und sie soll die Anzahl der Ergebnisse soweit begrenzen, dass nur wirklich vergleichbare Dokumente betrachtet werden. Darüber hinaus sollen die ermittelten Ergebnisse für einen Nutzer nachvollziehbar sein. Und die Lösung soll ebenso zum Vergleich mit einem in Begriffen der Dokumente formulierten Nutzerprofil wie auch zum Vergleich von Dokumenten untereinander genutzt werden können.

Bekannt ist das Prinzip des Word Embedding. Bekannte Verfahren, z. B. Word2Vec (inkl. seiner Varianten Paragraph2Vec, Doc2Vec etc.), GloVe und fastText, ermitteln die Semantik einzelner Worte/Terme und können somit explizit vorgegebene Begriffsmodelle ersetzen. Als Worte einer Sprache können zusammenhängende Zeichenketten (alphanumerische Zeichen, Bindestrich) verstanden werden. Ein Term kann als Obermenge von Wörtern aufgefasst werden, der noch weitere Satz- oder druckbare Sonderzeichen umfassen kann oder aus mehreren zusammengehörenden Worten und Termen bestehen kann.

Dazu sei auf folgende Quellen verwiesen.

Word2Vec: Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean , https://arxiv.org/abs/1301.3781 (letzter Zugriff 6.2.2019).

GloVe: Global Vectors for Word Representation, Jeffrey Pennington, Richard Socher, Christopher D. Manning, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, October 25-29, 2014, Doha, Qatar.),

fastText: Facebook's Artificial Intelligence Research lab releases open source fastText on GitHub, Mannes, John, https://techcrunch.com/2016/08/18/facebooks-artificial-intelligence-research-lab-releases-open-source-fasttext-on-github/ (letzter Zugriff 6.2.2019).

Diese Verfahren basieren auf kontinuierlichen – im Gegensatz zu diskreten – Termvektoren (A Neural Probabilistic Language Mode, Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin; Journal of Machine Learning Research 3 (2003) 1137–1155).

Bei diesen Verfahren werden Terme/Worte durch einen geringer dimensionalen numerischen Vektor repräsentiert, der in der Regel nur wenige hundert Dimensionen umfasst, der aber im Gegensatz zu einem diskreten Termvektor alle Vektorkomponenten nutzt. Während bei der diskreten Repräsentation die einzelnen Dimensionen der Orthonormalbasis des Vektorraums entsprechen, Begriffe somit symbolisch repräsentieren und Dokumente als Linearkombination der Orthonormalvektoren repräsentiert werden, werden bei der kontinuierlichen Repräsentation Worte als Punkte (oder Vektoren) in einem Raum repräsentiert, dessen Orthonormalbasis als subsymbolische Repräsentation latenter Bedeutungen interpretiert werden kann (die Worte werden quasi in den Raum latenter Bedeutungen eingebettet).

Worte und Dokumente der diskreten Repräsentation liegen durch die „sparseness“ auf den Hyper-Kanten und Hyper-Flächen eines hochdimensionalen Raums, bei der kontinuierlichen Repräsentation jedoch in der Regel mitten im Raum, respektive seinen niedrig-dimensionalen Unterräumen.

Um die Positionen der Worte in dem Vektorraum der kontinuierlichen Repräsentation zu bestimmen, nutzen die oben beschriebenen Word-Embedding-Verfahren Methoden des unüberwachten Maschinellen Lernens.

Diese Lernverfahren nutzen den Kontext der Worte in den Texten eines Textkorpus – sprich deren Umgebungsworte, um die Position des Wortes in dem Vektorraum zu ermitteln.

Dies hat den Effekt, dass Begriffe die in Texten in gleichen oder vergleichbaren Zusammenhängen auftreten, im Vektorraum in enger räumlicher Nähe zu liegen kommen (siehe Illustration in Fig. 1).

Aus den so trainierten Word Embeddings lassen sich inhaltlich ähnliche Begriffe über unterschiedliche Distanzmaße, wie z.B. euklidische Distanz oder Kosinus-Distanz, ermitteln.

Ein anderes Maß ist die sogenannte Kosinus-Ähnlichkeit (s. o. Manning et al.), mit der die Ähnlichkeit von Vektoren über deren Skalarprodukt ermittelt wird.

Mit der Kosinus-Ähnlichkeit A kann bestimmt werden, ob zwei Vektoren in die gleiche Richtung weisen ($A = 1$), in ähnliche Richtungen weisen ($0.7 < A < 1$), orthogonal sind ($A = 0$) oder in entgegengesetzte Richtungen weisen ($-1 <= A < 0$).

Während die Kosinus-Ähnlichkeit A von konventionellen Termvektoren lediglich im Intervall [0,1] liegen kann, kann sie bei Word Embeddings im Intervall [-1,1] liegen.

Doc2Vec bzw. Paragraph2Vec (Distributed Representations of Sentences and Documents, Quoc Le, Tomas Mikolov Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

https://cs.stanford.edu/~quocle/paragraph_vector.pdf (letzter Zugriff 6.2.2019)) erweitert den Ansatz von Word2Vec um die Berücksichtigung von Dokument-Identifikatoren, die wie separate Terme beim Training genutzt werden. Diese Identifikatoren werden wie andere Terme auch in den gleichen Vektorraum eingebettet und können von Worten lediglich anhand der Syntax ihrer Bezeichner unterschieden werden.

Im Gegensatz hierzu werden Dokumente und Anfragen im SIR Verfahren durch Linearkombination der Word Embeddings ihrer Worte repräsentiert und in einem separaten Dokumentenraum der gleichen Dimensionalität repräsentiert. Document Embeddings und Query Embeddings werden hierbei durch Addition aller Word Embeddings der Worte eines Dokuments resp. einer Anfrage und anschließende Normalisierung bzgl. der Dokument- bzw. Anfragelänge erzeugt. Ein Query Embedding Vector ist aus Zamani, Croft; Estimating embedding vectors for queries, in Proceedings of the 2016 ACM International Conference of the Theory of Information Retrieval, S. 123-132, DOI, 10.1145/2970398.2970403 bekannt.

Während der Word2Vec- und der Doc2Vec-Ansatz die zu repräsentierenden Worte als atomar ansehen, geht der fastText-Ansatz (s. o. Facebooks' Artificial Intelligence) einen Schritt weiter und repräsentiert Worte durch die Menge ihrer N-Gramme (die Menge aller Sequenzen von N aufeinander folgenden Teilzeichenketten des Wortes). Durch diese Erweiterung können auch morphologische Ähnlichkeiten von Worten (wie Präfixe, Suffixe, Flexionen, Pluralbildungen, Variation der Schreibweisen usw.) in die Berechnung der Position der Wortvektoren mit einfließen, so dass auch die Position bisher unbekannter Worte („out-of vocabulary" Terme) im Vektorraum ermittelt werden kann. Der fastText-Ansatz ist daher in begrenztem Maß tolerant gegenüber Schreibfehlern und unbekannten Worten.

Bedingt durch den fastText Ansatz, N-Gramme zu nutzen, besitzt ein darauf aufsetzender Ansatz zwar eine gewisse Toleranz gegenüber Schreibfehlern und unbekannten Worten. Er ist jedoch nicht sensitiv für wohlgeformte Worte und erlaubt es, selbst unsinnige Zeichenkombinationen noch vergleichbar zu machen,

solange in ihnen mindestens ein N-Gramm enthalten ist, welches auch in der Trainingsmenge auftritt.

Das Hauptproblem dieser Ansätze besteht jedoch darin, dass bei einer Repräsentation von Worten in einem kontinuierlichen, subsymbolischen Vektorraum jedes Wort zu allen anderen eine Entfernung besitzt und alle Worte zueinander ähnlich sind, wenn auch zu unterschiedlichen Graden.

Beispielsweise wird das Wort „Auto“ zu „Automobil“, „Kraftfahrzeug“ und „KFZ“ in enger räumlicher Nähe stehen bzw. deren Winkel klein und damit deren Kosinus-Ähnlichkeit groß sein, zu „Fahrzeug“, „Transportmittel“ und „Flugzeug“ wird die Distanz zunehmen, der Winkel größer und die Kosinus-Ähnlichkeit kleiner sein, jedoch wird dieses Wort auch eine Entfernung zu den Worten „Hühnerbrühe“, „hobeln“, „samtig“, „Kiel holen“ und „Ouagadougou“ besitzen und zu deren Vektoren einen sehr großen Winkel bilden.

Es fehlt somit ein Kriterium, mit dem die „ähnlichsten“ von den „unähnlichen“ Begriffen unterschieden werden können.

Kombiniert man die Word Embeddings der Worte einer Anfrage oder eines Dokuments wie beschrieben zu Query resp. Document Embeddings, überträgt sich dieses Problem: Alle Dokumente sind zu allen anderen Dokumenten ähnlich und eine Anfrage ist zu allen Dokumenten ähnlich, jeweils jedoch zu unterschiedlichen Graden.

In der Veröffentlichung Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". *Sidorov, Grigori; Gelbukh, Alexander; Gómez-Adorno, Helena; Pinto, David, Computación y Sistemas. 18 (3): 491–504.* doi:10.13053/CyS-18-3-2043. (letzter Zugriff 6.2.2019), wurde das Maß „soft cosine similarity“ vorgestellt, welches es erlaubt, in die Verrechnung der Vektorkomponenten der Kosinus-Ähnlichkeit einen zusätzlichen Gewichtungsfaktor einfließen zu lassen.

Dieser Gewichtungsfaktor kann dazu genutzt werden, die Ähnlichkeit von einzelnen Worten in die Berechnung der Ähnlichkeit von Dokumentenvektoren einfließen zu lassen. Im Prinzip könnte als Wortähnlichkeit die Kosinus-Ähnlichkeit von Word Embeddings genutzt werden. Dies verbietet sich jedoch für eine Suchfunktion zur Laufzeit aus Effizienzgründen, da bei der Berechnung des Skalarprodukts der Dokumentvektoren jedes Wort eines Dokuments oder einer Anfrage mit allen Worten eines anderen Dokuments verglichen werden müsste.

Durch eine Vorausberechnung der Wortähnlichkeiten kann dies zur Laufzeit zwar umgangen werden, dennoch bringt die Vorausberechnung einen quadratischen Aufwand mit n*(n-1)/2 Vergleichen mit sich.

Selbst wenn bei einem Vokabular von 100.000 Worten jede Berechnung nur eine Millisekunde benötigen würde, würde die Berechnung aller Ähnlichkeiten rund 57,9h benötigen. Eine Parallelisierung der Berechnung wäre zwar möglich, würde aber zusätzlich Hardware erfordern.

Auch ein auf „soft cosine similarity“ basierendes Verfahren würde unter dem Problem „alles ist zu allem in unterschiedlichen Graden ähnlich“ leiden.

Während eine rein Boole‘sche Retrieval-Funktion (s. o. Manning et al.) das harte Kriterium des Enthaltenseins eines Begriffs in einem Dokument nutzen kann, um die Anzahl der Dokumente auf diejenigen zu begrenzen, die den Begriff enthalten, stellt kein auf Begriffsähnlichkeiten basierender Ansatz ein analoges hartes Kriterium zur Verfügung.

Die KR102018058449A beschreibt zwar ein System und eine Methode zur semantischen Suche mittels Wortvektoren, das anscheinend auch auf einem zur Kosinus-Ähnlichkeit verwandten Ähnlichkeitsmaß basiert, es bleibt aber unklar, ob dieses Verfahren für diskrete Termvektoren oder kontinuierliche Word Embeddings ausgelegt ist. Die Vermutung liegt nahe, dass dieser Ansatz der beschriebenen Ähnlichkeitsproblematik unterliegt und alle Dokumente zurückliefert.

Die US20180336241A1 beschreibt ein Verfahren zur Berechnung der Ähnlichkeit von Suchanfragen zu Stellenbezeichnungen, welches Anfrage- und Dokumentvektoren aus Word Embeddings berechnet, und eine Suchmaschine, die eingeschränkt auf das Anwendungsgebiet der Stellentitelsuche dazu verwendet wird, ähnliche Stellenangebote zu ermitteln. Der konkrete Aufbau der Suchmaschine wird nicht beschrieben, auch wird weder auf die Ähnlichkeitsproblematik eingegangen noch beschrieben, wie die Menge der Suchergebnisse beschränkt werden kann.

Die WO2018126325A1 beschreibt einen Ansatz zum Erlernen von Document Embeddings aus Word Embeddings mit Hilfe eines Convolutional Neural Networks. Document Embeddings der präsentierten Erfindung werden hingegen durch lineare Kombination von Word Embeddings berechnet.

Die WO2017007740A1 beschreibt ein System, das kontextuelle und, im Gegensatz zu den strukturellen N-Grammen von fastText, morphologische Ähnlichkeiten in einer speziellen Form von „Knowledge powered neural NETwork“ (KNET) nutzt, um mit seltenen Worten oder Worten, die nicht im Dokumentkorpus auftreten, umzugehen. KNET kann als alternativer Ansatz zur Nutzung von Word2Vec, GloVe oder fastText in der vorliegenden Erfindung angesehen werden.

Die US20180113938A1 beschreibt ein auf Word Embeddings basierendes Recommender System für (semi-)strukturierte Daten. Die Ermittlung von Document Embeddings folgt einem anderen Prinzip. Auch hier wird die Ähnlichkeitsproblematik nicht adressiert.

Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 gelöst. Dabei werden Dokumente verwendet, die tokenisierte Zeichenketten aufweisen.

In einem ersten Schritt wird mit einem Indexierungsverfahren ein invertierter Index (auch inverser Index genannt) für mindestens eine Teilmenge der Dokumente berechnet. D.h. es wird eine Datei oder Datenstruktur angelegt, in der für jede tokensierte Zeichenkette angegeben ist, in welchen Dokumenten sie enthalten ist.

Anschließend werden für die mindestens eine Teilmenge der Dokumente Word Embeddings berechnet, d.h. die Zeichenketten werden auf einen Vektor mit reellen Zahlen abgebildet.

Dann wird für die mindestens eine Teilmenge der Dokumente jeweils ein Document Embedding berechnet, indem für jedes Dokument die Word Embeddings aller Zeichenketten, insbesondere Worte des Dokuments, addiert und mit der Anzahl der Zeichenketten, insbesondere Worten, normiert werden, wobei vorher, nachfolgend oder parallel mit den berechneten Word Embeddings mithilfe eines Clusteringverfahrens SimSet-Gruppen von ähnlichen Zeichenketten berechnet werden.

Anschließend wird in einer Anfragephase eine Anfrageexpansion vorgenommen, bei der

i) Anfrageterme, die in SimSet-Gruppen vorkommen, oder
ii) Anfrageterme, die nicht in den SimSet-Gruppen aber in den Dokumenten vorkommen, oder
iii) Anfrageterme, die nicht in den Dokumenten vorkommen, insbesondere auch falsch geschriebene Anfrageterme

für eine Vorselektion (insbesondere mittels des invertierten Indexes für die Teilmenge der Dokumente) der Dokumente verwendet werden, um eine Einschränkung der Treffermenge zu erreichen. Dann wird ein Query Embedding durchgeführt.

Danach wird ein Vergleich des Query Embeddings mit den Document Embeddings unter Verwendung der zuvor berechneten SimSet-Gruppen zur mengenmäßigen Einschränkung der Anzahl der zu vergleichenden Document Embeddings, vorselektierten Dokumente durchgeführt, um automatisch ein Ranking der Ähnlichkeit der Dokumente zu ermitteln und diese anzuzeigen und / oder abzuspeichern. Anhand dieses Rankings können z.B. die zur Anfrage oder einem anderen Dokument ähnlichsten Dokumente ermittelt werden.

Es sei angemerkt, dass SimSet-Gruppen keine Dokumente umfassen, sondern Worte.

In einer Ausführungsform wird für das Word Embedding ein CBOW Modell oder ein Skip-gram Modell verwendet.

In einer weiteren Ausführungsform wird ein nicht-parametrisiertes Clusteringverfahren verwendet, sodass keine A-priori-Annahmen gemacht werden müssen. Dabei können als Clusteringverfahren hierarchische Verfahren, insbesondere divisive oder agglomerative Clusteringverfahren eingesetzt werden. Es ist auch möglich, dass das Clusteringverfahren als dichtebasiertes Verfahren, insbesondere DBSCAN oder OPTICS, ausgebildet ist. Alternativ kann das Clusteringverfahren als graphenbasiertes Verfahren, insbesondere als Spectral Clustering oder Louvain, ausgebildet sein.

Zur Einschränkung des Suchraums kann in einer Ausführungsform eine Kosinus-Ähnlichkeit, eine Termfrequenz und/oder eine inverse Dokumentenfrequenz als Schwellenwert bei der Clusterbildung verwendet werden.

Die Aufgabe wird auch durch eine Vorrichtung mit den Merkmalen des Anspruchs 8 gelöst.

Ausführungsformen werden im Zusammenhang mit den folgenden Figuren dargestellt. Dabei zeigt

Fig. 1 ein Beispiel für Cluster von ähnlichen Begriffen in einer Menge von rd. 73.500 Dokumenten;

Fig. 2 eine schematische Darstellung einer Indexierungsphase in einer Ausführungsform des Verfahrens;

Fig. 2A Beispiele für ein Word Embedding und ein Document Embedding;

Fig. 3 eine schematische Darstellung der Ermittlung von SimSet-Gruppen;

Fig. 4A-C eine Bestimmung der ähnlichsten Word Embeddings zur Einschränkung eines Similarity Graphs;

Fig. 4D ein Beispiel für ein SimSet für das Beispiel aus Fig. 2A;

Fig. 5 eine schematische Darstellung der Erzeugung eines Similarity-Graphen im Rahmen eines Clustering-Verfahrens;

Fig. 6 eine schematische Darstellung einer Anfragevorbereitung;

Fig. 7 eine schematische Darstellung einer Fallunterscheidung einer Anfrageexpansion;

Fig. 8 eine schematische Darstellung eines Dokumenten-Retrievals.

Die Ausführungsformen, die im Folgenden beschrieben werden, machen sich das an sich bekannte Prinzip der Word Embeddings in Dokumenten zunutze.

Es wird davon ausgegangen, dass Dokumente und Anfragen bereits vorverarbeitet wurden und als tokenisierte Sequenzen von Zeichenketten in einer einheitlichen Zeichenkodierung vorliegen. Tokenisieren, heißt, einen Text in einzeln verarbeitbare Bestandteile (Worte, Terme und Satzzeichen) zu zerlegen.

Die Problemlösung erfolgt in zwei Phasen, der Indexierungsphase und der Anfragephase. Die Indexierungsphase dient dem Aufbau effizienter Datenstrukturen, die Anfragephase der Suche von Dokumenten in diesen Datenstrukturen. Diese zwei Phasen können optional durch eine dritte Phase, die Empfehlungsphase, ergänzt werden.

Indexierungsphase

Die Abfolge der Verarbeitungsschritte in der Indexierungsphase ist schematisch in Fig. 2 dargestellt.

Ausgangspunkt ist eine Menge von Dokumenten 101, die jeweils als tokenisierte Sequenzen von Zeichenketten vorliegen.

Mithilfe eines Indexierungsverfahrens 102 wird für diese Dokumente 101 ein invertierter Index 103 berechnet. Dieser inverse Index 103 ermöglicht anhand der in den Dokumenten 101 enthaltenen Zeichenketten, wie z. B. Worte und/oder Terme, den schnellen Zugriff auf alle Dokumente 100, in denen gegebene Zeichenketten enthalten sind.

Aus den Dokumenten 101 werden darüber hinaus mithilfe an sich bekannter Verfahren 104 zur Berechnung von Word Embeddings 105, wie Word2Vec, GloVe, FastText, Gauss2Vec oder Bayesian Skip-gram, Word Embeddings 105 für einen niedrig dimensionalen, kontinuierlichen Wortvektorraum errechnet.

Word Embedding 105 ist der Sammelbegriff für eine Reihe von Sprachmodellierungs- und Feature-Lerntechniken im Natural Language Processing (NLP), bei denen Zeichenketten aus einem Vokabular, insbesondere einem Wortschatz auf Vektoren reeller Zahlen abgebildet werden, die als Word Embeddings bezeichnet werden. Konzeptionell geht es um eine mathematische Einbettung eines Raumes mit vielen Dimensionen in einen kontinuierlichen Vektorraum mit einer geringeren Dimension.

Für die Berechnung der Word Embeddings 105 wird in der dargestellten Ausführungsform das CBOW Modell verwendet, welches die Vorhersage von Worten anhand von Kontextworten ermöglicht. In einer anderen Ausführungsvariante kann anstelle von CBOW auch ein Skip-gram Modell verwendet werden, mit dem für ein Wort die Vorhersage von Kontextworten möglich ist. Diese Berechnungsverfahren sorgen dafür, dass die Wortvektoren ähnlicher Begriffe (Begriffe, die häufig im gleichen Zusammenhang verwendet werden), im Wortvektorraum in räumlicher Nähe zueinander angeordnet werden.

Für die Dokumente der Dokumentenmenge 101 werden darüber hinaus Document Embeddings 107 berechnet 106, indem für jedes Dokument die Word Embeddings

105 aller Zeichenketten des Dokuments addiert und mit der Anzahl der Worte normiert werden.

Hierdurch werden numerische Überläufe und Abhängigkeiten der Document Embeddings 107 von der Dokumentenlänge vermieden, sodass auch unterschiedlich lange Dokumente noch sinnvoll miteinander verglichen werden können.

Da Dokumente, die die gleichen oder sehr ähnliche Worte (d. h. Zeichenketten) verwenden, mit hoher Wahrscheinlichkeit ähnliche oder verwandte Themen behandeln, führt die Addition ihrer Word Embeddings 105 dazu, dass deren Document Embeddings 107 in enger räumlicher Nähe zueinander im Dokumentenvektorraum angeordnet werden.

Im Fig. 2A werden Beispiele für ein Word Embedding 105 und ein Document Embedding 107 dargestellt.

Die zu untersuchende Dokumentenmenge weist in diesem Beispiel nur einen Satz auf: „Ein Polizeibeamter ist ein Beamter“.

Damit ergeben sich vier Vektoren für die Word Embedding 105 und ein Vektor für das Document Embedding 107.

In einem weiteren Schritt werden aus den Word Embeddings 105 mithilfe eines Clusteringverfahrens 108 Gruppen von sehr ähnlichen Zeichenketten/Worten, die im Folgenden als SimSet-Gruppen 109 bezeichnet werden, ermittelt. Dieser Schritt kann auch vorher, nachher oder parallel zu dem Schritt der Document Embedding 107 Bestimmung durchgeführt werden.

Da die Anzahl der potentiellen Gruppen von ähnlichen Worten unbekannt ist, wird ein nicht-parametrisiertes Clusteringverfahren 108 eingesetzt, bei dem die Anzahl der Cluster nicht vorgegeben werden muss. Zu den einsetzbaren Verfahren zählen hierarchische Verfahren, wie divisives Clustern, agglomeratives Clustern, und dichtebasierte Verfahren, wie DBSCAN, OPTICS und diverse Erweiterungen.

In einer Ausführungsvariante können auch graphenbasierte Verfahren, wie Spectral Clustering und Louvain, verwendet werden.

Diese Ausführungsvariante zur Berechnung von SimSets 109 ist in Fig. 3 dargestellt.

Für das graphenbasierte Clustern von Word Embeddings 105 werden die Ähnlichkeiten zwischen allen Word Embeddings 105 als gewichtete Kanten in einem Graphen – als Similarity Graph bezeichnet – 108.4 betrachtet, dessen Knoten durch die Word Embeddings 105 gebildet werden.

Die Gewichtung der Kanten entspricht hier dem Grad der Ähnlichkeit. In einer naiven Lösung würde dieser Graph vollständig verlinkt sein, da jedes Word Embedding eine Distanz resp. Ähnlichkeit zu allen anderen besitzt. Der Graph würde daher n*(n-1)/2 Kanten umfassen und beim Clustern müsste eine exponentielle Menge von Clustern (potentiell 2^n Teilmengen) durchsucht werden. Die Bestimmung der optimalen Cluster wäre daher NP-schwer.

Mit zwei Einschränkungen kann sowohl die Anzahl der zu betrachtenden Knoten des Similarity Graphen als auch die Menge der zu berücksichtigen Kanten drastisch reduziert werden.

Im Kontext einer Suche, die neben der eigentlichen Anfrage auch ähnliche Worte mitberücksichtigt, reicht es, die Zeichenketten/Worte zu betrachten, die in eine spezielle Form von Clustern – als SimSets 109 bezeichnet – fallen. Diese Zeichenketten/Worte sollten

a) häufig in der Textmenge auftreten (gemessen durch die Termfrequenz, TF, siehe Manning et al.),

b) einen hohen Informationsgehalt besitzen (gemessen durch die inverse Dokumentenfrequenz IDF, siehe Manning) und

c) zueinander sehr ähnlich sein.

Da Termfrequenzen in einem Korpus einer Potenzverteilung folgen, reicht es aus, dem Pareto-Prinzip zu genügen und diejenigen Terme zu wählen die z. B. 80 % - 95 % aller Terme mit der größten kombinierten TFIDF (Termfrequenz inverse Dokumentenfrequenz) (obige Fälle a und b kombiniert) des Korpus umfassen."

Der konkrete Wert kann als Wichtigkeitsschwellenwert genutzt werden, um die Anzahl der SimSets 109 zu steuern.

Die Ähnlichkeitsmessung von Word Embeddings 105 unter Verwendung der Kosinus-Ähnlichkeit (oben unter c) ist in Fig. 4A-C dargestellt.

Fig. 4A zeigt die Ähnlichkeit aller Word Embeddings 105 zu einem gegebenem Word Embedding (gestrichelter Referenzvektor).

Dann können z. B. alle Word Embeddings 105 mit negativer Ähnlichkeit – Kosinus-Ähnlichkeit < 0, Winkel > 90° – ausgeschlossen werden (Fig. 4B, schraffierte Halbebene).

Auch könnte man einen Ähnlichkeitsschwellenwert auf Basis der Kosinus-Ähnlichkeit für die Ähnlichkeit in einem Bereich von weniger 0,87 bis 0,7 ansetzen, und damit alle Word Embeddings mit einem Winkel zwischen 90° und 45° bis 60° als unähnlich ignorieren (Fig. 4C, fett schraffierte Segmente),

Übrig bleiben die Word Embeddings 105, die zum gestrichelten Referenzvektor am ähnlichsten sind, die mit einem Winkel von höchstens 30° - 45. Diese werden dann als Knoten des Similarity Graphen verwendet. Der konkrete Wert des Ähnlichkeitsschwellwertes steuert die Größe – im Sinne der Begriffszahl – der SimSets 109.

In der Fig. 4D ist die Berechnung der Kosinus-Ähnlichkeit für den Beispielsatz aus Fig. 2A dargestellt. Die Schattierung in den einzelnen Zellen korrespondiert dabei mit den Schraffierungen in den Fig. 4A – C.

In der Fig. 4D sind die Zahlenwerte für die Kosinus-Ähnlichkeit dargestellt, wobei eine symmetrische Anordnung vorliegt. Auf der Hauptdiagonalen betragen die Ähnlichkeitswerte naturgemäß 1.

In einem ersten Schritt können die negativen Ähnlichkeiten (z.B. Polizeibeamter – ist) aussortiert werden, was der Situation in Fig. 4B entspricht; d.h. es wird nur noch die positive Halbebene betrachtet.

Positive Zahlenwerte unterhalb eines Ähnlichkeitsschwellenwertes (hier 0,75) sind dunkelgrau hinterlegt und entsprechen dabei der Einengung des Winkelbereiches in Fig. 4C. Das Wort „ein“ hat demnach z.B. nur eine geringe Ähnlichkeit zu den Worten „Polizeibeamter“, „ist“ und „Beamter“.

Somit bleibt als relevanter Wert oberhalb eines Ähnlichkeitsschwellenwertes von 0,75 (und außerhalb der Hauptdiagonalen) mit einer Ähnlichkeit von 0,7533 die Wortpaarung „Polizeibeamter“ und „Beamter“ übrig. Diese beiden Worte bilden dann eine SimSet-Gruppe 109 für die Beispieldokumentenmenge.

Anhand dieser Betrachtung (und mit Bezug auf Fig. 3 und 5), kann der Similarity Graph 108.4 wie folgt konstruiert werden 108.3:

Für jedes Wort der Dokumentenmenge 101 wird das kombinierte TFIDF Maß berechnet und sortiert 108.1 und daraus eine nach absteigendem TFIDF sortierte, reduzierte Wortliste (d. h. Liste von Zeichenketten) 108.2 gewonnen.

Zur Extraktion des Similarity Graphen 108.3 werden diese Worte/Zeichenketten der Reihenfolge nach durchgegangen und für jedes Wort / jede Zeichenkette mit einer TFIDF über dem Wichtigkeitsschwellenwert, wird der in Fig. 5 dargestellte erste Entscheidungsprozess durchgeführt. Für den Fall eines negativen Ergebnisses bei einem der drei Vergleiche, wird die jeweilige Zeichenkette, das jeweilige Wort oder der jeweilige Term verworfen (in Fig. 5 nicht dargestellt).

Für jedes Wort / jede Zeichenkette werden zu dessen Word Embedding 105 die ähnlichsten Worte/Zeichenketten ermittelt, deren Kosinus-Ähnlichkeit den Ähnlichkeitsschwellwert übersteigt (zweiter Entscheidungsprozess in Fig. 5).

Für die entsprechenden Worte werden im Similarity Graphen entsprechende Knoten angelegt, sofern diese noch nicht existieren, und mit einer ungerichteten Kante versehen, deren Gewicht der konkreten Kosinus-Ähnlichkeit zwischen den Worten entspricht (Schritt 108.3 in Fig. 5).

Der so konstruierte Similarity Graph enthält alle Knoten mit hohen TFIDF Werten, die eine Ähnlichkeit zueinander größer/gleich dem Ähnlichkeitsschwellwert besitzen. Dieser Graph hat die Eigenschaft, dass alle Knoten, die in enger räumlicher Nähe im Wortvektorraum liegen, engmaschiger miteinander verbunden sind als mit Knoten, die weiter entfernt liegen.

In dem Similarity Graphen 108.4 können mit einem graphenbasierten Clusterverfahren, wie z. B. Louvain (*Fast unfolding of communities in large networks". Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne, Journal of Statistical Mechanics: Theory and Experiment. 2008 (10): P10008.* arXiv:0803.0476. Bibcode:2008JSMTE..10..008B. doi:10.1088/1742-5468/2008/10/P10008 (letzter Zugriff 6.2.2019)), Cluster von Worten/Zeichenketten identifiziert werden, die hohe Ähnlichkeit zueinander besitzen und von Clustern von Worten/Zeichenketten abgegrenzt werden, zu denen sie geringere Ähnlichkeiten besitzen. Diese Cluster von ähnlichen Worten werden als SimSets 109 für die weitere Verwendung gespeichert.

In einer Ausführungsvariante werden die SimSets 109 zum effizienten Retrieval über einen weiteren invertierten Index zugreifbar gemacht. Um schnell identifizieren zu können, ob ein gegebenes Wort in einem SimSet 109 enthalten ist und wenn ja in welchem. Hierfür kann der gleiche Mechanismus (ein invertierter Index) verwendet, wie bei der Bestimmung der Dokumente, die ein gegebenes Wort enthalten.

Anfragephase

Die Beantwortung einer Suchanfrage nach ähnlichen Dokumenten an die in der Indexierungsphase ermittelten Daten erfolgt in zwei Schritten.

Im ersten Schritt, der Anfragevorbereitung, wird eine Anfrage 201, die als tokenisierte Sequenz von Zeichenketten vorliegt, vorbereitet indem für sie, analog zu einem normalen Dokument, ein Query Embedding 205 berechnet wird.

Im zweiten Schritt, dem Retrieval, wird dieses Query Embedding 205 gegen die Document Embeddings 107 potentiell in Frage kommender, vorselektierter Dokumente 204 verglichen und diese anhand ihrer Ähnlichkeit sortiert, um dann insbesondere angezeigt und/oder abgespeichert zu werden. Dieser Vergleich erfolgt mit den im Clusteringverfahren gebildeten SimSet-Gruppen 109 zur mengenmäßigen Einschränkung der Anzahl der zu vergleichenden Document Embeddings 107. Damit wird dann automatisch ein Ranking der Ähnlichkeit der Dokumente ermittelt, angezeigt und/oder abgespeichert

Anfragevorbereitung

Der Ablauf der Anfragevorbereitung ist in Fig. 6 dargestellt.

Die Anfragevorbereitung besteht aus mehreren Teilen: der Berechnung des Query Embeddings 104 für eine Anfrage 201, die analog zur Berechnung der Document Embeddings 106 verläuft und ein Query Embedding 205 ergibt, einer Anfrageexpansion 202 und einer Dokumentenselektion 203.

Da im Dokumentvektorraum jedes Dokument zu allen anderen ähnlich ist (jedoch zu unterschiedlichen Graden), gilt dies auch für das analog konstruierte Query Embedding 104. Dies hätte jedoch zur Konsequenz, dass bei jeder Anfrage immer alle Dokumente gefunden werden würden, da ein hartes Auswahlkriterium fehlt.

Zur Konstruktion eines entsprechenden Auswahlkriteriums wird für die Anfrage 201 eine Anfrageexpansion 202 durchgeführt. Bei der Anfrageexpansion (siehe Fig. 7) wird unterschieden zwischen

a) Anfragetermen, die in SimSets 109 vorkommen,
b) Anfragetermen, die nicht in den SimSets aber im Korpus (d.h. der Dokumente 101) vorkommen,
c) Anfragetermen, die nicht im Korpus vorkommen. Dazu zählen auch falsch geschriebene Anfrageterme.

Im Fall a) besteht die Anfrageexpansion darin, dass für jedes SimSet 109, in dem ein Anfrageterm enthalten ist, die Dokumente vorselektiert werden, in denen wenigstens einer der SimSet Terme enthalten ist (202.1 in Fig. 7). Dieses Vorgehen hat zwar den Nachteil, dass Dokumente, in denen Terme mit einem geringeren Ähnlichkeitsgrad enthalten sind, ignoriert werden. Der Vorteil jedoch liegt in einer stark reduzierten Treffermenge (analog zu einer Boole'schen Suche) und der Erklärbarkeit der Treffer über die Begriffe der SimSets.

Im Fall c) können bei einer Umsetzung über eine Ausführungsvariante, die Word2Vec Word Embeddings verwendet, die vorselektierten Dokumente auf die leere Menge gesetzt werden (202.3 in Fig. 7).

Bei der Umsetzung über eine Ausführungsvariante, die fastText Word Embeddings verwendet, wäre in den Fällen b) und c) keine solche Vorselektion der Dokumente möglich, da diese Variante mit Schreibfehlern und „out-of-vocabulary" Termen umgehen kann. Um dennoch für diese Fälle eine Reduktion der Treffermenge zu erzielen, impliziert folgende Überlegung eine Lösung:

Wie beschrieben, bestehen SimSets 109 aus Termen, die

1) eine hohe TFIDF besitzen und
2) zueinander sehr ähnlich sind.

D. h., es kann einzelne Anfrageterme geben, die im Korpus enthalten sind, aber nicht in einem SimSet 109 und dennoch zu einem Anfrageterm eine Ähnlichkeit über dem Ähnlichkeitsschwellwert besitzen.

Im Fall b) und in der Ausführungsvariante, die fastText Word Embeddings verwendet, auch im Fall c), können für diese Anfrageterme über die Word Embeddings 105 die Dokumententerme bestimmt werden, die eine Ähnlichkeit oberhalb des Ähnlichkeitsschwellwerts besitzen, aber nicht in den SimSets enthalten sind (202.2 in Fig. 7). Diese Terme können dann ebenfalls zur Anfrageexpansion herangezogen werden, um eine Vorselektion 203 der Dokumente vorzunehmen.

Die vorselektierten Dokumente 204 werden zum Vergleich mit dem Query Embedding 205 an das Retrieval übergeben.

Es werden bei der Anfrageexpansion 202, wenn möglich, SimSets verwendet, um analog zu konventioneller semantischer Suche, Anfragen zu expandieren (siehe Fig. 7). Da die expandierten Anfragen zum Retrieval von Dokumentkandidaten aus dem invertierten Index verwendet werden, liefert das Verfahren analog zu einer konventionellen Suche eine erweiterte Ergebnismenge, ohne dabei jedoch in das beschriebene Problem des unbeschränkten Retrievals zu laufen, das ein rein auf Word Embeddings basierender Ansatz nach sich ziehen würde. Damit liefert dieses Verfahren gegenüber einer Volltext-Suche erweiterte, jedoch mengenmäßig begrenzte Ergebnisse.

Retrieval

Nachdem die Dokumente vorselektiert 204 wurden und das Query Embedding 205 berechnet wurde, erfolgt das Retrieval wie in Fig. 8 dargestellt.

Hierzu werden für die vorselektierten Dokumente 204 deren Document Embeddings 107 zu den selektierten Document Embeddings 302 zusammengestellt. Für jedes Document Embedding 302 wird mithilfe des Kosinus-Ähnlichkeitsmaßes die Kosinus-Ähnlichkeit zum Query Embedding berechnet und die Dokumente nach absteigender Ähnlichkeit zum Dokumentenranking 304 sortiert.

In einer Ausführungsvariante kann die Berechnung über eine bekannte Map-Reduce Architektur parallelisiert werden, um so auch sehr große Dokumentmengen effizient zu verarbeiten.

Da, wie beschrieben, die Kosinus-Ähnlichkeit einer kontinuierlichen Vektorraum-Repräsentation auch negative Werte annehmen kann, kann während des Dokumentenrankings 304 ein zusätzliches Filterkriterium verwendet werden, um die Menge der Suchtreffer weiter einzuschränken. Suchergebnisse, deren Document Embeddings eine negative Kosinus-Ähnlichkeit zum Query Embedding besitzen, können herausgefiltert werden, da sie – sozusagen – gegensätzlich zur Anfrage wären. Da auch kleine Kosinus-Ähnlichkeiten von Winkeln größer als 60° auf sehr unähnliche Vektoren hinweisen, ist es – in einer weiteren Ausführungsvariante von 303 – darüber hinaus zweckmäßig, die Dokumente in 302 anhand eines Mindestähnlichkeits-Schwellwertes zu filtern.

In einer weiteren Ausführungsvariante kann anstelle des Query Embeddings 205 auch ein Embedding von Nutzerprofilen verwendet wird, welches analog zu einem Query 205 oder Document Embedding 107 aus einer Beschreibung des Nutzers oder seiner Interessen konstruiert werden kann.

Empfehlungsphase

In einer weiteren Ausführungsvariante kann in einer optionalen Empfehlungsphase anstelle des Query Embeddings 205 auch ein beliebiges Document Embedding 107 zur Berechnung der Kosinus-Ähnlichkeit und zum Ranking der Dokumente untereinander verwendet werden, um so die zu einem Dokument ähnlichsten Dokumente zu bestimmen.

Durch die hier beschriebenen Ausführungsformen wird das technische Problem einerseits dadurch gelöst, dass Begriffsbedeutungen nicht wie bei herkömmlichen Suchverfahren durch ein Begriffsmodell vorgegeben werden müssen, sondern direkt aus dem Kontext der Worte/Zeichenketten innerhalb der Dokumente bestimmt werden können. Andererseits erlaubt die Ermittlung der SimSets 109 auf der Basis der bestimmten Begriffsbedeutung nicht nur die Menge der zu vergleichenden Dokumente zum Anfragezeitpunkt effizient einzuschränken,

sondern auch dem Benutzer Begründungen für die Trefferermittlung auf der Basis der in den SimSets vorausberechneten Begriffsähnlichkeiten zu geben, um so die Nachvollziehbarkeit der Suchergebnisse zu unterstützen.

Die Wirkung einer herkömmlichen semantischen Suche besteht in der Nutzung von Hintergrundwissen in Form von Begriffsmodellen, wie Taxonomien, Thesauri, Ontologien, Wissensgraphen, um bessere Suchergebnisse zu liefern als herkömmliche Volltextsuchmaschinen.

Der Vorteil der hier beschriebenen Ausführungsformen besteht darin, ohne dieses Hintergrundwissen auszukommen und die Bedeutung und Ähnlichkeit von Begriffen allein aus den Dokumenttexten zu erlernen.

Es ist damit auch in Anwendungsgebieten einsetzbar, in denen solch ein Hintergrundwissen nicht vorliegt oder dessen Erhebung zu kostspielig wäre.

Es ist nach Installation und Konfiguration ohne zusätzliche Informationen umgehend einsetzbar.

Entgegen einer naiven Verwendung von Word Embeddings zur Umsetzung einer semantischen Suchfunktion, wird es durch das Konzept der SimSets möglich, die Anzahl der Suchtreffer – analog zu einem rein Boole'schen Ausschlußkriterium – zu filtern und damit die Ergebnismenge für den Benutzer auf die „relevantesten" Dokumente einzuschränken.

Abwandlungen zur Umgehung der Erfindungen bestehen darin, vortrainierte Modelle von Word Embeddings zu verwenden. Allgemeine vortrainierte Modelle sind beispielsweise von Google, FaceBook und anderen bereits beziehbar.

Anstelle die Word Embeddings mit Word2Vec, GloVe oder fastText zu berechnen, könnte KNET zur Abwandlung der Erfindung genutzt werden.

Anwendungsmöglichkeiten der Ausführungsformen finden sich z. B. in Content- und Document-Management-Systemen, Informationssystemen, Information Retrieval Systemen von Bibliotheken und Archiven.

Bezugszeichenliste

101 Dokumente
102 Indexierungsverfahren
103 invertierter Index
104 Berechnung von Word Embeddings
105 Menge von Word Embeddings
106 Berechnung von Document Embeddings
107 Document Embedding
108 Clusteringverfahren
108.1 Berechnung und Sortierung der TFIDF
108.2 Sortierung der Worte absteigend nach TFIDF
108.3 Extraktion Similarity Graph
108.31 AnlagenAnlegen von Knoten und Kanten
108.4 Similarity Graph
108.5 Graph Clustering
109 SimSets / SimSet-Gruppe (Gruppe von ähnlichen Zeichenketten/Wörtern)

201 Anfrage
202 Anfrageexpansion
202.1 Alle Dokumente in SimSet-Gruppe enthalten
202.2 Dokumente, die Terme enthalten, deren Ähnlichkeit zur Anfrage größer ist als der Ähnlichkeitsschwellenwert
202.3 leere Menge
203 Vorselektion / Dokumentenselektion
204 vorselektierte Dokumente
205 Query Embedding

301 Doc Embedding Lookup
302 Selektierte Doc Embeddings
303 Ranking nach Kosinusähnlichkeit
304 DokumentenrankinDokumentenranking

Patentansprüche

1. Verfahren zur Vorauswahl und Ermittlung ähnlicher Dokumente aus einer Menge an Dokumenten (101), wobei die Dokumente (101) tokenisierte Zeichenketten aufweisen,

dadurch gekennzeichnet, dass

a) mit einem Indexierungsverfahren (102) ein invertierter Index für mindestens eine Teilmenge der Dokumente (101) berechnet wird,

b) für die mindestens eine Teilmenge der Dokumente (101) Word Embeddings (105) berechnet werden,

c) für die mindestens eine Teilmenge der Dokumente (101) für jedes dieser Dokumente (101) jeweils ein Document Embedding (107) berechnet wird, indem für jedes Dokument (101) die Word Embeddings (105) aller Zeichenketten, insbesondere Worte des Dokuments (101), addiert und mit der Anzahl der Zeichenketten, insbesondere Worten, normiert werden (106), wobei vorher, nachfolgend oder parallel

d) mit den berechneten Word Embeddings (105) mithilfe eines Clusteringverfahrens SimSet-Gruppen (109) von ähnlichen Zeichenketten berechnet werden, und

anschließend

e) in einer Anfragephase (200) eine Anfrageexpansion (202) vorgenommen wird, bei der

i) Anfrageterme, die in SimSet-Gruppen (109) vorkommen, oder
ii) Anfrageterme, die nicht in den SimSet-Gruppen (109) aber in den Dokumenten (101) vorkommen, oder

iii) Anfrageterme, die nicht in den Dokumenten (101) vorkommen, insbesondere auch falsch geschriebene Anfrageterme

für eine Vorselektion (203) der Dokumente verwendet werden, um eine Einschränkung der Treffermenge zu erreichen, und dann zunächst ein Query Embedding (205) bestimmt wird

und dann

f) ein Vergleich des Query Embeddings (205) mit den Document Embeddings (107) der, unter Verwendung der im Schritt d) mit dem Clusteringverfahren gebildeten SimSet-Gruppen (109) zur mengenmäßigen Einschränkung der Anzahl der zu vergleichenden Document Embeddings (107), vorselektierten Dokumente durchgeführt wird, um automatisch ein Ranking der Ähnlichkeit der Dokumente (101) zu ermitteln und diese anzuzeigen und / oder abzuspeichern.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, dass als Word Embedding Verfahren ein CBOW Modell oder ein Skip-gram Modell verwendet wird.

3. Verfahren nach Anspruch 1 oder 2, **dadurch gekennzeichnet**, dass ein nicht-parametrisiertes Clusteringverfahren (108) eingesetzt wird.

4. Verfahren nach Anspruch 3, **dadurch gekennzeichnet**, dass das Clusteringverfahren (108) als hierarchisches Verfahren, insbesondere ein divisives Clustern oder ein agglomeratives Verfahren ausgebildet ist.

5. Verfahren nach Anspruch 3, **dadurch gekennzeichnet**, dass das Clusteringverfahren (108) als dichtebasiertes Verfahren, insbesondere DBSCAN oder OPTICS, ausgebildet ist.

6. Verfahren nach Anspruch 3, **dadurch gekennzeichnet**, dass das Clusteringverfahren (108) als graphenbasiertes Verfahren, insbesondere als Spectral Clustering oder Louvain, ausgebildet ist.

7. Verfahren nach mindestens einem der Ansprüche 3 bis 6, **dadurch gekennzeichnet,** dass eine Kosinus-Ähnlichkeit, eine Termfrequenz und/oder eine inverse Dokumentenfrequenz als Schwellenwert bei der Clusterbildung verwendet werden.

8. Vorrichtung zur Vorauswahl und Ermittlung ähnlicher Dokumente aus einer Menge an Dokumenten (101), wobei die Dokumente (101) tokenisierte Zeichenketten aufweisen, mit

einem Mittel zur Ausführung eines Indexierungsverfahrens (102) zur Berechnung eines inversen Indexes für mindestens eine Teilmenge der Dokumente (101),

einem Mittel zur Berechnung von Word Embeddings (105) für die mindestens eine Teilmenge der Dokumente (101),

einem Mittel zur Berechnung von Document Embeddings (107), wobei für die mindestens eine Teilmenge der Dokumente (101) für jedes dieser Dokumente (101) jeweils ein Document Embedding (107) berechenbar ist, indem für jedes Dokument (101) die Word Embeddings (105) aller Zeichenketten, insbesondere Worte des Dokuments (101), addiert und mit der Anzahl der Zeichenketten, insbesondere Worten, normiert werden (106), wobei vorher, nachfolgend oder parallel mit den berechneten Word Embeddings (105) mithilfe eines Mittels zum Clustering SimSet-Gruppen (109) von ähnlichen Zeichenketten berechenbar sind,

einem Mittel zur Bestimmung eines Query Embeddings (205) und

einem Vergleichsmittel des Query Embeddings (205) mit den Document Embeddings (107) unter Verwendung der mit dem Clusteringverfahren gebildeten SimSet-Gruppen (109) zur mengenmäßigen Einschränkung der Anzahl der zu vergleichenden Document Embeddings (107), um automatisch ein Ranking der Ähnlichkeit der Dokumente (101) zu ermitteln und diese anzuzeigen und / oder abzuspeichern.

Fig. 1

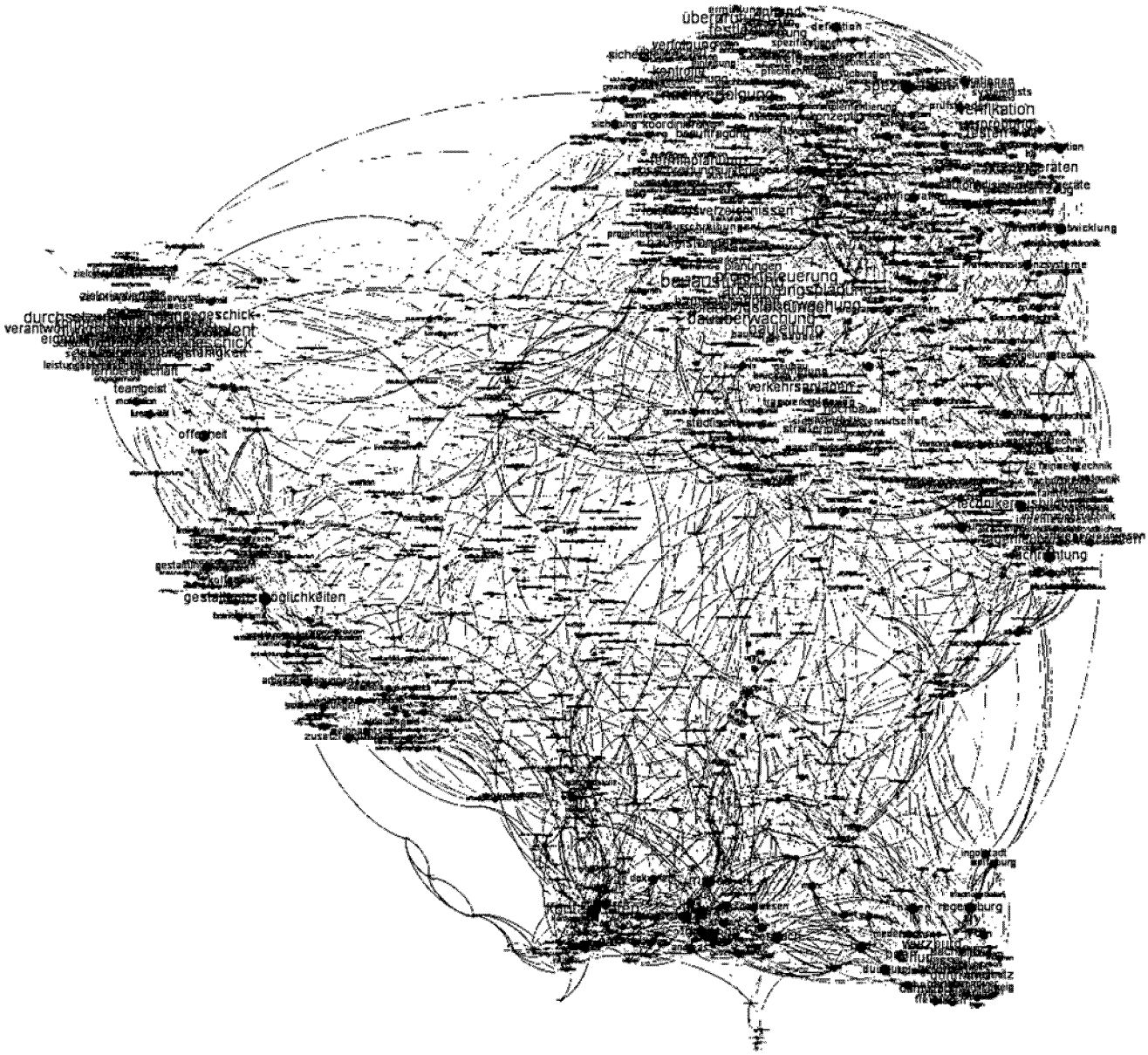

überprüfung
verifikation
planungen
verkehrsanlagen
offenheit
teamgeist
gestaltungsmöglichkeiten
regelung
bauüberwachung
bauplanung
verkehrsplanung
bauaufsicht
regelungen
verzeichnissen
durchsetzungsvermögen
verantwortungsbewusstsein
eigeninitiative
leistungsbereitschaft
selbstständigkeit
zuverlässigkeit
fachkompetenz
teamfähigkeit
kommunikationsfähigkeit
organisationsgeschick
verhandlungsgeschick
durchsetzungsvermögen
verantwortungsbewusstsein
eigeninitiative
leistungsbereitschaft
selbstständigkeit
zuverlässigkeit
fachkompetenz
teamfähigkeit
kommunikationsfähigkeit
organisationsgeschick
verhandlungsgeschick

Fig. 2

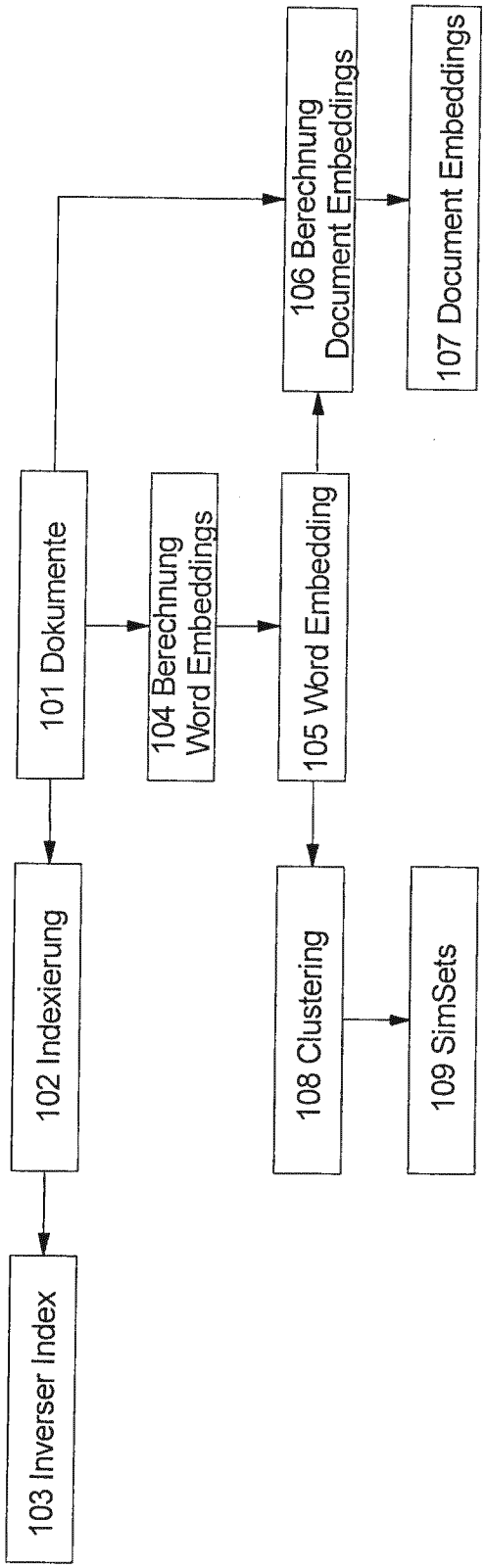


Fig. 2A

```
Dokument: 'Ein Polizeibeamter ist ein Beamter'

Word Embeddings
ein:              [1.018, 1.083, -3.313, 1.511, -0.017, 2.057, 1.152, 0.306, -3.382, -2.290]
Polizeibeamter: [0.561, -0.833, -0.017, -0.213, 0.106, 0.508, 0.432, -0.619, 0.558, -1.099]
ist:              [-0.785, 2.756, -3.063, 2.551, 2.918, -2.150, -4.965, -0.373, 1.646, -4.197]
Beamter:          [0.925, 0.063, 0.319, -0.203, -0.345, 0.546, 0.731, -0.185, 1.015, -0.881]

Document Embedding
'Ein Polizeibeamter ist ein Beamter'
                  [0.430, 0.767, -1.519, 0.912, 0.665, 0.241, -0.663, -0.218, -0.041, -2.117]
```

Fig. 3

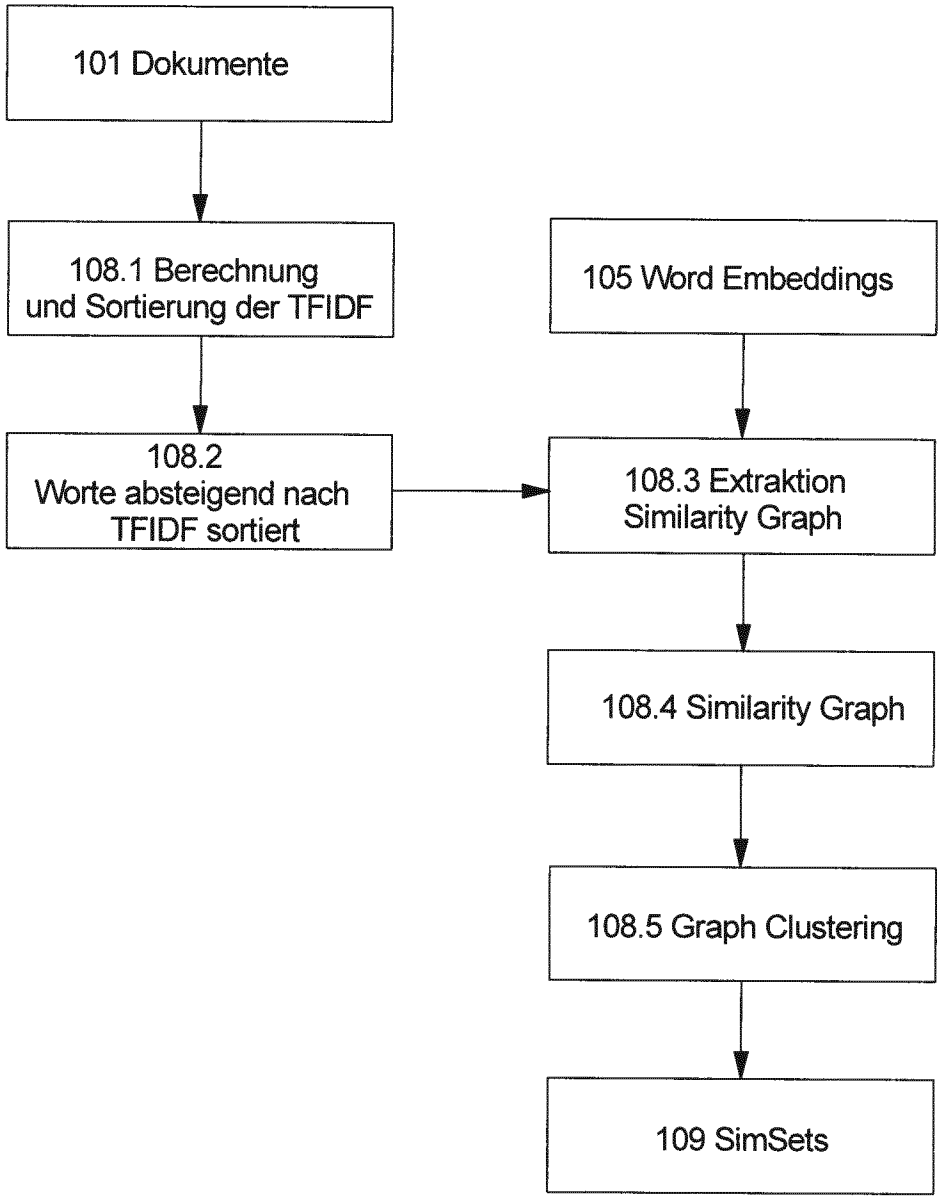

101 Dokumente
108.1 Berechnung und Sortierung der TFIDF
105 Word Embeddings
108.2 Worte absteigend nach TFIDF sortiert
108.3 Extraktion Similarity Graph
108.4 Similarity Graph
108.5 Graph Clustering
109 SimSets

Fig. 4

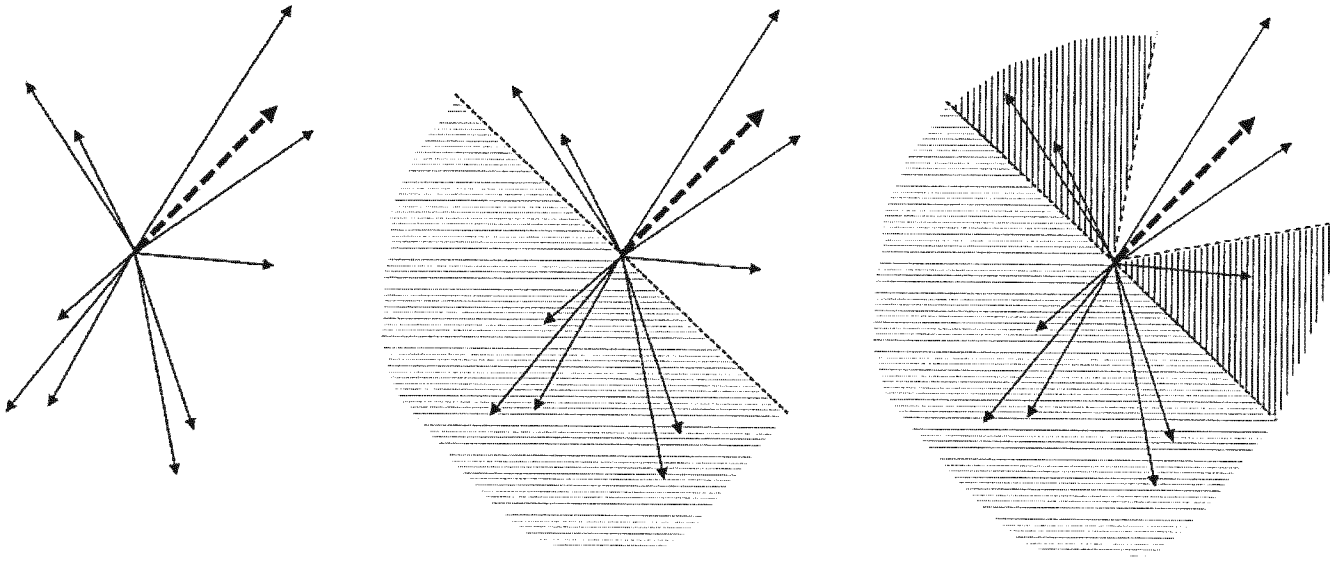
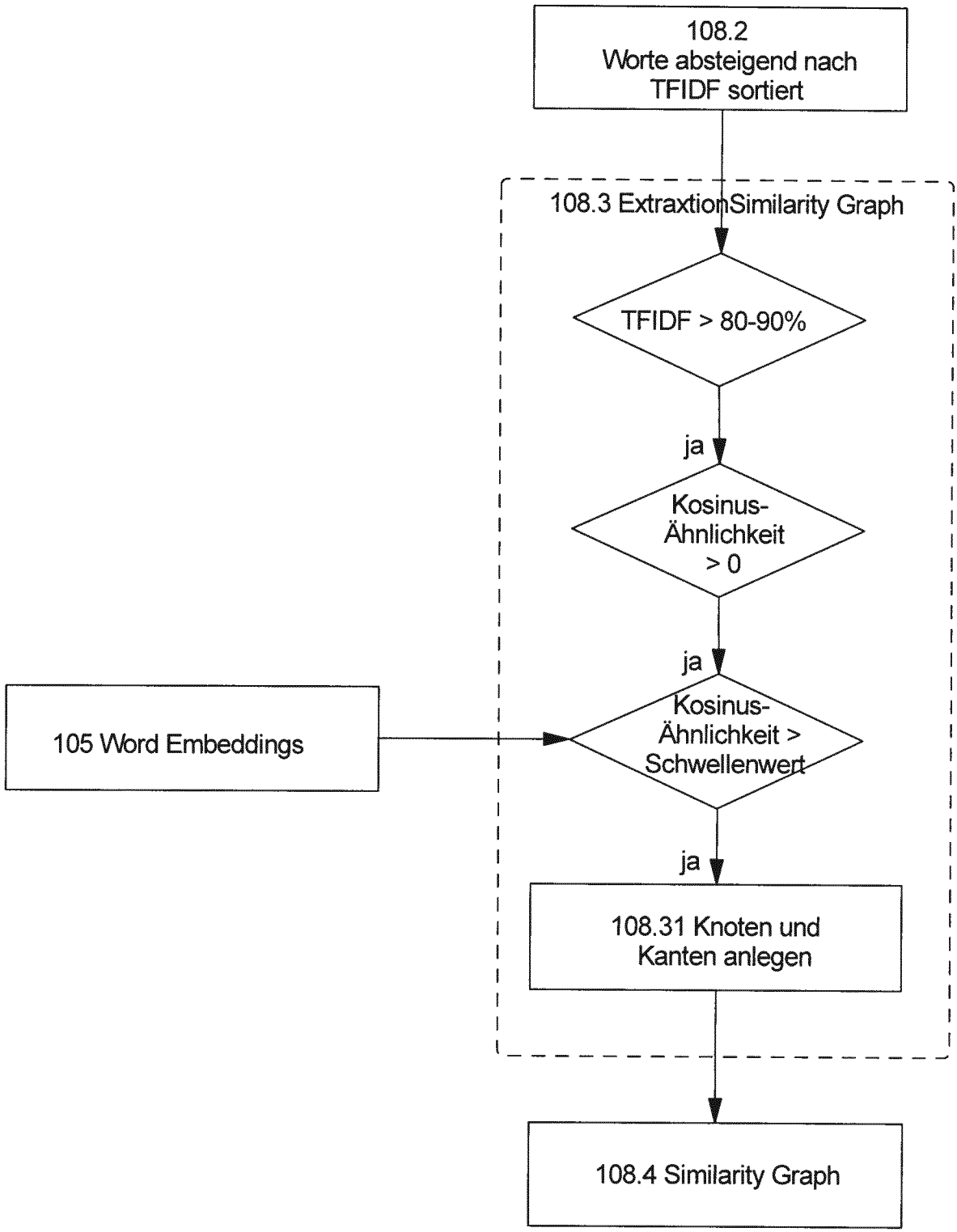


Fig. 4D

	ein	Polizeibeamter	ist	Beamter
ein	1	0.1222		
Polizeibeamter	0.1222	1	-0.0233	0.7533
ist	0.1778	-0.0233	1	-0.1368
Beamter	0.0122	0.7533	-0.1368	1

SimSet-Gruppe {Polizeibeamter, Beamter}

Fig. 5


108.2
Worte absteigend nach TFIDF sortiert
108.3 ExtraxtionSimilarity Graph
TFIDF > 80-90%
ja
Kosinus-Ähnlichkeit > 0
ja
105 Word Embeddings
Kosinus-Ähnlichkeit > Schwellenwert
ja
108.31 Knoten und Kanten anlegen
108.4 Similarity Graph

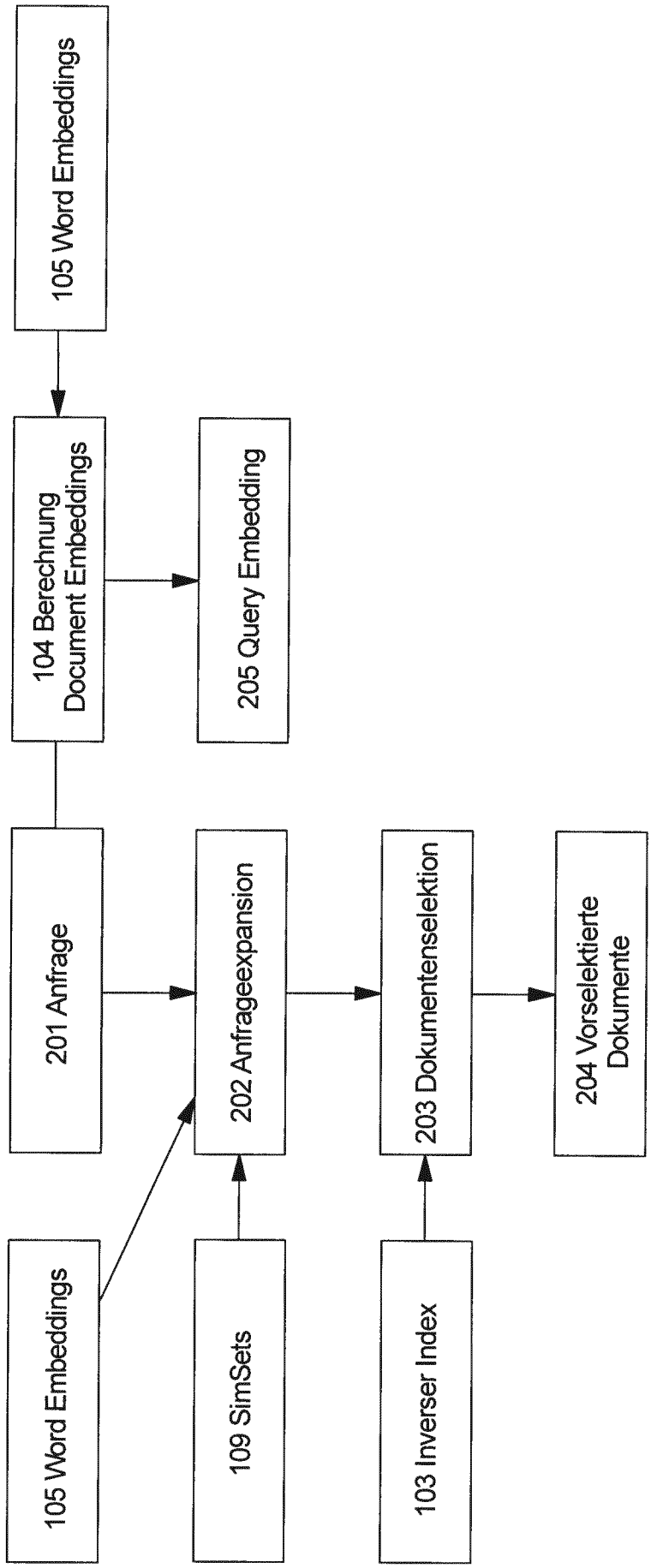

105 Word Embeddings
104 Berechnung Document Embeddings
205 Query Embedding
201 Anfrage
202 Anfrageexpansion
203 Dokumentenselektion
204 Vorselektierte Dokumente
105 Word Embeddings
109 SimSets
103 Inverser Index

Fig. 6

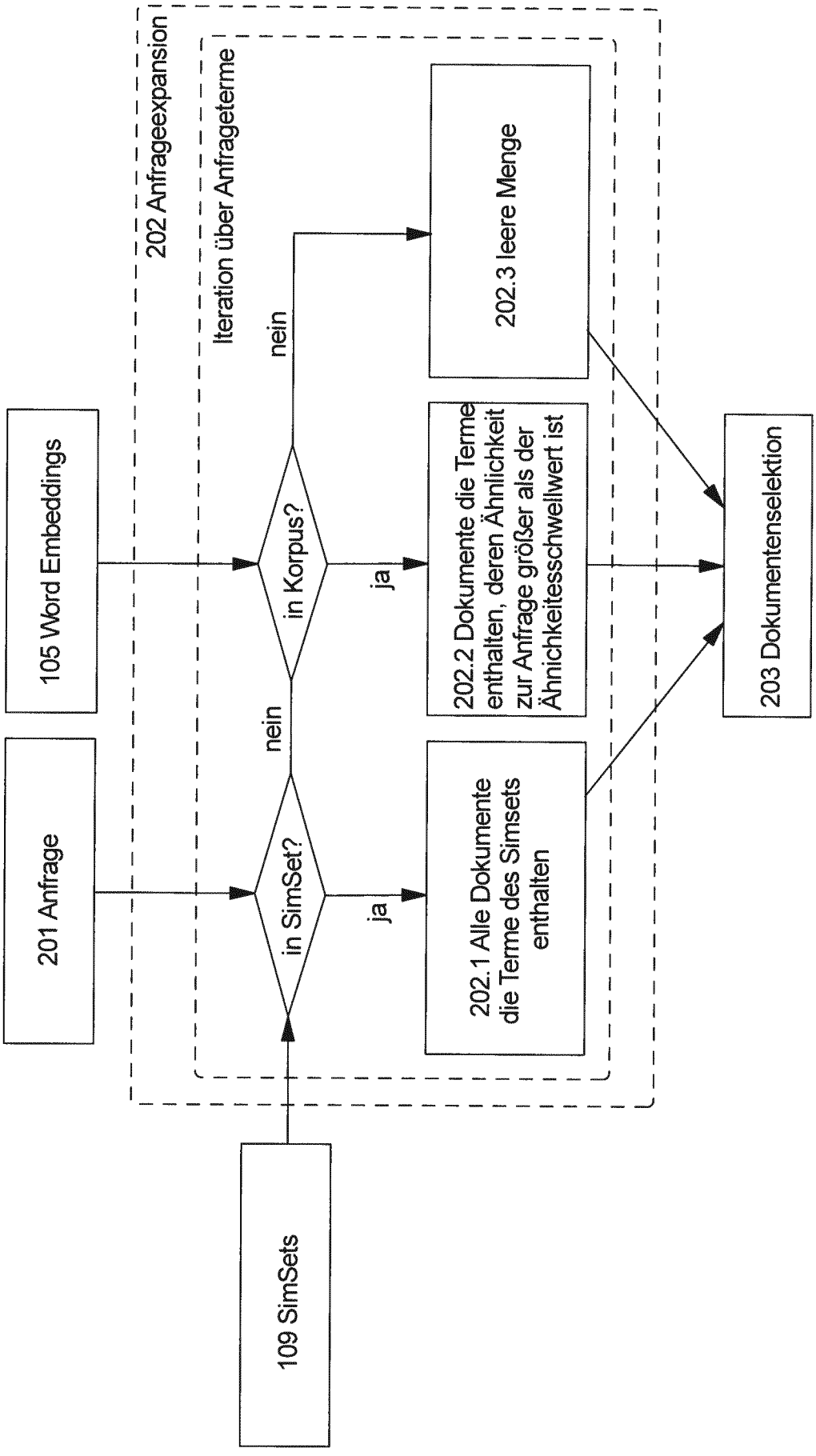
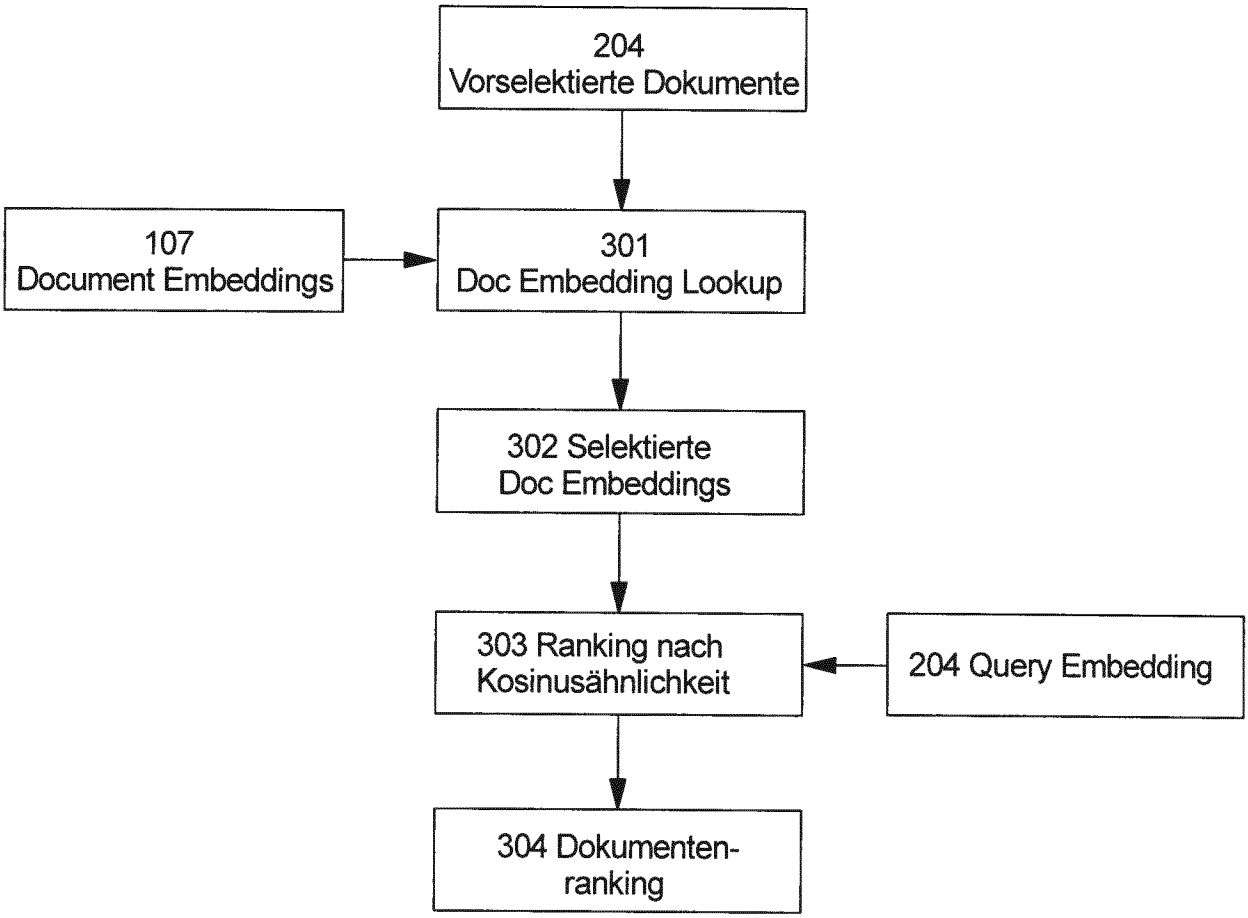

202 Anfrageexpansion
Iteration über Anfrageterme
105 Word Embeddings
201 Anfrage
109 SimSets
in Korpus?
in SimSet?
nein
ja
nein
ja
202.3 leere Menge
202.2 Dokumente die Terme enthalten, deren Ähnlichkeit zur Anfrage größer als der Ähnichkeitesschwellwert ist
202.1 Alle Dokumente die Terme des Simsets enthalten
203 Dokumentenselektion

Fig. 7

Fig. 8


204
Vorselektierte Dokumente
107
Document Embeddings
301
Doc Embedding Lookup
302 Selektierte
Doc Embeddings
303 Ranking nach
Kosinusähnlichkeit
204 Query Embedding
304 Dokumenten-
ranking

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2020/073304

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/33(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 108491462 A (UNIV KUNMING SCIENCE & TECH) 04 September 2018 (2018-09-04) paragraphs [0004], [0012] - [0016], [0019] - [0024]; claims	1-8
A	BHASKAR MITRA ET AL. "Neural Models for Information Retrieval" *ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853,* 03 May 2017 (2017-05-03), XP080945924 page 12; figures 5c, 5d page 23, paragraph 2 - paragraph 3 page 25, paragraph 3 - page 27, paragraph 3	1-8

☐ Further documents are listed in the continuation of Box C.

☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 November 2020	**03 December 2020**

Name and mailing address of the ISA/EP	Authorized officer
European Patent Office **p.b. 5818, Patentlaan 2, 2280 HV Rijswijk** **Netherlands** Telephone No. **(+31-70)340-2040** Facsimile No. **(+31-70)340-3016**	**Fournier, Christophe** Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/EP2020/073304

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108491462 A	04 September 2018	NONE	

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2020/073304

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
INV. G06F16/33
ADD.

Nach der Internationalen Patentklassifikation (IPC) oder nach der nationalen Klassifikation und der IPC

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
G06F

Recherchierte, aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	CN 108 491 462 A (UNIV KUNMING SCIENCE & TECH) 4. September 2018 (2018-09-04) Absätze [0004], [0012] - [0016], [0019] - [0024]; Ansprüche -----	1-8
A	BHASKAR MITRA ET AL: "Neural Models for Information Retrieval", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 3. Mai 2017 (2017-05-03), XP080945924, Seite 12; Abbildungen 5c, 5d Seite 23, Absatz 2 - Absatz 3 Seite 25, Absatz 3 - Seite 27, Absatz 3 -----	1-8

☐ Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen ☒ Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" frühere Anmeldung oder Patent, die bzw. das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
24. November 2020	03/12/2020

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Fournier, Christophe

1

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP2020/073304

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
CN 108491462 A	04-09-2018	KEINE	