Experiments for the design of a help desk system for the EHRI project – an Information Retrieval approach

Kepa J. Rodriguez
(Göttingen State and University Library)

DARIAH-DE Working Papers Nr. 5

Rodriguez, Kepa J.: Experiments for the design of a help desk system for the EHRI project : an Information Retrieval approach
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2014
(DARIAH-DE working papers 5)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2014-3
URN: http://nbn-resolving.de/ urn:nbn:de:gbv:7-dariah-2014-3-2

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Stefan Schmunk und Christof Schöch

Abstract: This paper describes the experiments realized for the conception of a helpdesk system for the EHRI project. The system will help researchers to find Collection Holding Institutions (CHI) suitable to support them to resolve questions about Holocaust relevant documentary sources. In this experiments we model descriptions of archival material in a vector space model, and we show that this model is able to give an answer to real user queries.

Keywords: Information Retrieval, Vektorraum, Metadaten, Archive

Information retrieval, Vector space model, metadata, archives

Experiments for the design of a help desk system for the EHRI project - an Information Retrieval approach

Kepa J. Rodriguez
(Göttingen State and University Library)

DARIAH-DE
Working Papers

Kepa J. Rodriguez: "Experiments for the design of a help desk system for the EHRI project - an Information Retrieval approach". *DARIAH-DE Working Papers* Nr. 5. Göttingen: DARIAH-DE, 2014. URN: urn:nbn:de:gbv:7-dariah-2014-3-2.

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm,
Stefan Schmunk und Christof Schöch herausgegeben.

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Abstract

This paper describes the experiments realized for the conception of a helpdesk system for the EHRI project. The system will help researchers to find Collection Holding Institutions (CHI) suitable to support them to resolve questions about Holocaust relevant documentary sources. In this experiments we model descriptions of archival material in a vector space model, and we show that this model is able to give an answer to real user queries.

Table of Contents

Introduction

Transnational research into the Holocaust rely more than other fields of historical research on an heterogeneous landscape of archives. Holocaust archives are fragmented and dispersed all over the world, making access to documentary sources time and resource consuming and complicate, if not impossible.

The main objective of the European Holocaust Research Infrastructure project (EHRI) is to support Holocaust research community setting a portal which provide researchers with online access to integrated documentary sources, and fostering collaborative research though a Virtual Research Environment (Blanke et al. 2014).

Surveys on Holocaust scholars and archival institutions realized at the EHRI project show that researchers find sources not only making use of online catalogs, finding aids or archival retrieval systems, but asking directly to collection specialists and reference archivists at collection holding institutions (Speck and Links, 2013).

In the complex and fragmented landscape of Collection Holding Institutions (CHI) with Holocaust related material the challenge for researchers is not to establish direct communication with archivists, but how to find the CHI that is most likely to have knowledge to help in his/her research question.

The integrated help desk will address this problem using Information Retrieval techniques to help researchers to find those CHIs whose holdings match the research questions.

In this paper we present the experiments realized for the conception and design of the help desk. The first step consisted of understanding the requirements of potential users of the system. More specifically, we needed to understand the kind of questions that users ask to collection specialists. The result of this analysis is presented in section 1. Section 2 presents the notion of relevance in a vector space model. We have choose this approach for the experimental settings and for the final implementation of the system. Section 3 describes the experiments and presents the results. For the experiments we have used datasets from 4 CHIs members of the EHRI project. The final section presents the conclusions of the experiments and discusses further work, especially the challenge of the multilingual character of the data.

1. Analysis of user requests

The first step to build the system consists in understanding the necessities of the user, the questions that users ask to collection specialists. To do it we have analyzed a set of e-mails send by different kinds of users (historians and not) to the Institute for War, Holocaust and Genocide Studies (NIOD).

The first set consists of 120 e-mails sent by users and researchers to the institution selected attending two criteria:

- The topics asked by the user refer to the content stored in the institution, not to administrative issues like appointments, permissions to access material, etc.
- The institution was able to give a response to the e-mail satisfying the request of the user.

The most important topics asked by the user are:

- Information requests about individual persons or families: most e-mails request this kind of information: 65 e-mails
- Information requests about concentration camps, ghettos and detention facilities: this group includes 12 e-mails
- Information requests about documents : 7 e-mails
- Information requests about places (all in Indonesia[1]) : 3 e-mails. These places are not related to detention facilities.
- Information requests about press, radios and underground press : 3 e-mails
- Information requests about forced labor : 2 e-mails asking only for this topic. We can find more questions about it in e-mails asking for other topics like persons and detention facilities.
- Other topics related to the Shoah and/or the Second World War: 16 e-mails.

The analysis was extended with e-mails provided by two institutions more: the Wiener Library, London, and Yad Vashem, the World Center for Holocaust Research, Documentation, Education and Commemoration in Jerusalem. Although the proportion of e-mails in each cluster changes in each institution, the topics covered are similar.

In order to get the desired information the user provides in the e-mails following information:

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

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 that can be incorporated in a vector of weighted features.

Kind of event

- Defined by verbs or nominalization: uprising, assassination, bombardment, etc.

[1] During the Second World War Indonesia was a Dutch colony. That motivates that some relevant collections of the NIOD contain information about South-East Asia and Japan.

2. Representation and relevance in a Vector Space Model

This section presents the concept of vector space model, and how it can be used to determine the relevance of the information stored in a CHI.

The vector space model

In a vector space model a document or text is represented by a vector of terms (or more generally features) in which each term becomes an independent dimension in a high dimensional space (Singhal 2001). For the representation of each document each term is associated to a value.

Given the index vectors for two documents or texts it is possible to compute similarity between them using using different measures which reflects the degree of similarity in terms and term weights (Salton et al. 1975).

Similarity measures are not inherent to the model (Singhal 2001) and can be computed in different ways, as measuring the distance between the extremes of the vectors, computing scalar product or computing the the angle between vectors and its cosine.

Collection Holding Institutions

For the conception of the helpdesk we define a collection holding institution by the knowledge stored in it. This knowledge involves the descriptions of archival material held in the CHI and potentially, its institutional profile.

Archival material and its descriptions, controlled vocabularies and other sources of knowledge can be represented as sets of words. This shallow representation cannot represent relations between collections or between parts of them, entities and events, but for the purpose of the helpdesk it is able to represent the enough knowledge stored in a CHI.

Extraction of features

EHRI project integrates collection descriptions. We use these descriptions as source of knowledge to extract the relevant features to build the vector space. The descriptions of archival material used for the experiments are described later in section 3.

There are different approaches to feature extraction from text. One of them is the extraction of just named entities or terms from a controlled vocabulary. Other systems use all words of the text (or their extracted lemmas) as possible features. More sophisticated approaches include natural language processing, paying attention to linguistic features like part of speech or grammatical role of the term in a sentence.

For our experiments we process the collection descriptions using the lemmatizer and part of speech tagger TreeTager (Schmid 1994). An example of the output of TreeTagger is presented in Figure 1. We extract from the processed text lemmas of words belonging to certain parts of speech: Verb, Noun and Proper Noun. The language processing is combined with a posterior filtering to remove stop words and to cluster some synonyms.

```
...
The        DT      the
list       NN      list
contains   VVZ     contain
the        DT      the
names      NNS     name
of         IN      of
all        PDT     all
those      DT      those
whom       WP      whom
the        DT      the
Nazis      NNS     Nazi
regarded   VVN     regard
as         IN      as
a          DT      a
potential  JJ      potential
threat     NN      threat
to         TO      to
...
```

Figure 1: Output of the Tree Tagger over data of Yad Vashem

Finally, all features of the model are stored in a list. Only features of the list will be used to extract vector representation of institutions and user queries.

Ranking of features

Features are ranked according to their specificity within a set of documents, and to their frequency. There are different metrics such as TF-IDF (Term Frequency – Inverse Document Frecuency; Wu et al. 2008) or Okapi-BM25 ("Okapi-BM25" 2014).

We performed experiments extracting and ranking features from descriptions of archival material delivered by selected CHIs (for details see annex), obtaining slightly better results using TF-IDF rather than the other ranking functions.

The TF-IDF function has two parts. The first one is the term frequency: the importance of a term in a document is proportional to the number of times that it appears in the document. To compute that first of all we find the term that appears more times in the document, and we divide the times that the term appears in the document by the number of times that the most frequent term appears. The value of TF is between 0 and 1.

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

Inverse Document Frequency (IDF) is based on counting the number of documents in which the term is indexed. The intuition behind it is that a term occurring in many documents is not a good discriminator, and it should be penalized given less weight than terms that occur in few documents.

$$idf(t,D)=\log\left(\frac{|D|}{1+(d\in D:t\in d)}\right)$$

IDF is computed dividing the number of documents by the number of documents in which the term appears, and taking the logarithm of the quotient. Terms appearing in all documents would have a IDF = log(1) = 0. To avoid it we add 1 to the denominator.

Finally TF-IDF is the product of the TF measure and the IDF measure.

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

An example of fragment of a vector after the ranking of features is presented in Figure 2.

```
...
fürnberg    0.048770845831336375
himmler     0.0070014866723226805
gruber      0.008558564246335389
viola       0.022194187649860805
dubská      0.01625694861044546
václav      0.01625694861044546
käuflerová  0.01625694861044546
florida     0.008558564246335389
movement    0.014002973344645361
jew         0.07701635339554948
babylon     0.011097093824930402
neuernová   0.01625694861044546
wantochová  0.01625694861044546
tricked     0.01625694861044546
marie       0.06301338005090412
fantlová    0.01625694861044546
raul        0.022194187649860805
...
```

Figure 2: Fragment of the vector representation of the Jewish Museum Prague

Representation of user queries

User queries are represented as vectors of the same vector space. The extraction of features is similar to the extraction of features of the institutions. After queries have been language processed, we check whether the lemma of each word is contained in the list of features mentioned earlier. If it is, it is extracted and ranked.

The ranking of features extracted from the user query is slightly different. Here we compute just the Term Frequency using the formula described in the previous section.

Relevance measure

Relevance of an institution to answer the query of the user is measured by the proximity of the vector of the institution to the vector representing the query in the vector space.

We represent the relevance of a CHI to answer the query of the user using cosine similarity. The similarity between two vectors is computed as the cosine of the angle between the vectors.

$$similarity(\vec{A},\vec{B})=\cos(\vec{A},\vec{B})$$

If both vectors are identical the value of the cosine is 1, and it represents the maximal possible relevance. If two vectors are orthogonal the value of the cosine is 0, and it represents that the institution is not relevant for the question of the user.

$$\cos(\vec{A},\vec{B}) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3. Experimental setting

For the experiments we build a vector space using collection and file level descriptions given by 4 institutions: Wiener Library, London, Institute for War, Holocaust and Genocide Studies. Amsterdam (NIOD), Yad Vashem National Authority for the Remembrance of the Martyrs and Heroes of the Holocaust, and Jewish Museum Prague.

- 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 Prague: 1 document / 10,825 words

The descriptions provided by NIOD were written in Dutch. Since the approach adopted for the experiments is not multilingual, they have been translated using an online commercial service. As we extract only words without paying attention to syntactic relationships and dependencies, the quality of machine translation is enough for our purposes.

The test set consists of e-mails send by:

- NIOD: 15 e-mails / 2759 words
- Yad Vashem: 21 e-mails / 3900 words
- Wiener Library: 30 e-mails / 2342 words

The e-mails were selected ensuring that the institution was able to answer the question of the user and that the question was about collections and content stored in the institution, rather than administrative issues like permissions to access and copy material or appointments with the archivists or researchers working at the CHI.

The task consisted of the mapping of the e-mails to the institutions that received them. Our assumption is that, if the system is able to find the receiver CHI, it will be able to find out which institution is able to help the user to find the desired information.

The first experiment was realized only with the e-mails sent to NIOD and Yad Vashem (we got the dataset of the Wiener Library later). The system was able to find the correct institution an 80% of the times.

After we incorporated the dataset of the Wiener Library the amount of e-mails assigned to the correct institution decreased to a 70%. The reason for this lower rate of success is that many of the e-mails provided by the Wiener Library are very short and unspecific, suitable to be answered by more than one institution. E-mails with more than 50 words sent to the Wiener Library were correctly classified.

When the user gives more details, the answers of the institutions are better. All e-mails with over 100 words would good classified.

Difficulties for the evaluation

The evaluation is not a trivial task. For some of the questions more than just an institution was able to answer to the questions of the user.

On the other hand, there are cases in which users give too generic questions, that can be answered by all the institutions[2]. In these cases if the system proposes another institution to help the user we cannot really consider the answer as incorrect.

Other experiments

We replied the experiments changing the knowledge used to build the vector space, the ranking function and the similarity measures.

Experiments combining different sources of knowledge:

- Use of controlled vocabularies and authority files increasing the weight of the terms pertaining to the vocabularies.
- Build a vector space model based only on terms of controlled vocabularies.
- Build a vector space model using only words annotated as proper name by the Part of Speech Tagger.

Ranking function:

- We replicated the experiment ranking the features with Okapi-BM25 (see "Okapi-BM25" 2014) instead of using TF-IDF

Experiments with different similarity measures. We replicated the experiments using two alternative similarity measures

- Euclidean distance (see "Euclidean Distance" 2014)
- Scalar product (aka Dot product; see "Scalar Product" 2014)

We got with all possible combinations a lower performance. Anyway the difference in performance between Okapi-BM25 and TF*IDF as ranking function is not very high.

Conclusions and further work

In this paper we have presented the experiments performed for the design of an automatic helpdesk system for the EHRI project.

The results show that the vector space approach is useful to help users to locate the CHIs that are the most likely able to answer user's questions in regard to Holocaust documentation.

One of the aspects which we need to address in the implementation of the helpdesk in the EHRI portal is the multilingual character of Holocaust related archival material and descriptions. That involves two different aspects:

1. Language detection: determine in which language is written each field of data
2. Machine translation of data from the original language into English.

[2] For instance, one of the e-mails asked just for "Information about transports between Bratislava and Auschwitz" without more details. That is an information that is partially answered in collection descriptions at last of three of the four institutions.

We have already built a model for language detection (Rodriguez 2013) using texts of the Leipzig Corpora Collection (Deutscher Wortschatz 1998-2014) and of the Proceedings of the European Parliament (EUROPARL, see Koehn 2005). For each language we have selected 200.000 lines of text, what corresponds to ca. 3.500.000 words. The performance of the model is enough for the task: for text written using Latin alphabet the performance is of 92.9% for samples of 10 words and 99.65% for samples of 30 words. For Cyrillic text the performance is of 97,3% for samples of 10 words and 100% for samples of 30 words.

For machine translation we plan to combine the use of self developed machine translation models with commercial services. We have already built models for German, French and Dutch using the tool delivered by the European project Moses Core. We are investigating how the quality of the translations can be improved extracting vocabularies from the translated material. For other languages we intend to use commercial machine translation services.

Annexes

Projects and Institutions

- EHRI: European Holocaust Research Infrastructure; http://www.ehri-project.eu.
- NIOD: Institute for War, Holocaust and Genocide Studies, Amsterdam, The Netherlands; http://www.niod.nl.
- The Wiener Library, London, UK; http://www.wienerlibrary.co.uk.
- Yad Vashem, World Center for Holocaust Research, Documentation, Education and Commemoration and National Authority for the Remembrance of the Martyrs and Heroes of the Holocaust, Jerusalem, Israel; http://www.yadvashem.org.
- Jewish Museum of Prague. Prague, Czech Republic. http://www.jewishmuseum.cz.
- Moses Core. Promoting Open-Source Machine Translation. http://www.mosescore.eu.

References

Blanke, Tobias; Vanden Daelen, Veerle; Frankl, Michal; Kristel, Conny; Rodriguez, Kepa J. and Speck, Reto (2014): "From fragments to an integrated European Holocaust Research Infrastructure." In: *Evolution der Informationsinfrastruktur: Forschung und Entwicklung als Kooperation von Bibliothek und Fachwissenschaft*, ed. Andrea Rapp, Norbert Lossau, Heike Neuroth. http://dx.doi.org/10.3249/webdoc-39006.

Deutscher Wortschatz (1998-2014). URL: http://corpora.uni-leipzig.de/.

"Euclidean Distance" (2013), Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Euclidean Distance. Version of December 22, 2013: http://en.wikipedia.org/w/index.php?title=Euclidean_distance&oldid=587199728.

Koehn, Philip (2005): "Europarl: A Parallel Corpus for Statistical Machine Translation", MT Summit. http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf. See also: European Parliament Proceedings Parallel Corpus 1996-2011. URL: http://www.statmt.org/europarl/.

"Okapi BM25" (2014), *Wikipedia, The Free Encyclopedia*, URL: http://en.wikipedia.org/wiki/Okapi_BM25. Version of February 10, 2014: http://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=594899794.

Rodriguez, Kepa J. (2013): "Building a 3-gram Model for Language Identification". *Research and Development Colloquium*. Göttingen: State and University Library. URL: http://www.slideshare.net/KepaJRodriguez/language-identification.

Salton, G.; Wong, A. and Yang, C.S. (1975): "A Vector Space Model for Automatic Indexing". *Communications of the ACM* 18 (11).

"Scalar Product" (2014), *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/wiki/Scalar_product. Version of February 13, 2014 : http://en.wikipedia.org/w/index.php?title=Dot_product&oldid=595360711.

Schmid, Helmut (1994): “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK. See also: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.

Speck, Reto and Links, Petra (2013): “The Missing Voice: Archivists and Infrastructures for Humanities Research”. *International Journal of Humanities and Arts Computing*, 7(1-2), 128-146.

Wu, H.C.; Luk, R.W.P.; Wong, K.F. and Kwok K.L. (2008): “Interpreting tf–idf term weights as making relevance decisions”. *ACM Transactions on Information Systems* 26 (3): 1–37. doi:10.1145/1361684.1361686.