Similarity-Based Clustering of Pre-Modern Arabic Names

Yousef, Tariq
yousef@saw-leipzig.de
Saxon Academy of Sciences and Humanities in Leipzig, Germany

Kinitz, Daniel
kinitz@saw-leipzig.de
Saxon Academy of Sciences and Humanities in Leipzig, Germany

Introduction

Data repositories must manage the identity of their entities. In the case of intellectual history, the challenge lies in premodern, and therefore non-standardised entity names. Our use case deals with Arabic persons related manuscripts (scholars, scribes, etc.). Thus, multiple occurrences of the same person with different spellings and name compositions must be identified and disambiguated. This paper presents a graph clustering approach that combines literal and numerical properties (name and year of event) with promising results. The particular challenge lies in the vast variability of name variants and sometimes unspecific dates.

Data

In this study, we used five different data sources – three manuscript catalogue data repositories and two reference works on figures from Arabic intellectual history [1]. The dataset of persons contains over 94,000 records, some of which are duplicates. However, classical Arabic names of persons have domain-specific characteristics. They can consist of up to six typical elements (prename, daughter/son of, known as, etc.). Premodern names are non-standardised, i.e. one and the same person can occur in different name variants, including transpositions of the name elements. Furthermore, since Arabic was widespread throughout the Islamicate world and included in non-Arabic Muslim cultures, names were adjusted to local standards of spelling and language conventions [2].

Pipeline

Figure 1 illustrates our processing pipeline, which consists of four main components. We started with normalising the names using diverse normalisation functions. Then, we built the similarity graph and employed clustering to group similar and related names in clusters. Further, we used a user interface to confirm or correct the resulting clusters.

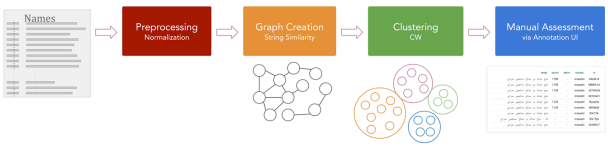

Figure 1: Processing pipeline

Preprocessing

The aim of normalisation is to reduce the heterogeneity among the names since they are collected from different resources. We applied normalisation on character and word levels. Besides NFC Unicode normalisation, we removed vocalisation and unified the charset, e.g. character equivalents in Persian and Arabic (ک vs. ك) and (Farsī ی → Arabic ي). Moreover, we removed honorary titles such as (ḥāǧǧ, šayḫ, mullá) and unified the writing of Arabic “son of” (نبا → نب).

Graph Creation

We calculated the similarity between every two normalised names in the corpus. For this purpose, we utilised a combination of several textual similarity metrics as well as numerical entity features, such as birth year and death year. Then, we modelled the obtained results as a graph by representing each name as a node and the combined similarity between two names as an edge between them, if it is greater than a predefined threshold. For textual similarity, we used a combination of **token-based metrics** such as the Jaccard Index and Sorensen Dice, **vector-based metrics** such as Term Frequency - Inverse Document Frequency (TF-IDF) with Cosine Similarity, and **edit distance-based metrics** such as Levenshtein, Hamming, Jaro-Winkler or Ratcliff-Obershelp. Our experiments showed that no single metric consistently achieves good performance, but combining metrics from different categories produces better results. For instance, token-based metrics work very well in the case of transpositions, but they fail to capture the actual similarity when there are minor differences among the tokens. However, edit distance metrics behave the opposite. TF-IDF with Cosine Similarity works very well with person names which contain rare elements.

Clustering

To group identical and similar entities, we employed the Chinese Whispers (CW) clustering algorithm. CW is a hard partitioning, randomised graph-clustering algorithm. CW does not require a predefined number of clusters, and can quickly find clusters in an extensive network, since its processing time increases linearly with the number of nodes (Biemann 2006). Moreover, domain experts can manually evaluate and correct the resulting clusters via a web-based user interface.

Evaluation

Due to the absence of a gold standard data set we used a similar, but more homogeneous dataset on premodern Arabic persons for quantitative evaluation [3]. We run the pipeline with different similarity metrics and similarity thresholds and used the traditional

clustering evaluation metrics such as Homogeneity, Completeness, and V-measure to compare the evaluation results. The results revealed that TF-IDF generally showed the best performance with 0.6 as threshold and achieved a significantly high Homogeneity (0.992), Completeness (0.958), and V-measure (0.975).

Acknowledgments

Our colleague Thomas Efer (Leipzig University) participated in a previous presentation. Funding from the German Academies' Programme and the Free State of Saxony.

Notes

1. Qalamos.net, Syrian National Library, Mar`ašī manuscript library in Qom/Iran; reference works: (Müller et al. 2018) and (Al-Ziriklī 2002). Data was engineered/refined at Bibliotheca Arabica (Brinkmann/Löhr 2021).
2. E.g. the famous Shi`ī scholar al-`Allāma al-Ḥillī (d. 1325 CE) could become `Allāme-ye Ḥillī in Persian (no definite article, genitive construction).
3. A subset taken from the virtual Hill Museum & Manuscript Library (hmml.org), containing 3000 names grouped into 893 clusters.

Bibliography

Biemann, Chris (2006): “Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems”. In Proceedings of TextGraphs: the first workshop on graph-based methods for natural language processing, pages 73–80, 2006.

Brinkmann, Stefanie / **Löhr, Nadine** (2021): “Bibliotheca Arabica – Towards a New History of Arabic Literature”. Comparative Oriental Manuscript Studies Bulletin: 7. pp. 197-206.

Müller, Christian / **Roiland, Muriel** / **Sublet, Jacqueline** (2018): “Onomasticon Arabicum Online”. https://onomasticon.irht.cnrs.fr, 2018.

A-Ziriklī, Ḫayr al-Dīn (2002): al-A`lām. 8 vols., Beirut: Dār al-`Ilm lil-Malāyīn, 2002. The initial full text is taken from shamela.ws.