M2 NLP

UE 901 EC3: Intelligent Systems and Recommendations

Project Report - Part 2

Authors:
Asmaa Demny
Fatima Habib
Cécile Macaire
Chanoudom Prach
Ludivine Robert

February 1, 2021

Contents

1 Introduction **2**

2 Data analysis and characterization **3**
2.1 Data structure & analysis 3

3 Concurrence study **5**
3.1 Approaches 5
3.2 Common Evaluation Metrics 6
3.3 Future work 7

References **8**

1 Introduction

In the context of a job offer from Amazon, we want to build an efficient recommender system which consist mainly of "*filtering information that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications*". This project aims to build a recommender system that will take a Wikipedia page as an input, and will recommend 10 new links based on information that will be defined in this report.

The following report will present the main points:

1. Data analysis and characterization,
2. Concurrence study (similar contexts),
3. Proposition of our recommender system strategy with the librairies,
4. The choosen approaches,
5. Analysis of the whole solution and conclusions.

2 Data analysis and characterization

From the ENWIKI DUMP PROGRESS ON 20210101[1] which is a Wikimedia dump service, we decided to take the data from:
`enwiki-20210101-pages-articles-multistream12.xml-p8554860p9172788.bz2`.
The following sections will provide information about the data structure, some statistics and the first ideas to answer the given problem.

2.1 Data structure & analysis

The data is a XML file of 166.3 MB. The XML file starts with a <mediawiki> tag which encompasses the metadata. The source language is given with the parameter *xml:lang*, here English. The recommender system will therefore only propose pages in this language.
The header of the file shows the site info in the tag <siteinfo> such as the site name ('Wikipedia'), the database name ('enwiki') and the namespaces. Each wiki page is included into a <page> tag.

```
  <page>
    <title>Colegio de Santa Cruz de Tlatelolco</title>
    <ns>0</ns>
    <id>8554864</id>
    <revision>
      <id>965714004</id>
      <parentid>934400346</parentid>
      <timestamp>2020-07-03T00:06:35Z</timestamp>
      <contributor>
        <username>GreaterPonce665</username>
        <id>30826712</id>
      </contributor>
      <minor />
      <comment>{{start date and age}}</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="16412" xml:space="preserve">{{Infobox university
|name              = Colegio de Santa Cruz de Tlatelolco
|image =
File:Iglesia_de_Santiago_Tlatelolco,_M%C3%A9xico_D.F.,_M%C3%A9xico,_2013-10-16,_DD_38.JPG
|native_name       =
|motto             =
|established       = {{start date and age|1536|1|6}}
|type              = [[Catholic education|Catholic]]
|city              = [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico City]]
|country           = [[Mexico]]
|campus            = [[urban area|Urban]]
}}
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 31.JPG|thumbnail|
Exterior of the church]]
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 46.JPG|thumb|
View of dome from below]]
The '''Colegio de Santa Cruz''' in [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico
City]], is the first and oldest European school of [[higher learning]] in the
[[Americas]]&lt;ref&gt;{{cite book|url=https://catalog.hathitrust.org/Record/101392426|
title=The first college in America: Santa Cruz de Tlatelolco.|location=Washington DC|
year=1936|author1=Steck|author2=Francis Borgia}}&lt;/ref&gt; and the first major school of
interpreters and translators in the [[New World]].&lt;ref&gt;{{cite book|chapter-
```

Figure 1: Extract from the XML file of a Wikipedia page.

[1] https://dumps.wikimedia.org/enwiki/20210101/

From Figure 1, the most important information that we find for each Wikipedia page are:

- the title (<title>),
- the ID (<id>),
- the parent ID (<parentid>),
- the contributor (<contributor>),
- the text (<text>) which contains for some pages the infobox, the different sections identified by '==History==', and references (<ref>).

From the title page, we can access the corresponding Wikipedia page by adding underscores instead of spaces. For example, to access 'Colegio de Santa Cruz de Tlatelolco' page, the URL is `https://en.wikipedia.org/wiki/Colegio_de_Santa_Cruz_de_Tlatelolco`.

The XML file has 171742 Wikipedia pages. Only 5 pages have an empty content from the <text> tag. They will therefore not be used to build the recommender system.

3 Concurrence study

3.1 Approaches

The following section presents the current approaches that exist to implement a recommender system on content-based filtering.

Doc2Vec [6] is a method to obtain content-based representations of document data to a vector space model. The algorithm is based on word2vec which represents word in a vector space model, and adopted to sentences, paragraphs, and documents. It uses neural network to save the context and the semantic words unlike traditional bag-of-words approaches which eliminate the context.

K-Nearest Neighbors (KNN) is a supervised classification algorithm which is based exclusively on the choice of the classification metric. It is a non-parametric machine learning method which assumes that similar things are closed to each other. The idea behind KNN is that, from a labeled database, we can estimate the class of a new data item by looking at what is the majority class of the k closest neighboring data (hence the name of the algorithm). The only parameter to set is k, the number of neighbors to consider. In the work of [1], they proved that a Real-Time recommendation engine powered by KNN classification model implemented with Euclidean distance method is capable of producing useful, quite good and accurate classifications. Also, competent of recommendations to the client at any time based on his immediate requirement rather than information based on his previous visit to the site.

TF.IDF short for **Term Frequency-Inverse Document Frequency**, is a statistical measure intended to evaluate how a word is relevant to a document among a collection of documents. A study in 2016 [2] showed that 83% of text-based recommender systems in digital libraries use tf-idf. It also has shown its efficiency in information retrieval and text mining, where variations of the tf-idf scheme are used by search engines as an essential tool in scoring and ranking a document's relevance given a use query. It can be calculated by multiplying two metrics: the term frequency and the inverse document frequency.

- **Term Frequency** (TF) measures the frequency of occurrence of the terms in the document.
- **Inverse Document Frequency** (IDF) measures how important a term is. While computing tf, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones.

Count-vectorizer is provided by scikit-learn library and converts a collection of text to a matrix of token counts.

Graph-Convolution-based approach, its acronym of GCN, is an approach for semi-supervised learning on graph-structured data. Handling directly on graphs, the GCN approach is relying on an efficient variant of Convolutional Neural Networks (CNN), and the alternative of the architecture of CNN is stimulated through localized first order approximation of spectral graph convolutions. Moreover, in the number of graph edges, its model scales linearly and learns hidden layer representations that encode both features of nodes and local graph structure [5].

Latent Dirichlet Allocation (LDA) is a generative statistical model which explains observation groups by unobserved ones. This is the reason why certain part of the data are similar.

It has already been used to generate Wikipedia links [3].

Bag of Words (BoW) is a method to represent text data with the algorithm of machine learning. It is easy to understand and implement. Moreover, it is known for its success in talking about the problem in language modeling and document classification. The method represents text which describes the occurrence of words in the documents which relates to a known word in the vocabulary and the presence of known words in the measurement.

The similarity matrix is to represent the distance between two sets of data, and its results of the distances are mapped to the value to create the representation.

In [7], they presented a **deep learning architecture with Recurrent Neural Networks** in a top-N content-based recommendation scenario. In more details, they proposed a deep architecture based on Long Short Term Memory (LSTM) networks to jointly learn two embeddings, the first one representing the items to be recommended, and the second one to encode the preferences of the user. From these two embeddings, they calculated the relevance score of each item for a specific user thanks to a logistic regression layer and were able to return the top-N items as recommendations. By comparing there work with two baselines based on neural networks, Word2Vec and Doc2Vec, and state-of-the-art algorithms for collaborative filtering, they showed the effectiveness of their approach. But such systems have some drawbacks. It requires a lot of data, extensive hyper-parameter tuning and the interpretability and explainability can be difficult.

3.2 Common Evaluation Metrics

Such systems require a proper evaluation. Many evaluation metrics are available for recommender systems and each has its own pros and cons. We present here a short summary of some of them.

Precision and Recall are binary metrics used to evaluate models with binary output. Table 1 present the results from a user that will be used to compute the precision and recall scores [4]. Precision is the ratio between the True Positives and all the Positives (TP / (TP+FP)).

	Recommended	Not recommended
Preferred	True-Positive (TP)	False-Negative (FN)
Not preferred	False-Positive (FP)	True-Negative (TN)

Table 1: Classification of the possible results of a recommendation of an item to a user.

Recall corresponds to the ratio between True Positives and True Positives + False Negatives (TP / (TP+FP)).

Mean Average Precision (MAP) computes the mean of the Average Precision (AP) over all the users. It takes in a ranked list of the N recommendations and compares it to a list of the true set of "correct" or "relevant" recommendations for that user.

Area Under The Curve - Receiver Operating Characteristics Curve AUC - ROC is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how

much the model is capable of distinguishing between classes.

Normalized Discounted Cumulative Gain (NDCG) is the metric of measuring ranking quality. It is mostly used in information retrieval problems such as measuring the effectiveness of the search engine algorithm by ranking the articles it displays according to their relevance in terms of the search keyword.

Mean Absolute Error (MAE) measures the errors between observed pairs expressing the same phenomenon.

Root Mean Square Error (RMSE) measures differences between values predicted (by a model) and the values observed.

Novelty and Diversity are the key dimensions of recommendation utility, and a fundamental research direction to keep making progress in the field.

Content-based filtering is useful when they do not exist any suggestions; and do not depend on the users. Moreover, the recommendations are all based on known features. Nevertheless, it is limited on the information available in the content and does not allow suggestions on new things.

3.3 Future work

Because we want to predict the 10 most related Wikipedia pages from a page, we will use the content-based filtering. In more details, we will take the content of the input page from the different tags and generate links by the future recommender system.

Based on the concurrence study, we plan to use Doc2Vec and KNN approaches. Indeed, Doc2Vec computes a representation of the document (here the page) in a vector space and, thanks to neural networks, will compute the similarity between the vectors. In the other hand, KNN is a powerful machine learning technique which tries to find the closest items. With this approach, we will have to define and extract the features to feed the algorithm. These features will be taken from the content of each page.

References

[1] David Adedayo Adeniyi, Zhaoqiang Wei, and Y Yongquan. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method". In: *Applied Computing and Informatics* 12.1 (2016), pp. 90–108.

[2] Joeran Beel et al. "Research-paper recommender systems : a literature survey". In: *International Journal on Digital Libraries* 17.4 (2016), pp. 305–338. ISSN: 1432-5012. DOI: 10.1007/s00799-015-0156-0.

[3] Karan Bhanot. *Building an Article Recommender using LDA*. https://towardsdatascience.com/lets-build-an-article-recommender-using-lda-f22d71b7143e. Online; accessed 31 January 2021. 2019.

[4] Asela Gunawardana and Guy Shani. "A survey of accuracy evaluation metrics of recommendation tasks." In: *Journal of Machine Learning Research* 10.12 (2009).

[5] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *arXiv preprint arXiv:1609.02907* (2017).

[6] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.

[7] Alessandro Suglia et al. "A Deep Architecture for Content-Based Recommendations Exploiting Recurrent Neural Networks". In: *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. UMAP '17. Bratislava, Slovakia: Association for Computing Machinery, 2017, pp. 202–211. ISBN: 9781450346351. DOI: 10.1145/3079628.3079684. URL: https://doi.org/10.1145/3079628.3079684.