A SURVEY ON EXPERT FINDING TECHNIQUES

AGENDA

Technology
Arts Sciences
TH Köln

- Hauptkomponenten
- Ressourcenauswahl
- Modelle
  - GENERATIVE PROBABLISTIC MODELS
  - VOTING MODELS
  - NETWORK-BASED-MODELS
- TEST DATA COLLECTIONS

KOMPONENTEN

Technology
Arts Sciences
TH Köln

- Es gibt drei **Hauptkomponenten** die in einem Experten-Retrieval-Model berücksichtigt werden müssen:

Kandidat

Dokument

Topic

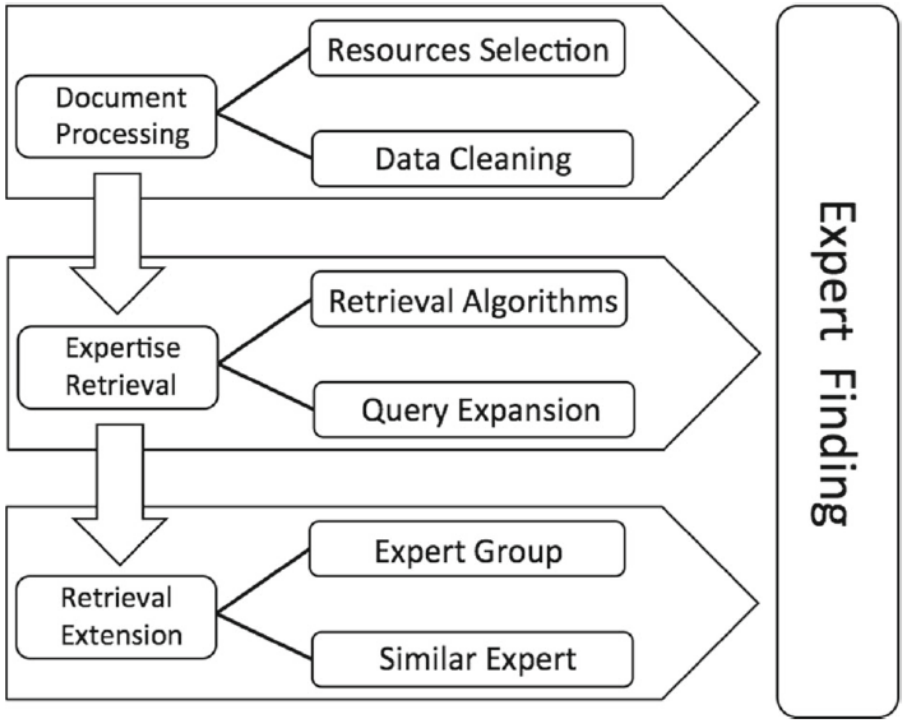
RESSOURCENAUSWAHL

Technology
Arts Sciences
TH Köln

- Ressourcenquellen für Daten in aktuellen Expertensuchsystemen
  - **Metadatenbanken** (Kontaktdaten, professionelle Fähigkeiten)
    - Daten müssen manuell eingetragen und geupdatet werden
    - Teuer/ Arbeitsaufwendig
  - **Dokument-Collections** (Publikationen, E-Mails, Webseiten)
    - Kandidaten können automatisch extrahiert werden
  - **Empfehlungsnetzwerke**
    - Experten können durch Verweise oder Empfehlungen gefunden werden

EXPERT FINDING

Technology
Arts Sciences
TH Köln



GENERATIVE PROBABLISTIC MODELS

Technology
Arts Sciences
TH Köln

- candidate generation models
  - Berechnen den Score eines Kandidaten anhand der Nennung in relevanten Dokumenten.
  - Zwei Stufen Model
  - Die Query wird nur in der Dokumentenauswahl berücksichtigt
  - $p(ca|q)$

- topic generation models:
  - Berechnen den Score eines Kandidaten anhand der Kandidaten Repräsentation
  - Kandidat-Term-Index oder Kandidat-Dokument Assoziation
  - $p(ca|q) = \frac{p(q|ca)p(ca)}{p(q)}$

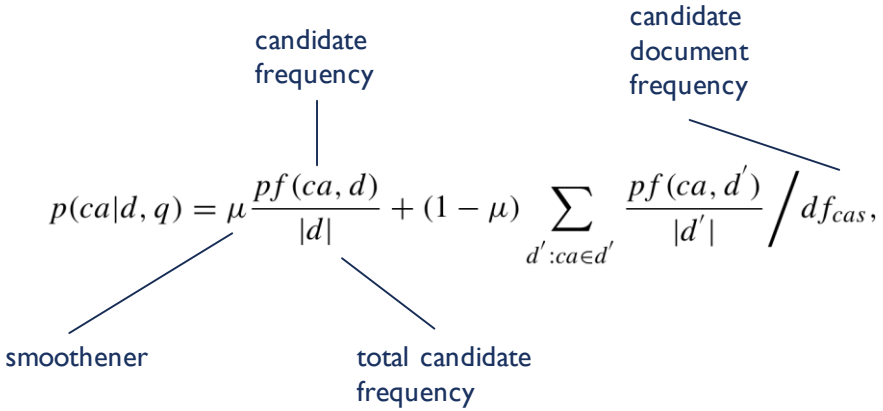
GENERATIVE PROBABLISTIC MODELS

Technology
Arts Sciences
TH Köln

Candidate Generation Models

Stufe I

- Welche Dokumente sind für eine Query relevant?
- Berechnung durch ein *Language Model* (Tf-Idf, bm25).



Stufe 2

- Wie oft wird ein in den Dokumenten genannt?
- Berechnung durch die *Experten Frequenz* (bereinigt und geglättet).

GENERATIVE PROBABLISTIC MODELS

Technology
Arts Sciences
TH Köln

Topic Generation Models

Candidate model

- Kandidaten werden durch einen Term Index aus den ihnen zugeordneten Dokumenten repräsentiert.
- Kandidaten Score wird berechnet aus der den **Query Termen** und ihrem **Kandidaten Index.**

- Benötigt Kandidaten Index

Document model

- Mit der Query werden die relevanten Dokumente ermittelt.
- Aus den relevanten Dokumenten werden die assoziierten Kandidaten ermittelt.

- Benötigt *document-candidate associations*

VOTING MODELS

Technology
Arts Sciences
TH Köln

- Voting Models aggregieren *votes* aus einem Dokumenten Ranking nach Kandidaten.
- Dies basiert auf der *document-candidate association*.
- Voting Models bevorzugen Kandidaten die mit vielen Dokumenten assoziiert sind.
  - Daher sollten die Dokumente gewichtet werden.
- Voting Models sind den *topic generation models* sehr ähnlich.

NETWORK-BASED-MODELS

Technology
Arts Sciences
TH Köln

- Netzwerkbasierte Modelle beziehen sich auf User-Netzwerke dies können unter anderem E-Mail Statistiken, Foreneinträge oder Ähnliches sein.

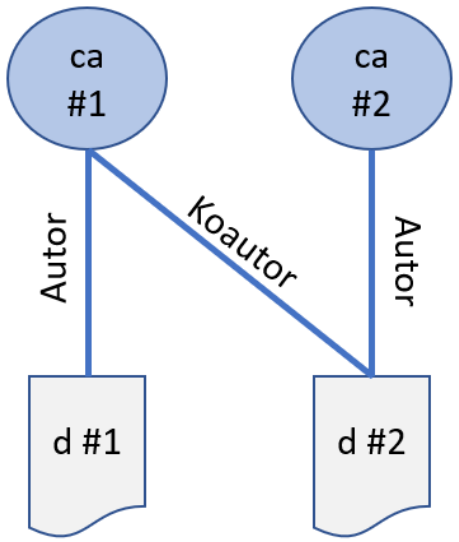
- Ein Beispiel:
  - Campbell u.a verwendete E-Mail-Statistiken als Repositorium von Kompetenznachweisen.
  - Die Idee dahinter ist: Dass Menschen intuitiv via E-Mail über ihren Fachbereich kommunizieren.
  - Je mehr E-Mails über ein bestimmtes Fachgebiet gesendet oder empfangen werden, desto höher ist die Wahrscheinlichkeit, dass die entsprechende Person Expertise in diesem aufweist.
  - Durch dieses Verfahren lassen sich genauere Vorhersagen als bei traditionellen Content-Based-algorithmen treffen, jedoch ist der Recall dieser Methode geringer durch die limitierten Ressourcen (E-mails)

NETWORK-BASED-MODELS

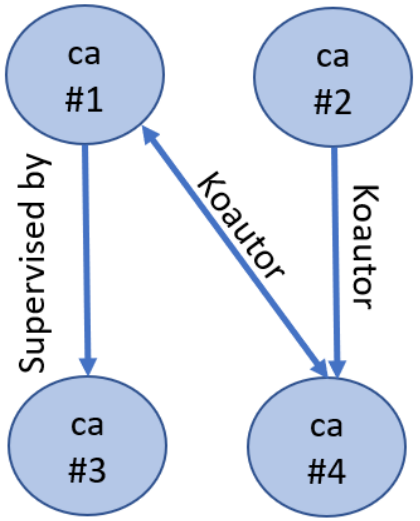
Technology
Arts Sciences
TH Köln

- Es gibt zwei Optionen zur Konstruktion von Grafen für Netzwerkbasierte Modelle:
  - 1) Dokumente und Kandidaten werden als Knoten betrachtet, die Assoziationen zwischen ihnen als Kanten
  - 2) Nur die Kandidaten werden als Knoten betrachtet und die Beziehungen zwischen ihnen als Kanten

1)



2)



ca = Kandidaten
d = Dokument

NETWORK-BASED-MODELS

Technology
Arts Sciences
TH Köln

- Algorithmen die in Network-based-models häufige Anwendung finden:
  - HITS
  - PageRank
- Experten Netzwerke weisen die selben Strukturen wie das Web auf.
  - Kandidaten oder Dokumente können als Webseite angesehen werden
  - Die **Kandidaten – Dokument Assoziation** und **die Kandidaten – Kandidaten Assoziation** können als Hyperlinks betrachtet werden

NETWORK-BASED-MODELS

Technology
Arts Sciences
TH Köln

- Vorteile:
  - Versteckte Informationen ("*hidden information*") können durch Netzwerk basierte Modelle extrahiert werden

TEST DATA COLLECTIONS

Technology
Arts Sciences
TH Köln

- UvT
  - Informationen von der Tilburg University.
  - Angestellte haben ihre Expertise selbst hinzugefügt. Das Themengebiet ist daher breit gefächert.
- DBLP
  - Informationen aus dem Bereich Informatik
  - Journal Artikel, Konferenzberichte
- CiteSeer
  - Informationen aus dem Bereich Informatik
  - Artikel Metadata, Hyperlinks zu Homepages