Design of a Helpdesk System for Archival Infrastructures – An Information Retrieval approach

Kepa J. Rodriguez (SUB-Göttingen)

Talk at the Göttingen Center for Digital Humanities (GCDH)

21/05/2013

CONNECTING COLLECTIONS

Outline

- The EHRI project
- Motivation
- Analysis of user requests
- Representation of archival institutions in the system
- Relevance of a institution to answer a query
- Experiments
- A first prototype: presentation and demo
- Further work

The EHRI project

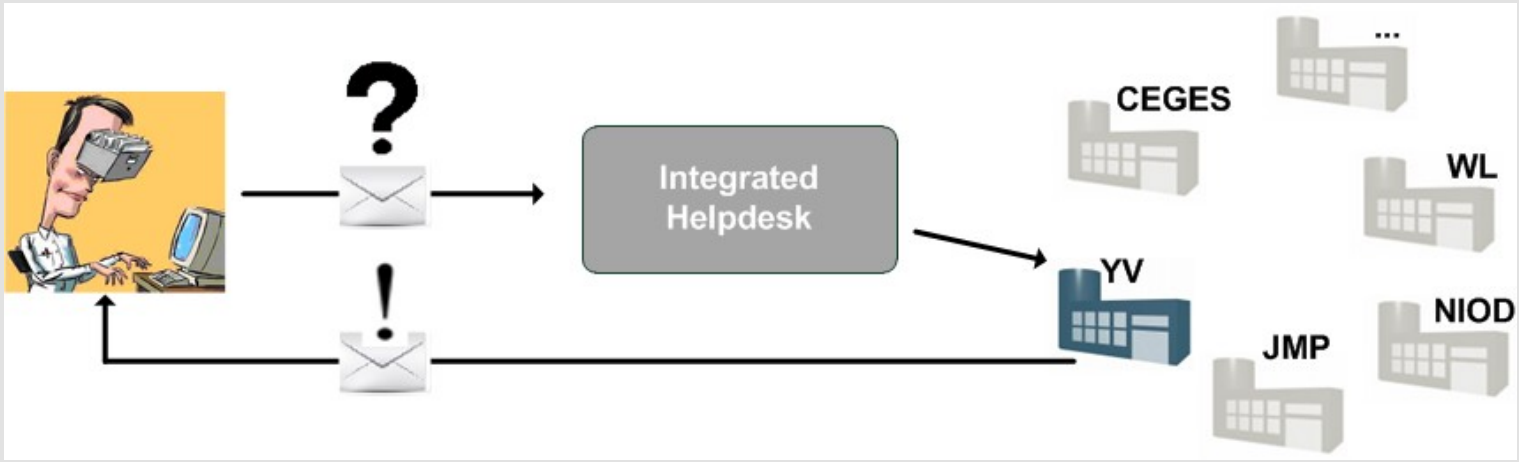
- EHRI – European Holocaust Research Infrastructure – aims to build a virtual research environment for Holocaust research
- 20 institutions of 13 countries are involved in the development
- EHRI aims to integrate collection descriptions of more than 1.000 archives, museums, memorials and other institutions
- More information: http://www.ehri-project.eu

Why a helpdesk?

- Often historians get useful information directly from archivists and not from catalogs and finding aids.
- In Shoah research the material is very dispersed: before archivists are contacted the researcher needs to identify the institution.
- One of the goals of EHRI and other European projects is to put in contact researchers and collection holding institutions.
- Partial solution: EHRI will provide an integrated help desk to support users in the identification of the institution.

What is a helpdesk?

- In the authomatic helpdesk:
  - The user writes this problem as an e-mail,
  - She/he sends it to the helpdesk
  - The system analyzes the problem and proposes suitable archival institutions.



Analyisis of user queries: user requests

Data set: 400 e-mails send to the NIOD-Amsterdam.
(Institute for War, Holocaust and Genocide Studies)

Most habitual user queries are:

- Information requests about individual persons or families.
- Information requests about concentration camps, ghettos, KZ and detention facilities. Forced labor.
- Information requests about documents.
- Information requests about places.
- Information requests about press, radios and underground press.

Analyisis of user queries: provided information (1)

Data about persons

- Name, family name, initials
- Biographic data: profession, position in companies, role in the army, resistance, antifascist and fascist organizations.
- Membership in political organizations, in resistance, etc.
- Dates of birth, travel, arrest, deportation, execution, death, exile.
- Places of birth, residence, travel, arrest, deportation, execution, death, exile.
- Citizenship

Analyisis of user queries: provided information (2)

Data about facilities and transports

- Place of KZ, detention facilities, ghettos.
- Route of transports, stations.
- Dates of transports.

Historical events:

- Described by groups of words.
- Often defined by verbs or nominalizations: uprising, assassination, bombardment, etc.

Proposed solution

The help desk:

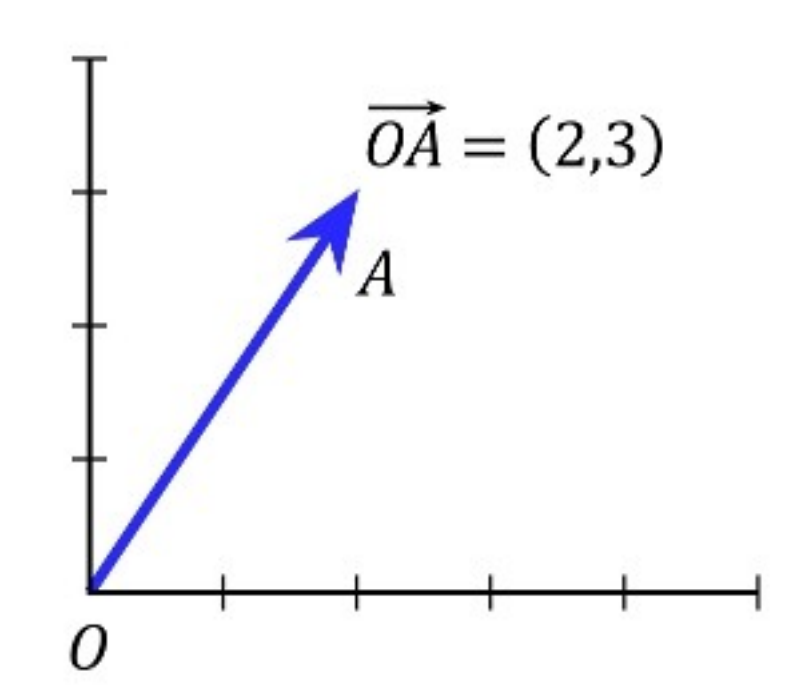
- Knows all necessary things about institutions.
  - How are institutions represented?
- Understands the needs of the user.
  - How is the query analyzed and represented?
- Gives feedback to the user about relevant institutions.
  - How is relevance computed?

How can we define an institution for our purposes?

- An institution is defined by:
  - Its collections, files, documents.
  - Its controlled vocabularies.
  - Its profile.

…. and represent it? … a short formal introduction

First of all an introduction to vector spaces



Example: bi-dimensional vector space

We use a larger space: 71.000 dimensions or more

Representation of institutions in vector space

- Collections, file descriptions, controlled vocabularies are sets of words.
- Features of the space are (selected) word lemmas
- Features are extracted using natural language processing
  - Text are tokenized and words identified
  - Each word is annotated with a Part of Speech
  - Each word is annotated with a lemma
  - Words with a selected list of POS tags are candidates for being a feature

Representation of institutions in vector space (1)

Example of NLP processed text

```
......
including   VVG include
circulars   NNS circular
from        IN  from
the         DT  the
Central     NP  Central
Refugee     NP  Refugee
Committee   NP  Committee
on          IN  on
the         DT  the
enlistment  NN  enlistment
of          IN  of
tradesmen   NNS tradesman
into        IN  into
the         DT  the
army        NN  army
......
```

Representation of institutions in vector space (2)

- Each feature is associated with a value/weight
- Two functions used for the experiments
  - TF-IDF: Term frequency / Inverse document frequency.
  - Okapi-BM25.
- User query: the weight is the term frequency

Representation of institutions in vector space (3)

Example of function: TF-IDF

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

Representation of institutions in vector space (4)

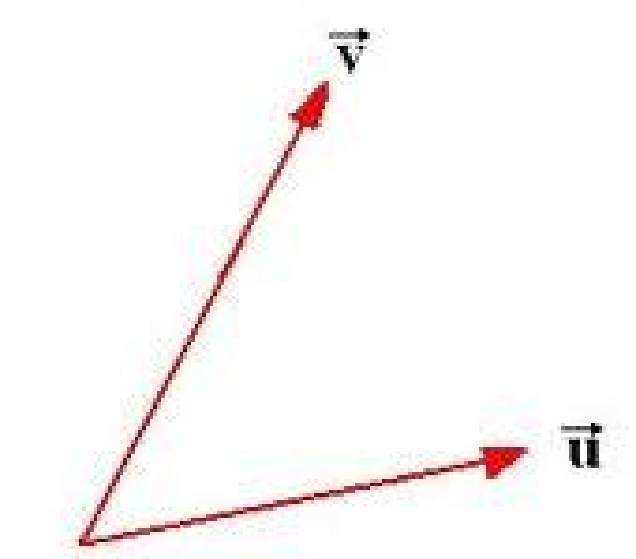
Example of vector (fragment)

```
......
activist   0.0729628611157319
answer  0.1446774839801297 5
welfare 0.0482258279933765 8
job 0.0482258279933765 8
number  0.0482258279933765 8
regard  0.5107400277810122
emigration 0.0729628611157319
create  0.0482258279933765 8
include 0.2553700138905061
soldier 0.0482258279933765 8
editor  0.0482258279933765 8
ownership  0.0729628611157319
poster  0.0482258279933765 8
branch  0.0482258279933765 8
disclosure 0.0482258279933765 8
underground0.0729628611157319
newspaper  0.4340324519403892 5
press  0.0482258279933765 8
......
```

Relevance: how can it be measured?

We can measure relevance as:

- Angle between the vectors (e.g. Cosine)
- Distance between the end-points of the vector
- Etc.



Experiments: task

- Collection descriptions were used to build a vector space model.
- Institutions provide us e-mails with information requests
  - The institution was able to give the requested information.
  - The information was related with their collections, not with administrative issues.
- Goal: Find out, to which institution an e-mail has been sent

Experiments: data sets

- Learning set: collection and file descriptions of
  - Wiener Library: 553 documents / 84,918 words
  - NIOD: 1929 documents / 2,095,402 words (English translation)
  - Yad Vashem: 17,086 documents / 1,491,858 words
  - Jewish Museum Prag: 1 document / 10,825 words
- Test set: e-mails sent to:
  - NIOD: 15 e-mails / 2759 words
  - Yad Vashem: 21 e-mails / 3900 words
  - Wiener Library: 30 e-mails / 2342 words

Experiments: Functions and metrics

- Features ranked using
  - TF-IDF
  - Okapi-BM25
- Similarity computed using
  - scalar product
  - cosine similarity
  - Euclidean distance
- Evaluation: Precision, Recall, F1

Experiments: evaluation

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Experiments: result

Using cosine similarity we obtain:

F1 =~ 0.7 – 0.8

Several cases are not real errors:

- More than one institution is able to answer the question

Difficulties

- Often more than one institution is able to answer a user request
  - That is good for the system, but makes the evaluation difficult
  - With more data we can do a relevance ranking based evaluation.
- Cases of very unspecific requests can be answered by almost all institutions.
  - Example: *“I am looking for transportations list from Bratislava to Auschwitz, happens on march 1942 witch concern my mother. May I ask you for help?“*

Difficulties

- Multilinguality
  - Collection and file descriptions are in different languages
  - Machine translation is necessary
    - For the experiments, this was done by hand with Google
  - A language identification modul will be necessary.
- We will integrate collection descriptions, but very interesting information is in levels not accesssible for EHRI
  - File level description
  - Documents

Prototype

- Implementation:
  - Web application writen in Java
  - Maven, Spring framework
- NLP based feature extraction
  - Tokenizer: self developed
  - POS-tagger and lemmatizer
    - TreeTagger

      http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
  - Extracted lemmas of words with POS Noun, Proper noun, Verb
- Ranking function: TF-IDF
- Similarity: Cosine

Next steps (1)

- Explore use of other sources of knowledge
  - Authority files: names of persons, places...
  - Institution descriptions

- Based on the data implement lists and dictionaries for:
  - Synonimy:
    - Nazi, nationalsozialist and national-sozialist are different features
  - Alternative spellings and misspellings
  - Text pre- and post- processing

Next steps (2)

- Multilinguality:
  - Use APIs for automatic translation
  - Implement a language identification modul
- Use the ranking of relevance
  - For evaluation
  - To present results to the user
  - Find a threshold to define when results are irrelevant and user needs to give more details

EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

NIOD Institute for War, Holocaust and Genocide Studies (NL)

CEGES-SOMA Centre for Historical Research and Documentation on War and Contemporary Society (BE)

Jewish Museum in Prague (CZ)

Institute of Contemporary History Munich – Berlin (DE)

YAD VASHEM The Holocaust Martyrs' and Heroes' Remembrance Authority (IL)

The Wiener Library – Institute of Contemporary History (UK)

Holocaust Memorial Center (HU)

HL-senteret Center for Studies of Holocaust and Religious Minorities (NO)

NAF National Archives of Finland (FI)

The Emanuel Ringelblum Jewish Historical Institute (PL)

King's College London (UK)

Georg-August-Universität Göttingen – Göttingen State and University Library (DE)

Athena RC/IMIS (GR)

DANS Data Archiving and Networked Services (NL)

Shoah Memorial, Museum, Center for Contemporary Jewish Documentation (FR)

ITS International Tracing Service (DE)

Memorial to the Murdered Jews of Europe (DE)

Terezín Memorial (CZ)

Beit Theresienstadt (IL)

VWI Vienna Wiesenthal Institute for Holocaust Studies (AT)

SEVENTH FRAMEWORK PROGRAMME

CONNECTING KNOWLEDGE

Design of a Helpdesk System for Archival Infrastructures – An Information Retrieval approach

21/05/2013

CONNECTING COLLECTIONS