RESEARCH-ARTICLE

Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning

ANIL OZDEMIR, Sabancı University, Tuzla, Istanbul, Turkey

BERKE ODACI, Sabancı University, Tuzla, Istanbul, Turkey

LORANS TANATAR BARUH

ONUR VAROL, Sabancı University, Tuzla, Istanbul, Turkey

SELIM BALCISOY, Sabancı University, Tuzla, Istanbul, Turkey

Published: 18 December 2025
Online AM: 01 July 2025
Accepted: 09 June 2025
Revised: 29 March 2025
Received: 20 October 2023

Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning

ANIL OZDEMIR, Sabancı University, Istanbul, Turkey
BERKE ODACI, Computer Science, Sabanci Universitesi Muhendislik ve Doga Bilimleri Fakultesi, Istanbul, Turkey
LORANS TANATAR BARUH, Salt Istanbul, Istanbul, Turkey
ONUR VAROL, Sabancı University, Istanbul, Turkey
SELIM BALCISOY, Computer Science, Sabanci Universitesi Muhendislik ve Doga Bilimleri Fakultesi, Istanbul, Turkey

Recent efforts to digitize textual, visual, and physical forms of cultural heritage require advanced tools for preservation and analysis. The availability of extensive online data creates a need for intelligent systems to help users and archivists understand latent relationships in these collections. A major challenge in cultural heritage studies is the labor-intensive process of analyzing these materials. Inconsistent linguistic terms and ambiguous concepts in digital documents make it difficult to uncover relationships without expert supervision. Moreover, while advanced models based on large-scale pretraining demonstrate strong performance in extracting semantic relationships, they depend on extensive pretraining on large external datasets, limiting their applicability for smaller or specialized collections. We propose a system that combines natural language processing for entity extraction with graph representation learning to model relationships among documents, categories, and n-grams, resulting in a fully connected network representation. Unlike methods requiring large-scale pretraining, our approach operates effectively using only the information available in the dataset itself, making it particularly suited for smaller cultural heritage document collections. The system extracts significant terms from document metadata, produces embeddings for each document, and uses these embeddings to build a recommendation system for entity discovery. We tested the system on a collection of early 20th-century documents from Crete, evaluating its performance against alternative methods in collaboration with experts from the archival research organization SALT. This approach not only facilitates deeper insights into smaller, specialized collections but also reduces dependency on vast external training resources, enhancing its practical utility in cultural heritage studies.

CCS Concepts: • **Computing methodologies** → **Machine learning**;

Additional Key Words and Phrases: Natural Language Processing, Machine Learning, Graph Representation Learning, Recommendation Systems

Authors' Contact Information: Anil Ozdemir, Sabancı University, Istanbul, Turkey; e-mail: aozdemir@sabanciuniv.edu; Berke Odaci, Computer Science, Sabanci Universitesi Muhendislik ve Doga Bilimleri Fakultesi, Istanbul, Turkey; e-mail: berkeodaci@sabanciuniv.edu; Lorans Tanatar Baruh, Salt Istanbul, Istanbul, Turkey; e-mail: lorans.baruh@saltonline.org; Onur Varol, Sabancı University, Istanbul, Turkey; e-mail: onur.varol@sabanciuniv.edu; Selim Balcisoy (corresponding author), Computer Science, Sabanci Universitesi Muhendislik ve Doga Bilimleri Fakultesi, Istanbul, Turkey; e-mail: balcisoy@sabanciuniv.edu.

ACM 1556-4711/2025/12-ART60
https://doi.org/10.1145/3746658

ACM Reference format:
Anil Ozdemir, Berke Odaci, Lorans Tanatar Baruh, Onur Varol, and Selim Balcisoy. 2025. Enhancing Cultural Heritage Archive Analysis via Automated Entity Extraction and Graph-Based Representation Learning. *ACM J. Comput. Cult. Herit.* 18, 4, Article 60 (December 2025), 25 pages.
https://doi.org/10.1145/3746658

1 Introduction

Today, most **Cultural Heritage (CH)** institutions have started to digitize parts of their collections and archives to improve accessibility, preservation of originals, publicity, and visibility of the institution on the Internet. With this recent development, digital document collections have been multiplying. These collections include contents from libraries, galleries, museums, and archives [86].

However, the increasing demand for digitalization in the CH domain brings several challenges for analyzing documents in this field. One of the primary challenges is that reading, analyzing, and processing these documents are time-consuming and require trained personnel and adequate funding. During such processes, critical entities within the documents may not be consistently expressed in the same linguistic terms, or they may contain ambiguous concepts. As a result, uncovering these relationships without careful examination by experts is challenging, and much information may be overlooked in manual analysis. These challenges necessitate approaches that enable users to understand latent meanings in collections, discover and investigate relationships, and extract necessary information from collections in CH.

In response to these challenges, various studies have been conducted to assist archivists by solving the mentioned problems in the field of CH. These studies have utilized methods from multiple disciplines. Most of these methods employ **Natural Language Processing (NLP)** techniques, such as **Named Entity Recognition (NER)** [56], to identify and classify notions, events, and concepts within the collections. Ehrmann et al. [24] provide a comprehensive survey of NER techniques tailored for historical documents, addressing the unique challenges posed by noisy and domain-specific text. Topic modeling techniques like **Latent Dirichlet Allocation (LDA)** [6] and Latent Semantic Analysis [23] are also used to understand underlying themes, time periods, and temporal expressions in textual collections. For instance, Andresel et al. [2] utilized LDA to cluster and curate multilingual and heterogeneous historical collections, addressing challenges in repositories like Europeana. Additionally, visualization tools and interfaces are employed to tackle these issues. Examples include MindMap [74], which uses visualization and taxonomies to interpret textual collections, and InfoTouch [40], an exploratory visualization interface for photo collections. These advancements in collection and document research prompt us to focus on improving corresponding techniques within the CH domain.

In this work, we introduce a novel system utilizing a collection of digitized documents in the field of CH. To access such a collection, we collaborated with archive professionals from SALT (saltonline.org), a cultural institution dedicated to public service through research-based exhibitions, publications, and digitization projects. After discussions with the professional archivists at SALT, we decided to use the Waqfs of Crete archive, which consists of official records of Muslim inhabitants of Crete who relocated to Turkey in the 1920s due to the population exchange between Turkey and Greece. The documents are in Modern Turkish, but many words have Ottoman origins. The archive also includes summaries of the documents, their dates, and metadata about their categories.

We developed a document recommendation system that combines various NLP techniques and heterogeneous network embedding methods (sumgrams[1] [59], Metapath2vec [21], skip-gram [52]) using only the textual summaries and metadata of the documents. First, the system extracts n-grams that describe linguistically meaningful phrases and combines them with temporal and categorical information to represent them in a heterogeneous

[1]https://github.com/oduwsdl/sumgram.

network. Then, the system produces vector representations for the documents in the collection by utilizing graph representation learning algorithms. These representations are used for visualization and in designing a document recommendation system. The vector embeddings reveal relationships between the documents.

Finally, we conducted experiments with professional archivists to test the performance of the proposed recommendation system. The results showed that our system provided more relevant recommendations compared to baseline models such as TF-IDF and Doc2Vec. Domain experts classified the model's recommendations as either "relevant" or "irrelevant" based on their professional judgment. While the experts have deep knowledge in their respective domains, their different opinions on the relevance of recommended documents highlight the real-world complexity of document recommendation systems.

2 Related Works

There has been a significant effort in the digitization, exploration, and preservation of CH. Most of this work has been focused on creating digital representations of cultural artifacts and the creation of metadata and documentation associated with this. This effort's significant consequence is an enormous collection of documents digitally available in text, images, and other representations. Three examples include the Biodiversity Heritage Library [29], which is an open-access digital library of 12 natural history collections; a digital archive of the Ottoman/Turkish serial novel that is part of a broader resource that preserves, classifies, and analyzes the CH of Ottoman/Turkish society [70]; and the Arts and Humanities Data Service, [11] a UK national service aiding the discovery, creation, and preservation of digital resources covering areas such as archaeology, history, literature, and linguistics.

The increase in the number of these archives and collections affects the number of works in exploration, visualization, and relationship extraction projects. Using such digital collections has increased considerably in the last 20 years. Examples of these include EPOCH [62], which is a multimodal interface for interacting with digital heritage artifacts by using virtual reality, and CULTURA [34], which is a project that covers topics such as finding entity, event, and relationship extraction within unstructured text obtained from digital CH collections using various NLP techniques. In addition, Dou et al. [22] propose a knowledge graph to find ontology and relationships for intangible CH. Moreover, Aviles Collao et al. [4] describe tools that produce similar cluster representations in CH content using spatial vector-based computation technique, and Salisu et al. [67] introduce visual techniques to depict development and changes over time in CH collections.

At the same time, with the spread of recommendation systems technology, the number of studies in the field of the CH of the recommendation system began to increase gradually [48, 60]. However, despite the increase in the number of studies in this area, there are limited resources specifically addressing the application of recommendation systems on historical documents. Recently, Pavlidis [60] reviewed the examples of recommendation system technology in the field of CH and gave examples of the recommendation system studies aimed at increasing user engagement in museum visits and mentioned their limits and assumptions. In this study, a document recommendation system has been designed to explore and understand the documents in the field of CH. The proposed system uses the representation learning method on the graph network structure.

2.1 Representation Learning for Graph Networks

Representation learning approaches extract beneficial information for various entities presented in a machine learning model [5, 77]. Good representations capture the posterior distribution of the underlying explanatory factors for the observed input. Recently, the use of representation learning has grown rapidly across different fields, including speech recognition, signal processing [14, 16, 19], object recognition [25, 83], and NLP [15, 44, 53, 78].

Several approaches have been proposed to learn representations that encode structural information in graphs [12, 32, 33]. The principal idea is to learn a mapping function f that embeds nodes, links, or entire (sub)graphs as

points in a low-dimensional vector space $\mathbb{R}^d$. This function optimizes geometric relationships reflecting the original graph's topology. The optimized embeddings can be used in many downstream tasks, including classification, clustering, and similarity search [33]. Representation learning on graph-structured information includes latent and observable relational learning models [58], latent space approaches to social network analysis [36], and geometric deep learning on graphs [9, 55].

Traditional methods for extracting latent space for graph-structured information are based on Eigendecomposition, which becomes increasingly complex with larger graphs. This inefficiency led to the development of random walk strategy-based representation learning algorithms, such as metapath2vec [21].

DeepWalk [61] was the first method to combine random walk strategy and skip-gram [52] for learning latent representations, treating walks as the equivalent of sentences in Word2vec [54]. DeepWalk optimizes for high probabilities of vertices in the neighborhood, but LINE [76] argued that it lacked a clear objective for network properties. Node2vec [26] generalized DeepWalk and LINE, introducing parameters to control random walk behavior and explore neighborhoods using Breadth-First and Depth-First sampling.

While these methods focused on homogeneous networks, Metapath2vec [21] addressed representation learning in heterogeneous networks. It modified the random-walk strategy with a meta-path-based concept, constructing a heterogeneous neighborhood and using a heterogeneous skip-gram model to extract embeddings. Metapath2vec captures both semantic and structural relations in heterogeneous networks, making it the most accurate option for the heterogeneous graph structure we constructed from the documents.

2.2 Representation Learning for Recommendation Systems

This work focuses on developing document recommendation systems in the field of CH, using representation learning techniques. A recommendation system predicts the "rating" or "preference" a user would give to a specific item [65]. The increase in studies on representation learning (Section 2.1) has led to their use in recommendation systems across various industries [92], including video and music services [80], online shopping [13, 43], social media [84], tourism [88], news [93], and document recommendation [85].

Common approaches include **Collaborative Filtering (CF)**, which uses shared user characteristics [68], and content-based (CB) methods [81], which rely on content and user profiles. Both face difficulties with the cold-start problem, where there is insufficient initial user data, reducing recommendation effectiveness [45]. Hybrid approaches have been proposed to improve recommendations [45].

2.2.1 Document Recommendations. Early works include the Fixit system [35], which provides query-free search benefits, and a system by Budzik and Hammond [10] that recommends relevant documents during web browsing. Weng and Chang [85] used ontology and spreading activation models for research paper recommendations, addressing the cold-start problem (see Section 2.2). Nagori and Aghila [57] proposed a hybrid model using LDA [7] and CF techniques. Shaparenko and Joachims [72] used language modeling and convex optimization for top-N document recommendations based on cosine similarity.

2.2.2 Representation Learning in Document Recommendation Systems. Representation learning has become common in document recommendation systems. Guan et al. [27] introduced a graph-based algorithm for social tagging services, outperforming traditional methods. Yang et al. [87] incorporated text features into graph representation learning using matrix factorization. De Boom et al. [18] proposed a framework for short texts using weighted word embedding aggregation. Gupta and Varma [28] combined network and semantic embeddings for scientific paper recommendations. Zhang et al. [91] proposed a framework for top-N recommendations using heterogeneous information, capturing word sequence and local semantics. Studies by Kong et al. [37] and Brochier et al. [8] also achieved state-of-the-art results with network and textual embedding methods.

3 Background

3.1 The Waqfs of Crete Archive

The Waqfs of Crete archive, managed by the Directorate of the Pious Foundations, is a significant collection that includes official records of the Muslim inhabitants of Crete who relocated to Turkey during the 1920s population exchange between Turkey and Greece. This archive, spanning from 1825 to 1928, provides a comprehensive view of the island's multi-layered social structure, especially from cultural and economic perspectives. The documents offer insights into the communal issues faced by the Muslim population of Crete and trace the impact of political and administrative changes over the years [38, 95].

This extensive archive is categorized by the cities where the Pious Foundations operated, depicting their role in managing both urban and rural endowments, such as mosques, shop quarters, water deposits, fountains, and mansions. By participating in the local economic life of Cretan Muslims and establishing relationships within the community and with the island's government, the documents enable us to reconstruct the socio-economic history of the island at the turn of the 20th century. The archive includes approximately 100,000 digitized images that bring to light intriguing data regarding the island's bi-communal dynamics, illustrating the changes in societal relations following the establishment of the autonomous regime and the unification with Greece [95].

These 15,200 documents in Ottoman Turkish and Greek have been accessible online at SALT Research since 2013.[2] They were cataloged through a joint project with the Department of History and Archeology of the University of Crete and the FORTH foundation between 2005 and 2009. Greek documents were read and cataloged by students from the University of Crete, while scholars from Istanbul worked on the Ottoman script documents, with the Department of History of Boğaziçi University as the local partner. The titles of the documents in Ottoman Turkish were transcribed into modern Turkish, and those in Greek were translated into English [89, 90].

As a result of this collaboration, three theses and several articles were published [39, 89, 90]. Some copies of the documents were exhibited in the History Museum of Hania during the "Economy and Society in Crete" exhibition in 2013 and 2014. However, the specificity of the material limited its use to academicians and experts. To bring this archive to a broader audience, this research reported in this article was initiated in 2021.

The metadata was analyzed, focusing on documents written in Ottoman Turkish, totaling 10,200 records. These records were categorized by cities, dates, and thematic subcategories and visualized using graphics, providing archivists with valuable numeric data on the archive's composition and its temporal and categorical variations. Additionally, a small number of records from other Greek centers such as Rhodes, Lesbos, Thessaloniki, Kavala, and Alexandroupolis were identified.

3.2 Graphs and Networks Science

Graphs are a popular data structure and a universal model for describing many complex systems in physical, biological [49, 71], social, and information systems [1]. From the broadest perspective, a graph represents a collection of objects that are called nodes (vertices), along with interactions that are called edges between pairs of these objects. The term network is occasionally described as a graph in which attributes correspond to set of nodes and edges, but networks can capture more detailed properties of connections and nodes.

The fundamental terms and concepts related to networks are described below, along with their formal definitions and explanations:

— *Graph Structure.* Formally, a graph G is represented as $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges connecting these nodes.
— *Nodes and Edges.* A graph is defined by a non-empty set of nodes (vertices), denoted as $u \in V$, and a non-empty set of edges, denoted as $e \in E$, which connect these vertices.
— *Endpoints of an Edge.* Each edge $e \in E$ is associated with one or two vertices, referred to as its endpoints.

[2] https://archives.saltresearch.org/handle/123456789/496.

— *Adjacency and Neighbors.* If node u is adjacent to node v, they are connected by an edge e. Two adjacent nodes are also called *neighbors*.
— *Adjacency Matrix Representation.* A graph can be represented as an *adjacency matrix*, $\mathbf{A} \in \mathbb{R}^{\|V\|\times\|V\|}$, where $\mathbf{A}[u, v] = 1$ if $(u, v) \in E$, and $\mathbf{A}[u, v] = 0$ otherwise. For a simple undirected graph, the adjacency matrix is symmetric.

Although there are different types of graphs, undirected simple graphs were chosen for this study because co-occurrence networks are inherently undirected, representing mutual associations between entities without implying directionality. Additionally, using simple graphs ensures clarity by avoiding multiple edges or self-loops, which are unnecessary for analyzing co-occurrence relationships.

3.2.1 Multi-Relational Graphs. An important factor for distinguishing different graphs is the type of interactions between nodes or edges. For example, graphs can have multiple node types as in bipartite graphs or multiple edge types as in multigraphs. Edge types can specify different relations between nodes. The network formulations described earlier for simple graphs can be extended to capture more nuanced relations. At the same time, different adjacency matrices can be created for different relation types. Such graphs are called multi-relational and can be expressed by an adjacency matrix $\mathbf{A} \in R^{\|V\|\times\|R\|\times\|V\|}$, where R is the set of relations [33].

3.2.2 Heterogeneous Graphs. In heterogeneous graphs, multiple types of vertices and edges exist. In other words, one can partition the set of nodes U into disjoint sets, such property can be expressed as $V = V_1 \cup V_2 \cup V_3 \cdots \cup V_n$ where $V_i \cap V_j = \emptyset$ and $\forall i \neq j$. Then, any heterogeneous network can be expressed as a graph in the form of $G = (V, E)$ consisting of a non-empty object set V and a non-empty set of link E. Each node $v \in U$ and each link $e \in E$ is associated with a mapping function $\phi(v) \rightarrow T_v$ and $\phi(e) \rightarrow T_e$, where T_v and T_e denotes the type of v and e, $|T_v| + |T_e| > 2$.

In this study, the constructed network is a specialized heterogeneous information network categorized as a *Multi-partite* graph, where edges can only connect vertices of different types, as described in [33].

3.3 The Metapath2vec Framework

The Metapath2Vec [21] algorithm is frequently used to capture complex relationships between different types of entities and connections. It employs a deep learning approach inspired from word2vec [54] model. This approach (i) determines a meta-path that follows a distinct set of entities, (ii) generates a corpus using random walks following meta-paths, and (iii) trains a skip-gram model that incorporates heterogeneous network structures. As discussed in Sun and Han [75], meta-path scheme ρ can be expressed as the following:

$$v_1 \xrightarrow{R_1} v_2 \xrightarrow{R_2} .. \xrightarrow{R_{n-1}} v_n, \tag{1}$$

where $v_n \in V$ denotes the type of nodes and R denotes the set of composite relations along v_1 to v_n. In addition, as stated in [21], meta-paths are commonly used in a symmetric way, which means first node type v_1 should be the same with the node type last v_n in p.

Recently proposed Meta-path-based random walk strategy in [21] states that the semantic relationships between different types of nodes can be properly incorporated into skip-gram. Metapath2vec explains the meta-path random walk traversal protocol as given heterogeneous Network $G = (V, E, T)$ with $|T_V| > 1$ and meta-path scheme ρ. Transition probability at step i is described as follows:

$$p(v^i + 1|v_t^i, \rho) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) = t + 1 \\ 0 & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) \neq t + 1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases}, \tag{2}$$

where $v_t^i \in V_t$ and $N_{t+1}(v_t^i)$ indicate the V_{t+1} type of neighborhood of node v_t^i.

3.4 Skip-Gram

Recently, Mikolov et al. [52] proposed Skip-Gram, an unsupervised learning method designed to find the most related words for a given word in natural language texts. The Word2vec algorithm employs the Skip-Gram method, which aims to predict the surrounding words in a sentence based on a focus word and a context window that includes the surrounding words. This process generates embedding vectors for words. Mikolov et al. [54] stated that the objective of the Skip-Gram model is to maximize the average log probability of a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, where c represents the size of the training context.

$$\arg\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le j \le c, J \ne 0} \log P(w_{t+j} | w_t; \theta). \tag{3}$$

3.4.1 Skip-Gram Application on Homogeneous Networks. DeepWalk [61] and Node2vec [26] took advantage of random walks and the skip-gram model to learn the representation of a vertex that facilitates the prediction of its structural context—local neighborhoods—in a homogeneous network. Afterward, the objective becomes to maximize the network probability in terms of learning local structures, given a network $G = (V, E)$ as follows:

$$\arg\max_{\theta} \prod_{v \in V} \prod_{c \in N(v)} p(c|v; \theta), \tag{4}$$

where $N(v)$ is the neighborhood of node v in the network G, and $p(c|v; \theta)$ denotes the conditional probability of having a context node c given a node v.

3.4.2 Skip-Gram Application on Heterogeneous Networks. Finally, Dong et al. [21] develop the core idea behind Word2Vec by applying the Skip-gram algorithm to heterogeneous networks and introduce the heterogeneous skip-gram model to model the heterogeneous neighborhood of a node. As described by Dong et al., the objective of the proposed skip-gram is to maximize the probability of having the heterogeneous context $N_t(v), t \in T_V$ given a node v in heterogeneous Network $G = (V, E, T)$ with $|T_V| > 1$ as follows:

$$\arg\max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} \log p(c_t|v; \theta), \tag{5}$$

where $N_t(v)$ denotes neighborhood of v's with the tth type of nodes, and $p(c_t|v; \theta)$ is commonly described as a softmax function [21].

3.4.3 Generating Node Embeddings. As the last step, the generated heterogeneous graph is traversed according to the pre-defined meta path scheme ρ to generate sequences of nodes, including a node itself and its neighboring nodes. Such corpus of sequences of nodes is then given as input for the heterogeneous skip-gram model to produce embeddings for each node with the primary objective that nodes with similar node neighborhoods should produce similar vector embeddings. The extracted vector embeddings for nodes in a heterogeneous network are in the same vector space, even if they represent different node types.

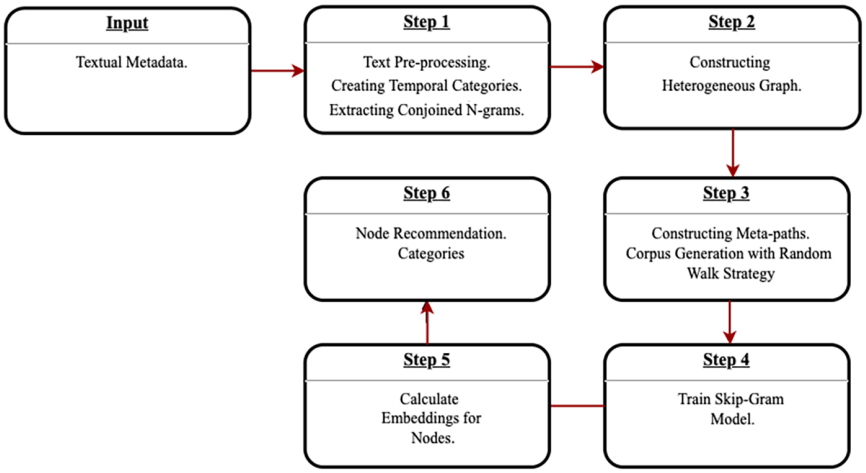
Formally, node embeddings are obtained by solving the problem described below,

Given a heterogeneous network G, the task is to learn d - dimensional latent representations $\boldsymbol{X} \in R^{|V| \times d}$, $d \ll |V|$.

The output of the problem described above is a low-dimensional matrix $\mathbf{X}$, where the vth row, a d-dimensional vector X_v, represents the embedding of node $v \in V$ in the network.

4 Methods

We construct a heterogeneous network structure using raw textual data and features extracted from metadata. To analyze this network, we employ Metapath2Vec, a state-of-the-art graph representation learning model. Representation for each node is obtained in the form of embeddings, which are later utilized for document



(a)

Fig. 1. System architecture of the proposed system. The flow follows the structure of the article, starting with preprocessing and feature extraction (Section 4.1), followed by the construction of the heterogeneous document network (Section 4.2), document representation learning (Section 4.3), and document recommendation (Section 4.4). The figure illustrates the high-level workflow without referencing specific subsections to maintain visual clarity.

recommendation and exploratory analysis. In this section, we present the approach we used for data preprocessing, transformation, and network representation learning.

The general architecture of the proposed system and the processes that transform data can be observed in Figure 1. Our proposed system uses textual data and the structured metadata accompanied with it as input. In the first step, textual pre-processing is applied to transform data into a more standard format. These preprocessing steps ensure that the text is structured in a consistent manner suitable for further analysis. Then, feature extraction, which includes extracting sumgrams and temporal categories, is performed on the processed texts. The extracted features are subsequently used in two key tasks: heterogeneous graph construction and representation learning for entities using the Metapath2Vec method. Finally, the produced representations are used for data exploration and document recommendation.

4.1 Pre-Processing and Feature Extraction

Previous experiments in the literature have demonstrated that different pre-processing methods could have a significant effect on the performance of many machine learning models [50]. These methods can enhance the performance of a model by reducing the dimensionality of data and also reduce the noise in the inputs [30]. In this study, we followed standard preprocessing steps such as converting text to lowercase, removal of symbols and punctuation, and tokenization.

4.1.1 Temporal Categories. Constructing a heterogeneous network is one of the major requirements to take advantage of the Metapath2vec method and the constructed graph in this study described in Section 4.2. In graph construction, the design of the node types to be used in graph generation is a very critical point. One method to be able to describe the concept of time on graphs was to create different graphs for different time intervals, but with this method, it would not be possible to create representations in the same vector space for nodes in different time zones.

Based on this motivation, we have produced new features that we called “temporal category” as a way of expressing both temporal information and categorical information in the same node type. What is meant by

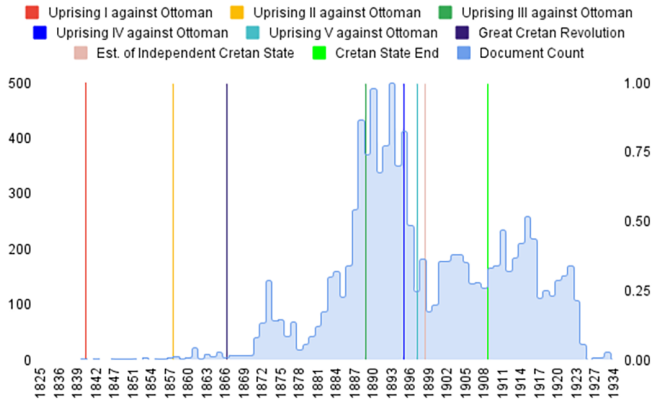


Fig. 2. Temporal characterization of document collection in the "Waqfs of Crete Dataset." Important events in Crete history have been represented as color-coded vertical lines.

"temporal category" is the name we give to obtain a new attribute by combining the attributes that contain temporal information and also categorical information in the data used. Combining temporal information and categorical would allow, for example, to filter documents on both categorical and temporal properties. To explain the applied technique with an example, Figure 2 illustrates the distribution of the years to which the documents belong, providing insight into the temporal characteristics of the dataset. As a result, years were divided into n different sets so that each set includes numerically close amounts of documents, and these were encoded according to categories (e.g., For the finance category, n distinct features have been created, such as *finance_1840_1860* or *finance_1860_1880*, which represent documents belonging to the finance category and written between 1840–1860 and 1860–1880, respectively.).

4.1.2 Conjoined N-Grams. In computational linguistics and probability, an n-gram refers to a contiguous sequence of n items from a given sample of text. Such items can be letters, words, phonemes, syllables, or base pairs, according to the application type [73]. In this study, n-grams refer specifically to words; that is, the smallest units that make up paragraphs and sentences. Latin numerical prefixes have been used to express n-gram types in the literature, such as an n-gram of size 1 is called "*unigram,*" n-gram of size 2 is referred to as "*bigram,*" size 3 is called "*trigram,*" and n-gram of size 4 is "*quad-gram,*" and so on.

The term *conjoined n-gram*, also known as *sumgrams,*[3] corresponds to the most frequent *n-grams* in the text collections. In other words, sumgrams refer to higher-order n-grams (e.g., "world health organization") generated by conjoining lower-order n-grams (e.g., "world health" and "health organization"). In this study, we generate conjoined n-grams of different n-gram classes (bigrams, trigrams, k-grams, etc.) as part of the document summaries, instead of limiting the summary to a single n-gram class (e.g., bigrams).

4.2 Constructing Heterogeneous Document Network

In this work, the constructed heterogeneous document network is a special type of heterogeneous graph categorized as a *multi-partite* graph, as discussed in Section 3.2.2. This type of graph can only contain edges that can connect to different node types. The proposed graph consists of three types of nodes and two types of edges, as shown in Figure 3. According to the specifications in Section 3.2.2, a heterogeneous network containing three different node types was created. Constructed network includes 13,547 conjoined n-grams, 7,336 documents, and

[3] https://github.com/oduwsdl/sumgram.

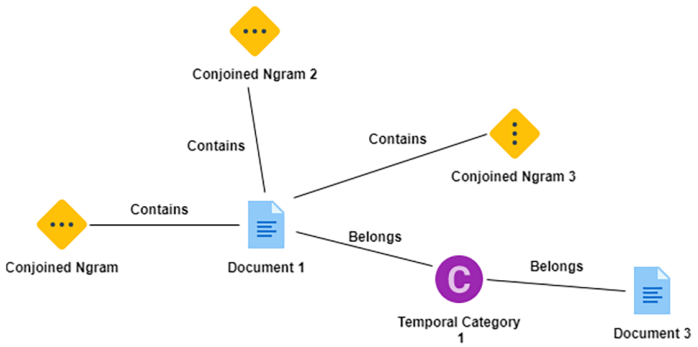


Fig. 3. Illustration of proposed undirected graph structure. Documents and their relationships extracted from the metadata and raw documents. Metapath2Vec approach learns representations from the illustrated network structure.

74 temporal categories, for a total of 20,957 nodes and 30,395 edges. Detailed node type and edge descriptions and their corresponding objects can be seen below:

— *Document.* Represents documents in the document collection used as inputs.
— *Conjoined N-Grams.* Represents n-grams extracted from texts (see Section 4.1.2).
— *Temporal Categories.* Generated from metadata of documents (see Section 4.1.1).

As can be observed in Figure 3, edge types are the following:

— *Document—Temporal Category.* Edge is constructed if the document is written in the category and time interval represented by the temporal category node.
— *Document—Conjoined N-gram.* Edge is constructed if document contains particular n-gram.

Figure 3 includes an example of how the proposed heterogeneous graph is constructed utilizing the previously mentioned node types and relations coupling them.

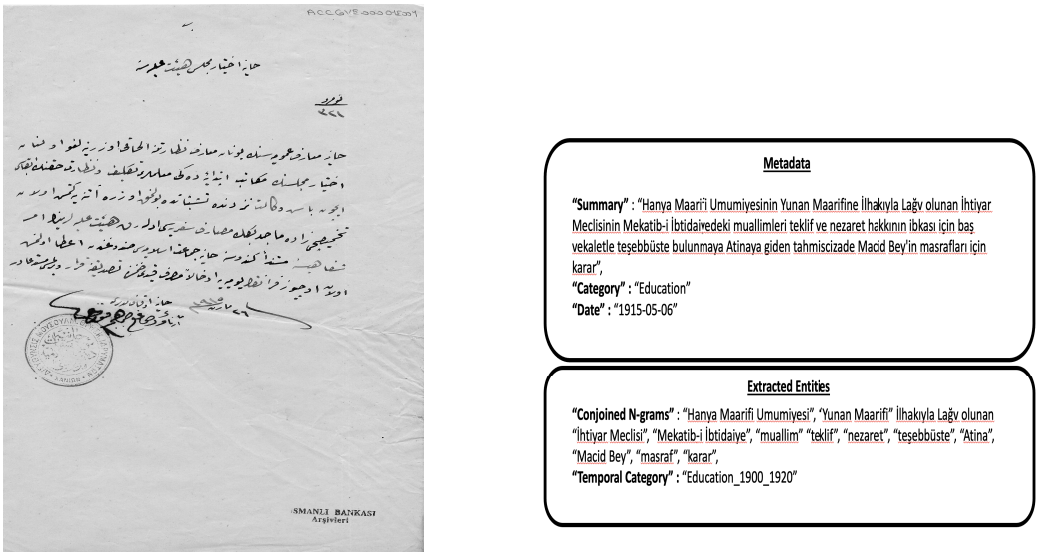
4.3 Document Representation

In this work, proposed system utilizes Metapath2vec to find the node representations corresponding to the documents in the data. In addition, to compare our system, we also obtained vector representations for nodes using different methods, including TF-IDF [47], Doc2Vec [42], and **Bidirectional Encoder Representations from Transformers (BERT)** [20].

4.3.1 Document Representation Using Metapath2vec. In order to obtain embeddings for different node types on the constructed heterogeneous network, the methodology presented by Dong et al. [21] and explained in the Section 3.3 was followed and implemented using Networkx [31], Gensim [63], and Stellar Graph[4] libraries.

Metapath2Vec algorithm offers significant advantages for our task. It not only generates vector embeddings for individual entities like documents and n-grams but also captures relationships between these entities through its meta-path-based approach. This ensures that contextual embeddings from document content are combined with structural associations derived from metadata, resulting in a fully connected network representation. These capabilities make Metapath2Vec particularly effective in modeling the many-to-many correspondences between documents, categories, and n-grams, addressing the sparsity and variability in CH datasets.

[4]https://github.com/stellargraph/stellargraph.



(a) (b)

Fig. 4. Example document from the dataset and its metadata. Original documents are coming from archives and mostly in handwriting form (left). Metadata partly created by experts for translations and categorization (right).

The meta-path scheme that we have defined in our system by following Equation (1) is as follows: Equivalent of the elements in the pattern in the text can be observed in Figure 4.

$$v_{N-gram} \xrightarrow{R_1} v_{Document} \xrightarrow{R_2} v_{Temporalcategory} \xrightarrow{R_2} v_{Document} \xrightarrow{R_1} v_{N-gram}.$$

During skip-gram and random walk implementation, we used the following hyper-parameters: number of length 30 walks ($l = 30$) per root node $w = 2$, neighborhood size $k = 10$, negative sample size 5, and vector dimensionality $d = 128$. These parameters are suitable parameters suggested by the Stellar Graph[5] library according to the structural properties of our data, such as corpus size.

4.3.2 Document Representation Using TF-IDF. TF-IDF is a numerical statistic that refers to the term frequency-inverse document frequency and accounts for the relative frequency of words in a particular document through an inverse proportion of the word along with all documents in a corpus. *TF* represents term frequency of term i in a document j, while *IDF* refers to inverse document frequency of term i [3, 69]. TF-IDF score of a word in documents expressed by Manning et al. [47] is as follows:

$$\mathrm{tf}(t,d) \equiv \frac{f_{t,d}}{\sum_{t' \in d}^{f_{t',d}}}, \tag{6}$$

where $f_{t,d}$ is the raw count of a term in a document.

To learn vector representation of document embeddings, we followed the same procedure expressed by Dai et al. [17] and treat the classical bag-of-words model where each word is represented as a one-hot vector weighted by TF-IDF. Finally, the document is represented by a constructed vector.

$$\mathrm{idf}(t,D) \equiv \log \frac{N}{|d \in D : t \in d|}, \tag{7}$$

[5] https://github.com/stellargraph/stellargraph.

where N represents the total number of document in the corpus $N = |D|$, and $|d \in D : t \in d|$ denotes the number of documents where term t appears.

$$\text{tfidf}(t, d, D) \equiv \text{tf}(t, d) \times \text{idf}(t, D). \tag{8}$$

4.3.3 Document Representation Using Doc2vec. Le and Mikolov [42] proposed Doc2vec, a generalized extension of the Word2vec method [54]. The authors seek to extend embedding learning from individual words to word sequences (e.g., paragraphs) by introducing a paragraph vector—an unsupervised learning method designed to generate vector embeddings for variable-length text segments, such as sentences and documents. Doc2vec proposes Distributed Bag-of-Word in the same way as skip-gram (Section 3.4), except that input is altered by token that represents the particular document. To learn vector representation of document embeddings, we followed the procedures that described by Le and Mikolov [42] and Lau and Baldwin [41].

4.3.4 Document Representation Using BERT. BERT model [20, 82] is a deep bidirectional encoder network based on a purely transformer architecture. BERT is capable to capture the bidirectional representations of texts by jointly conditioning on both the left and right context in all layers. Furthermore, a pre-trained BERT model can be applied to a broad range of tasks by fine-tuning with just one additional output layer. We used the publicly available BERTurk-Base cased model[6] that was pre-trained on unlabeled Turkish corpus, which has a size of 35 GB text. BERTurk model consists of 12 transformer layers and 768 hidden states size. We fine-tuned the model by adding one extra dropout and a fully connected layer that has 768 input and 2 output units at the end of the network and jointly trained it with the BERTurk model. We applied the sentence-transformation framework proposed by [64] on the top of the pre-trained BERTurk-Base cased model to obtain document embeddings. The proposed sentence-transformation technique enables the BERT model to compute dense vector representations for sentences, paragraphs, and images.

4.4 Document Recommendation

To perform document recommendation for a query document, the document embedding for the query was first calculated. Next, the closest Top-K document embeddings were identified using cosine similarity, and the corresponding documents were suggested as recommendations. We have shown in previous sections that the Metapath2vec algorithm is used to calculate these embeddings. To compare the proposed system with baselines, we also calculated embeddings and performed recommendation for documents using TF-IDF, Doc2vec, and BERT (see Sections 4.3.2, 4.3.3 and 4.3.4).

4.4.1 Top-K Document Recommendation Using Metapath2vec. The description of our proposed algorithm that makes top-k recommendations for documents using the Metapath2vec model can be seen in Algorithm 1. To summarize, a heterogeneous graph structure was constructed using the features we extracted using the summaries of the documents and the metadata of the documents. Then, we generated a corpus of node sequences using the random walk strategy over this graph structure and obtained the vector representations for the nodes using the heterogeneous skip-gram algorithm. Finally, we used cosine similarity to determine the distance (similarity) between the query node and the remaining others and suggested the top-k similar nodes according to the type of query node.

5 Experiments and Results

In this section, we discuss a systematic evaluation process performed to assess the performance of our proposed system.

[6] https://github.com/stefan-it/turkish-bert.

Algorithm 1: Metapath2vec Document Recommendation

Input: The training documents $d^T = \{n, t, s\}$. Sumgram n, temporal category t, document summary s, a meta-path scheme ρ, walks per node W, walk length l, embedding dimension d, neighborhood size k, query node q_n, top k element k
Output: First k element in $S \in R^{\|V\| \times d}$
Initialize X, empty list S ;
Construct heterogeneous graph $G = (V, E, T)$ using d^T;
X = Metapath2vec(G,ρ,v,l);
S = Recommend(q_n,**X**);
return S;

Metapath2vec(G,ρ,v,l)
for $w \in W$ **do**
  for $v \in V$ **do**
    update **X** according to the steps in 3.3 ;
  end for
end for
return $\mathbf{X} \in R^{\|V\| \times d}$;

Recommend(q_n,**X**)
for each $x \in \mathbf{X}$ **do**
  calculate cosine similarity θ between x and q_n;
  insert (x,θ) to S;
end for
sort S in increasing θ;
return S;

5.1 Dataset Description

As a result of our conversations with professional archivists in the SALT team, we decided to use Waqfs of Crete, which are an archive consisting of official records of Muslim inhabitants of Crete who moved to Turkey during the 1920s due to the population exchange between Turkey and Greece. Crete is the largest and most populous of the Greek islands and the fifth-largest island in the Mediterranean, located approximately 160 km south of the Greek mainland.

Crete remained under Ottoman rule until 1898, after which it briefly existed as an independent state from 1898 to 1908 before officially becoming part of Greece. In the 19th century, Crete was home to sizable Muslim and Christian communities, with tensions occasionally escalating into uprisings between the two groups [39].

Documents spanning the period from 1825 to 1928 in Ottoman Turkish and Greek provide an opportunity to examine the multi-layered social structure on the island, especially from a cultural and economic perspective. The metadata of Waqfs of Crete provided by a SALT Research team comprises a specialized library and an archive of physical and digital sources and documents on visual practices, the built environment, social life, and economic history in Turkey. The metadata contains information for approximately 10 thousand documents and includes the summary of those documents, the year they were published, the location, the language used, and the documents' picture. An example document and its metadata are depicted in Figure 4. More information about the data and description of underlying metadata are explained below. Figure 5 presents a visual overview of the tool used to explore this metadata during analysis and evaluation.

— *Image of Document*. Pictures of documents in the collection that were originally written in Ottoman. An example *document*[7] picture can be seen in Figure 4.

[7] https://archives.saltresearch.org/handle/123456789/63175?locale=en.

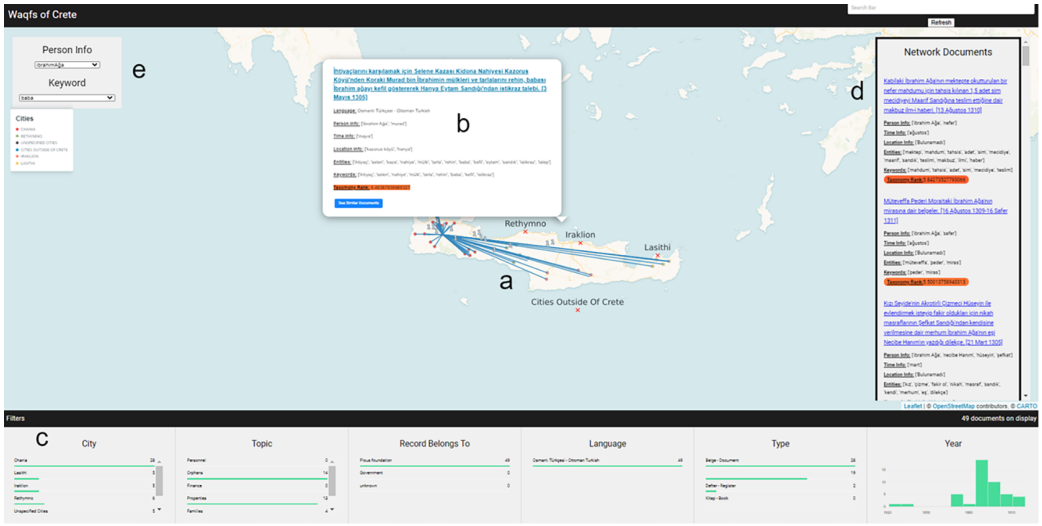


Fig. 5. Overview of the tool used for metadata analysis and visualization in the "Waqfs of Crete" dataset. The tool provides insights into the archive's structure, including document summaries, date, person info, and locations.

— *Summary*. Text summarizing the event depicted in the pictorially available documents. These texts were written by archive professionals. The average number of tokens in these summaries is calculated as 30.
— *Date*. The date the documents were written. Year distribution is shown in Figure 2.
— *Language*. The language of the documents is Modern Turkish, but most of the words are of Ottoman origin, which is difficult for a normal person to read.
— *Category*. Documents are organized according to the 14 categories they belong to by archive professionals. These categories include Aid, Complaints, Education, Families, Finance, Inheritance, Military Service, Miscellaneous, Orphans, Personnel, Population Exchange, Properties, Request for Protection, and Verdicts. Category distribution are shown in Table 1.

5.2 Manual Annotation Setup

To evaluate our document recommendation system trained on the Waqfs of Crete dataset, we conducted an experiment with five archive professionals and five non-professional annotators on the Waqfs of Crete dataset. The rules of the organized experiment can be seen below:

— Two query documents were randomly selected for each of the five most popular categories. For the query documents that we selected randomly, we suggested the 10 documents that our proposed metapath2vec (see Algorithm 1) model recommends as the most relevant to the query document.
— Each participant was provided with a total of 10 query documents, consisting of 2 documents from each category. For every pair of query documents, our model suggested an additional 10 documents that were expected to be relevant to the content or topic of the query documents. This process resulted in a total of 50 suggested documents per participant. During the experiment, participants were asked to annotate the suggested documents as "Relevant" (indicating relevance to the query documents), "Irrelevant," or "Language is not clear."
— In order to measure the agreement among the participants, we enforce 50% (5 out of 10 documents) overlap between query documents that are given to each annotator.

Table 1. Category Distribution of the "Waqfs of Crete Dataset"

Category	Document Count
Orphans	2,041
Finance	1,802
Properties	1,288
Families	632
Inheritance	498
Miscellaneous	399
Aid	335
Verdicts	238
Education	129
Complaints	101
Military Service	44
Population Exchange	12
Request for Protection	8
Total	10,145

Documents are classified into 13 categories and majority of them covers family and financial issues.

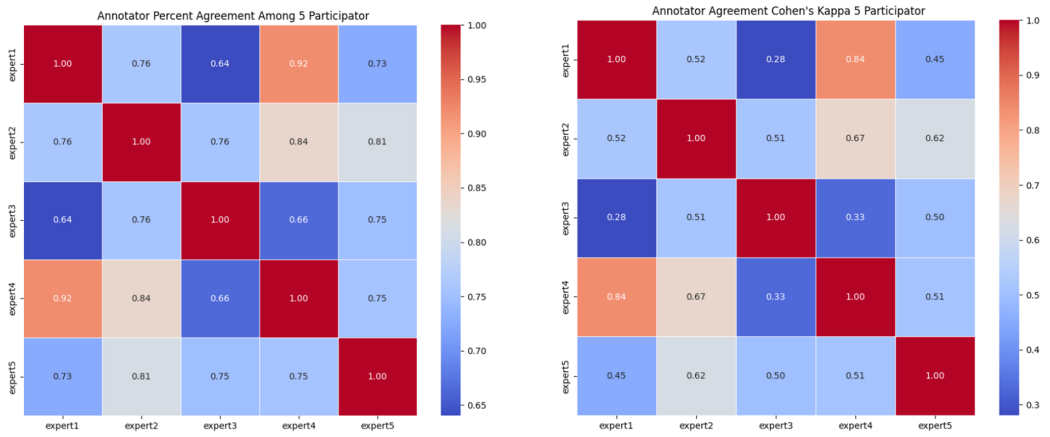
—Documents were selected from the categories of "Family," "Property," "Finance," "Inheritance," and "Orphans." These are the five most common categories in the data.

5.3 Inter-Rater Reliability (IRR)

The IRR (also called an inter-annotator-agreement) is the degree of agreement among different raters as described by Saal et al. [66]. It is a score of how much homogeneity or consensus exists in the ratings given by various judges. Different machine learning tasks require labeled data that are frequently annotated by humans. We conducted two different IRR techniques to measure how well multiple annotators can make the same annotation decision for a certain category on different objects. The annotation scheme used in this study was to annotate the recommended document for a query document as "Relevant," "Irrelevant," or "Language is not clear." As we have stated in the experiment specifications (Section 5.2), half of the documents we provide to users to measure the annotator agreement are mutual to everyone participating in the experiment. We used Cohen's kappa and Percent Agreement to measure Agreement. The methodologies underlying the techniques and the results can be seen in the following section.

5.3.1 Percent Agreement. The percent agreement refers to the joint probability of agreement and is the simplest and the least IRR measure. It is computed as the percentage of the time the annotators agree in a nominal or categorical rating system. It is insensitive to the fact that agreement may happen only based on chance [79]. To calculate the measure of percent agreement between two raters, we count the number of ratings in agreement and calculated the fraction of agreement over all the rated documents.

5.3.2 Cohen's Kappa Coefficient κ. This statistic measures the inter-rater agreement for categorical items [51]. Cohen's kappa is thought to be a more reliable measure than simple percent agreement calculation, as κ considering the possibility of the agreement occurring by chance. The κ measures the agreement between two raters who each classify N objects into C mutually exclusive categories. The definition of Cohen's kappa is the



(a) (b)

Fig. 6. Analysis of agreements of expert annotators. (a) Percent agreement and (b) Cohen's kappa scores of the annotations collected from five domain experts. Experiment yielded average pairwise agreement of 82% and substantial inter-annotator agreement (Cohen's $\kappa = 0.61$). The results indicate that there is a moderate to substantial agreement among the answers of the participants.

following:

$$\kappa \equiv \frac{p_o - p_e}{1 - p - e} = \frac{1 - p_o}{1 - p_e}, \tag{9}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement.

5.3.3 Evaluation of Inter-Rater Reliability. Our experiment yielded an average pairwise agreement of 75.2% and moderate inter-annotator agreement (Cohen's $\kappa = 0.49$) between all 10 annotators while average pairwise agreement of 82% and substantial inter-annotator agreement (Cohen's $\kappa = 0.61$) between 5 professional annotators [51]. The results of pairwise agreement scores and Cohen's kappa scores among professional participants are shown in Figure 6.

5.4 Model Performance Evaluation

The similarity score between the vector embeddings of documents recommended by our model (10 documents per query document) for the query documents and the embeddings for query documents (10 in total) was re-calculated for models including TF-IDF, Doc2vec, and BERT to perform a detailed comparison. In order to compare the proposed recommendations in the most accurate and objective way, we have followed the following procedures:

(1) Participants had annotated the recommended documents for query documents as "Relevant," "Irrelevant," or "Language is not clear." An important point here is that the documents recommended for the query document are the Top-N recommendation produced by our proposed model (see Section 4.3.1), that is, the 10 documents with the highest similarity for the query document, as stated in Section 4.4.
(2) To construct a benchmark, we calculated the similarity scores between the query documents and the documents recommended by our model. Additionally, we performed the same process using three other models for comparison. Similarities calculated between queries and documents with these extra models are solely used for performance comparison and only the outcomes of our proposed approached used for recommendation and annotation experiments. In other words, similarity scores and embeddings are

re-obtained between the recommended documents and the query documents with three extra different models, including TF-IDF, Doc2vec, and BERT.

(3) We evaluate the performance of models using the Mann–Whitney U test [46] according to the answers given by the participants in the experiment. Mann–Whitney U test is a statistical test that can be used to characterize the degree of separation between two frequency distributions. Using this test, we wanted to prove that there is a significant difference between similarity score distributions of differently labeled documents.

For an accurate evaluation, we applied the Mann–Whitney U test [46] which is a non-parametric statistical technique that analyzes the differences between the medians of two sets. It tests the null hypothesis that is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from the second sample [46].

According to our hypothesis, documents annotated as "*Relevant*" yield higher similarity scores than documents marked as "Irrelevant." In this way, we applied the hypothesis tests in two different ways. First, we calculated the test statistics between the similarity scores of the documents marked "relevant" and "irrelevant" for each user. Notice that, documents marked as "irrelevant" are under 5% and are not removed from the testing phase since they add equal noise to each model. This resulted in 10 different test statistics for 10 participants per model. and the distribution of user test statistics can be seen in Figure 7(a). In addition, the methodology we followed while obtaining test statistics for users is illustrated in Algorithm 2.

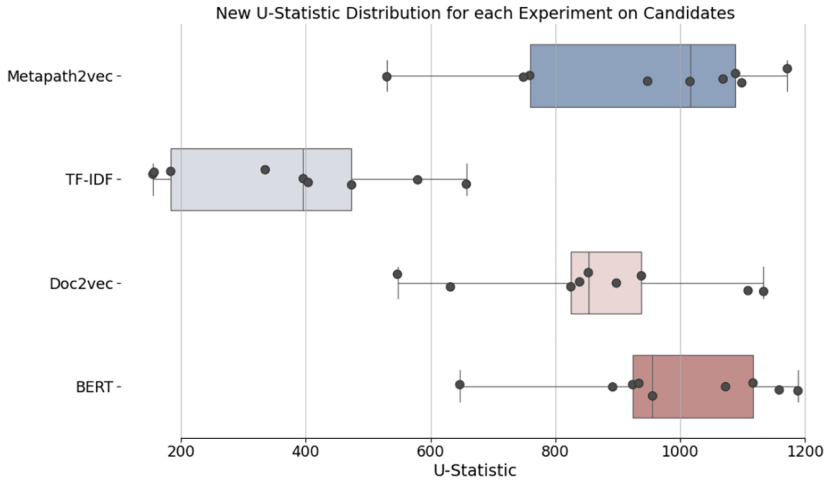
Second, we calculated U-statistics for document category, that is, we extracted test statistics between the similarity scores of annotated documents for each document category. In this way, five test statistics for five categories per model were extracted, and their distribution is shown in Figure 7(b). In addition, the methodology we followed while obtaining test statistics for categories is illustrated in Algorithm 3. According to the results in Table 2, the model we proposed achieves a U-Statistic score of 936.7 in the "Among Participants" experiment and 921.1 in the "Among Categories" experiment, both of which are higher than the scores of TF-IDF (371.2 and 389.1, respectively) and Doc2Vec (863.8 and 798.9, respectively). While our model performs slightly behind BERT, which scores 987.8 and 1035.5 in the respective experiments, it still demonstrates competitive performance. To summarize, U-Statistic scores of the proposed Metapath2vec model and BERT are higher than the remaining models (TF-IDF and Doc2vec). In other words, our proposed model and BERT are better approach for distinguishing the context of documents compared to other models.

6 Conclusions

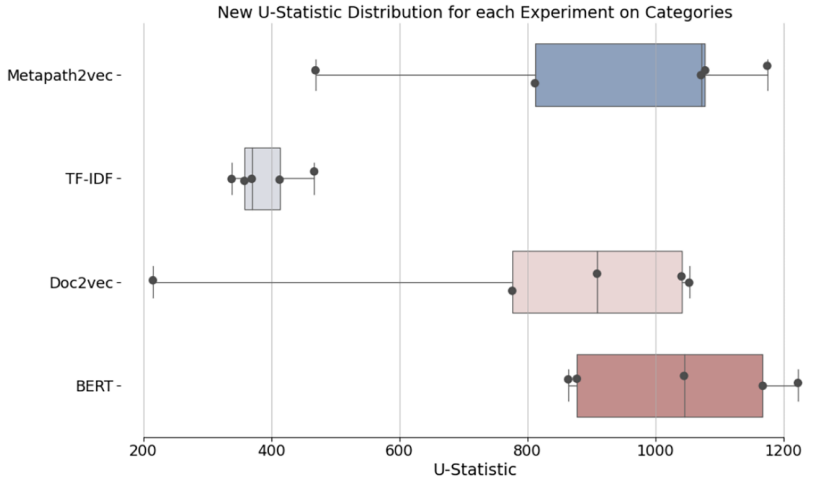
In this work, a system was developed and tested on a collection of documents in the field of CH. Thus enabling us to provide valuable insights in terms of inner and inter-communal relations within a society at the turn of the 20th century and to examine more deeply the CH of this Mediterranean island at the crossroads of East and West [39, 89, 90].

In the document recommendation system, we have introduced constructs of heterogeneous document networks using novel feature extraction procedures, such as using conjoined n-grams and creating temporal categories to describe the data best. In the meantime, we only benefited from the textual summaries of the documents and the documents' metadata. Then, using the graph structure created, the system adopts heterogeneous skip-gram and random walk strategies, which are the techniques used by the recently proposed Metapath2vec algorithm, generates embeddings for documents, and uses these embeddings in the recommendation system. The extracted embeddings were primarily used to design the document suggestion system, but visible results were also obtained by visualizing these embeddings.

We designed an experiment with total of 10 participants including 5 archive professionals and 5 normal users to test the performance of the recommendation system we propose. During the experiment, we supplied query documents and recommended the top 10 results suggested by our proposed model for query documents.



(a)



(b)

Fig. 7. Evaluation of the recommendation system. (a) Distribution of U-Statistics between the similarity scores of annotated documents for experiment participants. We calculate test statistics for 10 participants, 1 per participant for each model. Test statistics distribution for each model are presented. (b) Distribution of U-Statistics between the similarity scores of annotated documents for each document category. We calculate test statistics for 5 categories, 1 per category for each model.

Finally, we asked them to annotate these suggested documents as “Relevant,” “Irrelevant,” and “Language is not clear.” This is a demanding task since the language of the documents was difficult for an average person to understand. To compare the results, we used the Mann–Whitney U test [46], a non-parametric hypothesis test, and we hypothesized that documents marked as “Relevant” had higher similarity scores than documents marked “Irrelevant” for a query document. In this framework, we compared the models with each other according to their significant test statistics of the experimental results. As a result of the experiments, the model we proposed gave

Algorithm 2: U-Statistic-Users[8]

Input: model m, participant list **P**
Output: non-empty list of U-Statistics U_{users}
initialize empty list U_{users};
for each $p_i \in$ **P do**
  Retrieve $l_{relevant}$ and $l_{irrelevant}$ [8] for p_i from m;
  Calculate $U_{pi} \equiv$S(l$_{relevant}$, $l_{irrelevant}$);
  Insert U_{pi} into U_{users};
end for
return U_{users};

Table 2. Performance Evaluation of the Proposed and Baseline Models

Model	Experiment Type	Mean U-Statistics
Our Model	Among Participants	936.7
TF-IDF	Among Participants	371.2
Doc2Vec	Among Participants	863.8
BERT	Among Participants	987.8
Our Model	Among Categories	921.1
TF-IDF	Among Categories	389.1
Doc2Vec	Among Categories	798.9
BERT	Among Categories	1035.5

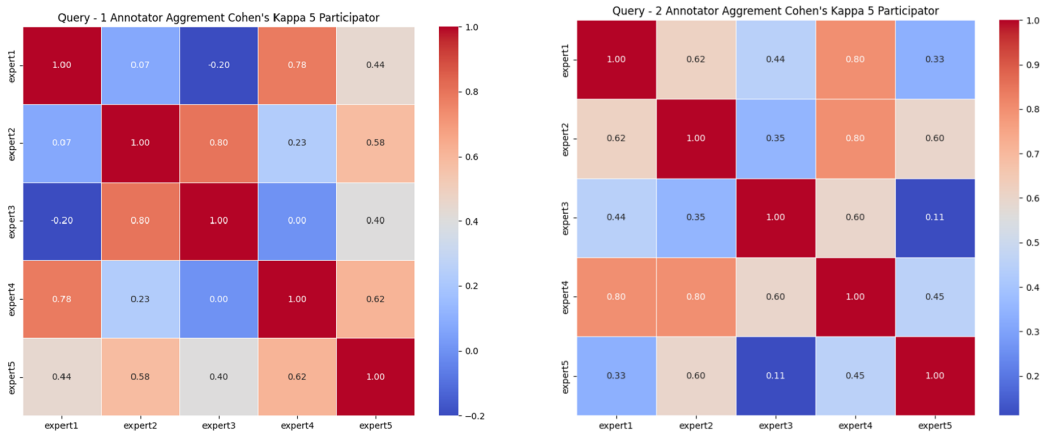
Experiments for category- and participant-level analysis presented and average test statistics of Mann–Whitney U test study performance of the recommendation system with respect to user annotations.

Algorithm 3: U-Statistic-Categories[9]

Input: model m, category list **C**
Output: non-empty list of U-Statistics $U_{categories}$
initialize empty list $U_{categories}$;

for each $c_i \in$ **C do**
  Retrieve $l_{relevant}$ and $l_{irrelevant}$ [9] for c_i from m;
  Calculate $U_{ci} \equiv$S(l$_{relevant}$, $l_{irrelevant}$);
  Insert U_{ci} into $U_{categories}$;
end for
return $U_{categories}$;

better results than the alternative document embedding methods, these are TF-IDF and Doc2vec, but gave a very similar result to the BERT. In conclusion, considering that BERT is a model trained with huge amounts of data and is too parametric and very complex, our result is reasonable. Based on the experiment results, we observed that the percent agreement and Cohen's kappa scores of the experts varied significantly depending on the query document. This result points out to complex nature of the subjectivity of document recommendation systems. While all five experts possess deep knowledge in their respective domains, they offered different answers on the relevance of recommended documents. In some query documents, the agreement was higher, while some



(a) (b)

Fig. 8. Analysis of agreements of expert annotators. (a) Cohen's kappa scores of the annotations based on the query-1 document. (b) Cohen's kappa scores of the annotations based on the query-2 document.

query documents resulted with experts having no agreement among themselves at all. These results can be seen in Figure 8. Beyond the technological aspects, it is essential to consider the critical and cultural dimensions of the interaction between users and IT systems. Digital tools, no matter how advanced, are often limited in their ability to fully understand the cultural and contextual nuances embedded in CH documents. This limitation highlights the importance of involving professional domain experts who can interpret these nuances and provide context that automated systems might overlook. Their expertise ensures that the analysis and recommendations made by the system are both accurate and relevant, reflecting the true significance of the documents.

The role of professional domain experts is particularly crucial when dealing with complex and culturally rich datasets. Their supervision can significantly enhance the effectiveness of digital tools, ensuring that critical cultural aspects are not missed and that the interaction between users and IT systems is as productive as possible. This collaboration between advanced digital tools and expert knowledge is essential for achieving the best outcomes in the field of CH.

Our experiment results revealed that agreement among experts varies significantly based on the query document, indicating the subjective nature of document recommendation systems. While all five experts possess deep knowledge in their respective domains, their different answers on the relevance of recommended documents underscore the complexity and subjectivity inherent in this task. This observation highlights the necessity of involving domain experts in the evaluation and interpretation processes to ensure the accuracy and relevance of the recommended documents.

In conclusion, this work demonstrates how machine learning, network science, and NLP techniques can be combined to assist archivists in discovering relevant documents when searching for specific entities, timeframes, or societal events. Building upon the foundational ideas and methodologies for combining metadata with graph-based representations first explored in Özdemir [94], this study extends and refines these approaches to address the unique challenges of CH collections. Such systems can save significant time for experts and help them uncover unseen patterns of interactions otherwise not possible. As natural language understanding techniques develop and resources for rare languages increase, similar approaches can further improve and address the needs that remain open for the CH community. Nevertheless, the supervision of professional domain experts remains an essential component in achieving the most accurate and contextually meaningful results, ensuring that both technological and cultural aspects are adequately addressed.

While this study primarily focuses on a single collection, its methods lay the groundwork for broader applications in CH. Future work could explore integrating external datasets and leveraging Linked Data vocabularies to reveal connections across archives. For example, projects like Europeana[10] and ARIADNE+[11] illustrate the potential for uncovering spatial and temporal relationships on a larger scale. Expanding this workflow to include such elements would provide richer contextual insights, further enhancing the value of archival systems for CH studies.

References

[1] Tülay Adalı and Antonio Ortega. 2018. Applications of graph theory [scanning the issue]. *Proceedings of the IEEE* 106, 5 (2018), 784–786. DOI: https://doi.org/10.1109/JPROC.2018.2820300

[2] Medina Andresel, Sergiu Gordea, Srdjan Stevanetic, and Mina Schütz. 2022. An approach for curating collections of historical documents with the use of topic detection technologies. *International Journal of Digital Curation* 17, 1 (Sep. 2022), 12. DOI: https://doi.org/10.2218/ijdc.v17i1.819

[3] Ignacio Arroyo-Fernández, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, Juan-Manuel Torres-Moreno, and Grigori Sidorov. 2019. Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Computer Speech and Language* 56 (2019), 107–129.

[4] J. Aviles Collao, L. Diaz-Kommonen, M. Kaipainen, and J. Pietarila. 2003. Soft ontologies and similarity cluster tools to facilitate exploration and discovery of cultural heritage resources. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, 2003, 1–5.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[6] David Blei, Andrew Ng, and Michael Jordan. 2001. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*. T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Vol. 14, MIT Press. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[8] Robin Brochier, Adrien Guille, and Julien Velcin. 2019. Global vectors for node representations. In *Proceedings of the World Wide Web Conference*, 2587–2593.

[9] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.

[10] Jay Budzik and Kristian J. Hammond. 2000. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI '00)*. ACM, New York, NY, 44–51. DOI: https://doi.org/10.1145/325737.325776

[11] Lou Burnard and Harold Short. 1994. An Arts and Humanities Data Service. Report of a Feasibility Study commissioned by the Information Systems Sub-Committee of the Joint Information Systems Committee of the Higher Education Funding Councils.

[12] Fenxiao Chen, Yun-Cheng Wang, Bin Wang, and C.-C. Jay Kuo. 2020. Graph representation learning: A survey. *APSIPA Transactions on Signal and Information Processing* 9, 1 (2020), e15.

[13] Yankai Chen, Quoc-Tuan Truong, Xin Shen, Jin Li, and Irwin King. 2024. Shopping trajectory representation learning with pre-training for E-commerce customer understanding and recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. ACM, New York, NY, 385–396. DOI: https://doi.org/10.1145/3637528.3671747

[14] Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord. 2019. Unsupervised speech representation learning using WaveNet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 12 (Dec. 2019), 2041–2053. DOI: https://doi.org/10.1109/TASLP.2019.2938863

[15] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12 (2011), 2493–2537.

[16] George Dahl, Marc'Aurelio Ranzato, Abdel-Rahman Mohamed, and Geoffrey E. Hinton. 2010. Phone recognition with the mean-covariance restricted boltzmann machine. *Advances in Neural Information Processing Systems* 23 (2010), 469–477.

[17] Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors. arXiv:1507.07998. Retrieved from https://arxiv.org/abs/1507.07998

[18] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters* 80 (2016), 150–156.

[10] https://www.europeana.eu/en.

[11] https://ariadne-infrastructure.eu.

[19] Li Deng, Michael L. Seltzer, Dong Yu, Alex Acero, Abdel-Rahman Mohamed, and Geoff Hinton. 2010. Binary coding of speech spectrograms using a deep auto-encoder. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, 1–10.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[21] Yuxiao Dong, Nitesh, V. Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 135–144.

[22] Jinhua Dou, Jingyan Qin, Zanxia Jin, and Zhuang Li. 2018. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. *Journal of Visual Languages and Computing* 48 (2018), 19–28.

[23] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88)*. ACM, New York, NY, 281–285. DOI: https://doi.org/10.1145/57167.57214

[24] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys* 56, 2 (Sep. 2023), Article 27, 47. DOI: https://doi.org/10.1145/3604931

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=S1v4N2l0

[26] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864.

[27] Ziyu Guan, Can Wang, Jiajun Bu, Chun Chen, Kun Yang, Deng Cai, and Xiaofei He. 2010. Document recommendation in social tagging services. In *Proceedings of the 19th International Conference on World Wide Web*, 391–400.

[28] Shashank Gupta and Vasudeva Varma. 2017. Scientific article recommendation by using distributed representations of text and graph. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 1267–1268.

[29] Nancy E. Gwinn and Constance Rinaldo. 2009. The biodiversity heritage library: Sharing biodiversity literature with the world. *IFLA Journal* 35, 1 (2009), 25–34.

[30] Emma Haddi, Xiaohui Liu, and Yong Shi. 2013. The role of text pre-processing in sentiment analysis. *Procedia Computer Science* 17 (2013), 26–32.

[31] Aric Hagberg, Pieter Swart, and Daniel S. Chult. 2008. *Exploring Network Structure, Dynamics, and Function Using NetworkX*. Technical Report. Los Alamos National Lab. (LANL), Los Alamos, NM (United States).

[32] William L. Hamilton. 2020. *Graph Representation Learning*. Morgan and Claypool Publishers.

[33] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin* 40, 3 (2017), 52–74. Retrieved from http://sites.computer.org/debull/A17sept/p52.pdf

[34] Cormac Hampson, Maristella Agosti, Nicola Orio, Eoin Bailey, Seamus Lawless, Owen Conlan, and Vincent Wade. 2012. The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In *Proceedings of the Euro-Mediterranean Conference*. Springer, 668–675.

[35] Peter E. Hart and Jamey Graham. 1997. Query-free information retrieval. *IEEE Expert* 12, 5 (1997), 32–37.

[36] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association* 97, 460 (2002), 1090–1098.

[37] Xiangjie Kong, Mengyi Mao, Wei Wang, Jiaying Liu, and Bo Xu. 2018. VOPRec: Vector representation learning of papers with text information and structural identity for recommendation. *IEEE Transactions on Emerging Topics in Computing* 9, 1 (2018), 226–237.

[38] Elektra Kostopoulou. 2009. The Muslim Millet of Autonomous Crete: An Exploration into Its Origins and Implications. Ph.D. Thesis. Bogazici University, Institute for Graduate Studies in Social Sciences, Istanbul, Turkey. Advisor(s) Eldem, Edhem. Retrieved from https://digitalarchive.library.bogazici.edu.tr/handle/123456789/17807

[39] Elektra Kostopoulou. 2016. The island that wasn't: Autonomous Crete (1898–1912) and experiments of federalization. *Journal of Balkan and Near Eastern Studies* 18, 6 (2016), 550–566. DOI: https://doi.org/10.1080/19448953.2016.119

[40] Per Ola Kristensson, Olof Arnell, Annelie Björk, Nils Dahlbäck, Joackim Pennerup, Erik Prytz, Johan Wikman, and Niclas Åström. 2008. InfoTouch: An explorative multi-touch visualization interface for tagged photo collections. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, 491–494.

[41] Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, 78–86. DOI: https://doi.org/10.18653/v1/W16-1609

[42] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*, 1188–1196.
[43] Kun, Chang Lee, and Soonjae Kwon. 2008. Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: A causal map approach. *Expert Systems with Applications* 35, 4 (2008), 1567–1574.
[44] Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2020. *Representation Learning for Natural Language Processing*. Springer Nature.
[45] David Maltz and Kate Ehrlich. 1995. Pointing the way: Active collaborative filtering. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 202–209.
[46] H. Mann and D. Whitney. 1947. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60.
[47] Christopher D. Manning, Prabhakar Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
[48] Mohammed Maree, Amjad Rattrout, Muhanad Altawil, and Mohammed Belkhatir. 2021. Multi-modality search and recommendation on Palestinian cultural heritage based on the holy-land ontology and extrinsic semantic resources. *Journal on Computing and Cultural Heritage* 14, 3 (Jul. 2021), Article 29, 1–23. DOI: https://doi.org/10.1145/3447523
[49] Alireza R. Mashaghi, Abolfazl Ramezanpour, and Vahid Karimipour. 2004. Investigation of a protein complex network. *The European Physical Journal B–Condensed Matter and Complex Systems* 41, 1 (2004), 113–121.
[50] Viera Maslej-Krešňáková, Martin Sarnovský, Peter Butka, and Kristína Machová. 2020. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences* 10, 23 (2020), 8631.
[51] Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22, 3 (2012), 276–282.
[52] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781
[53] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. Retrieved from https://aclanthology.org/L18-1008/
[54] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*, 3111–3119.
[55] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M. Bronstein. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5115–5124.
[56] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.
[57] Rohit Nagori and G. Aghila. 2011. LDA based integrated document recommendation model for e-learning systems. In *Proceedings of the 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*. IEEE, 230–233.
[58] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104, 1 (2015), 11–33.
[59] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Bootstrapping web archive collections from social media. In *Proceedings of the ACM Conference on Hypertext and Social Media (HT '18)*. ACM, New York, NY. DOI: https://doi.org/10.1145/3209542.3209560
[60] George Pavlidis. 2019. Recommender systems, cultural heritage applications, and the way forward. *Journal of Cultural Heritage* 35 (2019), 183–196.
[61] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710.
[62] Panagiotis Petridis, Daniel Pletinckx, Katerina Mania, and Martin White. 2006. The EPOCH multimodal interface for interacting with digital heritage artefacts. In *Proceedings of the International Conference on Virtual Systems and Multimedia*. Springer, 408–417.
[63] Radim Rehurek and Petr Sojka. 2011. *Gensim–Python Framework for Vector Space Modelling*. *NLP Centre, Faculty of Informatics, Masaryk University*, *Brno, Czech Republic* 3, 2 (2011).
[64] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084
[65] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender Systems Handbook*. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.), Springer, 1–35.
[66] Frank E. Saal, Ronald G. Downey, and Mary A. Lahey. 1980. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin* 88, 2 (1980), 413–428.
[67] Saminu M. Salisu, Eva Mayr, Velitchko, Andreev Filipov, Roger A. Leite, Silvia Miksch, and Florian Windhager. 2019. Shapes of time: Visualizing set changes over time in cultural heritage collections. In *EuroVis (Posters)*, 45–47.

[68] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, 285–295.

[69] Craig W. Schmidt. 2019. Improving a tf-idf weighted document vector embedding. arXiv:1902.09875. Retrieved from https://arxiv.org/abs/1902.09875

[70] Ali Serdar and Reyhan Tutumlu. 2018. Building a digital Ottoman/Turkish serial novel archive. *AIUCD* 2015 (2018), 33.

[71] Preya Shah, Arian Ashourvan, Fadi Mikhail, Adam Pines, Lohith Kini, Kelly Oechsel, Sandhitsu R. Das, Joel M. Stein, Russell T. Shinohara, Danielle S. Bassett, et al. 2019. Characterizing the role of the structural connectome in seizure dynamics. *Brain* 142, 7 (2019), 1955–1972.

[72] Benyah Shaparenko and Thorsten Joachims. 2009. Identifying the original contribution of a document via language modeling. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 350–365.

[73] Advaith Siddharthan. 2002. Christopher D. Manning and Hinrich Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 2000. ISBN 0-262-13360-1. 620 pp. 64.95/£ 44.95 (cloth). *Natural Language Engineering* 8, 1 (2002), 91.

[74] Scott Spangler, Jeffrey T. Kreulen, and Justin Lessler. 2002. Mindmap: Utilizing multiple taxonomies and visualization to understand a document collection. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. IEEE, 1170–1179.

[75] Yizhou Sun and Jiawei Han. 2013. Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology* 18, 4 (2013), 329–338.

[76] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077.

[77] Yijun Tian, Chuxu Zhang, Ronald Metoyer, and Nitesh V. Chawla. 2021. Recipe Representation Learning with Networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. ACM, New York, NY, 1824–1833. DOI: https://doi.org/10.1145/3459637.3482468

[78] Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. arXiv:1812.05069. Retrieved from https://arxiv.org/abs/1812.05069

[79] John S. Uebersax. 1987. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 101, 1 (1987), 140–146.

[80] Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Proceedings of the 26th Neural Information Processing Systems Conference (NIPS '13)*. Neural Information Processing Systems Foundation (NIPS), 2643–2651.

[81] Robin Van Meteren and Maarten Van Someren. 2000. Using content-based filtering for recommendation. In *Proceedings of the 30th Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, 47–56.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, 5998–6008.

[83] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2021. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (Oct. 2021), 3349–3364. DOI: 10.1109/TPAMI.2020.2983686

[84] Zhi Wang, Wenwu Zhu, Peng Cui, Lifeng Sun, and Shiqiang Yang. 2013. Social media recommendation. In *Proceedings of the Social Media Retrieval*. Springer, 23–42.

[85] Sung-Shun Weng and Hui-Ling Chang. 2008. Using ontology network analysis for research document recommendation. *Expert Systems with Applications* 34, 3 (2008), 1857–1869.

[86] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. 2018. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE Transactions on Visualization and Computer Graphics* 25, 6 (2018), 2311–2330.

[87] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang. 2015. Network representation learning with rich text information. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2111–2117.

[88] Chien-Chih Yu and Hsiao-Ping Chang. 2009. Personalized location-based recommendation services for tour planning in mobile tourism applications. In *Proceedings of the International Conference on Electronic Commerce and Web Technologies*. Springer, 38–49.

[89] Elif Yılmaz. 2014. Girit arşivi, archive of Crete. *Bilgi Dünyası* 15 (2014), 217–223. 1

[90] Yılmaz Elif. 2017. Cretian records held in public and private archives in Turkey. *Mavi Atlas* 5, 1 (2017), 224–237.

[91] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W. Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1449–1458.

[92] Yin Zhang, Ziwei Zhu, Yun He, and James Caverlee. 2020. Content-collaborative disentanglement representation learning for enhanced recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20)*. ACM, New York, NY, 43–52. DOI: https://doi.org/10.1145/3383313.3412239

[93] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, 167–176.

[94] Anıl Özdemir. 2021. *Data-Driven Exploration of Document Collection to Understand Underlying Social Fabric Using Graph Representation Learning*. Master's Thesis. Sabanci University, Istanbul.

[95] Pınar Şenışık. 2007. *The Transformation of Ottoman Crete: Cretans, Revolts and Diplomatic Politics in the Late Ottoman Empire*, 1895–1898. Ph.D. Thesis. Bogazici University, Institute for Graduate Studies in Social Sciences, Istanbul, Turkey. Retrieved from https://digitalarchive.library.bogazici.edu.tr/handle/123456789/17831

Received 20 October 2023; revised 29 March 2025; accepted 9 June 2025