ENRICHING VISUAL HISTORY ARCHIVE WITH EXTERNAL KNOWLEDGE FROM WIKIPEDIA

1. Introduction

The USC Shoah Foundation Institute for Visual History and Education has nearly 52,000 videotaped interviews with survivors and witnesses of the Holocaust and other genocides in their archives. Researchers are allowed to search through and view these testimonies via the online portal VHA Online[1]. These audiovisual testimonies have not been transcribed. Instead, the videos are indexed by keywords, personal names, interview codes, and geolocation. Since the indexing terms are manually selected by editors, the limited selection of terms cannot fully describe and represent the content of the indexed videos. Moreover, to retrieve segments of testimonies for users, lexical matching are performed between their free-form search queries and the limited indexing terms, which often leads to mismatch between user search intent and retrieved videos. To remedy these problems, we would like to integrate the external knowledge derived from Wikipedia[2] into the visual history archive.

2. Potential Applications

Enriching the testimony archive with Wikipedia knowledge base enables a couple of potential applications:

(1) Exploiting Wikipedia knowledge enables the system to supplement segments of testimonies with relevant information derived from the external knowledge base. For example, as shown in Figure 1, the segment is indexed by the term *Israel Independence Day*. By integrating the Wikipedia data relevant to this particular concept, the system will be able to provide users with additional information about the independence of Israel. In Figure 1, the supplement information is provided as a snippet together with a Wiki link next to the segment.

[1] `http://vhaonline.usc.edu`
[2] `http://www.wikipedia.org`

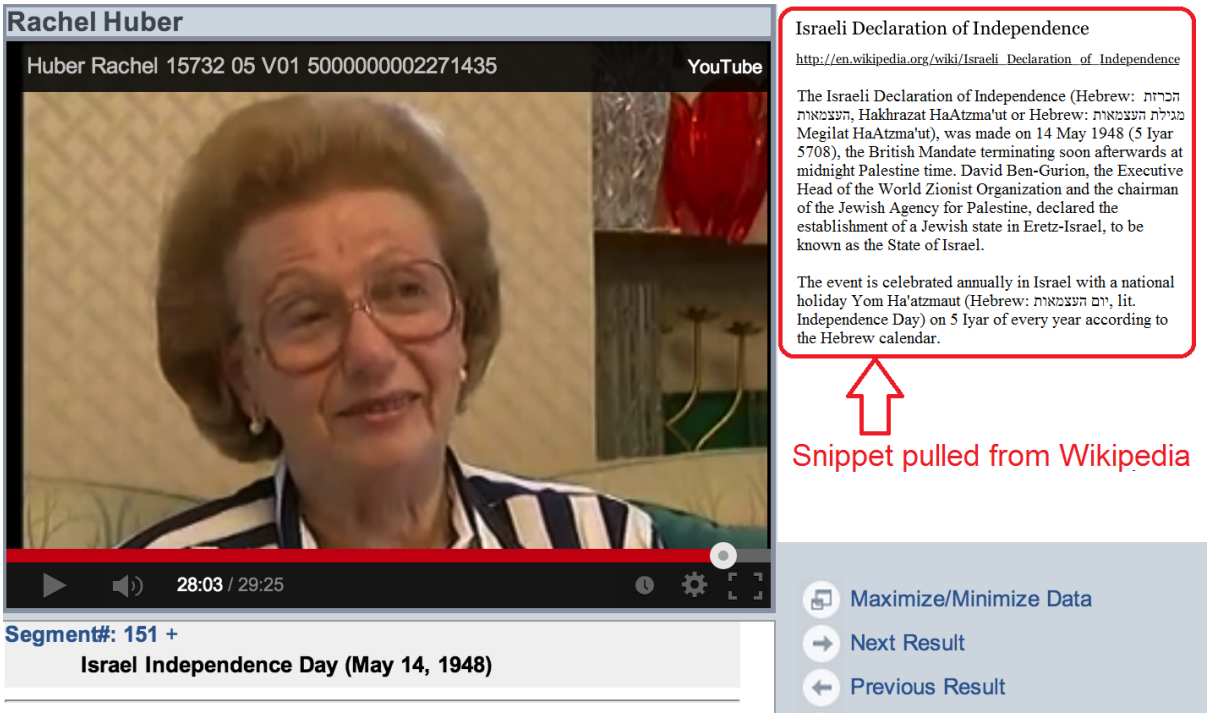

FIGURE 1. A sample segment with a supplementary snippet pulled from Wikipedia

(2) Indexing terms can be enriched with the semantic knowledge extracted from Wikipedia. More specifically, as opposed to indexing videos with individual keywords, the system will be able to build the index at the concept level, where the concepts are distilled by latent semantic analysis of Wikipedia data. Indexing with concepts remedies the problem of mismatch between user search intent and retrieved videos.

(3) To further address the mismatch problem, the rich content in Wikipedia can be utilized for *query expansion*, which is a process of expanding search queries to match additional relevant testimonies. In particular, by leveraging the Wikipedia corpus, the system will be able to reformulate a search query by adding new terms in the same concepts to the query, so that more relevant videos can be retrieved.

3. PROJECT DESCRIPTION

The objective of the RIPS summer project is to build a system which integrates semantic knowledge extracted from Wikipedia into the visual history archive. There are two primary approaches to establish connections between the different data sources:

(1) The system generates a vector of terms to represent each document in Wikipedia. Each entry of the vector indicates the weight of a specific term. The most popular term weighting method is the TF-IDF weighting scheme. On the testimony side, indexing terms and query terms are transformed to bag-of-words representations in a similar way. In this *Vector Space Model* (VSM), the degree of relationship between testimonies and Wikipedia documents can be measured with a proper similarity metric, e.g., cosine similarity.

(2) As opposed to measuring relationship between the two data sources at the term level, we can establish connections between them at the concept level. To this end, a probabilistic topic model is used to map data from both the testimony archive and Wikipedia to the same latent concept space. More specifically, we can employ a typical topic model, *Latent Dirichlet Allocation* (LDA) [1], on both testimonies and Wikipedia documents to extract latent topics (concepts). Advanced topic models can also be used to fuse together the data in two different modalities.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.