Experimenting with Arabic Text Visualizing

Nadim Obeid, Israa Huzayyen and Bassam Hammo
King Abdullah II School for Information Technology,
The University of Jordan, Amman, Jordan
b.hammo@ju.edu.jo

Abstract—The amount of unstructured textual data on the Internet has increased dramatically. Text visualization becomes a significant tool that facilitates knowledge discovery and insightful presentation of large amounts of data. In this paper, we present a technique of the visual exploration of Arabic text documents. We apply Latent Semantic Indexing (LSI) as a dimensionality reduction technique that helps in extracting some knowledge from text documents. We present an Arabic Visualization System. The experiments were conducted on several datasets extracted from an Arabic corpus. The system can create views from different perspectives based on the user’s needs. It is successfully used for visualizing different kinds of document corpora. The system is very helpful for data analysis offering quick insight into the structure of the visualized corpus.

Keywords—Latent Semantics, Visualization, Arabic Text

I. INTRODUCTION

Current search engines attempt at retrieving data that matches the user’s need in a fast way. Retrieved results are basically based on matching the query’s “bag of words” without paying much attention to the semantic of the query. To better understand and analyzing textual data, we need to get insight the documents to overview and extract the hidden semantic concepts from textual data. One way to do this is through text visualization.

Much of the text information comes in the form of unstructured text, which makes it hard to understand. Text visualization can help in many aspects such as: (1) to have a top level view of the topics in the corpora (2) to get insight into the corpora to explore relationships between the topics (3) to show the highly structured textual contents in a simplified way and (4) to show the main dimensions of a highly dimensional space of textual data.

In information retrieval (IR), textual documents are usually represented using the bag-of-words representation, where each word from the document vocabulary stands for one dimension of the multidimensional space of documents. This leads to a very high dimensionality which requires a lot of processing to analyze and search. Reducing the dimensionality [11] is important for different aspects of automated text processing including document visualization.

In our research, we apply Latent Semantic Indexing (LSI) as a dimensionality reduction technique that helps in extracting some knowledge from corpus of text documents LSI employs a technique from linear algebra called Singular Value Decomposition (SVD) and the bag-of-words representation of text documents for extracting words with similar meanings. The next step in the process is to use the multidimensional scaling technique to position the documents on a 2-dimensional plane which can be plotted as a graph on a computer screen [3].

In this paper, we present an implementation and we show the results of experimenting with an Arabic Visualization System (AVS). The experiments were carried on an experimental Arabic corpus and due to the limited size of this paper, we make all experiments and results available through http://nlp.ju.edu.jo. The paper is organized as follows. Section 2 shows a previous work on text visualization. Section 3 provides a short description of LSI and multidimensional scaling. Section 4 elaborates on the Arabic visualization system (AVS). The experimental results are given in Section 5, while Section 6 provides conclusions and future work.

II. PREVIOUS WORK

In [6], a visualization of Shakespearean scenes using the SPIRE Galaxies visualization approach is presented. Each dot represents a single document. Related documents are shown close together, while documents shown apart are less related. It proposed a visualization tool which they called a “Connex”. It allows the user to categorize the various relationships, hide or show particular relationships, highlight asymmetric relationships and reorder the document display.

In [10], ThemeView (formerly known as ThemeSpace) is described as a visualization system where themes within the document spaces appear on the computer screen as a relief map of natural terrain. The mountains indicate the dominant themes where the valleys indicate the weak themes. Themes close in content will be close visually based on the many relationships within the text spaces. The purpose of this system is to help the analyst to understand the clustered document collection.

In [5] a method for visualization of document corpus based on LSI to find the main topics being discussed in these documents is presented. It used the standard Bag-of-Words (BOW) presentation together with Term Frequency-Inverse Document Frequency (TF.IDF) weighting. In the BOW representation there is a dimension for each word; a word is encoded as a feature vector with word frequencies as elements. The space contains these vectors is called a semantic space. It proposed combining two methods which are linear subspace and multidimensional scaling. The first one likes using LSI and the second one aims at reducing the number of dimensions. The result is visualized documents in a 2-dimensions plan.
In [12], Kartoo is described as a visual meta search engine which produced an image representation of the results. Nodes are used to represent the different web pages, while links between nodes indicating the keywords that link the pages. The size of each node enabled the user to determine the relevance of each of the sites to the search terms, the larger the size the greater the match to the search criteria.

III. LATENT SEMANTIC ANALYSIS (LSA)

LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse [7]. This technique maps each term from a vector space representation to a vector in the reduced space. The advantages of the reduced term space are:

1. The dimensions in the space are uncorrelated (they are orthogonal).
2. The representations are less noisy.
3. The representations incorporate higher-order (latent) association structure among terms and documents.

Latent Semantic Indexing (LSI) uses a truncated Singular Value Decomposition (SVD) applied to the term-document matrix $X$. The SVD is a technique that based on a linear algebra. It is the foundation for the latent semantic analysis technique for document indexing and retrieval [4]. LSI applies the SVD method to the conventional term-document matrix $X$ with size $m \times n$ to decompose it to three matrices of special forms (see Eq. 1).

$$X = U [m \times r] . S [r \times r] . V^T [r \times n]$$

where $U$ and $V$ are orthogonal matrices that contain the left and right singular vectors of $X$, respectively, $U$ describes the original row entities (terms) and $V$ describes the original column entities (documents). $V^T$ is the transpose of $V$. $S$ is the diagonal matrix that contains the singular values of $X$, and the subscript $r$ denotes the number of singular values (the rank of $X$). If the singular values are sorted in descending order, SVD can project the data onto a lower, $k$ dimensional space spanned by their singular vectors corresponding to the $k$ largest singular values [9, 2, 8]. The new decomposition is shown in Eq. 2:

$$X^* = U^*[m \times k] . S^* [k \times k] . V^{T*} [k \times n]$$

In semantic spaces, the left and right singular vectors specify the locations of terms and documents respectively. Singular values are used to scale terms and documents in the space allowing clusters of terms and documents to be more organized. In the reduced space, the related terms or documents appear near each other’s after applying the SVD to the matrix [1].

A. Multidimensional scaling

Multidimensional scaling is a set of statistical techniques that often used in information visualization. It aims to explore the similarities or dissimilarities in data. the input of MDS is a proximity matrix (dissimilarities between terms). This matrix is a symmetric. We used the non metric MDS with the stress function (see Eq. 3). So the better the distances between points on the plane approximate real similarity between terms, the lower the value of the stress function. The stress equals zero only when distances between points match exactly with similarity between terms [5].

$$\text{Eq. 3: } \text{Stress} = \sum_{i \neq j} (\delta_{ij} - d(X_i, X_j))^2$$

where $X_i, X_j$ are points with two dimensions. $\delta_{ij}$ represents the similarity between two vectors (in our case, these are the terms $i$ and $j$).

IV. THE ARABIC VISUALIZATION SYSTEM

In this section, we describe the implementation and the major functions of AVS depicted in Fig. 1. The major components and functions of AVS include:

1. Pre-processing module: responsible for stemming and tokenizing Arabic documents.
2. Vector space model module: responsible for converting each document into a vector in a multidimensional space vectors.
3. LSI module: applies latent semantic indexing to reduce space dimensionality.
4. MS module: applies multidimensional scaling to convert information in a 2-dimensional space.
5. Visualization module: visualizes the document(s) in a 2-dimensional space.

The processing procedure in AVS is as follows:

**Input:** Corpus of documents to visualize in form of TF-IDF vectors.

**Output:** Set of 2-dimensional points representing the terms.
Steps:

1. Reducing the number of dimensions by calculating $k$-dimensional semantic space generated by input corpus of documents.

2. Projecting the extracted terms into the semantic space which is a low dimensional Space.

3. Applying multidimensional scaling using stress function on terms with Euclidian distance in semantic space as similarity measure.

4. Representing the terms in order to the two dimensions coordinates correspond to each term results in an understanding and meaningful visual graphs.

The output results from AVS are different visual graphs which reflect different aspects of the visualization and can include the following graph options:

1. Viewing the different corpus categories as zones of colors where each color corresponds to a specific category. This visualization gives the user a general view about what type of documents is contained in the corpus.

2. Viewing colored clusters with the most meaningful terms with the highest weights among the other terms. This type of visualization helps the user to understand what a document is about.

Viewing all extracted terms and showing how these terms are close to each others. This will emphasize the terms that come together frequently so the relations between the terms are highlighted.

V. PROCESSING ARABIC TEXT

We tested AVS on Arabic corpus available for researchers at [http://nlp.ju.edu.jo](http://nlp.ju.edu.jo). The corpus varies in its resources and categories. First, we selected the categories and documents under which AVS will be tested. These documents went under preprocessing to prepare them for visualization. The preprocessing step includes stemming of words to extract roots. For this purpose we used a stemmer to remove common affixes from words (affixes such as prefixes, suffixes and postfixes).

The next step in preparing the documents involves removing the most common stop words. Stop words are functional words with high frequency but has no effect on documents meaning. Stop words have been predetermined in a list and used to remove similar words from documents under processing. Similar to English, Arabic stop words list include among other words, prepositions, particles, conjunctions, etc.

A. Representation of Arabic text

Text representation is based on representing the words as vectors in a multidimensional space through the vector space model. Each term is a dimension in the space so the coordinate values assigned to document vectors are given by terms weights computed using the standard TF-IDF. The $i_{th}$ element of the vector containing frequency of the $i_{th}$ word is multiplied with $\text{IDF}_i = \log(N/df_i)$, where $N$ is the total number of documents and $df_i$ is document frequency of the $i_{th}$ word (the number of documents from the whole corpus in which the $i_{th}$ word appears). After assigning a weight to each unique term that has been extracted, the term-document matrix is constructed. The rows of the matrix are the extracted terms while the columns are the documents. Each cell contains the weight with which the term of its row appears in the document denoted by its column. Then for each document vector, we normalize its term weights by calculating the document length (vector length) and divide each weight by this length. Normalization is important for converting all the vectors into a standard length. At this point, we have the conventional vector space that each extracted term occupies a particular position in the term-document space.

For reducing the number of dimensions, latent semantic analysis (LSA) is applied. It reveals the latent relations between terms and documents. Multidimensional scaling (MDS) method is another dimension reduction method that is used to visualize the terms in a 2-dimensions space.

VI. EXPERIMENTAL RESULTS

For our experiments, we use a corpus of Arabic documents which contains 4 categories that cover the following fields: Bahaeyen, Saheeh al Bukhari, medical science and sports. We employ the MATLAB software to perform the visualization process (cf. Fig. 2).

A. Experimental results I

The first visualization results are shown in Fig. 3. Each field is assigned a color: Bahaeyen (green), Saheeh al Bukhari (red), medical science (blue) and sports (black). This gives a conceptual landscape and gives the user a feel of the corpus content. In this way, the system can be employed as a clustering system where the points in the space represent term are the terms of documents. Each term takes its position according to its category so all the terms that are considered to be about sports are close to each other and they are away from the terms that are considered to be about Saheeh Bukhari.
B. Experimental results II

The previous result was too abstract and of little help to users. To provide more details, we offer to visualize the terms in their respective positions in the conceptual landscape. This enables a user to visualize the terms that occur in her/his document(s) and their relationships and degree of spatial closeness (cf. Fig. 4).

For instance, the words "قدم" (foot) and "كرة" (ball) are represented close to each other in the conceptual space because they are used together frequently in the documents and they appeared together (cf. Fig. 5).

C. Experimental results III

While experimental results II may be helpful, they are not sufficiently clear because when the number of terms is too large, the space will be crowded with a large number of terms. This muddles up the visualized terms. To make visualization more understandable and clear, we opted to provide levels of visualization as follows: at the top level, we only allow the most powerful terms to be viewed (i.e., available for visualization) in a particular cluster. Fig. 6 shows the top level visualization (i.e., the most three representative terms that have the highest weight) in each cluster.

The system allows the user to look deeply at each category in a single graph. The user can choose to go into more details where more terms in a cluster can be viewed. We have implemented three visualization levels which the user can invoke (cf. Fig. 7).

VII. CONCLUSION AND FUTURE WORK

We have, in this paper, employed Latent Semantic Indexing (LSI) as a dimensionality reduction technique that helps in extracting some knowledge from text documents. We have presented an Arabic Visualization System. The experiments were carried on several datasets extracted from an Arabic corpus.

The system can create views from different perspectives based on the user’s needs. It is successfully used for visualizing different kinds of document corpora. The system is very helpful for data analysis offering quick insight into the structure of the visualized corpus. It works as a clustering system for a corpus of documents. It allows a user a visualization of the most significant key terms and in-depth visualization of a particular category. This feature can be used to find the some relationships and similarities between documents.

This system can be more informative if it is enhanced to be a visualization dictionary system. This can be achieved by filtering the extracted terms and using only those terms that are more representative of their categories.

It is important to note that there are plans to perform more experiments on different kinds of document corpora in Arabic and other Languages. Furthermore, it is worthwhile investigating different methods for creating LSI spaces and to perform cross-language visualization.
Fig. 4. Experimental results II

Fig. 5. representation of words foot "قدم" and ball "كرة".

Fig. 6. Experimental results III.1
REFERENCES


