KG4Vis: A Knowledge Graph-Based Approach for Visualization Recommendation

Haotian Li
HKUST, SMU

Yong Wang
SMU

Songheng Zhang
SMU

Yangqiu Song
HKUST

Huamin Qu
HKUST

Problem Statement

Given a table

Name	Math Score	Physics Score
Alice	90	84
Bob	75	80
Bill	92	95

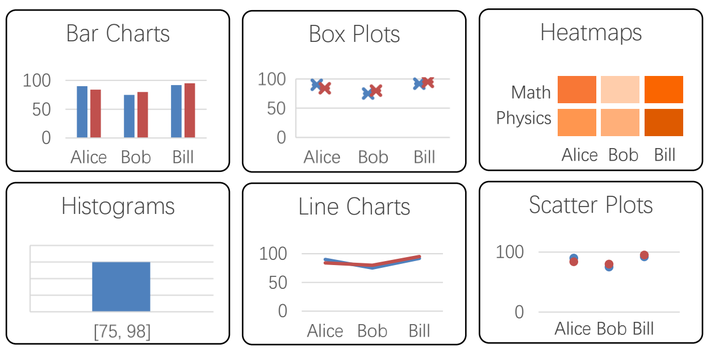
Knowledge Graph

Generate Visualization Rules

Entropy of values is large → Box Plot

Values are monotonic → Line Chart

Recommend how to visualize the data



Related Works

Rule-based approaches

- An explicit and manual list of rules and heuristics by expert judgment.
- Difficult and tedious to compile a complete rule list.
- Rules may not be generalizable to different datasets or visualization choices.

Machine-learning-based approaches

- Train a deep learning model to learn from dataset-visualization examples.
- Do not need to manually specify the rules.
- Work as a black box, difficult to tell why such visualization is recommended.
- Users may not trust in the recommended visualizations.

Overall Workflow and Details

Step 1 Feature Extraction

Continuous Features

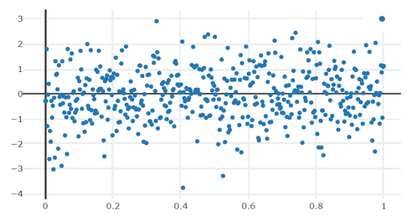
	Feature Name	Explanation
Values	num_unique_elements	This feature counts the number of unique values.
	sortedness	This feature measures the how sorted the values are.
	%_outliers_1.5iqr %_outliers_3iqr %_outliers_3std %_outliers_1_99	These features measure the proportion of outliers according to 1.5IQR/3IQR/3Std/(1%, 99%) rule.
	unique_percent	This feature measures the proportion of unique values in a data column.
	entropy gini	These features measure how disordered the values are.
	normality_p normality_statistic	These features reflect how much the values are distributed normally.
	normalized_range range	These features reflect the range of values in a data column.
	q25 q75 normalized_median normalized_mean min max mean median length	These features are some basic statistical measurements of a data column.
	mode_percent	This feature measures the proportion of mode values in a data column.
	#_missing_elements missing_percent	These features reflect the number/ proportion of missing values in a data column.
	mean_value_length median_value_length	They measure the length of mean/ median values in a data column.
	min_length_of_value std_length_of_value max_length_of_value	They measure the minimum/ standard deviation/maximum length of values in a data column.
Names	#_of_words_in_name	This feature measures how many words in the column name.
	name_length	This feature measures the length of the column name.
	#_uppercase_char	This feature measures how many upper case in the column name.
	field_name_length	This feature measures how many characters in the column name.

Categorical Features

	Feature Name	Explanation
Types	specific_string specific_integer specific_decimal specific_datetime	The specific data type of this data column is string/integer/decimal/datetime.
	general_temporal general_quantitative general_categorial	The general data type of this data column is temporal/quantitative/categorial.
Values	has_none	The data column contains missing value.
	is_monotonic	Values are monotonic.
	is_lin_space	Values are in linear space.
	is_log_space	Values are in log space.
	is_unique	All values are unique.
	is_sorted	All values are sorted.
	has_outliers_1.5iqr has_outliers_3iqr has_outliers_3std has_outliers_1_99	Outlier exists according to 1.5IQR/3IQR/ 3Std/(1%, 99%) rule.
	is_normal_1 is_normal_5	Values are normally distributed with $p < 0.01/p < 0.05$.
	is_only_field	The data column is the only one within the dataset.
Names	1st_uppercase	The column name starts with an upper case.
	x_in_name y_in_name id_in_name time_in_name digit_in_name whitespace_in_name dollar_in_name pounds_in_name euro_in_name yen_in_name	A word or symbol "x", "y","id", "time", digit, whitespace, "$","€", "£", "Y" is in the column name.

Step 2 Knowledge Graph Construction

x	0	0.01	...	0.96	0.98	1
y	-0.28	0.41	...	-0.32	-0.02	1.12



Dataset-visualization pairs

Construct Entities (Nodes in KG)

Data Column
Categorical Features
Continuous Features
Visualization Design

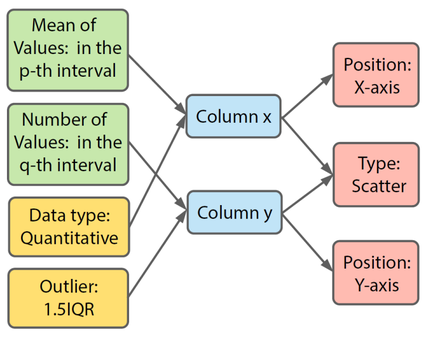
Entities from column x
Mean of Values: in the p-th interval
Type: Scatter
Column x
Position: X-axis
Data type: Quantitative

Entities from column y
Number of Values: in the q-th interval
Type: Scatter
Column y
Position: Y-axis
Outlier: 1.5IQR

Construct Relations (Edges in KG)

CF, DF → Data Column
(Feature) is a data feature of (Data Column)

Data Column → Visual Design
(Data Column) is visually encoded with (Visualization Design)

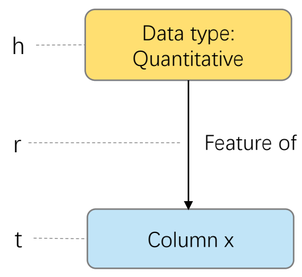


Knowledge Graph

Step 3 Embedding Learning

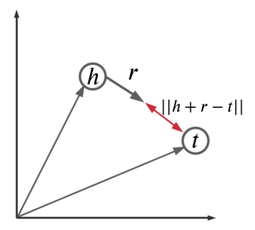
TransE

Each edge is represented as
(head entity, relation, tail entity)
i.e., a triplet for short



Each edge with two entities and a relation has a corresponding embedding vector:
$(\vec{h}, \vec{r}, \vec{t})$

Train the embeddings to achieve:
$\vec{h} + \vec{r} \approx \vec{t}$



Score for a relation (h, r, t)

$$g(h,r,t) = -\left\|\vec{h} + \vec{r} - \vec{t}\right\|_2$$

When $\vec{h} + \vec{r}$ is closer to $\vec{t}$, score is higher

Original TransE Loss Function for Training Embeddings

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'} ReLU\left(\gamma + g(h',r,t') - g(h,r,t)\right)$$

TransE-adv (with Self-adversarial Negative Sampling) Loss Function for Training Embeddings

$$L = -log\sigma\left(\gamma + g(h,r,t)\right) - \sum_{(h',r,t')\in S'} w(h',r,t') log\sigma(-g(h',r,t') - \gamma)$$

- γ : Margin parameter, $\gamma > 0$
- S : Triplets (h,r,t) in the graph
- S': Generated negative triplets (h',r,t')
- σ : Sigmoid function
- w: The triplet's probability of being true

Step 4 Embedding-based Inference

Bar	Values in a column is not sorted
Box	Values are not evenly distributed
Line	The column is not the only one in dataset
Scatter	Values in a column are numerical
Histogram	Outlier exists in a column (3Std)

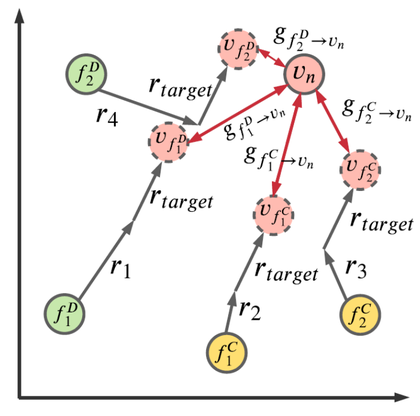
Data Feature → Visual Design Choice

For a feature f_i ,
$\vec{f_i} + \vec{r_i} = \vec{d_{im}}$
where r_i is a relation connecting to f_i
$\vec{d_{im}}$ represents an imaginary data column

Relation r_{target} maps data to a vector $\vec{v_c}$
$\vec{d_{im}} + \vec{r_{target}} = \vec{v_c}$

Given f_i , for each visual choice v_n ,
define a score indicating how much v_n is preferred

$$g_{f_i \to v_n} = -\left\|\vec{f_i} + \vec{r_i} + \vec{r_{target}} - \vec{v_n}\right\|$$



New Data → Visual Design Choice

Given a new data d_{new} ,
extract all its features F_{new}
For each visual choice v_n ,
define a score indicating which v_n is recommended

$$g(d_{new}, v_n) = \frac{1}{|F_{new}|} \sum_{f_i \in F_{new}} g_{f_i \to v_n}$$

$$= \frac{-1}{|F_{new}|} \sum_{f_i \in F_{new}} \left\|\vec{f_i} + \vec{r_i} + \vec{r_{target}} - \vec{v_n}\right\|$$

Visualization choice v_n with a higher score is better

Evaluation Setup

Visualization Corpus

- VisML corpus
- 88,548 dataset-visualization pairs

About Knowledge Graph

216,851 Entities	9,679,463 Triplets	1,000 Embedding Dimension

Quantitative Evaluation

Two Inference Tasks

- Inference of visualization types
- Inference of visualization axis

Try different embedding learning models

	Axis	Visualization Type	
	Accuracy	MR	Hits@2
TransE-adv	**0.7350**	**1.9567**	**0.7489**
TransE	0.7214	1.9718	0.7445
RotatE	0.7193	1.9608	0.7458

Qualitative Evaluation

Rules and recommendation results – Case study

Scatter	Values in a column are numerical
Scatter	Outlier exists in a column (1.5IQR)
Box	Values are not evenly distributed
Box	The entropy of values is large
Line	Values are monotonic
Line	The data type of a column is decimal
Line	The column is not the only one in dataset
Bar	All values in a column are unique
Bar	Values in a column is not sorted

Participants and Procedure of Expert Interviews

- 12 researchers who have conducted research in data visualization for at least 1 year.
- Experts were asked to finish the three tasks through online meetings

Tasks of Expert Interviews

Task 1
Provided top-5 rules of each visualization type.
Give each generated rule a score ranging from 1 (the least reasonable) to 5 (the most reasonable).

Feedbacks
Overall, the generated rules are appreciated by experts.

Task 2
Provided 30 datasets and corresponding top-2 recommended visualizations by our approach.
Give each recommended visualization a score ranging from 1 (the least reasonable) to 5 (the most reasonable).

Feedbacks
Average score is 3.7944, thought to be of high quality.

Task 3
Provided 30 datasets, ask the experts to select top-2 visualization types, for collecting their preferred design choices.

Acknowledgment:
This poster is originally made by Mr. Yuzheng Cai for course "Data Visualization: Application and Implementation" instructed by Dr. Siming Chen, Fudan University, October 2021.