Development of the Recommender System of Arabic Books Based on the Content Similarity

Shaykhah Hajed Alotaibi and Muhammad Badruddin Khan,

Information Systems Department,
College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, KSA

Abstract
This research article develops an Arabic books' recommendation system, which is based on the content similarity that assists users to search for the right book and predict the appropriate and suitable books pertaining to their literary style. In fact, the system directs its users toward books, which can meet their needs from a large dataset of Information. Further, this system makes its predictions based on a set of data that is gathered from different books and converts it to vectors by using the TF-IDF system. After that, the recommendation algorithms such as the cosine similarity, the sequence matcher similarity, and the semantic similarity aggregate data to produce an efficient and effective recommendation. This approach is advantageous in recommending previously unrated books to users with unique interests. It is found to be proven from the obtained results that the results of the cosine similarity of the full content of books, the results of the sequence matcher similarity of Arabic titles of the books, and the results of the semantic similarity of English titles of the books are the best obtained results, and extremely close to the average of the result related to the human assigned/annotated similarity. Flask web application is developed with a simple interface to show the recommended Arabic books by using cosine similarity, sequence matcher similarity, and semantic similarity algorithms with all experiments that are conducted.

Keywords:
Recommendation system, books, cosine similarity, sequence matcher, semantic similarity, content-based algorithm.

1. Introduction

Recommender systems represent an information filtering system for interacting with large and complex information spaces. They comprise personalised views, prioritising items that are likely to be of interest to the user, predicting the ratings or preferences by which a user would provide to an item, targeting the individuals who do not have sufficient knowledge and personal experience to evaluate the valued alternatives by which a website may possibly offer.

Recommender systems use several filtering techniques that can be divided into two categories Content-Based (CB), and Collaborative Filtering (CF). CB systems recommend items by matching item contents and features that are automatically extracted from these items such as the term frequency and genre. In contrast, CF systems recommend items to users based on the opinions and tastes of other users [9]. Individuals usually rely on recommendations in making daily decisions, such as when selecting a book to read or a movie to watch where it is common to rely on recommendations that are provided by others or by the system based on the similarity of the books' or movies' contents. .

To study these behaviours and benefit from them, the recommender system uses the algorithms to create a prototype of recommendations produced by users or based on the similarity of the contents for providing recommendations to other users.

These recommendations are designed to make predictions for every item that has a similar content based on a set of gathered data, such as punctuation profile, part-of-speech profile, sentiment analysis, author, average sentence length, average word length. Following that, the recommendation algorithm aggregates the data to produce effective recommendations. In fact, this approach is commonly known as the content-based algorithm where its aim is to ensure that if the contents of the items are compatible with other previously available items, the recommendations that are derived from these similar items should be also relevant and of interest to other items.

2. Literature review

2.1 Recommender Systems using Collaborative Filtering Method

According to Bhatnagar (2017), information retrieval and filtering technique represent the background of every recommender system, which gives a clear nature of the way it operates.

Manuscript received August 5, 2022
Manuscript revised August 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.8.23

Additionally, it is used to support the collaboration among users and to monitor their behaviours in order to recommend items with clear reviews. The characteristics of each item are collected along with all users' recommendations, ratings and comments to provide a clear view of the products being searched. The entire data is sent to the recommender systems to act as a filter for users in order to meet their needs.
Linden et al. (2003) review that Amazon.com is one of the pioneers of the recommender systems, particularly, in the commercial setting. Amazon.com uses an item-based collaborative filtering, which recommends various items for every customer. This implies that Amazon builds a list of related items for each item whenever an item is purchased or looked at. Following that, Amazon.com recommends items from that item's lists. The item-to-item collaborative filtering algorithm is compared with three conventional approaches, which comprise the collaborative filtering, cluster models and search-based approaches. It is found that the item-to-item collaborative filtering algorithm depends on titles that the user purchases or rates. This process is rapid for large datasets and provides recommendations with excellent quality. Due to the highly correlated similar items, this algorithm produces recommendations in real-time, scales to massive datasets, generates high quality recommendations and provides recommendations to users on the main webpage of the site. Once users sign in to Amazon.com and look at the front page, their recommendations are mostly into the form "You viewed... Customers who viewed this also viewed...".

Liu et al. (2010) present a research on developing an effective information filtering mechanism for news and recommendations on a large-scale website such as Google News. They first conduct a log analysis on the change of user's interests in news topics over time. The log analysis demonstrated variations in users' interests in news, and shows that these interests are influenced by the local news' trend. Based on the analytical results, Liu et al. decompose the interests of users' news into two parts, which comprise: the genuine interests and the influence of local news trends. They proposed a Bayesian framework to model a user's genuine interests by using a past click history and predicting current interests by combining genuine interests and the local news' trend. They combine the method that is used with the existing collaborative filtering method to generate different personalised news' recommendations. Moreover, they conduct an experiment with the news recommender by using the combined method on a part of the live traffic through Google News website. In comparison with the existing collaborative filtering method, their experiment show that the combined method improves the quality of news' recommendations and increases the number of visits into the Google News website. Furthermore, Liu et al. attempt to study several advanced methods in the future research to combine the information filtering and collaborative filtering mechanisms for gaining the advantages of both mechanisms.

Okon et al. (2018) design and develop a recommendation model that uses an object-oriented analysis and design methodology, including an improved collaborative filtering algorithm and an efficient quicksort algorithm to filter, prioritise and efficiently deliver relevant information by using relevant recommender systems. Accordingly, they achieve that by implementing a model based on the use of the python Model-View-Controller (MVC) framework, which is known as the Django Framework that applies a real-time cloud-hosted NOSQL database called the FireBase database. Okon et al. find out that the speed and scalability of book recommendations are improved within the range of 90% to 95% with a performance record that includes several recommendations, which are obtained from the system.

2.2 Recommender Systems Using Content Based Method

Abu Samra (2017) proposes an Arabic language tag recommender system by using Arabic Wikipedia as a source of information. Further, Abu Samra uses the analysis of the latent semantics to detect the encountered similarities between short text and Wikipedia articles. Moreover, Abu Samra uses the Apache Spark in order to deal with the huge volume of Wikipedia and complex calculations pertaining to the latent semantic analysis that is used to analyse the content of Wikipedia articles into three matrices. In the evaluation process, Abu Samra determines the number of the top articles the system must use in order to result in a qualified and considerable number of tags. After that, Abu Samra assesses the system by runnuing the tag recommender onto the dataset where the obtained results are recorded accordingly. Abu Samra also uses two experts to check the outputs of the system. The system is assessed by over 100 tweets and it achieves 84.39% of the mean average precision, and 96.53% of the mean reciprocal rank, while it faces difficulties that are related to the Arabic language and repetitions of rare words.

Al-Malahmeh (2014) proposes a semantic Malaysian tourism recommender system to provide a personalised information, particularly, in the field of tourism that incorporates the semantic technology with a recommender system for delivering the information that is more related to the interests of the tourists. Furthermore, Al-Malahmeh constructs the system based on the content and integration of the Natural Language Interface and Content-based Recommender system and incorporates the semantic technology for the Malaysian tourism web service.

Additionally, Al-Malahmeh evaluates the validity of the system by using Pellet and Fact++, to evaluate the Malaysia Tourism Ontology inference. The information retrieval performance is achieved by using the precision and recall for measuring the retrieval effectiveness, including the usability by using a questionnaire for measuring users' satisfaction.

2.3 Recommender Systems Using Hybrid Method

FERNÁNDEZ (2018) provides a useful suggestion of products to online users in order to increase their consumption on websites and build a movie recommendation mechanism within the Netflix. FERNÁNDEZ covers the most popular recommendation system algorithms, collaborative filtering, content-based filtering and hybrid approaches. The aim of his research is to understand the pros and cons of the entire related algorithms, and decide which algorithm is the best that can be a suitable fit for the dataset. Based on his discussion, popularity and collaborative filtering approaches are implemented for collaborative filtering that uses the memory-based and model-based approaches. The problem with popularity is that all recommendations are the same for every single user, and hence, FERNÁNDEZ do not focus on these results. The memory-based approach is based on the similarity that occurs among users or items. The user-based collaborative filtering approach is not implemented due to the large ratio between the number of users and items within the system. After that, the accuracy of the system is not considered the best and is computationally inefficient. The item-based collaborative filtering approach is implemented by using the cosine and the Pearson correlation as the distance function.

Ali et al. (2016) presents a hybrid book recommender system that combines Content-Based (CB) filtering and the Collaborative Filtering (CF) approaches that are accompanied with the association rule mining and books, in order to search for relevant books that are of interest to its readers and to deal with the limitations in quality, accuracy and precision of recommendations' criteria. Additionally, Ale et al. propose a recommender system, which proceeds through different filtering approaches that form the collaborative filtering recommendation for extracting books' features. In fact, the content-based recommendation that is used by the TF-IDF to deal with book contents, and association rule-mining are combined to obtained from the content-based and collaborative filtering approaches.

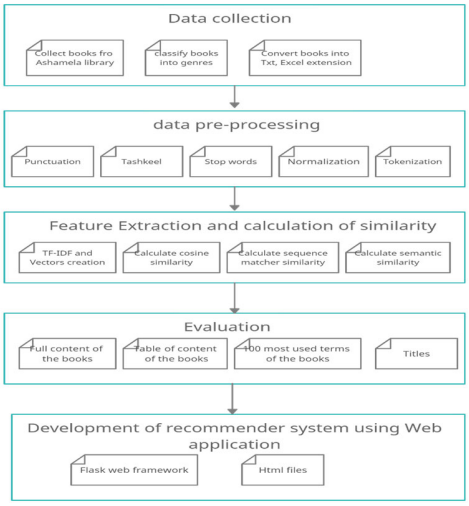
Ali et al. construct their recommendation system database from the information of users and books. They ask users to provide their interests during their registrations where their behaviours and ratings are automatically monitored. Moreover, Ali et al. extract the information of the given books that include the books' titles, chapters, parts and table of contents. Moreover, Ali et al. evaluate their system for the impact of the association rule mining on the generated results of the system where the accuracy of the generated results is checked by using several metrics that comprise; precision, recall and f-measure. Finally, Ali et al. conclude that the recommender system of the hybrid book provides recommendations' results that are better and more accurate in comparison with the results, which are generated by the CB and CF approaches. In fact, these results are further refined based on the association rule mining.

Mathew et al. (2016) state that a Book Recommendation System (BRS) apply the content based filtering, collaborative filtering and association rule mining approaches to produce a recommendation that is based on the purchaser's interest. Further, Mathew et al. use the collaborative filtering approach for creating a model based on different behaviours and items, which are purchased by the user. In fact, rating for these items assists in predicting the item that may be of interest to a user. Furthermore, the content based filtering approach is applied for creating a model that is based on the description or the content of a particular item. Mathew et al. analyse and process the extracted by data using different data mining methods and techniques.

The major problem that is faced by the researchers is to develop a new website application for book selling and to implement their recommendation system according to the content based filtering and collaborative filtering approaches whether the obtained feedback and rating of the users are believable or not.

3. Methodology

The methodology is divided into different sub-sections. The data collection shows how the dataset is developed. The data pre-processing provides the steps of the pre-processing phase. The feature extraction and the calculation of the similarity among the documents are used. The TF-IDF, cosine similarity and semantic similarity are evaluation measures of performance, which highlight the demonstration of developing the recommender system based on the use of the web application and its parts.



3.1 Data Collection

The first step of the proposed method is to collect the data through the following steps:

Books' data collection: 250 books are collected from the Ashamela library, which are divided into five genres where each genre includes 50 books. These books are converted into a .txt extension, and hence, all of the books are converted into a single excel sheet.

Data Collection Challenges: due to the lack of the Arabic contents, collecting the books is difficult and is time consuming. There is no efficient platform for Arabic books, and if there is any, it lacks of arrangement and flexibility. Ashamela's library consists of a reasonable amount of books, but with limited extensions (.epub) that are dedicated to phones and (.bok), which only works with the Ashamela application. Therefore, 250 books that are downloaded from the Ashamela library have the .epub extension. Consequently, each book is converted into .txt file where all of these books are compiled into an excel sheet to be easily read by the Python programming language.

3.2 Data Preprocessing

The second step of the proposed method is the preprocessing section, which is a process of converting data to something a computer can understand and improve. Arabic books are preprocessed by using the Python programming language. In particular, the Python preprocessing functions that are implemented for Arabic books comprise the punctuation removal, remove tashkeel, stop words removal, normalization and tokenization.

3.3 Feature Extraction and the Calculation of the Similarity among Documents

Books' content needs to be converted to vectors in order to use a particular technique for retrieving the feature, such as the cosine similarity that works on vectors. These vectors are implemented via the TF-IDF.

The TF-IDF stands for the Term Frequency – Inverse Document Frequency where it is one of the most important techniques that is used to retrieve the information for representing the importance of a particular word or phrase to a given book. The TF-IDF uses two statistical methods. The first method is the Term Frequency method, which refers to the total number of times a particular term appears in a book against the total number of all words in the book. The second method is the Inverse Document Frequency method that measures the amount of information a word provides and the weight of a given word in the entire document.

The sklearn.feature_extraction module can be used to extract features in a format that is supported by different machine learning algorithms when using a number of particular datasets. The TfidfVectorizer converts books to a matrix of TF-IDF features. The vectors that are resulted from the TF-IDF are used in the cosine similarity and semantic similarity to recommend books based on a selected book. The cosine similarity calculates the similarity between the vectors, which are produced from the TF-IDF, while the semantic similarity that calculates the distance among items is based on their meanings or semantic contents.

3.4 Evaluation

To evaluate the results of the recommendation system and test the performance of the work, 10 books are chosen from different genres. Following that, three persons are asked to compare these books and decide on which books are similar to each other and rank them in the matrix from 5 to 1, where 5 denotes the high similarity and 1 denotes the not similar. After that, the results of the average is taken from the human assigned/annotated similarity and are compared with the results of the cosine similarity and semantic similarity.

The cosine similarity and semantic similarity are applied to books in multiple cases where the full contents of the books, the table of content, the most 100 used terms of the books and the titles of books, and then the results are ranked from 5 to 1 by using the IF condition. In each case, the results are compared with the average of the human assigned/annotated similarity to determine the closest result and the most similar book.

3.5 The Development of the Recommender System Using the Web Application

The web of the Arabic books recommender system is intended to prove and display the results of different operations performed on the used data. The web is simply created to apply the entire experiences into a single platform. The web application is created by using the Flask web framework, which is written in Python and HTML files.

The first page displays all the experiments that are performed on the system. This implies that the experiment

is chosen based on the recommendation of the system. Moreover, another page opens by selecting the menu of all the books, which are used in the system to choose from them for further recommendations. When pressing the 'predict' button, the result page appears and presents the name of the chosen book where it shows five recommended books based on the cosine similarity and five recommended books based on the semantic similarity. Finally, the 'home' button returns to the home page so that it can choose a new experience.

4. Experimental Analysis and Results

4.1 Discussion of the results of the experiments:

For the first experiment, the recommendation based on the full contents of the books, three persons are asked to read 10 books with full contents by considering that the human assigned/annotated similarity is correct, and the nominated results are the closest to the similarity. The average of the results of similarity among these books are ranked in the matrix from 5 to 1, where 5 denotes the high similarity and 1 denotes the not similar (see Table 4.1). After that, the results are compared with the results of the

	أحاديث إسماعيل بن جعفر-متون الحديث	أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	أخبار الدولة العباسية-التاريخ	أصول بلا أصول-الفرق والردود	الإبانة عن أصول الديانة-العقيدة	السنن المأثورة للشافعي-متون الحديث	السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	الفقه الأبسط-العقيدة	تاريخ بيت المقدس-التاريخ	فضائل القرآن للقاسم بن سلام-علوم القران
أحاديث إسماعيل بن جعفر-متون الحديث	-	5	1	1	1	5	1	4	1	5
أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	5	-	1	1	1	5	1	4	1	5
أخبار الدولة العباسية-التاريخ	1	1	-	3	2	1	3	1	2	1
أصول بلا أصول-الفرق والردود	1	1	3	-	3	1	2	1	2	1
الإبانة عن أصول الديانة-العقيدة	1	1	2	3	-	1	3	1	3	1
السنن المأثورة للشافعي-متون الحديث	5	5	1	1	1	-	1	5	1	5
السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	1	1	3	2	3	1	-	1	2	1
الفقه الأبسط-العقيدة	4	4	1	1	1	5	1	-	1	5
تاريخ بيت المقدس-التاريخ	1	1	2	2	3	1	2	1	-	1
فضائل القرآن للقاسم بن سلام-علوم القران	5	5	1	1	1	5	1	5	1	-

cosine similarity

Table 4.1: The average of the human assigned/annotated similarity matrix for the full contents of books

of the full contents of books, and the results are ranked into a matrix from 5 to 1 by using the IF condition (see Table 4.2) along with the sequence matcher similarity and the semantic similarity of the full contents of books.

The discussed results below are the closest results obtained from the experiments to the average of the human assigned/annotated similarity results.

	أحاديث إسماعيل بن جعفر-متون الحديث	أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	أخبار الدولة العباسية-التاريخ	أصول بلا أصول-الفرق والردود	الإبانة عن أصول الديانة-العقيدة	السنن المأثورة للشافعي-متون الحديث	السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	الفقه الأبسط-العقيدة	تاريخ بيت المقدس-التاريخ	فضائل القرآن للقاسم بن سلام-علوم القران
أحاديث إسماعيل بن جعفر-متون الحديث	-	5	1	1	1	5	1	4	1	5
أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	5	-	1	1	1	5	1	4	1	5
أخبار الدولة العباسية-التاريخ	1	1	-	3	2	1	3	1	2	1
أصول بلا أصول-الفرق والردود	1	1	3	-	3	1	2	1	2	1
الإبانة عن أصول الديانة-العقيدة	1	1	2	3	-	1	3	1	3	1
السنن المأثورة للشافعي-متون الحديث	5	5	1	1	1	-	1	5	1	5
السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	1	1	3	2	3	1	-	1	2	1
الفقه الأبسط-العقيدة	4	4	1	1	1	5	1	-	1	5
تاريخ بيت المقدس-التاريخ	1	1	2	2	3	1	2	1	-	1
فضائل القرآن للقاسم بن سلام-علوم القران	5	5	1	1	1	5	1	5	1	-

Table 4.2: The cosine similarity matrix for the full contents of books

The results of the first three books that are obtained from the cosine and semantic similarity are discussed and compared with the results of the average of the human assigned/annotated similarity.

For the first book, namely, أحاديث إسماعيل بن جعفر- متون الحديث, the most similar books to it are based on the average of human assigned/annotated similarity, which are the السنن المأثورة للشافعي-متون الحديث and فضائل القرآن للقاسم بن سلام-علوم القران where these are rated with a scale of 5. After that, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران and الفقه الأبسط-العقيدة books are rated with a scale of 4. In the cosine similarity, the most similar books include the السنن المأثورة للشافعي-متون الحديث and فضائل القرآن للقاسم بن سلام-علوم القران and أحكام القرآن للجصاص تحقق قمحاوي-علوم القران, which are rated with a scale of 5 and the الفقه الأبسط-العقيدة book is rated with a scale of 4.

The second book is the أحكام القرآن للجصاص تحقق قمحاوي- علوم القران where the most similar books to it are based on the average of the human assigned/annotated similarity, which is the فضائل القرآن للقاسم بن سلام-علوم القران book that is rated with a scale of 5. Following that is the أحاديث إسماعيل بن جعفر-متون الحديث and السنن المأثورة للشافعي-متون الحديث books, which are rated with a scale of 4 and the الفقه الأبسط-العقيدة book, which is rated with a scale of 3. In the cosine similarity, the most similar books comprise the أحاديث إسماعيل بن جعفر-متون الحديث and السنن المأثورة للشافعي-متون الحديث and فضائل القرآن للقاسم بن سلام-علوم القران books, which are rated with a scale of 5, and the الفقه الأبسط-العقيدة book, which is rated with a scale of 4.

The third book, namely, the أخبار الدولة العباسية-التاريخ, does not include much similar books to it based on the average human assigned/annotated similarity where the highest rating of the books reaches two for the books أصول بلا أصول-الفرق والردود, السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود and- تاريخ بيت المقدس-التاريخ. In the cosine similarity, the results are extremely close to the results of the average of the human assigned/annotated similarity where the books أصول بلا أصول-الفرق والردود, السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود, and تاريخ بيت المقدس-التاريخ are rated with scales of 3, 3 and 2, respectively..

The results of the cosine similarity in the first experiment are better than the results of the sequence matcher similarity, and the semantic similarity and the closest results to the average of the human assigned/annotated similarity results.

In the second experiment, the obtained results of the cosine similarity, the sequence matcher similarity, and semantic similarity pertaining to the books' table of contents vary from the results of the average of the human assigned/annotated similarity, and hence, none of these results are considered the best in this experiment.

In the third experiment, the obtained results of the cosine similarity, the sequence matcher similarity, and the semantic similarity pertaining to the most 100 used terms of the book vary from the results of the average of the human assigned/annotated similarity. Therefore, none of these results are considered the best in this experiment.

In the fourth experiment, the recommendation based on the titles of the books, the results of the average of the human assigned/annotated similarity matrix (see Table 4.1) are used to be compared with the results of the cosine similarity, sequence matcher similarity (see Table 4.3) and semantic similarity of the Arabic titles of the books. The cosine similarity, sequence matcher similarity and semantic similarity are applied to the title of each book where the results are ranked into a matrix from 5 to 1 by using the IF condition.

The discussed results below are the closest results obtained from the experiments to the average of the human assigned/annotated similarity results.

	أحاديث إسماعيل بن جعفر-متون الحديث	أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	أخبار الدولة العباسية-التاريخ	أصول بلا أصول-الفرق والردود	الإبانة عن أصول الديانة-العقيدة	السنن المأثورة للشافعي-متون الحديث	السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	الفقه الأبسط-العقيدة	تاريخ بيت المقدس-التاريخ	فضائل القرآن للقاسم بن سلام-علوم القران
أحاديث إسماعيل بن جعفر-متون الحديث	-	1	2	2	2	3	1	2	2	3
أحكام القرآن للجصاص تحقق قمحاوي-علوم القران	1	-	2	2	2	2	2	2	2	4
أخبار الدولة العباسية-التاريخ	2	2	-	2	2	2	2	3	3	2
أصول بلا أصول-الفرق والردود	2	2	2	-	2	2	3	2	2	2
الإبانة عن أصول الديانة-العقيدة	2	2	2	2	-	2	2	3	2	2
السنن المأثورة للشافعي-متون الحديث	3	2	2	2	2	-	2	2	2	3
السيوف المشرقة ومختصر الصواقع المحرقة-الفرق والردود	1	2	2	3	2	2	-	2	2	2
الفقه الأبسط-العقيدة	2	2	3	2	2	3	2		-	2
تاريخ بيت المقدس-التاريخ	2	2	2	2	2	2	2	2	-	2
فضائل القرآن للقاسم بن سلام-علوم القران	3	4	2	2	2	3	2	2	2	-

Table 4.3: The sequence matcher similarity matrix for the titles of books

The first book, namely أحاديث إسماعيل بن جعفر-متون الحديث, in the sequence matcher similarity, all books are rated with a scale of 2 and 1 except the السنن المأثورة للشافعي-متون الحديث book is rated with a scale of 3, while in the average of the human assigned/annotated similarity is rated with a scale of 5. The فضائل القرآن للقاسم بن سلام-علوم القران book is rated with a scale of 3, while in the average of the human assigned/annotated similarity is rated with a scale of 5. It can be seen that the highest scales in the average of the human assigned/annotated similarity are those that are rated with a scale of 3 in sequence matcher similarity, and the book of the same genre is rated with a higher scale.

For the second book, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران book, in the sequence matcher similarity, all books are rated with a scale of 2 and 1 except the فضائل القرآن للقاسم بن سلام-علوم القران book is rated with a scale of 4, while in the average of the human assigned/annotated similarity is rated with a scale of 5. It can be seen that, only the book of the same genre, which is علوم القران, is rated with a higher scale.

For the third book, the أخبار الدولة العباسية-التاريخ book, in the sequence matcher similarity, all books are rated with a scale of 2 except the تاريخ بيت المقدس-التاريخ, الفقه الأبسط-العقيدة books are rated with a scale of 3.
It can be seen that, the book of the same genre is rated with a higher scale.
The semantic similarity depends on the meaning of the text, and from the obtained results, it appears that the Bert code cannot read Arabic words correctly, so the titles of the books are translated into English and the Bert code are applied to it. Google translate is used to translate the books titles and then the result of semantic similarity of English titles is presented in the below table (See Table 4.4) and compared with the average of the human assigned/annotated similarity (Table 4.1). For the first book namely, the hadiths of Ismail bin Jaafar - the text of the hadith book, the high rated book with scale of 4 is The Sunnahs of al-Shafi'i - the text of the hadith book, in the average of the human assigned/annotated similarity, it is rated with a scale of 5. The Provisions of the Qur'an for Jasas Check Qamhawi - Qur'an Sciences, Simple Fiqh – Creed, and the Virtues of the Qur'an by Al-Qasim bin Salam - Qur'an Sciences books are rated with a scale of 3, while in the average of the human assigned/annotated similarity are rated with a scale of 4 and 5.

For the second book, the Provisions of the Qur'an for Jasas Check Qamhawi - Qur'an Sciences book, the high rated book with scale of 4 is The Virtues of the Qur'an by Al-Qasim bin Salam - Qur'an Sciences book, in the average of the human assigned/annotated similarity, it is rated with a scale of 5.
The hadiths of Ismail bin Jaafar - the text of the hadith, The Sunnahs of al-Shafi'i - the text of the hadith books are rated with a scale of 3, while in the average of the human assigned/annotated similarity are rated with a scale of 4.
For the third book, Abbasid state news-history book, all books are rated with a scale of 2 and 3, and in the average of the human assigned/annotated similarity all books rated with a scale of 1 and 2.

5. Analysis of the results of the experiments

In each experiment, the results of cosine similarity and semantic similarity are subtracted from human assigned/annotated similarity to find out how much

The Virtues of the Qur' an by Al-Qasim bin Salam - Qur'an Sciences	The History of Jerusalem - History	Simple Fiqh - Creed	Bright swords and a summary of the burning reality -	The Sunnahs of al-Shafi' i - the text	Evidence for the origins of religion -	Origins without origins - the	Abbasid state news-history	Provisions of the Qur' an for Jasas Check Qamhawi - Qur'an Sciences	The hadiths of Ismail bin Jaafar -	
3	2	3	2	4	2	2	2	3	-	The hadiths of Ismail bin Jaafar - the text of the hadith
4	2	2	2	3	3	2	3	-	3	Provisions of the Qur'an for Jasas Check Qamhawi - Qur'an Sciences
2	2	2	2	3	2	2	-	3	2	Abbasid state news-history
2	2	1	2	2	2	-	2	2	2	Origins without origins - the difference and responses
4	3	2	2	2	-	2	2	3	2	Evidence for the origins of religion - creed
3	2	3	2	-	2	2	3	3	4	The Sunnahs of al-Shafi'i - the text of the hadith
3	2	1	-	2	2	2	2	2	2	Bright swords and a summary of the burning reality - the difference and responses
2	1	-	1	3	2	1	2	2	3	Simple Fiqh - Creed
4	-	1	2	2	3	2	2	2	2	The History of Jerusalem - History
-	4	2	3	3	4	2	2	4	3	The Virtues of the Qur'an by Al-Qasim bin Salam - Qur'an Sciences

Table 4.4: The semantic similarity matrix for the English titles of books

there are differences between them and what results are closest to the human assigned/annotated similarity.

5.1 Analysing the results of the cosine similarity, sequence matcher similarity, and semantic similarity of the full contents of books

For the first book, the أحاديث إسماعيل بن جعفر-متون الحديث. The results of differences between cosine similarity and human assigned/annotated similarity are less than the differences between sequence matcher similarity and human assigned/annotated similarity and the differences between semantic similarity and human assigned/annotated similarity and very small in the entire books which is between 0.00267 and 0.15333. In sequence matcher-human difference, the difference with some books reached to 0.795787, 0.996281, 0.798139, 0.996902 and 0.996902 which makes the results of similarity between them little to zero. In semantic-human difference, the difference with some books reached to 0.774937, 1.048501, 0.768674 and 1.09756 which makes the results of similarity between them little to zero.

For the second book, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران. The results of differences between cosine similarity and human assigned/annotated similarity are less than the differences between sequence matcher similarity and human assigned/annotated similarity and the differences between semantic similarity and human assigned/annotated similarity and very small in the entire books which is between 0.016 and 0.15416. In sequence matcher -human difference, the difference with some books reached to 0.795787, 0.797192, 0.597923 and 0.99878 which makes the results of similarity between them little to zero. In semantic-human difference, the difference with some books reached to 0.774937, 0.79385 and 1.061997 which makes the results of similarity between them little to zero.

For the third book, the أخبار الدولة العباسية-التاريخ. The results of differences between cosine similarity and human assigned/annotated similarity are less than the differences between sequence matcher similarity and human assigned/annotated similarity and the differences between semantic similarity and human assigned/annotated similarity and very small in the entire books which is between 0.00299 and 0.17822. The differences resulting from the sequence matcher-human differences and the semantic-human differences are not very high like other books due to the small results of similarity that have been assigned by humans with all compared books. The sequence matcher -human difference reached to 0.398191, 0.396543, and 0.398191 in some books. The semantic-human difference reached to 0.58601, 0.455358, and 0.420849 in some books, which makes them high differences compared to the cosine-human difference results.

The cosine similarity results are better than the sequence matcher similarity and semantic similarity results. The results of cosine similarity are closest to human assigned/annotated similarity, while the results of the sequence matcher similarity and the semantic similarity are vary.

The sequence matcher similarity depends on the similarity of the input sequences or strings and find the longest contiguous matching subsequence.

The semantic similarity calculation depends on the meaning of the text and the used books are contain a huge content up to a million words, so the results of the sequence matcher similarity and the semantic similarity are very small, not accurate and the differences cannot be considered as a significant value. The difference when comparing it with human assigned /annotated similarity will decrease if it is small and will increase if the book is more similar.

5.2 Analysing the results of the cosine similarity, sequence matcher similarity, and semantic similarity for the table of contents of the books

For the first book, the أحاديث إسماعيل بن جعفر-متون الحديث. In the results of the cosine-human difference, there are high differences reached to 0.78438, 0.99834, 0.78391 and 0.98722. The results of differences between sequence matcher similarity and human assigned/annotated similarity are disparate, some of the differences reached to 0.747612, 0.94375, 0.761661, and 0.934884, which are considered high differences, and there are small differences that cannot take into account. In the semantic-human difference, there are high differences reached to 0.5992, 0.452125, 0.4149 and 0.398646.

For the second book, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران. In the results of the cosine-human

difference, there are high differences reached to 0.78438, 0.79326, and 0.96815. The results of differences between sequence matcher similarity and human assigned/annotated similarity are disparate, some of the differences reached 0.747612, 0.734579, 0.558599, and 0.935085, which are considered high differences, and there are small differences that cannot take into account. In the semantic-human difference, there are high differences reached to 0.427142, 0.44311, and 0.42811.

For the third book, the أخبار الدولة العباسية-التاريخ. The results of differences between sequence matcher similarity and human assigned/annotated similarity are less than the differences between cosine similarity and human assigned/annotated similarity and the differences between semantic similarity and human assigned/annotated similarity in the most of the books. The differences resulting from the semantic-human difference are not very high like other books due to the small results of similarity that have been assigned by humans with all compared books. The results of the sequence matcher-human difference reached to 0.339698, and 0.329114 in some books.

In the results of the cosine-human difference, the differences reached to 0.4, 0.3603, and 0.38889. In the semantic-human difference, there are high differences reached to 0.5992, 0.465641, 0.419775, 0.469616, and 0.419443.

All the books compared are Islamic and had a lot of similar words and sentences, and their tables of contents may contain similar words and identical meanings. The results of calculating the sequence matcher similarity and semantic similarity of the books' table of contents are close to each other and considered that all of the books are similar to the comparative book.

5.3 Analysing the results of the cosine similarity, sequence matcher similarity, and semantic similarity for the most 100 used terms of books

For the first book, the أحاديث إسماعيل بن جعفر-متون الحديث. The results of the cosine-human difference reached to 0.62603, 0.65908, and 0.7509. In the sequence matcher-human difference, there are high differences reached to 0.714894, 0.933071, and 0.739271. In the semantic-human difference, there are high differences reached to 0.706885, 0.706975, 0.673411, 0.656146 and 0.688619.

For the second book, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران. The results of the cosine-human difference reached to 0.62603, 0.64206, and 0.85739, which are still considered high differences. In the sequence matcher-human difference, there are high differences that reached to 0.714894, 0.732939, 0.547262, and 0.925781. In the semantic-human difference, there are high differences that reached to 0.659661, 0.673554, 0.68445, 0.619445 and 0.618806. For the third book أخبار الدولة العباسية-التاريخ. The results of the cosine-human difference and sequence matcher-human difference are not very high like other books due to the small results of similarity that have been assigned by humans with all compared books. The results of the cosine-human difference reached to 0.24435, 0.22066, and 0.279 in some books. In the sequence matcher-human difference, the differences reached to 0.352381, 0.309465, and 0.338998. In the semantic-human difference, there are high differences reached to 0.706885, 0.659661, 0.665268, and 0.667719.

The results obtained when finding the 100 most used terms that are similar and identical for all books (e.g. الله، قال، بن، صلي، وسلم، الحديث، رسول، عبد، حدثنا ،عمر so ،دينار، سمع، انس، يقول النبي، الانصار، الناس، فقالوا) calculating cosine similarity and semantic similarity for the most 100 used terms of each book gives similar and not accurate results.

5.4 Analysing the results of the cosine similarity, sequence matcher similarity, and semantic similarity for the titles of books

For all of the discussed books, It can be seen that all the results that appeared from calculating the cosine similarity are 0 except the books of the same genre that show very poor results, which indicates that they were calculated based on the genre mentioned in each title. Therefore, the results of the differences of cosine similarity and human assigned/annotated similarity did not show any useful results.

For the first book, the أحاديث إسماعيل بن جعفر-متون الحديث. The results of the sequence matcher-human differences reached to 0.615789 and 0.535849 in three books only and the difference in other books is small. The results of the semantic-human differences reached to 0.55, and 0.45.

For the second book, the أحكام القرآن للجصاص تحقق قمحاوي-علوم القران. The results of the sequence matcher-

human differences reached to 0.583333 and 0.615789 in two books only and the difference in other books is small. The results of the semantic-human difference, there are high differences that reached to 0.31, 0.39, and 0.31.

For the third book أخبار الدولة العباسية-التاريخ. The results of the sequence matcher-human difference are not very high like other books due to the small results of similarity that have been assigned by humans with all compared books. The results of the semantic-human difference, the differences reached 0.31, 0.36, and 0.33.

Clearly, the obtained results of the sequence matcher similarity are considered the best in this experiment and better than the results of cosine similarity and semantic similarity, but the results of the semantic similarity differences are small in all compared books. In an attempt to understand Bert code that used to calculate the semantic similarity, and how it reads each word in the title, a code to display the words with their indices or tokens was written. The code of bert to display the words with their tokens and the results of comparing the أحاديث إسماعيل بن جعفر-متون book with the أحكام القرآن للجصاص تحقق قمحاوي-الحديث book is shown in the figure below 5.1).

```
10,273          احاديث
5,068          اسماعيل
1,802              بن
6,233            جعفر
17                  -
1,891             مت
1,733           ##ون
2,746          الحديث
6,007          احكام
2,930          القران
3,843             للج
4,500          ##صاص
7,991          تحقق
4,479             قم
29,890         ##حاوي
17                  -
4,977          علوم
2,930          القران
Vector similarity for  *similar*  meanings:  0.56
```

Figure 5.1: Result of bert code to display the words with their tokens

The result of the bert code in figure 4.14 shows each word in the first title أحاديث إسماعيل بن جعفر-متون الحديث with their tokens, and each word in the compared title أحكام القرآن للجصاص تحقق قمحاوي-علوم القران with their tokens. Obviously, the Bert code cannot read all the words in the Arabic language, and the semantic similarity is based on the meaning of the text so if the word is not read correctly, the meaning will be completely different. For example, the word متون has not been read and is divided into two parts مت and ون, the word للجصاص has not been read and is divided into two parts للج and صاص, the word قمحاوي has not been read and is divided into two parts قم and حاوي. The meaning of these words are changed to different meanings which are unable to compared in this way so the obtained results of similarity 0.56 cannot be considered correct.

Many Arabic words were read in this way using Bert code, so all results obtained cannot be considered correct. In an attempt to make Bert code read all the words correctly, all book titles are translated into English and applied semantic similarity to them using Bert code.

5.5 Analysing the results of the semantic similarity for the English titles of books

For the first book, the hadiths of ismail bin jaafar - the text of the hadith. The results of the semantic-human differences in most of the results did not exceed 0.3, which can be considered very small.

For the second book, the provisions of the qur'an for jasas check qamhawi - qur'an sciences. The results of the semantic-human differences in all the results did not exceed 0.3, which can be considered very small.

For the third book, the abbasid state news-history. The results of the semantic-human differences in all the results did not exceed 0.3, which can be considered very small.

All results are very close to the average of the human assigned/annotated similarity, so it is clear that the result of semantic similarity in English titles is better than the result in Arabic titles.

Clearly, the obtained results of the semantic similarity of the English titles are considered the best in this experiment and better than the results of cosine similarity and semantic similarity of the Arabic titles.

5.6 Comparisons between the cosine similarity, the sequence matcher similarity and the semantic similarity

The cosine similarity is a mathematical function that is used to calculate the distance between two vectors, which are produced from the TF-IDF function to find the similarity among different texts based on the similar words that are available in both texts.

The sequenceMatcher is an available class in the difflib Python package. SequenceMatcher used to compare the similarity of two input sequences or strings and produce information about file differences in various formats. In other words, this class is useful to use when finding similarities between two strings on the character level.

On the other hand, the semantic similarity is a metric that is defined over a set of texts or terms where the idea of distance among these items is based on the likeness of their meanings or semantic contents. Furthermore, this type of similarity is a broader term for applying different similarity calculation techniques, which attempt to find out the similarity among different texts based on the meaning of the text. In the full contents of the books experiment, the doc2vec model is used to find the semantic similarity while BERT model is used in all other experiments. Bidirectional Encoder Representations from Transformers or BERT has been a popular technique in NLP. It is designed to help computers understand the meaning of ambiguous language. BERT model can output 512 tokens to apply our similarity measures to it, and the full contents of the used books are huge and reached up to million word so for the full contents experiment doc2vec model is used to calculate semantic similarity. Doc2Vec is a model that represents each document as a vector. It is implemented using Python and Gensim..

6. Conclusion and Future Research

This chapter presents an overview of the thesis's approaches and results that are revealed in the previous chapters. The main objective pertaining to this thesis is to build a recommendation system for Arabic books by following these approaches:

1. Manually collect books and convert them into an excel file.
2. Mapping each book into its genres based on the Alshamela library.
3. Proposing a number of mechanisms that facilitate various preprocessing techniques for alleviating the texts' noise with punctuation, stop words' removal, and normalization.
4. Developing a method for converting texts into vectors by using TF-IDF features in the Python programming language.
5. Comparing the performance of different classification results and applying them to a simple interface.

The proposed method is used to validate the obtained results of the recommendation system by comparing them with the results, which are obtained from the cosine application and semantic similarity for 10 selected books of each genre, with the average of the human assigned/annotated similarity.

In terms of recommending books based on the book's full content, the obtained results of the cosine similarity are found better than the obtained results of the sequence matcher similarity and semantic similarity and are extremely approaching from the obtained results of the human assigned/annotated similarity's average.

In terms of recommending books based on the book's table of content, the obtained results of the cosine similarity, the sequence matcher similarity and semantic similarity pertaining to a book's table of contents vary from the obtained results of the average of human assigned/annotated similarity, and hence, none of these results are considered the best for this experiment.

In terms of recommending a book according to the most 100 used terms of books, the obtained results of the cosine similarity, the sequence matcher similarity and semantic similarity of the most 100 used terms of books vary from the obtained results of the human assigned/annotated similarity's average, and hence, none of these results are considered the best for this experiment.

In terms of recommending a book according to the books' titles, the obtained results of the cosine similarity pertaining to the titles of the books vary from the results of the average of the human assigned/annotated similarity. The results of semantic similarity of English titles are better than the results of Arabic titles and are close to the human assigned/annotated similarity's average. The results of sequence matcher similarity have a slight similarity rate with the average of the human assigned/annotated similarity based on the title found in the name of each book. Therefore, the results of sequence matcher similarity and the results of semantic similarity of English titles are considered the best in this experiment.

6.1 Recommendations for the Future Research

This dissertation can be further enhanced in order to deliver further effective and accurate results. In particular, the below points highlight some of the tasks that can be further improved for the future research:

- Adding more books of different genres such as stories and novels to realize the differences and similarities among these books as they use all Islamic books that are not simply differentiated among them.
- Using the word embedding vectors rather than the TF-IDF vectors, which carry further information in comparison with the TF-IDF vector. However, they are more memory intensive.
- Using a cloud solution to increase the memory usage, adding more books, increasing the speed and reducing computer's freezing and hanging issues.

References

[1] Burke, Robin & Felfernig, Alexander & H. Göker, Mehmet. (2011). Recommender Systems: An Overview. Ai Magazine. 32. 13-18. 10.1609/aimag.v32i3.2361.

[2] Khusro, Shah & Ali, Zafar & Ullah, Irfan. (2016). Recommender Systems: Issues, Challenges, and Research Opportunities. 10.1007/978-981-10-0557-2_112.

[3] Bhatnagar, V. (2017). Collaborative filtering using data mining and analysis. Hershey, PA: Information Science Reference, An imprint of IGI Global.

[4] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-ite m collaborative filtering. IEEE Internet Computing, 7(1), 76-80. doi:10.1109/mic.2003.1167344

[5] Gómez-Uribe, Carlos & Hunt, Neil. (2015). The Netflix Recommender System. ACM Transactions on Management Information Systems. 6. 1-19. 10.1145/2843948.

[6] FERNÁNDEZ, L. E., M. (2018). Recommendation System for Netflix. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-fernandez_tcm235-874624.pdf

[7] Liu, Jiahui & Dolan, Peter & Pedersen, Elin. (2010). Personalized news recommendation based on click behavior. International Conference on Intelligent User Interfaces, Proceedings IUI. 31-40. 10.1145/1719970.1719976.

[8] Okon, Emmanuel & Eke, Bartholomew & Oghenekaro Asagba, Prince. (2018). An Improved Online Book Recommender System using Collaborative Filtering Algorithm. 10.13140/RG.2.2.24240.46086.

[9] Ali, Z., Khusro, S., & Ullah, I. (2016). A Hybrid Book Recommender System Based on Table of Contents (ToC) and Association Rule Mining. Proceedings of the 10th International Conference on Informatics and Systems - INFOS 16. doi:10.1145/2908446.2908481

[10] Mathew, P., Kuriakose, B., & Hegde, V. (2016). Book Recommendation System through content based and collaborative filtering method. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). doi:10.1109/sapience.2016.7684166

[11] Abu Samra, Y. K. (2017). Tag Recommendation for Short Arabic Text by Using Latent Semantic Analysis of Wikipedia. , , غزة, - الجامعة الإسلامية. http://hdl.handle.net/20.500.12358/20061

[12] ALMALAHMEH, T. M. (2014). SEMANTIC RECOMMENDER SYSTEM FOR MALAYSIAN TOURISM INDUSTRY. Retrieved from http://studentsrepo.um.edu.my/4646/1/TIRAD_MOHAMMED_AREF_ALMALAHMEH.pdf

[13] Dou, Y., Yang, H., & Deng, X. (2016). A Survey of Collaborative Filtering Algorithms for Social Recommender Systems. 2016 12th International Conference on Semantics, Knowledge and Grids (SKG). doi:10.1109/skg.2016.014

[14] Chandak, M., Girase, S., & Mukhopadhyay, D. (2015). Introducing Hybrid Technique for Optimization of Book Recommender System. Procedia Computer Science, 45, 23-31. doi:10.1016/j.procs.2015.03.075

[15] Worked out example: Item based Collaborative filtering for Recommender Engine. (2014, December 30). Retrieved from https://ashokharnal.wordpress.com/2014/12/18/worked-out-example-item-based-collaborative-filtering-for-recommenmder-engine/

[16] Algorithms. (n.d.). Retrieved from http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/itembased.html

Shaykhah Hajed Alotaibi has obtained her bachelor's degree in information studies from Princess Nourah Bint Abdulrahman University and is currently preparing for her MSc. degree in Information systems at Al-Imam Muhammad Ibn Saud Islamic University.

Dr. Muhammad Badruddin Khan obtained his doctorate in 2011 from Tokyo Institute of Technology, Japan. He is a full-time a professor in department of Information Systems of Al-Imam Muhammad Ibn Saud Islamic University since 2012. The research interests of Dr. Khan lie mainly in the field of data and text mining. He is currently involved in number of research projects related to machine learning and Arabic language including pandemics prediction, Arabic sentiment analysis, improvement of Arabic semantic resources, Stylometry, Arabic Chatbots, trend analysis using Arabic Wikipedia, Arabic proverbs classification, cyberbullying and fake content detection, and violent/non-violent video categorization using YouTube video content and Arabic comments, and has published number of research papers in various conferences and journals. He is also co-author of a book on machine learning.