Summarization of Book annotation with videos using cross-media retrieval

As the rapid progress of the e-learning platform and the bloom of on-line educational resources, there are more and more needs from different aspects to improve the quality and feeling of the study in this age of Internet and multi-media. In this context, this thesis proposes and creates an approach to enable the annotation of the traditional books in PDF format with abundant educational videos on YouTube.

Here the meaning of the word annotation is that when a student or reader has some doubts on some keywords or concept in a PDF book, then the relative chapter title has already been linked to a list of relevant YouTube videos which may help to understand the doubted keywords comprehensively.

As a result, in order to achieve the target descripted above, the techniques from a field of computer science called cross-media retrieval should be considered and utilized. More in details, the field cross-media retrieval actually belongs to a bigger field called information retrieval, which can be briefly summarized as the process of documents retrieval according to a specific query. And one of the most famous instances of the implementation of information retrieval is the web search engine, such as google search engine or baidu search engine.

Thus, the thesis first investigates and compare different types of methods that do cross-media in different ways, such as deep learning methods, learning to rank methods and graph-based methods.

There are many tackles along the way to achieve the mentioned goal, first, There isn't any pipeline or system that link the text in PDF books and videos from an online platform before, thus it should first be done by ourselves, which means we should first recognize the chapter titles (keywords or important sentences) in PDF format books, and then retrieve video results from YouTube using the extracted chapter titles as queries, and finally improve and select only best ones to be linked as the final results.

Thus, along the journey of this thesis, many different kinds of tools and methods have been utilized. First, in order to extract text contents from PDF books, a python library called pyPDF has been utilized. Then to limit the range of the search or do a customized search on YouTube, google custom search engine has been utilized. And to extract subtitles of videos, pyTube has been utilized. After all the previous mentioned steps, feature engineering work has been done and Xgboost models have been trained to classify whether a video result is matched or not with the specific text query automatically.

In the end, the cross-media retrieval has been done through all the steps mentioned above, and the quality of the ranking lists of the prepared queries from PDF books has been globally improved by moren than 10%.

A little bit more in details, there are mainly two different contributions of what we've done. The first one is from engineering aspect, the concrete engineering steps are:

1. Collection of books as resource of text query.
2. Randomly select 10 chapter titles from each book as queries.
3. Retrieve top 10 video results from YouTube automatically.
4. Implement a front-end representation which enable the direct representation to future users to see the relevant videos of text in books.

And the significant meaning of the engineering part is the automation of the whole process, which means it is feasible to design a similar but more end-to-end and robust system to provide the same kind of information service with high quality. And that's a requirement must be fulfilled if an engineering prototype project can be generalized as an educational or industrial service.

The second contribution is from cross-meida retrieval or application of machine learning algorithms perspective, the concrete steps are:

1. Extract sets of named entities from the text query in book and subtitles of returned result videos.
2. Construct subsets of named entity sets called instance sets.
3. Design and construct 24-dimensional feature vectors for each pair of a text snippet of the chapter titles in PDF books and a YouTube video.
4. Train Xgboost models to classify labeled datasets to determine whether the pair is matched or not.
5. Predict with test sets and filter out predicted not matched video results.
6. Calculate mean average precision of all the queries and compare it with the mean average precision of original ranking list from YouTube to check whether the quality of the returned ranking list is improved.

And the significant meaning of the information retrieval part is the feasibility of continuously improvement of search quality, which means besides the service has an end-to-end and robust system, and the searched relevant video results can be more and more matched to the specific text queries from books. As a result, this kind of service can be probable more and more helpful for the students or any kinds of readers since whenever they have some doubts on some words or concepts in a PDF or online book, this kind of service will provide relevant videos directly when they click the text in that book.

And in the machine learning model part, although there was a strong disadvantage on lacking of labeled data, **we could still get a lower bound above 0.7 of f1-score and obvious increases at least 0.1 on all mean average precision @k (k=1, 2, ..., 10).** And in our experiment case, there was sometimes a problem that the machine learning model will filter out most of the returned video results which lead to lack of video results for some specific query. But whenever we don't only consider the top 10 results of YouTube, for example, we make the pool of choices top 20 results of YouTube or even not only search in YouTube, also in some other video platforms such as Coursera, Udacity, then this problem will be solved properly.

As a conclusion, this work contributes both on the engineering aspect to provide a prototype of an automatic annotation system of text in books with videos and on the cross-media retrieval aspect to provide a way to extract features and train machine learning models which filter out not matched results so that the quality of the result lists can be improved.