Digital Libraries Zusammenfassung

February 10, 2019

Kapitel 1 - Einführung

Bibliothek eine Einrichtung, die unter archivarischen, ökonomischen und synoptischen Gesichtspunkten publizierte Information für die Benutzer sammelt, ordnet und verfügbar macht

Organisationsprinzipien in Bibliotheken

- Sortierung der Bücher nach Signatur (eindeutiger Schlüssel, der Fachgebiet codiert)
- Inhaltliche Erschließung der vorhandenen Dokumente
  - **Formalerschließung**: Autor, Titel, Verlag, etc.
  - **Sacherschließung**: inhaltliche Beschreibung, z.B. Schlagwörter, Zusammenfassung oder Klassifikation
- Einfaches Auffinden durch Suche in Index nach vordefinierten Schlagwörtern

Mögliche Dienste einer DL

- Suche einer bestimmten Publikation
- Suche nach "passenden" Publikationen auf Basis von
- Exakten oder inexakten Metadaten
- Inhaltlichen Beschreibungen
- Suche nach ähnlichen Publikationen
- Suche nach "guten" Publikationen zu einem Thema
- Suche nach "passenden" Publikationen zu einem Benutzer
- Alles im lokalen Bestand oder bibliotheksübergreifend
- Erweiterte Metadaten (z.B. eingehende und ausgehende Zitate, Lesefrequenz)
- Automatische Zusammenfassung und Aufbereitung von Information aus Publikationen
  - Beschreibung einer Publikation/einer Reihe mit Schlagwörtern oder als textuelle Zusammenfassung
  - Beantwortung einer konkreten Frage
  - Zusammenstellung von Argumente für und gegen eine These

Mögliche weitere Aufgaben einer DL

- **Erschließung** von Dokumentbeständen
- **Digitalisierung** bestehender Dokumentbestände (aber auch von anderen Artefakten, insbesondere in den Geisteswissenschaften)
- **Langzeitarchivierung** von Dokumentbeständen

Unterschiede der DL

- **Abdeckung** der Publikationen
  - Fokus auf einen Verlag
  - Fokus auf ?wichtige? Publikationen
  - Fokus auf online verfügbare
- **Zugriffsrechte**
- **Volltext** vs. **Verweis** zur Online-Publikation
- Mächtigkeit des **Suchinterfaces**
- **Aufbereitung** der Metadaten, Mehrwertdienste
  - Keyphrases
  - Zitate ein- und ausgehend
  - bibliometrische Maße
  - Einfluss von/auf Autoren und Publikationen

Kapitel 2 - Wissenschaftliches Publizieren

Ich lasse hier ein gutes Stück der VL weg weil es scheiße langweilig ist

Publikationshierarchie in der Informatik

- Workshops:
  - Publikation erster Ideen und Ergebnisse, 6 Seiten, informell
  - Oft zu Themen, die gerade aktuell werden
  - Oft Schwerpunkt bei Diskussion statt Präsentation
- Konferenzen:
  - Publikation aktueller Forschungsergebnisse, 12 Seiten
  - Strenge Anforderungen an Neuheit und Qualität
  - Oft thematisch relativ breit, viele Teilnehmer
  - Zusätzlich Demos, Poster, Panels, Tutorials, ...
  - Pausen zur Interaktion, ?community-building?
- Zeitschriften:
  - Oft archivierender Charakter
  - Publikation erweiterter Fassungen von Konferenzbeiträgen, Surveys, ... 10-40 Seiten
  - Möglichkeit zur Revision auf Basis von Gutachten

Qualitätssicherung: Peer Review Fachkompetente **Gutachter** erstellen **Gutachten** über Einreichungen

- Auswahl der Gutachter durch Editor, unabhängig von Autoren
- Empfehlung zu Annahme, Überarbeitung oder Ablehnung
- Idealerweise aussagekräftige inhaltliche Kommentare, Wünsche, Anregungen
- Typisch Gutachter anonym gegenüber Autoren
- Oft auch Autoren anonym gegenüber Gutachtern (double blind)
- Normalerweise 2-3 Gutachten pro Beitrag
- Begutachtungszeiten:
  - Konferenzen: typisch 6-16 Papiere à 12 Seiten in 4 Wochen
  - Zeitschriften: typisch 4-6 Wochen pro Beitrag à 30 Seiten
- Begutachtung in der Regel kostenlos (hoffentlich nicht umsonst)

Typischer Ablauf für Konferenzen

1. Call for Papers durch Organisatoren
2. Einreichung von fertig formatierten Beiträgen durch Autoren
3. Begutachtung durch Wissenschaftler, gesteuert durch Organisatoren
4. Zusammenstellung des Tagungsbands durch Organisatoren
5. Veröffentlichung:
   - Selbstverlag, online, etc.
   - Durch Fachgesellschaften (ACM, IEEE, VLDB, GI, etc.)*
   - Durch wissenschaftliche Verlage*

   *Zugriff oft nur gegen $$$
6. Zusammenstellung des Programms durch Organisatoren
7. Registrierung für Konferenz durch Autoren ($$$)
8. Vortrag etc. bei Konferenz durch Autoren

(Ablauf für Workshops analog, für Zeitschriften bis Schritt 5)

Wichtige Verlage? To learn or not to learn, that is the question.

Probleme des traditionellen Systems:
Anzahl der wissenschaftlichen Arbeiten wächst exponentiell, je nach Gebiet verdoppelt sich die Anzahl der Publikationen alle 10-15 Jahre
Gründe:

- Weltweit mehr Forscher (Asien, Afrika) und mehr Projektmittel (DFG, BMBF, EU, NSF, DARPA, ...)
- Beurteilung hängt praktisch immer von Publikationen ab (Einstellung, Verdauerung, Beförderung, Projekte)
- Oft zählt Anzahl, nicht Qualität

Elektronische Zeitschriften ? Lösung?

- Billiger, da kein Druck und keine Lieferung
- Keine Begrenzung der Seitenzahl
- Schnelle Verbreitung
- Aber: Begutachtung bleibt Engpass

Identifiers

- Digital Object Identifiers
- Personen-Identifier

Mögliche Features zur Autordisambiguierung
Ähnlichkeit von zwei (Mengen von) Publikationen mit ähnlichen Autornamenkann abhängen von

- Ähnlichkeit der **Autornamen**: Wei Wang vs. Wang Wei vs. W. Wang vs. Wei X. Wang
- Ähnlichkeit der A**utor-IDs**, wenn vorhanden
- Ähnlichkeit der **Publikationstitel**: Issues with author disambiguation? vs. ?Methods for author disambiguation? vs. ?Methods for Virus Recognition?
- Ähnlichkeit der **Publikationsorte**: gleiches Journal, thematisch ähnliche Venues (aber wie bestimmt man das?)
- Ähnlichkeit der **Publikationszeiten**: Publikationen nur von 1990-2000 vs. Publikation im Jahr 2016
- Ähnlichkeit der **Affiliations**: MPI Informatik vs. Max-Planck-Institute for Computer Science
- Ähnlichkeit der **Co-Autoren**: Annahme: Dieselbe Menge von Co-Autoren publiziert immer mit dem gleichen Autor eines Namens (aber: Co-Autoren können selbst nicht eindeutig sein)

Bewertung von Autoren

- Anzahl von Publikationen
- Anzahl von Zitaten
- **Hirsch-Index (h-index)**: größte Zahl h, so dass mindestens h Publikationen des Autors mindestens h mal zitiert wurden
- **Hirsch-Index mit Zeitconstraint**, z.B. h5: wie Hirsch-Index, aber zeitliche Beschränkung der betrachteten Publikationen (z.B. bei h5 auf die letzten 5 Jahre)
- Werte hängen stark von Datenbasis ab, z.B. für Ralf Schenkel:
  - Google Scholar: h-index=32
  - CiteSeer: h-index=13
- Hirsch-Index analog für Journals definierbar
- i10-Index: Publikationen, die mindestens 10mal zitiert wurden
- Weitere Varianten, um potentielle Probleme des h-Index zu umgehen: g-index, e-index, c-index, s-index, Normalisierung der Coautor-Zahl, ...

Kapitel 3 - Einführung in Information Retrieval

Information Retrieval
Information Retrieval beschäftigt sich mit der Repräsentation, Speicherung und Organisation von Informationen und dem Zugriff auf Informationen. Im Regelfall bestehen die gespeicherten Informationen aus Texten

Herausforderungen an IR-Systeme

- Speicherung und effizienter Zugriff auf riesige Datenmengen
- effiziente und effektive Suche
- komplexe Suchanfragen (Queries)
- Bewertung und Vergleich von Anfragen und Suchergebnissen
- (visuelle) Aufbereitung von Suchergebnissen, Navigation, Benutzerführung
- einfaches Textverstehen
- automatische Textsynthese

Begriffsbildung
Suche in Dokumentkollektionen kann auf verschiedenen Abstraktionsstufen stattfinden Vergleiche hierzu die Ebenen der Semiotik:

- **Syntax** - Ein Dokument wird als Folge von Symbolen betrachtet. Beispiele:
  - Zeichenkette in Texten
  - Histogramm oder Kontur in Bildern
- **Semantik** - Ein Dokument wird auf der Ebene seiner Bedeutung betrachtet. Semantik hat immer etwas mit Interpretation zu tun.
- **Pragmatik** - Ein Dokument wird hinsichtlich seines Verwendungszusammenhangs betrachtet. Beispiele:
  - Enthält ein Dokument eine Lösung meines Problems?
  - Was ist die Absicht des Autors des Textes?

Daten, Information, Wissen

- Daten -¿ syntaktische Ebene
- Informationen -¿ semantische Ebene
- Wissen -¿ pragmatische Ebene

Relevanz ist abhängig

- vom aktuellen Wissen des Benutzers
- vom aktuellen Problem des Benutzers
- von der subjektiven Erwartung des Benutzers

Precison

- erfordert nur die Analyse des Retrieval-Resultats
- kann vom Endbenutzer eingeschätzt werden
- ist ein subjektives Maß

Recall

- erfordert die Analyse der gesamten Dokumentenbasis
- ist dem Endbenutzer nicht zugänglich
- ist ein subjektives Maß

Systemorientierte Prozess-Sicht

- Crawl - strategies for crawl schedule and priority queue for crawl frontier
- Extract & clean - handle dynamic pages, detect duplicates, detect spam
- Index - build and analyze web graph, index all tokens or word stems
- Match - fast top-k queries, query logging, auto-completion
- Rank - scoring function over many data and context criteria
- Present - GUI, user guidance, personalization

Methoden und Techniken des IR

- Modellierung von Dokumenten und Text - DL+IR
- (approximatives) String-Matching - IIR
- Textvorverarbeitung und Indexing - DL+IR IIR
- Benutzerinteraktion und Visualisierung - DL+IR
- Benutzermodellierung und Personalisierung - DL+IR
- Relevanzanalyse - DL+IR
- verteilte und Peer-to-Peer Softwaretechnik - IIR
- Kategorisierung, Klassifikation - Data Mining
- Natural Lange Processing (NLP) - (Computer-) Linguistik DL+IR
- Web-Technologie - DL+IR
- Datenstrukturen, effiziente Symbolverarbeitung - IIR

Kapitel 4 - Boolesches Retrieval

Anfragen und einfache Datenstrukturen

Dokumente sind die Einheiten des Datenbestandes bezeichnet, die durch das jeweilige Information Retrieval System bearbeitet werden

Dokumentkollektion oder Korpus - Die Grundmenge an Dokumenten, für die Information Retrieval durchgeführt wird

Terme oder Index-Terme - Im Information Retrieval diejenigen Einheiten der Dokumente, die Gegenstand der logischen Repräsentation sind

Informationsbedarf und Ad-hoc-Anfragen
Die Formulierung und Beantwortung von **Ad-Hoc-Anfragen** ist eine Standardaufgabe des Information Retrieval:

- Gesucht: Dokumente aus der Dokumentkollektion, die für eine Anfrage relevant? im Hinblick auf den jeweiligen Informationsbedarf sind. Relevanz durch denjenigen definiert, der Anfrage gestellt hat.
- Algorithmen zur Anfragebeantwortung sollen effizient (d.h. schnell ihre Ergebnisse liefern) und effektiv (d.h. möglichst genau die Menge der ?relevanten? Dokumente auffinden) sein.

Informationsbedarf:

- Sachverhalt, über den ein Nutzer etwas in Erfahrung bringen möchte.
- Nicht exakt definiert
- unterscheidet sich von einer Anfrage (die Benutzer an das Retrieval-System richtet, um Informationsbedarf zu formulieren).
- Oft Folge von Anfragen notwendig, um Informationsbedarf zu erfüllen

Boolesches Retrieval-Modell - ein Information-Retrieval-Modell der folgenden Art:

- Die logische Repräsentation betrachtet die Dokumente als Menge von Wörtern.
- Anfragen werden aus Index-Termen zusammen mit den Booleschen Operatoren AND, OR und NOT gebildet.

Obwohl ein linearer Scan (grepping) bei kleineren Datenvolumina sehr effizient durchführbar ist, gibt es Gründe, nach weiterführenden Methoden zu suchen:

1. Sehr große Datenbestände, z.B. das Web;
2. Komplexe Suchbedingungen, z.B. ?Brutus? und ?Caesar?, aber nicht ?Calpurnia?
3. Flexibleres Matching, z.B. Ähnlichkeit zu Suchbegriffen oder Nachbarschaft von Wörtern als Kriterium;
4. Ranking der Antwortmenge, um die beste oderrelevanteste Antwort zu erhalten

Term-Dokument Inzidenzmatrix - enthält eine Zeile für jeden betrachteten Term t und eine Spalte für jedes im Grundbestand vorkommende Dokument d

Tritt t in dem Dokument d auf, so enthält das Matrixelement (t, d) eine 1, sonst eine 0:

$$M(t,d) = \begin{cases} 1, & \text{Falls } t \text{ in } d \text{ vorkommt} \\ 0, & \text{sonst} \end{cases}$$

Die Inzidenzmatrix erlaubt verschiedene Sichtweisen:

- Jede Zeile $(t, \cdot)$ stellt einen Vektor dar, der angibt, in welchen Dokumenten der Term t vorkommt.
- Jede Spalte $(\cdot, d)$ bildet einen Vektor, der angibt, welche Terme in dem Dokument d auftreten.

Hoher Speicherbedarf: siehe 4-12

Invertierter Index - besteht aus einem Vokabular (Dictionary) und den Positionen (Postings). Das Vokabular enthält alle Index-Terme zu Dokumentenkollektion D. Die Positionen-Tabelle enthält zu jedem Term aus dem Vokabular alle Dokument-IDs und ggf. weitere Informationen, z.B. Positionen innerhalb von Dokumenten, an denen er auftritt. Die Positionsliste eines Terms heißt auch invertierte Liste des Terms.

Es gelten folgende weitere Vereinbarungen:

- Jedes Dokument $d \in D$ besitzt einen (auf D) eindeutigen Identifikator DocId, beispielsweise eine eindeutige Dokumentnummer oder seine URI im Web.
- Jedem Index-Term t wird seine Dokumenthäufigkeit oder Document Frequency zugeordnet, die angibt, in wie vielen Dokumenten t vorkommt. Im betrachteten Fall ist die Dokumenthäufigkeit eines Terms t gleich der Länge seines Positionsvektors.
- Die Positionen jedes Terms werden nach DocId sortiert.

Anfrage Operationen

- Konjunktion (AND)
- Disjunktion (OR)
- Negation (NOT)

Vorverarbeitung von Dokumenten und Indexierung

Indexierung Die klassischen Dokumentmodelle abstrahieren ein Dokument auf eine Menge von sogenannten Indextermen oder Deskriptoren. Idealerweise sollten Indexterme so gewählt sein, dass sie

1. den Inhalt der einzelnen Dokumente adäquat repräsentieren,
2. eine möglichst klare Abgrenzung der einzelnen Dokumente gewährleisten,
3. die Verknüpfung von thematisch ähnlichen Dokumenten ermöglichen.

Token - die Instanz einer begrenzten Zeichenreihe (Character-String), die in dem gegebenen Dokument auftritt und zu einer für die Weiterverarbeitung semantisch sinnvollen Einheit gruppiert ist. Ein Token kann in einem Dokument mehrfach auftreten.

Typ - die Klasse aller Token, die dieselbe Zeichenreihe enthalten

Term - ein (ggf. ?normalisierter?) Typ, der in das Vokabular aufgenommen werden kann. Die Normalisierung kann z.B. hinsichtlich Groß-/ Kleinschreibung, Morphologie (Wortart, Flexionsform etc.), Rechtschreibung erfolgen.

Problembereiche des Tokenisierung

- Satzzeichen
- Binde- bzw. Trennstriche
- Groß-/Kleinschreibung
- Ziffern/Zahlen
- Wortlänge
- zusammengesetzte Wörter
- Umlaute etc.
- Schreibfehler
- Sprache

Vorverarbeitung

Normalisierung - der Prozess der Kanonisierung von Token, damit irrelevante Abweichungen nicht ins Gewicht fallen. Terme sind also die ?Normalformen? von Token.- Es muss gelten: *Die Normalisierung muss für Dokumente und Anfragen auf die gleiche Weise erfolgen und zu den gleichen Normalformen führen.*

Reduktion auf Wortstämme und Lemmatisierung
Durch die Reduktion von Wörtern auf eine Grundform oder auf einen Wortstamm können Äquivalenzklassen gebildet werden. Dadurch lässt sich die Größe von Indexen und die Komplexität von Anfragen stark reduzieren.
Im Wesentlichen gibt es zwei Ansätze

- **Lemmatisierung:**
  - **Lemma** - Die Grundform eines Wortes, wie man sie beispielsweise in Lexika findet
  - **Lemmatisierung** - dieReduktion von Wörtern auf ihre Grundform nach linguistisch gültigen Regeln
- **Stemming** - eine heuristische Methode zur Reduktion von Wörtern auf einen Wortstamm

Im Gegensatz zu Lemmatisierung wird Stemming von Linguisten nicht als gültiges Verfahren akzeptiert

- Die zugrunde liegende Methode folgt keinen linguistisch abgesicherten Regeln, sondern ist rein heuristisch begründet.
- Stemming ist sprachabhängig.
- Stemming mischt beugungs- und ableitungs-induzierte Reduktion

TODO: 4-46 to 4-70

Weitere wichtige Retrievaloperatoren

Phrasenanfragen oder Phrase Queries suchen nach Auftreten von Phrasen in Dokumenten. Separatoren und Stoppwörter werden dabei (oft) nicht betrachtet

Wortpaarindexe indexieren jedes aufeinanderfolgende Paar von Termen in einem Dokument als Phrase

Probleme mit Wortpaarindexen

- Falsch-positiv e Ergebnisse, die eine Filterung der Ergebnisse erforderlich machen
- Index kann sehr groß werden, da das Vokabular sehr groß werden kann

Positionsindex - besteht wie ein invertierter Index aus einem Vokabular und einer Positionsliste. Er speichert dabei zusätzlich für jeden Term t aus dem Vokabular seine Positionen für jedes Dokument, in dem er auftritt

Proximity-Queries (Nachbarschaftsanfragen) - stellen eine verallgemeinerte Form der Phrase Queries dar. Bei Proximity Queries wird nicht die genaue Wortsequenz gesucht, sondern Textstellen, in denen die angegebenen Einzelwörter einen bestimmten Maximalabstand nicht überschreiten

Methodische Ansätze der Rechtschreibkorrektur

Korrektur isolierter Terme

- Jeder einzelne Anfrageterm wird separat behandelt.
- Diese Methode kann keine fehlerhafte Anfrage, die aus korrekten Termen besteht, erkennen
- Grundannahmen
  - Es existiert eine Liste der korrekten Wörter bzw. Terme
  - Es gibt eine Methode zur Berechnung der Distanz zwischen einem korrekten und einem fehlerhaften Wort (**Edit-Distanz**)

Kontext-sensitive Korrektur
Arten:

- **Hit-basierte** - untersucht mögliche Ersetzungen für die einzelnen Anfrageterme durch Terme mit geringer Edit-Distanz und zählt die Anzahl der mit der so modifizierten Anfrage gefundenen Dokumente (**sehr ineffizient**)
- **Spelling Correction mit Wortsequenzen** - Wir generieren wie bei der hitbasierten Methode alle Kombinationen von Alternativen für die Terme der Anfrage und wählen die Sequenz mit der höchsten geschätzten Wahrscheinlichkeit
- **Phonetische Korrektur** - die Korrektur von Fehlern, die aufgrund des gleichen Klangs zweier Schreibweisen entstehen

SOUNDEX-Algorithmen - Korrekturalgorithmen, die auf dem sog. Phonetic Hashing basieren

Schwachpunkte des Booleschen IR-Modells

- Boolesche Anfragen werden schnell recht komplex.
- Die Retrieval-Strategie basiert auf einer binären Entscheidung, lässt also kein Ranking zu

Stemmer

Kapitel 5 - Retrievalmodelle

Das Vektorraum-Modell

Sichten auf ein Dokument

- **Layout-Sicht** - Darstellung eines Dokuments auf einem zweidimensionalen Medium.
- **Strukturelle bzw. logische Sicht** - Definiert den Aufbau bzw. die logische Struktur eines Dokuments(LaTeX)
- **Semantische Sicht** - Betrifft die Aussage eines Dokuments und ermöglicht dessen Interpretation.

Modelle - Sei D eine Menge von Dokumenten und Q eine Menge von Anfragen. Ein **Dokument-Modell** für D, Q ist ein Tupel $(D, Q, \rho_{\mathcal{R}})$, dessen Elemente wie folgt definiert sind:

- D ist die Menge der Repräsentationen der Dokumente $d \in D$. In $d \in D$ können Layout-, logische und semantische Sicht codiert sein.
- Q ist die Menge der formalisierten Anfragen.
- $\mathcal{R}$ ist ein Retrieval-Modell und formalisiert ein Prinzip, ein Paradigma oder eine linguistische Theorie.

Klassische Retrieval-Modelle
Die klassischen Retrieval-Modelle abstrahieren ein Dokument $d \in D$ zu einer unstrukturierten Menge von Indextermen, die sich quasi unmittelbar und automatisch aus d gewinnen lassen.
Die Dokumentrepräsentation d eines Dokumentes d besteht aus gewichteten Indextermen, die aus d stammen. Unterscheidung der klassischen Retrieval-Modelle:

- Art und Weise, wie sich Gewichte w_i für die Indexterme t_i berechnen.
- Art und Weise, wie formalisierte Anfragen q konstruierbar sind.
- Art und Weise, wie sich die Retrieval-Funktion $\rho_{\mathcal{R}}(q, d)$ berechnet.
- Art und Weise, wie die Menge relevanter Dokumente $R(q)$ konstruiert wird.

Boolesches Modell
Vorteile:

- Mächtigkeit: Prinzipiell kann mit einer Bool?schen Anfrage jede beliebige Teilmenge von Dokumenten aus einer Kollektion selektiert werden.
- einfache und genaue Implementierbarkeit

Nachteile

- die Schwarz-Weiß-Aufteilung in die Menge R (bzw. $D \backslash R$) der als relevant (bzw. nicht-relevant) geschätzten Dokumente ist zu streng
- keine Ordnung auf der Antwortmenge R hinsichtlich der geschätzten Relevanz
- die Größe der Antwortmenge ist schwierig zu kontrollieren
- keine Möglichkeit zur Gewichtung von Fragetermen
- umständliche Formulierung von Anfragen
- schlechte Retrieval-Qualität

mehr zum Vektorraummodell

Termhäufigkeit - die Anzahl der Auftreten eines Terms in einem Dokument, ist für die Relevanz eines Dokuments bzgl. einer Anfrage wichtig.
Die Termverteilung und -häufigkeit sollten deshalb bei der Indizierung aufgezeichnet werden.

Termgewichte

1. Die Term-Dokument-Inzidenzmatrix (vgl. Kapitel 2) des Booleschen Retrieval-Modells enthält für jeden Term t und jedes Dokument d einen Eintrag $m(t, d)$. Sein Wert ist 1, falls t in d auftritt und 0 sonst.
2. Wir ersetzen nun die binäre Angabe, ob ein Term in dem Dokument auftritt oder nicht, durch ein Gewicht $w(t, d)$, das im Zusammenhang mit der Anzahl der Auftreten des Terms in dem jeweiligen Dokument steht.
3. Term Frequency oder Term-Häufigkeit bezeichnet das Gewichtungsschema, in dem direkt die Anzahl $tf_{t,d}$ der Auftreten des Terms t in Dokument d als Gewicht verwendet wird.

Term-Dokument Häufigkeitsmatrix - enthält eine Zeile für jeden Term $t \in V$ aus dem Vokabular V und eine Spalte für jedes in der Dokumentkollektion D vorkommende Dokument $d \in D$. Es sei:

$$M(t, d) = \begin{cases} k, & \text{Falls } t \text{ in } d \text{ an } k \text{ Stellen vorkommt} \\ 0, & \text{sonst} \end{cases}$$

Bag-of-Words-Modell - Repräsentiert man Dokumente durch Term-Häufigkeitsvektoren, so wird die Wortordnung innerhalb eines Dokuments nicht berücksichtigt.

Absolute Termhäufigkeit
Ziel:

- Einsatz der Termhäufigkeit $tf_{t,d}$ zur Bestimmung von Scores für das Retrieval.
- Die absolute Anzahl von Auftreten eines Terms t in dem Dokument d ist als Maß nicht geeignet.
- Ein Dokument d_1, in dem t 10-mal auftritt, ist nicht unbedingt 10-mal so relevant, wie ein Dokument d_2, in dem t nur einmal auftritt.
- In bestimmten Kontexten sind ?typische? und deshalb häufig vorkommende Termini wie Stoppwörter zu behandeln.

Eine Möglichkeit, den Einfluss der absoluten Termhäufigkeit etwas abzumildern, ist der Einsatz eines logarithmischen Häufigkeitsmaßes als Termgewicht im Dokumentvektor

Dokumenthäufigkeit - df_t die Anzahl der Dokumente $d \in D$, in denen t auftritt.

Bei der Gewichtung mehrerer Terme in einer Anfrage spielt es eine Rolle, wie häufig jeder Term in der Dokumentkollektion vorkommt:

- "Seltene" Terme werden als signifikanter, genauer: trennschärfer, angesehen als "häufige".

  Ein Dokument, das einen seltenen Term enthält, ist für diesen Term mit großer Wahrscheinlichkeit relevant. Seltene Terme sollten also ein hohes Termgewicht erhalten.
- Ein Dokument, das einen häufigen Term enthält, ist sicherlich relevanter für diesen Term als ein Dokument, das den Term nicht enthält.

  Andererseits ist der häufige Term weniger trennscharf. Häufige Terme sollten daher berücksichtigt werden, aber ein geringeres Termgewicht bekommen als seltene Terme.

Inverse Dokumenthäufigkeit - Für einen Term t des Vokabulars ist die inverse Dokumenthäufigkeit (inverse document frequency) idf_t von t in der Kollektion D definiert durch

$$idf_t := \log \frac{N}{df_t}$$

Da $df_t \leq N$ gilt, ist $0 \leq idf_t$

- Hohe inverse Dokumenthäufigkeit idf_t bedeutet hohe Trennschärfe von t.
- Geringe inverse Dokumenthäufigkeit idf_t bedeutet geringe Trennschärfe von t.
- Der Einfluss der Dokumenthäufigkeit wird durch das logarithmische Maß gedämpft.

Kollektionshäufigkeit (collection frequency) - Anzahl der Auftreten von t in der gesamten Dokumentkollektion.

tf?idf-Gewichtung - ordnet einem Term t des Vokabulars das wie folgt definierte Gewicht $tf \cdot idf_{t,d}$ in dem Dokument $d \in D$ der Dokumentkollektion zu:

$$w_{t,d} := tf \cdot idf_{td} := tf_{t,d} \cdot idf_t := tf_{t,d} \cdot \log \frac{N}{df_t}$$

Falls t in d vorkommt, ist das Termgewicht $w_{t,d}$ also

- **am höchsten**, wenn t häufig in d, aber insgesamt in einer geringen Zahl von Dokumenten der Kollektion auftritt;
- **geringer**, wenn t seltener in d oder insgesamt in einer größeren Zahl von Dokumenten der Kollektion auftritt;
- **am geringsten**, wenn t in praktisch allen Dokumenten auftritt.

Cosinus-Ähnlichkeit von Dokumenten (cosine similarity) - $sim(d, d')$ zweier (Dokument-)Vektoren d und d' ist definiert durch

$$sim(d, d') := \frac{d^T d'}{||d||||d'||}$$

Ähnlichkeit von Anfrage und Dokument
Der Score eines Dokuments d für die Anfrage q wird als Cosinus-Ähnlichkeit $sim(d, q)$ der entsprechenden Vektoren berechnet:

$$sim(d, q) := \frac{d^T q}{||d||||q||}$$

Probabilistische Modelle

Wahrscheinlichkeiten im IR? Die grundlegende Idee ist es, Dokumente nach absteigender Relevanzwahrscheinlichkeit zu ordnen.

Probability Ranking Principle Sei d ein Dokument aus der Kollektion. Die binäre Zufallsvariable R beschreibe die Relevanz eines Dokuments: $R = 1$ bedeutet also relevant, $R = 0$ bedeutet nicht relevant.

Fragen für lernen

Welche Probleme gibt es bei der Digitalisierung traditioneller Bibliotheken?

- Scanqualität
  - Auflösung
  - Format
- Schrifterkennung
  - Buchstabenerkennung (Schwierig bei "exotische Schriftarten")
  - Worterkennung ()Schwer bei seltene Sprachen/ Spezialvokabular

Welche Publikationsmöglichkeiten gibt es? (Ablauf)

- Workshop
- Konferenzen
- Zeitschriften

Nennen sie ein paar mögliche Retrieval-Szenarien

- Empfehlungen
- Bildersuche
- Question answering
- Desktop-Suche
- Ortsabhängige Suche
- Adhoc-Suche

Erklären Sie Proximity-Queries

Welche Ansätze gibt es zur Rechtschreibkorrektur?

- Korrektur isolierter Terme
- Kontext-sensitive Korrektur
- Phonetische Korrektur

Wozu werden Sprachmodelle verwendet?

Ein Sprachmodell modelliert die in einer Sprache auftretenden Sätze statistisch. Es erlaubt, die Wahrscheinlichkeit zu bestimmen, mit der eine vorgegebene Wortfolge vorkommt

Erklären sie das Probability Ranking Principle