Interactive Visual Search System Based onMachine Learning Algorithm

Anas Al-Fayoumi and Mohammad Hassan
Department of Computer Science, Zarqa University, Jordan

Abstract: *This paper presents a tool that enables non-technical end-users to use free-form queries in exploring a large scale datasets with simple and interactive direct technique. The proposed approach is based on effective integration of different techniques, such as data mining, visualization and Human-Computer Interaction (HCI). The proposed model has been incorporated into a prototype developed as a web-based application using different programming languages and software tools. The system has been implemented based on a real dataset, whereas the obtained results indicate the efficiency of such approach.*

Keywords: *Visualization, feature extraction, data mining, visual data mining, machine learning.*

Received September 17, 2015; accepted October 18, 2015 published on line January 28, 2016

1. Introduction

Visual Data Mining (VDM) is a young and emerging discipline that combines knowledge and techniques form several different areas. The ultimate goal of VDM is to devise visualizations of large amounts of data that facilitate the interpretation of the data. Thus, VDM tools should be expected to provide informative but not necessarily nice visualizations [13]. VDM techniques have been proven to be of high value in exploratory data analysis, and they also have a high potential for mining large datasets [1].

Involving the user within the query process, such as: Enabling them to change the mining constraints by adding or ignoring query constrains is a considered benefit. In our approach, the interaction between the user and the proposed model occurs during the search process. Such approach is not only facilitating the production of a higher quality model, but also increasing the user satisfaction.

Traditional search techniques that looking for find items according to predefined features are commonly used techniques. Such techniques start with the user's input query, and then the search algorithm should discover the user intent, matching documents, and sort them by relevancy. Matching based search has some drawbacks, such as: The user intent is hard to be identified using a small amount of information in the query, or the search results could be huge, which contradict the fact that the working memory of a human is limited and can hold up to 7±2 items only [9]. In addition, the match based search usually has textual results which make it hard for users to find the required items without reading the whole text, while redundant results are always expected in such approaches. To overcome the match based search drawbacks, the search system should imply the following characteristics: It should maintain a reasonable number of search results in each iteration, also it should initiate the search with a generic matching intent that converges the actual user intent. In the meanwhile, it should present the user with representative items that summarize a larger set of results and using visual representations of results to help the user drill down fast.

Our goal is to develop a search system that helps users searching and exploring large datasets along with satisfying results, by integrating data visualization with discrete optimization. Such system should incrementally detect the user intent and iteratively narrowing down the search space to a satisfactory set of results.

The rest of the paper is organized as follows. An overview is represented in section 2. Related works are reported in section 3. The proposed approach is described in section 4. Data visualization is reported in section 5. Experimental evaluation is presented in section 6. Section 7 includes the conclusions and the future work.

2. Overview-VDM

Visualization is a process that data, information and knowledge representation is converted to visible, which provides an interface between human and computer information processing systems. The use of effective visualization techniques can quickly and efficiently deal with large amounts of data to find the hidden features, relationships, patterns and trends that can guide a new predictable and more efficient decision-making [2].

Data mining techniques and algorithms make decision-making difficult to understand and use. On the other hand, visualization can make it easier to understand the mining results; it used to guide data mining algorithms and allows users to participate in the process of decision making [5].

Since, there is a huge amount of patterns generated by data-mining algorithm in textual form and it is almost impossible for the human to interpret and evaluate the patterns in details and extract interesting knowledge. VDM techniques aim to involve the human in the data-mining process, and applying human perceptual abilities to the analysis of datasets. Visualizing and presenting data in an interactive graphical form often fosters new insights, encouraging the formation and validation of new hypotheses for better problem-solving and enhancing deeper domain knowledge.

Data visualization allows the data analyst to gain insight into the data and come up with new hypotheses or options [12]. The verification of the hypotheses is also followed by data visualization and then accomplished by machine learning, generating new options for the user. As a result, VDM usually allows faster data exploration by using visualization techniques and often provides better results to enhance the degree of user's happiness by promoting a much higher degree of user satisfaction and confidence regarding the query results.

VDM is based on an automatic part, the data mining algorithm, and an interactive part, the visualization technique. There are three common approaches to integrate the human in the data exploration process: Preceding Visualization (PV), Subsequent Visualization (SV) and Tightly Integrated Visualization (TIV) [13].

In PV approach, data is visualized in some visual form before running a data-mining algorithm. By interaction with the raw data, the data analyst has full control over the analysis in the search space. Interesting patterns are discovered by exploring the data.

While in SV approach, an automatic data-mining algorithm performs the data mining task by extracting patterns from a given dataset. These patterns are visualized to make them interpretable for the data analyst. The SV approach enables the data analyst to specify feedbacks. Based on the visualization, the data analyst may want to return to the data-mining algorithm and use different input parameters to obtain better results.

Using TIV approach, an automatic data-mining algorithm performs an analysis of the data but does not produce the final results. A visualization technique is used to present the intermediate results of the data exploration process. The combination of some automatic data-mining algorithms and visualization techniques enables specified user feedback for the next data mining run. Then, the data analyst identifies the interesting patterns in the visualization of the intermediate results based on his domain knowledge.

In our proposed approach and in order to achieve independence of data-mining algorithms from the application, a given automatic data-mining algorithm can be very useful in some domains but may have drawbacks in other domains. Since there is no automatic data mining algorithm (with one parameter setting) suitable for all application domains, TIV leads to a better understanding of data and the extracted patterns, which is adopted by our approach.

3. Related Works

Ribler and Abrams [11] introduced a number of general methods for visualizing commonality in sets of text files. Each visualization technique was simultaneously comparing one file in the set to all other files in that set.

Soukup and Davidson's monograph on VDM [12] has taken a business perspective and practice-based approach to the process and application of VDM, with a limited coverage of visual data representations and interactive techniques.

Ertek [6] developed a model based on VDM through a novel information visualization scheme, namely square tiles visualization. A hybrid visualization scheme is proposed and implemented to represent data with categorical and numerical attributes.

In [7], Apolo system was introduced, which used a mixed-initiative approach combining visualization. Apolo system provides a rich user interaction and machine learning to guide the user to incrementally and interactively explore large network data and make sense of it.

Another business enhancement using VDM was introduced in [15]. Such model performed sentiment analysis using VDM on stocks.

4. The Proposed Approach

The proposed model starts with data collection. Data was downloaded from a commercial site. The following steps summarize the steps of our VDM model as shown in Figure 1:

1. Data Collection.
2. Data Cleaning and Integration.
3. Information Extraction.
4. Infer Preferences (Searching and find best Matches).
5. Visualization.
6. Cluster on Demand.

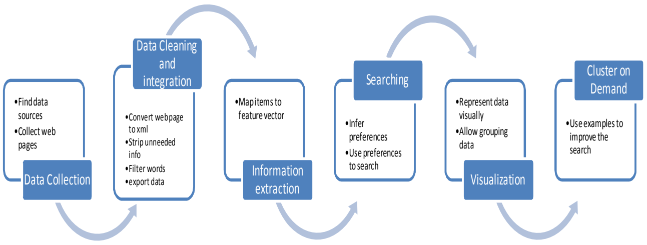


Figure 1. Interactive visual search system workflow.

In our approach some parameters were specified first to restrict the search space, and then the data-mining is performed automatically using the lazy greedy algorithm [10], while finally the patterns found by the data-mining algorithm are presented to the user. In order to, make data-mining more effective; it is important to involve the human in the data mining process and combine flexibility, creativity, and the general knowledge of human.

4.1. Datasets

Data extractions from web pages for information processing and crawling web pages using Selenium web driver tool for finding data are both important tools to find information on the Internet. In this section, we describe web crawling, extracting data from a set of hyperlinked HTML pages, converting data into XML, and presenting it for further processing by our interactive search system. In the meanwhile, it is an important issue for the data analyst to have a prior knowledge about the nature and features that should be extracted from crawled web pages; this will help the data analytic to define the features that should be extracted from the textual data.

To implement our proposed model, data was downloaded from jo.opensooq.com including two months of vehicle advertisements (March, and April, 2014), while each month used to construct a stand-alone dataset containing about 7500 records.

This phase included three major steps; first, crawling cars text advertisements from jo.opensooq.com which containing Unicode text (Arabic text) and English as well. Second, filtering and extracting features of these web pages, so a definition for covered features and supported items should be combined together, while the third step is dataset discretization.

As a result of the above three steps a dataset is constructed, but with two formats. The first, format is discretized dataset which is a result of the three steps and used for searching, while the second, format is detailed dataset, which is a result for the first two steps and used for displaying results for the user. Figure 2 shows the required steps to build the datasets according to our model.

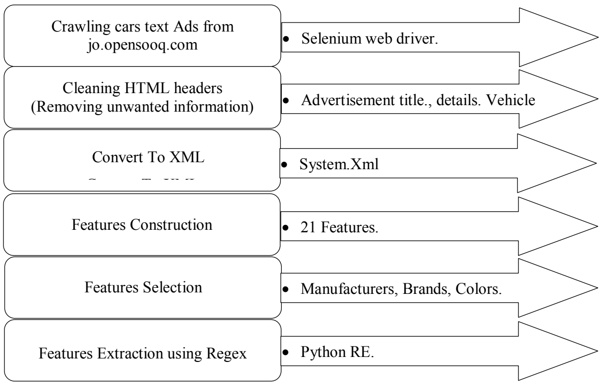

Crawling cars text Ads from jo.opensooq.com
• Selenium web driver.
Cleaning HTML headers (Removing unwanted information)
• Advertisement title., details. Vehicle
Convert To XML
• System.Xml
Features Construction
• 21 Features.
Features Selection
• Manufacturers, Brands, Colors.
Features Extraction using Regex
• Python RE.

Figure 2. The required steps to build the datasets.

Data crawling is the process by which we collect pages and hypertext documents from websites. In web search engines, the web search engine needs to gather as many web pages as possible and make them available for searching [3].

Data crawling includes the following three steps, which is developed using the C# scripting:

- Creating a web client variable to handle the web page.
- Creating a string variable handling the web address and the unique advertisement ID.
- Downloading the web page using Selenium web driver and save it to a local storage drive.

The upshot of data crawling is 160,000 advertisements, while those advertisements were divided into two datasets, each dataset holds 80,000 advertisements.

It's important to mention that converting from HTML to XML is needed, where many unwanted details could be eliminated (such as photos), where only specific tags are needed in the search process (Title, Details, Price and Phone number of the advertisement). Another important issue is that; most advertisements are written in Arabic language, so XML file is saved in UTF8 encoding in order to support Arabic language processing in features extraction step. As a result of such converting to XML and data cleaning, about 21,000 advertisements were remaining for each dataset.

4.2. Feature Extraction

Feature extraction is the process of transforming input data to produce a set of features or features vector [14]. In Feature extraction process, the extracted features are expected to contain relevant information from input data i.e., vehicle specifications.

The problem of feature extraction has been studied extensively under the topic of entity disambiguation [8]. In our proposed model, a set of specified features were extracted from the title, details, price, and mobile number tags in the XML file using regular expressions.

Feature extraction was performed using a pre-trained model using Python. If no such expert knowledge is available, general dimensionality reduction techniques may help. Human expertise, which is often required to convert raw data into a set of useful features, can be complemented by automatic feature construction methods. In our approach, the feature construction was integrated with the modeling process. In other approaches, feature construction is a preprocessing [14].

The feature is synonymous of input variable or attribute and data can be represented by a fixed number of features which can be binary, categorical or continuous

In our proposed model, data is text and constructed in categorical form. Twenty one features are extracted for each vehicle advertisements, features are extracted from Arabic and English advertisements as well. Table 1 shows a portion of these constructed features.

Table 1. Portion of the constructed features.

Feature ID	Feature	Description
1	Manufacturer	Defines the car manufacturer i.e., BMW, Kia.
2	Brand	Defines the car brand under their manufacturer i.e., Accord, Vectra.
3	Model	Defines year of production i.e., 2000, 1990.
: :	: :	: :
19	Color	Defines the color of the car, 12 colors were defined.
20	Contact Number	Defines the contact number of the Advertiser.
21	Price	Defines the price of the car.

Another sub-step within the feature extraction is feature selection. Feature selection is the process of choosing interesting features among the selected ones for further processing [8]. For the twenty one features were mentioned above, some features need to be bound or selected. Since, we cannot define all vehicle manufacturers, brands and colors, only a specific manufacturers, brands and colors were specified. So, we specify fourteen cars manufacturers, eighty eight brands, and thirteen colors (including unknown). Table 2 shows a portion of manufacturers and brands.

Table 2. Portion of vehicle manufacturers and brands.

Manufacturer	Brand	Manufacturer	Brand
Nissan	Sunny Altima Tida Maxima Morano	Mitsubishi	Lancer Galant Pajero Colt
Honada	Civic Accord CRV City	Mercedes	200 Series C Series S Series

4.3. Dataset Discretization

Discretization algorithms have played an important role in data mining and knowledge discovery. Data discretization techniques are used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals [4]. Interval labels can then be used to replace actual data values. Discretization techniques are typically applied as a preprocessing of data mining step.

Table 3. Portion of features discretization criteria.

Feature	Old Value	Discretized Value	Feature	Old Value	Discretized Value
Manufacturer	Hyundai	10	Model	1960-1969	120
	Kia	11		1970-1974	121
	Daewoo	12		1975-1979	122
	Toyota	13		1980-1984	123
	Nissan	14		1985-1989	124
	Mitsubishi	15		1990-1994	125
	Honda	16		1995-1999	126
	Opel	17		2000-2004	127
	Mercedes	18		2005-2009	128
	BMW	19		2009-2014	129
	Volkswagen	20	Engine Size (CC)	<1300	130
	Citroen	21		1300-1500	131
	Peugeot	22		1501-1800	132
	Ford	23		1801-2000	133
Brand	Avante	30		2001-2500	134
	Accent	31		2501-3500	135
	Sonata	32		>3501	136
	Elentra	33	Price	Undifined	140
	Verna	34		<3000	141
	Tuscani	35		3000-3499	142

Discretization techniques can be classified into many categories. In our approach, we used a static, global, direct, supervised and top-down discretization. Where, text attributes were converted to numeric integer values, Table 3 above shows a portion of the features discretization criteria, in which the values of the textual features and the continuous numeric values were converted into numeric discrete values.

5. Data Visualization

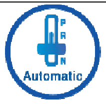
Data visualization is all about understanding ratios and relationships among numbers. Not about understanding individual numbers, but about understanding the patterns, trends, and relationships that exist in groups of numbers. So creating data visualization is more than simply translating a table of data into a graph. Data visualizations should communicate data in the most effective way; to truly reveal the data they should be quick, accurate, and powerful. Creating visuals can easily summarize and represents data to users, making complicated sets of data more understandable and memorable [16].

5.1. Icon-based Visualization Technique

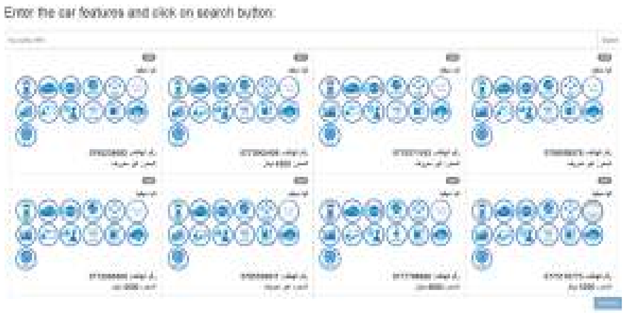
Icon-based visualization or iconographic techniques represent each data entry individually, allowing verification of rules and behavior patterns of the data. Icons with similar properties can be recognized and thus form groups and it can be analyzed in particular [16]. Using multiple icons located in one position is an effective and efficient method for large high dimensional data set visualization. Summary icons can help display local data details and overall context at the same time [2].

In this model a set of icons represents the features of each entity were introduced, where Table 4 shows a portion of these Iconic-based features.

Table 4. portion of the Iconic-based features.

Feature	Description
	Automatic gear transmission.
R 1 3 5 2 4 6 Manual	Manual gear transmission.

In our approach, a special framework was used from Python called Flask web framework. Flask web framework provides us with tools, libraries and techniques that allow developing a web application; Figure 3 shows the base webpage of the proposed model.



Enter the car features and click on search button:

Figure 3. The base webpage.

Index web page allows the user to type a text query which is converted to a set of features, while base web page shows the results of the query using iconic-based visualization and allows the user to interact with the system by choosing the best matching items.

The following scenario explains how the proposed Interactive Visual Data-Mining (IVDM) system works. The user types the query text in the search textbox, and then the query is translated to features vector. The most representative items are found (between 7 and 9) in the discretized dataset using the lazy greedy algorithm as a data-mining technique. In the next step, the most representative items are transformed to their JSON format, and using Flask library the system sends an AJAX post to the web interface along with a string which containing the query JSON format results. Then, the query result will be displayed using iconic-based visualization technique as shown in Figure 3 using Java scripting. The user interacts with the system by selecting the most preferred items, where the IDs of the selected items send as a returned value of the AJAX post. The system receives the selected items IDs, and finds the common feature between the selected items and generates a new query combining the original query and the common features between the selected items. The IVDM system repeats the previous steps and displays new results on base web page. Finally, the user continues interacting with IVDM system till he finds the desired targeted item.

6. Experimental Evaluation

In this section, we describe the approach that is adopted in the process of performance evaluation. There are several performance metrics that can be measured such as evaluating the user's happiness or the time complexity for searching. Nevertheless, these metrics are applicable, but they need a live system to evaluate the performance. To overcome this problem i.e. publishing the system then evaluate it, we perform our experiments using the existing datasets and comparing the IVDM system with SQL, which is commonly used in advertisement websites.

The experimental subset consists of 30 randomly chosen items from each dataset to evaluate IVDM system. Two features were selected for each item (manufacturer and brand) and start searching for the item itself or a similar item which has the same features.

SQL users will explore all items that satisfied the chosen features. On the other hand, IVDM system users will have 8 options in each iteration and select among them according to the user preferences.

We continued evaluating the proposed system by using three features (manufacturer, brand, and model)), four features (manufacturer, brand, model, and color) and five features (manufacturer, brand, model, Color, and gear transmutation type) features using the same criteria. We compare the number of items that the user explores to find the desired target item.

Figure 4 shows the number of advertisements that have been read (for two features), using SQL and IVDM to find exact or similar items in the first dataset for 30 randomly chosen items.

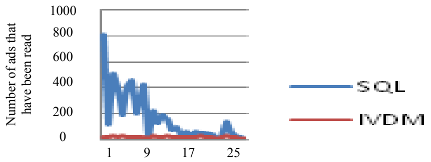


Figure 4. The number of advertisements that have been read using two features.

Figure 5 below shows the average number of advertisements required to find a similar item for the 30 random selected items in the two datasets. It's obvious that the minimum number of items that the user read, the best performance can achieve. So, IVDM system has a better performance in all cases and especially with less number of features.

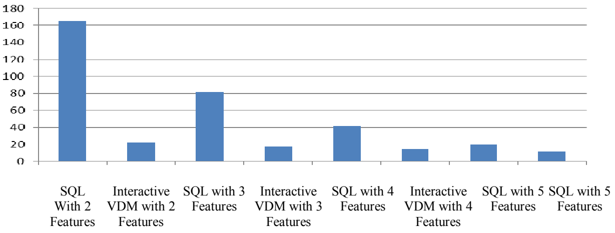


Figure 5. The average number of advertisements that have been read for the first dataset.

7. Conclusions and Future Works

The research scope for this paper focused on the development of interactive visual search system large scale datasets and databases. The proposed approach is unique in integrating several aspects. We used discretization algorithm during dataset construction and feature extraction phase. We used lazy greedy algorithm as a data mining technique. We used visualization to assist the search process by integrating user interactions with search process. The proposed model has been incorporated in a prototype developed as a web-based application using different programming languages and software tools. The system has been implemented based on a real dataset, whereas the obtained results indicate the efficiency of such approach.

Currently, we are concentrating on the following extensions to the proposed approach. The First, improving the feature extraction technique to work with general datasets. As another improvement, we are working on adding more datasets and increasing the performance of the data mining algorithm in terms of time and space complexity. Finally, we want to investigate and gather further requirements to improve the usability and friendliness of our proposed system.

References

[1] Alazmi A., "Data Mining and Visualization of Large Databases," *International Journal of Computer Science and Security*, vol. 6, no. 5, pp. 295-314, 2012.

[2] Aparicio M. and Costa C., "Data Visualization," *Communication Design Quarterly Review*, vol. 3, no. 1, pp. 7-11, 2014.

[3] Bergman M., "White Paper: The Deep Web: Surfacing Hidden Value," *The Journal of Electronic Publishing*, vol. 7, no. 1, 2001.

[4] Chao-Ton S. and Jyh-Hwa H., "An Extended Chi2 Algorithm for Discretization of Real Value Attributes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 437-441, 2005.

[5] Chen S., Tang L., Liu W., and Li Y., "An Improved Method of Discretization of Continuous Attributes," *Procedia Environmental Sciences*, vol. 11, pp. 213-217, 2011.

[6] Ertek G., "Developing Competitive Strategies in Higher Education through Visual Data Mining," *Faculty of Engineering and Natural Sciences*, Sabanci University, Istanbul, Turkey, pp. 1-14, 2009.

[7] Horng D., Kittur A., Hong I., and Faloutsos C., "Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning," *in Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, Canada, pp. 167-176, 2011.

[8] Liu X. and Wang H., "A Discretization Algorithm based on a Heterogeneity Criterion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1166-1173, 2005.

[9] Miller G., "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information," *Psychological Review*, vol. 101, no. 2, pp. 343-352, 1994.

[10] Minoux M., "Accelerated Greedy Algorithms for Maximizing Submodular set Functions," *in Proceedings of the 8th IFIP Conference on Optimization Techniques*, Springer, pp. 234-243, 1978.

[11] Ribler R. and Abrams M., "Using Visualization to Detect Plagiarism in Computer Science Classes," *in Proceedings of IEEE Symposium on Information Visualization*, USA, pp. 173-178, 2000.

[12] Soukup T. and Davidson I., *Visual Data Mining*, New York, Wiley, 2002.

[13] Stahl F., Gabrys B., Gaber M., and Berendsen M., "An Overview of Interactive Visual Data Mining Techniques for Knowledge Discovery," *WIREs Data Mining Knowledge Discovery*, vol. 3, no. 4, pp. 239-256, 2013.

[14] Su C. and Hsu J., "An Extended Chi2 Algorithm for Discretization of Real Value Attributes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 437-441, 2005.

[15] Velu C. and Kashwan K., "Performance Analysis for Visual Data Mining Classification Techniques of Decision Tree, Ensemble and SOM," *International Journal of Computer Applications*, vol. 57, no. 22, pp. 65-71, 2012.

[16] Ware C., *Information Visualization*, Waltham, Mass, Morgan Kaufmann, 2013.

Mohammad Hassan received his BS degree from Yarmouk University in Jordan in 1987, the MS degree from University of Jordan, in 1996, and the PhD degree in Computer Information Systems from Bradford University, UK in 2003. He is working as an associate professor in the Department of Computer Science at Zarqa University in Jordan. His research interest includes information retrieval systems and database systems.

Anas Al-Fayoumi received his BS degree in Computer Engineering from Al-Balqa Applied University-FET in Jordan in 2009, and the MS degree in Computer Science from Zarqa University in Jordan in 2015. He is working in ISD at UNRWA in Jordan. His research interest includes information retrieval, data mining, and computer security.