Multifaceted Feature Sets for Information Retrieval

Melanie Imhof

Zurich University of Applied Sciences

University of Neuchatel

Zurich University of Applied Sciences

zh aw

Université de Neuchâtel

une

Overview

Motivation – Application Examples

- Newsfeed – Sort news based on user-preferences

- Tweet search – Sort tweets based on relevance criteria

GeoCLEF 2008

- Collection:
  - The Glasgow Herald (1995)
  - The Los Angeles Times (1994)
  - Tagged with geographical coordinates of the locations in the news article
- Topics: 24 geographically challenging topics
  - "Nobel prize winners from Northern European countries"

Location Weighting

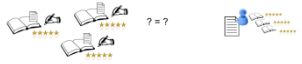
Idea: Characteristic locations should have a high weight. Adapt term weighting idea to locations.

$$s(D,Q) = \sum_{l \in D \cap Q_{area}} \frac{tf(l,D) \cdot (k_1 + 1)}{tf(l,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot \underbrace{log \frac{N - df(l) + 0.5}{df(l) + 0.5}}_{idf(l)}$$

- l = Location
- $df(l)$ = Number of documents that contain l
- $tf(l,D)$ = Location frequency (number of occurrences of l in D)

Social Book Search 2012

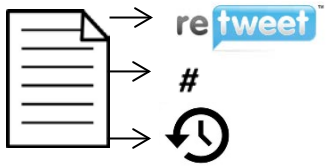
- Collection:
  - 2.8 Mio. Books from Amazon with professional Meta-Data (ISBN, Title, Publisher, Binding, #Pages, Authors, Listprice, Browsenodes,…)
  - User-generated Meta-Data from LibraryThing (Reviews, Ratings, tags,…)
  - User Profiles (Username, Library (title, author, rating), friends, groups)
- Topics: Generated from user information need on LibraryThing
  - Title, group, username, narrative, type (subject, author, edition), genre



Motivation – Application Examples

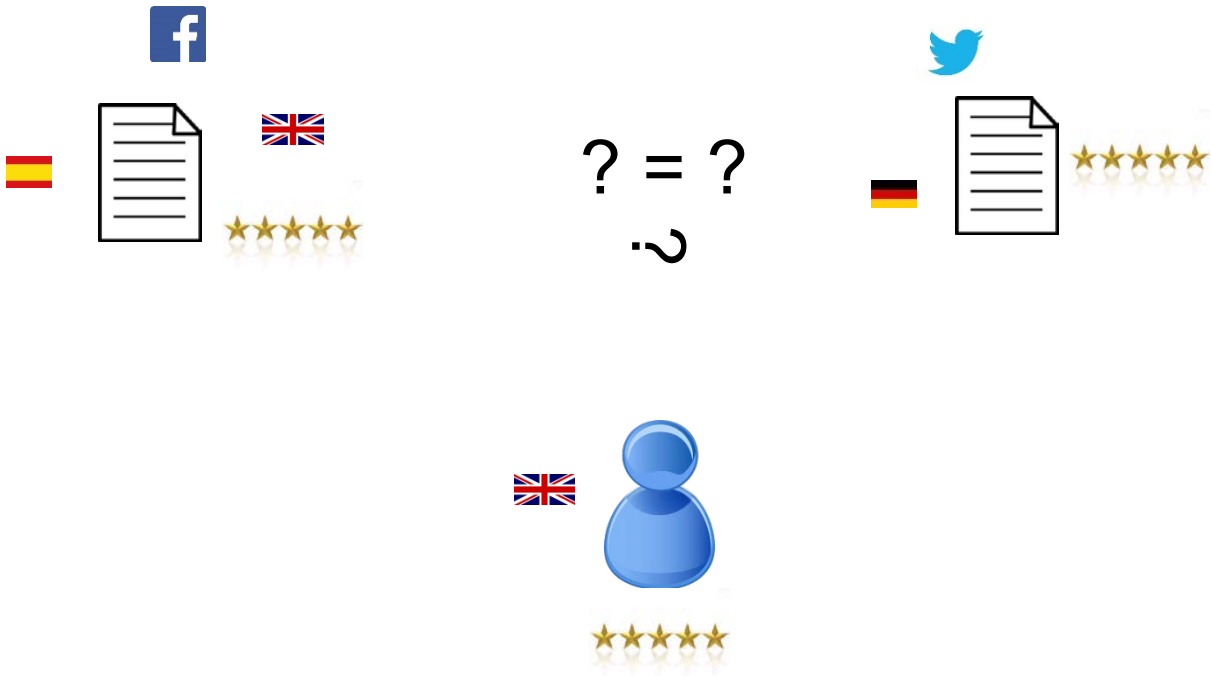
- Newsfeed – Sort news based on user-preferences

- Tweet search – Sort tweets based on relevance criteria

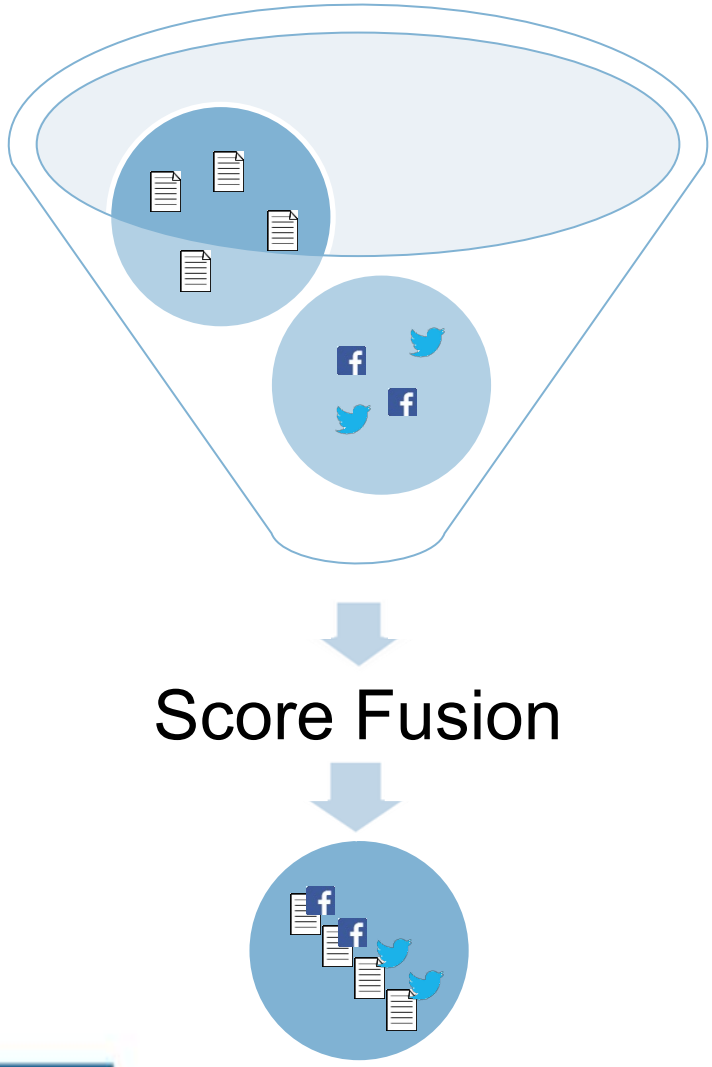


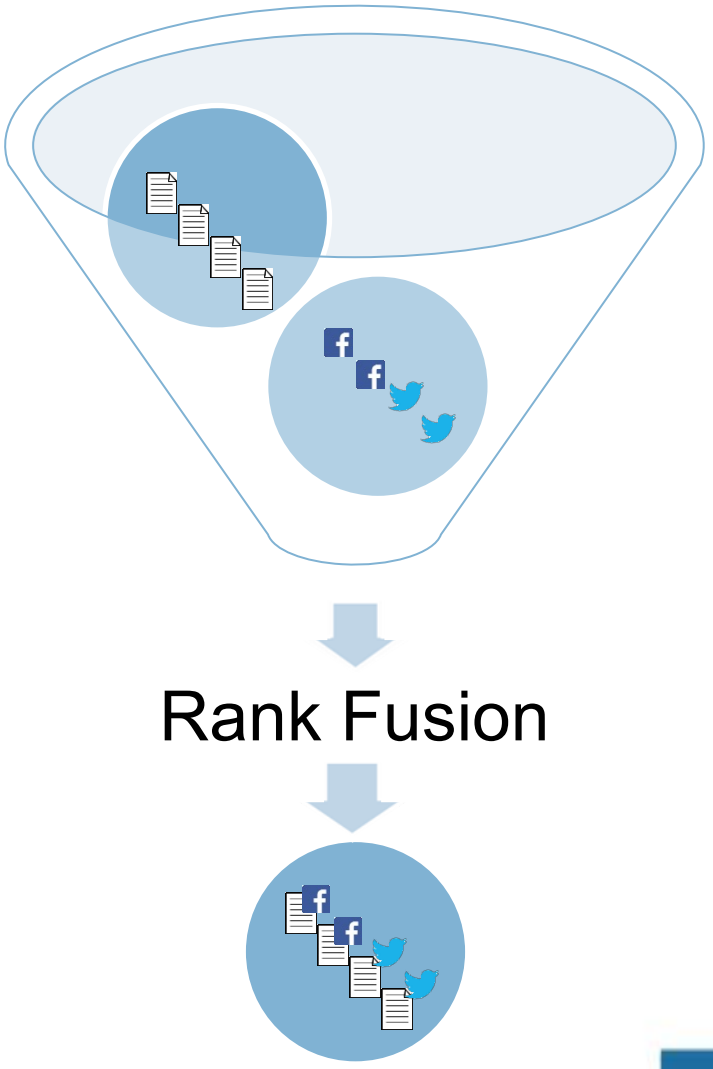
Motivation - Tasks

- Ranking, clustering, filtering and recommendation usually require a comparison between items (documents, users, queries)



Related Work – Fusion Methods





Related Work

- Score fusion: Scores are not normalized → Weighting needed
- Rank fusion: What if features are not of the same importance?

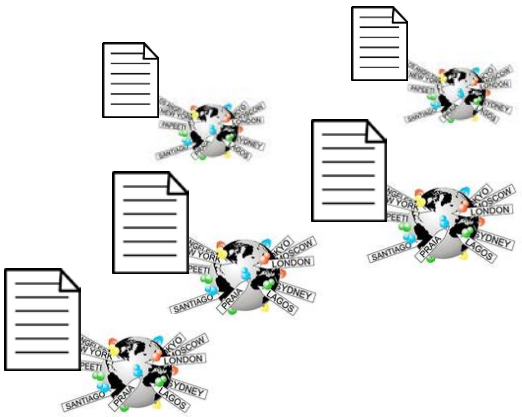
1. Learn weights by logistic regression
2. Smart Guessing

Limitations

- Learn the weights
  - Needs a lot of data to learn from
    - relevance assessments
    - implicit feedback from users
  - For most of the CTI project partners to little data available
  - How much data is enough?
- Smart guessing
  - Is very difficult if the scores are not normalized

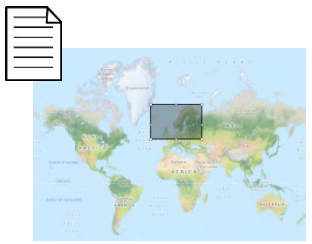
GeoCLEF 2008

- Collection:
  - The Glasgow Herald (1995)
  - The Los Angeles Times (1994)
  - Tagged with geographical coordinates of the locations in the news article
- Topics: 24 geographically challenging topics
  - "Nobel prize winners from Northern European countries"



Collection

? = ?



Topic

GeoCLEF 2008

TOPIC

<identifier>77-GC</identifier>
<title>Nobel prize winners from Northern European countries</title>
<location>-24.87,54.18,32.21,71.28</location>
<TOPONYMS>
<TOPONYM>Northern Europe</TOPONYM>
<TOPONYM>Norway</TOPONYM>
<TOPONYM>Sweden</TOPONYM>
<TOPONYM>Finland</TOPONYM>
<TOPONYM>Iceland</TOPONYM>
<TOPONYM>Denmark</TOPONYM>
</TOPONYMS>
<description>Documents mentioning Noble prize winners born in a Northern European country.</description>
<narrative>Relevant documents contain information about the field of research and the country of origin of the prize winner. Northern European countries are: Denmark, Finland, Iceland, Norway, Sweden, Estonia, Latvia, Belgium, the Netherlands, Luxembourg, Ireland, Lithuania, and the UK. The north of Germany and Poland as well as the north-east of Russia also belong to Northern Europe.</narrative>
</topic>

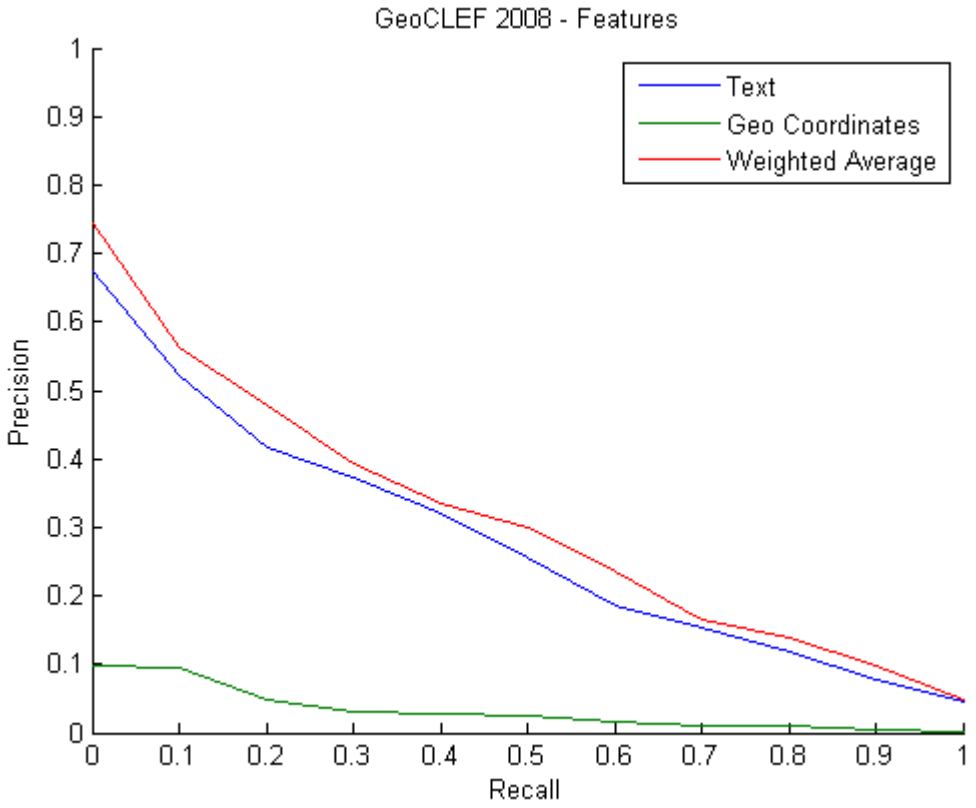
DOCUMENT

<DOC>
<DOCNO>LA042094-0021</DOCNO>
<TEXT>
Authorities on Monday identified a San Fernando man as the motorist who was fatally shot during a car chase last weekend near Hansen Dam Park. Louie Herrera, 23, died ...
</TEXT>
<TOPONYMS>
<TOPONYM **lat="34.05" lon="-118.24"** >Los Angeles </TOPONYM>
<TOPONYM **lat="34.05" lon="-118.24"** >Los Angeles </TOPONYM>
<TOPONYM **lat="36.27" lon="-118.38"** >Hansen Dam Park </TOPONYM>
<TOPONYM **lat="34.28" lon="-118.43"** >San Fernando </TOPONYM>
</TOPONYMS>
</DOC>

Discrete values

Importance of Geographical Features

- Can we improve the search result using the geographical features?



Related Work – Term weighting: BM25

Idea: Characteristic terms should have a high weight.
Characteristic terms are locally frequent tf, but globally rare df.

$$s(D,Q) = \sum_{t \in Q} \frac{tf(t,D) \cdot (k_1 + 1)}{tf(t,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot \underbrace{log \frac{N - df(t) + 0.5}{df(t) + 0.5}}_{idf(t)}$$

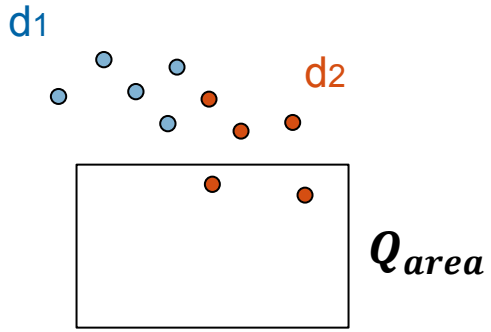
- t = Term
- D = Document
- N = Number of documents in the collection
- $df(t)$ = Number of documents that contain t
- $tf(t,D)$ = Term frequency (number of occurrences of t in D)
- idf = Inverse document frequency
- $avgdl$ = Average document length
- k_1, b = constants

Location Weighting

Idea: Characteristic locations should have a high weight. Adapt term weighting idea to locations.

$$s(D,Q) = \sum_{l \in D \cap Q_{area}} \frac{tf(l,D) \cdot (k_1 + 1)}{tf(l,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot \underbrace{log \frac{N - df(l) + 0.5}{df(l) + 0.5}}_{idf(l)}$$

- l = Location
- $df(l)$ = Number of documents that contain l
- $tf(l,D)$ = Location frequency (number of occurrences of l in D)



Multifaceted Feature Weighting

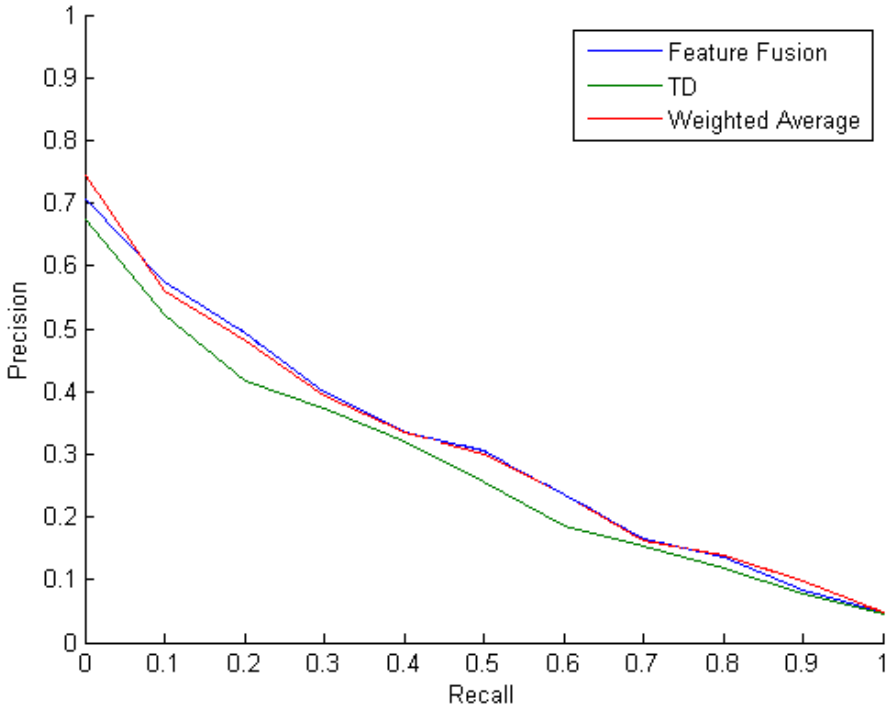
$$s(D,Q) = \sum_{t \in Q} \frac{tf(t,D) \cdot (k_1 + 1)}{tf(t,D) + k_1 \cdot (1 - b + b \cdot \frac{|\boldsymbol{D_t}|}{\boldsymbol{avgdl_t}})} \cdot idf(t) +$$

$$\sum_{l \in D \cap Q_{area}} \frac{tf(l,D) \cdot (k_1 + 1)}{tf(l,D) + k_1 \cdot (1 - b + b \cdot \frac{|\boldsymbol{D_l}|}{\boldsymbol{avgdl_l}})} \cdot idf(l)$$

- Scores are not normalized, but:
  - Lower bounded → 0
  - Average is 1
  - For both scores.

Multifaceted Feature Weighting – GeoCLEF 2008 - Result

- Location Feature Weighting performs nearly as good as learned weighted average!



Open Questions

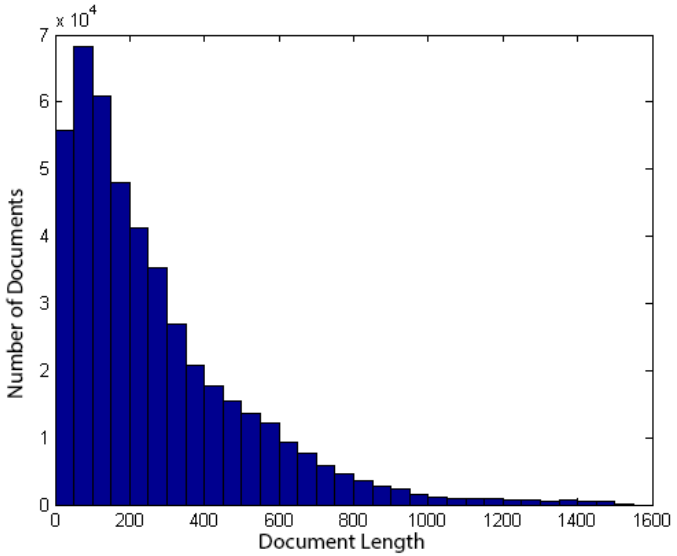
- How can we normalize the scores?
- What document length should be used?

$$\frac{|D_t|}{avgdl_t} + \frac{|D_l|}{avgdl_l} \; ?$$

- Is the term weighting scheme to such modifications?
  - Weighting Scheme Robustness

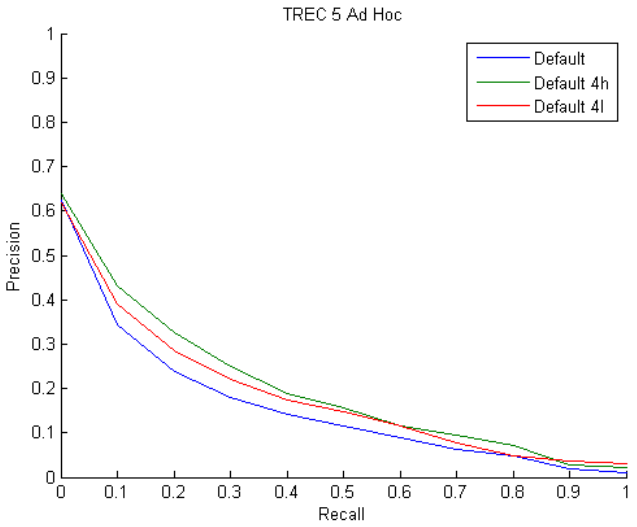
Weighting Scheme Robustness to Variance in Document Length

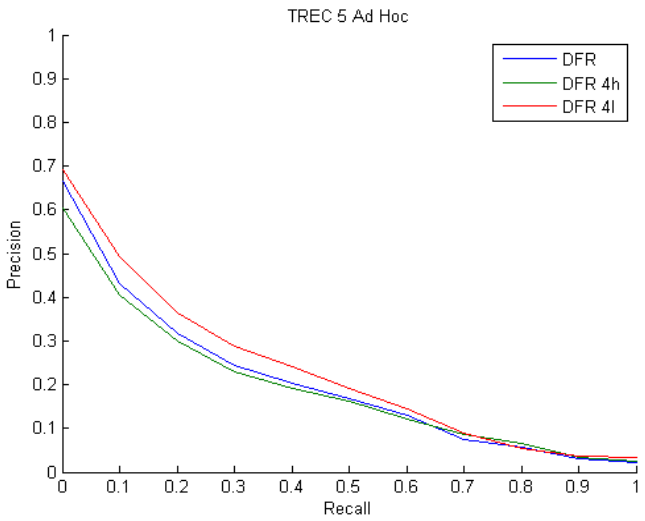
- Tipster Document Collection
  - High variance in document length
  - Remove relevant documents of the longest document bin
  - Remove relevant documents of the shortest document bin
  - Check stability of the retrieval performance



Weighting Scheme Robustness to Variance in Document Length

- BM25 and Language Model (LM) show little variance
- Tf-Idf performance increases when removing long documents
  - angle between query and topic, doesn't work with long documents, chances are smaller for a small angle in high dimensional space
- Divergence from randomness (DFR) increases when removing short documents





Social Book Search 2012

- Collection:
  - 2.8 Mio. Books from Amazon with professional Meta-Data (ISBN, Title, Publisher, Binding, #Pages, Authors, Listprice, Browsenodes,)
  - User-generated Meta-Data from LibraryThing (Reviews, Ratings, tags,)
  - User Profiles (Username, Library (title, author, rating), friends, groups)
- Topics: Generated from user information need on LibraryThing
  - Title, group, username, narrative, type (subject, author, edition), gerne

Collection

? = ?

Topic incl. User

Status

- Best approach until now: Index all textual fields
- Apply location weighting to authors
- Apply location weighting to rating
  - What's the query? The higher the better?
- How can we apply idea to continuous features, such as the number of pages?
  - Someone only reading books with less then 100 pages will prefer a book suggestion with a small number of pages

Questions, Hints, Leads?