http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.1, February 2020, pp.83-107
https://doi.org/10.7838/jsebs.2020.25.1.083

상황인식형 비즈니스 차트 추천기 개발을 위한 개방형 온라인 텍스트로부터의 시각화 규칙 추출 방법 연구

A Method of Mining Visualization Rules from Open Online Text for Situation Aware Business Chart Recommendation

Qingxuan Zhang*, 권오병(Ohbyung Kwon)**

초 록

데이터의 성격과 시각화의 목적에 따라 비즈니스 차트를 선택하는 것은 비즈니스 분석에 유용한 지식이다. 그러나 현재 시각화 도구에는 상황에 맞는 비즈니스 차트를 선택할 수 있는 기능이 부족하다. 또한 매번마다 시각화 방법에 대해 전문가의 도움을 요청하는 것은 비용과 시간이 소요된다. 따라서 본 연구의 목적은 온라인으로 게시된 문서로부터 비즈니스 차트 선정 규칙에 대한 지식을 추출하여 비즈니스 차트 생산성을 향상시키는 방법을 제안하는 것이다. 이를 위해 인터넷에서 비즈니스 차트를 묘사하는 한국어, 영어 및 중국어 비정형 데이터를 수집하고 TF-IDF를 사용하여 컨텍스트와 비즈니스 차트 간의 관계를 계산했다. 또한 Galois 래티스를 사용하여 비즈니스 차트 선택 규칙을 생성했다. 제안된 방법으로 생성된 규칙의 품질을 평가하기 위해 실험군과 대조군에 대해 실험을 수행했다. 그 결과 제안된 방법으로 의미 있는 규칙이 추출되었음을 확인했다. 본 연구의 결과물로 시각화 전문가의 도움 없이도 사무직 직원들이 비즈니스 차트를 효율적으로 선택할 수 있을 것으로 기대된다. 또한 작업 중인 문서를 기반으로 비즈니스 차트를 추천함으로 직원 교육에 유용할 것이다.

ABSTRACT

Selecting business charts based on the nature of the data and the purpose of the visualization is useful in business analysis. However, current visualization tools lack the ability to help choose the right business chart for the context. Also, soliciting expert help about visualization methods for every analysis is inefficient. Therefore, the purpose of this study is to propose an accessible method to improve business chart productivity by creating rules for selecting business charts from online published documents. To this end, Korean, English, and Chinese unstructured data describing business charts were collected from the Internet, and the relationships between the contexts and the business charts were calculated using TF-IDF. We also used a Galois lattice to create rules for business chart selection. In order to evaluate

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03036394).

* First Author, MS, Department of Management, Kyung Hee University(bbandft@khu.ac.kr)
** Corresponding Author, Professor, School of Management, Kyung Hee University(obkwon@khu.ac.kr)

Received: 2019-11-19, Review completed: 2019-12-11, Accepted: 2019-12-24

the adequacy of the rules generated by the proposed method, experiments were conducted on experimental and control groups. The results confirmed that meaningful rules were extracted by the proposed method. To the best of our knowledge, this is the first study to recommend customizing business charts through open unstructured data analysis and to propose a method that enables efficient selection of business charts for office workers without expert assistance. This method should be useful for staff training by recommending business charts based on the document that he/she is working on.

키워드 : 비즈니스 차트, 데이터 시각화, 텍스트 마이닝, TF-IDF, 갈로아 래티스, 실증 연구
Business Chart, Data Visualization, Text Mining, TF-IDF, Galois Lattice, Empirical Test

1. Introduction

The phrase “a picture is worth a thousand words” refers to the importance of visualization, a way of “bringing out… meaning in data” [39]. Data visualization in the context of big data helps decision-makers discover information inherent in the data Tegarden [42] and has the advantages of pinpointing new ideas and delivering content in an intuitive and timely manner [32]. In addition, the information processed through visualization can become a tool for understanding and insight that significantly influences the decision-making process. Data visualization is also an important technique for data mining that aids cognitive functions and processes; it helps us interpret large amounts of data, summarize that data, understand overall patterns, and provide intuition for analysis [27]. In particular, the visualization of management information through business charts can contribute to the quality of decision-making [5, 42, 45], satisfaction Enzenhofer [12], and immersion Shin [37] Hence, visualization contributes positively to the performance of a data analysis Al-Kassab [1] and eventually the performance of the organization.

The creation of business charts can be made easier and more accurate with data visualization tools. MS Excel, for example, offers a variety of chart types, including the bar chart, pie chart, line chart, area chart, and scatter chart, and it supports wizards with built-in capabilities, making it easier for users to create business charts. Such data visualization support tools can be used for training new employees, interns, and trainees to make good decisions they are known to enhance learning effectiveness and improve educational decision-making [19, 40].

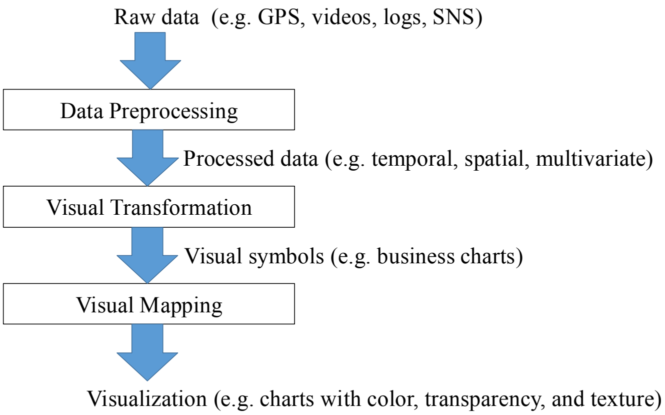
The selection of business charts for business data visualization must involve consideration of the user context, including the nature of the problem, the direction of storytelling [24], and the characteristics of the equipment on which the chart will be displayed [26]. In order to visualize data successfully and select a business chart that fits the situation, it is first necessary to understand the structure and characteristics of the dataset to be visualized. However, existing data visualization tools do not create or propose charts on behalf of the user in an autonomous and intelligent manner. Time and effort are required to become familiar with these visualization tools, multiple repetitions of the creation process may be necessary, and the results may not be satisfactory. Unfortunately, little research exists on how to automate or support decision-making in the selection of business charts [11].

The purpose of this study is to propose an intelligent method of autonomously determining the best type of data visualization for creation of business charts according to the context in terms of both individual and problem characteristics. To this end, we propose a method of generating a rule for selecting a business chart based on documents published online using a method that combines the term frequency-inverse document frequency (TF-IDF) statistic and a Galois lattice, which is a graphic method of representing knowledge structures. The TF-IDF method is used to identify keywords for the rule, and the Galois lattice is used for rule generation. A Galois lattice has the advantage of finding rules even if the conditional expressions are not complete, so users can find business charts even if the situation has not been declared intact. In order to evaluate the quality of the business chart selection rules generated by the proposed method, we conducted experiments on an actual user group.

2. Related Work: Data Visualization

Data visualization involves use of visual channels to represent datasets [9, 15], transforming various types of data into appropriate visual representations so that understanding can be assured and analysis of data can be completed efficiently [10]. Uses of data visualization range from publishing media [33] to traffic information systems [10], medical information systems [6], and corporate data analysis [24]. Accordingly, various data visualization tools have been developed to support data visualization tasks for diverse types of users. Visualization libraries are available for professional visualization tools such as Gephi and Weave [7], data analysis tools such as Matlab, and spreadsheets such as Microsoft Excel. There are many visualization functions that can be added in programming languages that favor data analysis, such as R and Python.

Data visualization generally follows the sequence represented in <Figure 1>. First, the raw data are collected from various sources. After preprocessing, the data are processed there are various types of processed data such as temporal, spatial, spatio-temporal, and multivariate. Then, based on the types of visualization and processed data, the best visual


Raw data (e.g. GPS, videos, logs, SNS)
Data Preprocessing
Processed data (e.g. temporal, spatial, multivariate)
Visual Transformation
Visual symbols (e.g. business charts)
Visual Mapping
Visualization (e.g. charts with color, transparency, and texture)

〈Figure 1〉 Data Visualization Process

symbol is selected. The final step is to decide whether to use color, animation, video, or some other visualization elements to determine the appearance of the selected chart.

The choice of the optimal visual symbol (i.e., the type of business chart in this study) is a decision that requires a lot of knowledge. If the system does not choose correctly, and the built-in charts are not appropriate, then it is quite difficult for users to replace them with the desired pictures. Typically, many complex dialog boxes and commands must be used. This is a significant, recognized problem with all current visualization programs, which so far has not been solved. In fact, users of Microsoft Excel have found this process so difficult that the Windows version provides a so-called “Wizard” interface that takes the user through a set of question-and-answer dialogs. However, this solution can be tedious, and it still does not provide the user with sufficient flexibility to specify desired displays easily. Creating custom displays is also difficult with scientific visualization systems like the IBM Visualization Data Explorer. In these systems, code must be written using either conventional or visual programming languages [31].

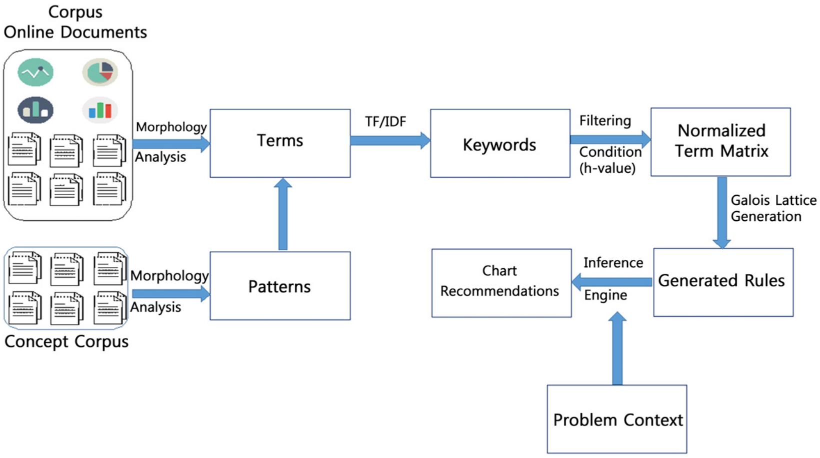
Users selectinga business chart should consider a few things: First, each business chart has its own parameters. For example, bar charts include position, height, width, and color line charts include position, color, and line width and pie charts include percentage of whole and color. Users must determine which business chart is optimal based on which characteristics they want and which parameters to display-based on those characteristics. Second, the choice of visualization method or chart depends on the type of task.For example, in geographic information, time and location, type of vehicle, destination, and speed of movement are the determinants of visualization methods [35]. Visualization also depends on the nature of the work, situational awareness [13], type discovery or clustering [34], and traffic congestion monitoring [4].

However, with existing visualization tools, selecting visual symbols or business charts still depends entirely on the user's prior knowledge or, in cases where that knowledge is inadequate, the help of a professional. Although existing visualization tools support creation of business charts with widget-like features, they do not customize business charts for the task at hand. This is where recent research on semantic data visualization comes in. Semantic data visualization is an emerging field of research offering new perspectives on natural language processing and in-depth analysis of text data [18, 20]. Ontological research on how to visualize semantics is also underway [30]. This research focuseson visualizing knowledge discovery in XML, RDF, and OWL formats, especially from multiple text repositories [2]. Alvarado-Uribe et al. [2] summarized discovery as "the process and infrastructure required for a user to find an appropriate item" (p. 132). Deciding how to visualize with consideration of user context is an important research challenge. For example, OntoAIV, a representative visualization tool that implements semantic data visualization, also provides graph sets (e.g., bar charts and heat maps) for visualizing semantics, but the choice of which graphs should be visualized is still left to the user [2]. Other knowledge discovery tools, such as IntelliSearch [28], Sindice [43], and Knowledge Graph [38] do not yet have visualization capabilities. In sum, we need a way to reduce the cost and time of selection for creation of business charts by extracting knowledge from various sources of information such as online data without relying on prior human knowledge.

3. Proposed Method

3.1 Overall Process

The overall process for extracting rules to aid in selecting of business charts based on data from online text is shown in <Figure 2>. First, the concept corpus recognizes sentence patterns (e.g., "* refers to *", "* is used when *") that represent the definitions, concepts, or characteristics of an object. The concept corpus is a collection of statements that contain a description or definition of a concept. Then, a search is performed for portal texts, document sites, and corporate management reports according to business chart names, and sentences belonging to the already acquired sentence patterns are extracted from the previously obtained online texts. These sentences and sentence patterns correspond to particular business charts expressing certain concepts and features. After that, each sentence is extracted using terms of morphological analysis and the stop words are deleted. The importance of each keyword is then calculated for each business



〈Figure 2〉 Overall Process

chart by calculating TF-IDF around terms. Next, value of threshold, namely h-value (0~1), is set to generate a matrix having a value of 1 when a TF-IDF value exceeds a pre-determined threshold, and 0, otherwise. Then, the algorithm for generation of a Galo is lattice is applied to the 0/1 matrix to generate the keyword-business chart selection rule. Once a set of selection rules has been generated, the rule set is used to infer the execution time of the problem context entered by the user in the form of a sentence or a vector of a specific word, resulting in an optimal business chart recommendation.

3.2 Search by Patterns

First, we recognize sentence patterns (such as "* refers to *", "* is used when *") that express the definitions, concepts, and characteristics of an object and help identify the business chart that will generate the rule. To do this, we use Wikipedia, which defines business chart definitions, concepts, and features. Wikipedia's content is managed by the portal, which has its own name space. A Wikipedia namespace is a set of Wikipedia pages whose names begin with a particular reserved word recognized by the MediaWiki software (followed by a colon). Wikipedia consists of a total of 13 update glossaries. Next, on each page of Wikipedia, we find a sentence with a subject that matches the title of the page. For example, if CD-RW is the title of the page, the following content column looks for sentences that begin with CD-RW. It recognizesthree sentences: "CD-RW is a *", "CD-RW must be *" and "CD-RW can *". Examples of concept-specific explanatory patterns obtained in this way are shown in <Table 1>, and full sets are shown in <Appendix A>.

〈Table 1〉 Extracted Definition Patterns

Definition Patterns
(a/an/the) * allow(s) *
(a/an/the) * comprise(s) *
(a/an/the) * consist(s) of *
(a/an/the) * describe(s) *
(a/an/the) * enable(s) *
(a/an/the) * include(s) *
(a/an/the) * is (an) (generic) term/mean of *
(a/an/the) * is a (specific) class of *
(a/an/the) * is a chart/method that/for *
(a/an/the) * is a concept (in) *
(a/an/the) * is a reference to *
(a/an/the) * is a set/kind/feature/type/sort/piece/portion/form/collection/family/representation/member of *
(a/an/the) * is a specification for *
(a/an/the) * is an effort to *
(a/an/the) * is/are ((most) widely) used to *
(a/an/the) * is/are a *
(a/an/the) * is/are defined *
(a/an/the) * is/are made up of *
(a/an/the) * is/are useful *
(a/an/the) * is/was/are/were/has been/have been used *
(a/an/the) * provides (a/an/the) *
(a/an/the) * refer(s) to *
(a/an/the) * require(s) *
A*chart is a *

〈Table 2〉 Sample Corpus Created by Crawling Online Documents on Business Charts

Chart Name	Description
Radar Chart	A radar chart is commonly used by consultants to demonstrate how a client organization compares to its competitors in a given industry.
Area Chart	An area chart is a line chart where the area between the line and the axis are shaded with colors. These charts are typically used to represent cumulated totals over time and are the conventional way to display stacked lines.
⋮	⋮
Line Chart	Line charts are used for connection of individual data points in a data view. They are especially useful when it comes to displaying trends over time. The trajectory of the line, up or down, gives you clear information on how a specific data point progressed over time.

3.3 Data Collection

To build a learning set for creating business chart selection rules, we first determine the list of business charts. In this study, 17 business charts were selected, including the most frequent bar and column charts. Next, a corpus (corpus from online document), other than the concept corpus, is constructed by selecting only sentences that contain the same patternsand business chart names as those extracted from the Internet. <Table 2> provides an example of a corpus constructed in this study.

3.4 Rule Generation by Galois lattice

To construct a Galois lattice, we first use the TF-IDF method to recognize keywords for each business chart from the corpus. TF-IDF is calculated using the following equation:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of keyword i indocument j

df_i = number of documents containing keyword i

N = total number of documents

When the TF-IDF value of each chart is determined for each keyword, a chart-keyword map is constructed. An example of a chart-keyword map is shown in <Table 3>. With the h-value set to 0.05, the filtered, reduced chart-keyword map is shown in <Table 4>.

In the actual problem, since neither the keyword vector value nor the x-axis value is recognized, it is necessary to discriminate with very limited input values. Therefore, a discrimination algorithm that deals with input features containing a large number of null values must be used. In addition, we should be able to recommend multiple charts as only partially defined input feature values. In other words, there is

〈Table 3〉 Sample Chart-Keyword Map

Chart_name	Keyword	n	TF-IDF
candlestick chart	bullish	102	0.159709
candlestick chart	bearish	101	0.098174
candlestick chart	dozens	96	0.083749
candlestick chart	patterns	107	0.079742
tree map chart	tree	45	0.076502
candlestick chart	live	100	0.074516
histogram chart	visualizes	42	0.068795
calendar chart	calendar	82	0.044454
⋮	⋮	⋮	⋮
area chart	swimlane	3	0.005228

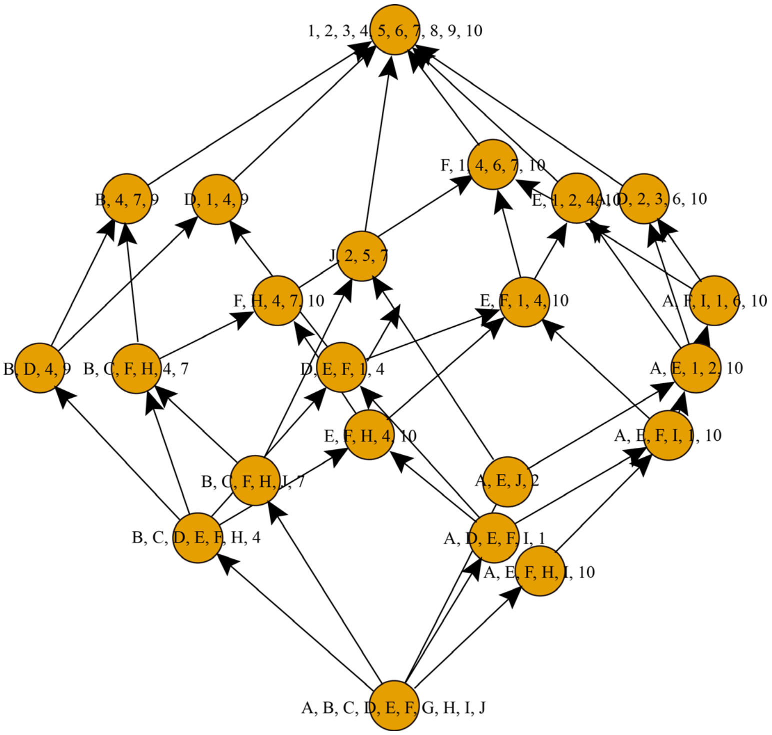
〈Table 4〉 Reduced Chart-Keyword Map (h-value = 0.05)

Chart_name	Keyword	n	Value
candlestick chart	bullish	102	1
candlestick chart	bearish	101	1
candlestick chart	dozens	96	1
candlestick chart	patterns	107	1
tree map chart	tree	45	1
candlestick chart	live	100	1
histogram chart	visualizes	42	1
calendar chart	calendar	82	0
⋮	⋮	⋮	⋮
area chart	Swimlane	3	0

a many-to-many correspondence between input feature values and output values. In this situation, the existing discrimination algorithm is difficult to apply. Therefore, we use the Galois lattice construction method, which is very useful for the task of knowledge discovery in databases [14] and automatic formation of concepts [8].

Galois lattice, or G-lattice, is a graphic method of representing knowledge structures. The nodes of G-lattices represent all the possible concepts in a given body of knowledge in addition, a notion defines a set of individuals or properties with no exceptions or idiosyncrasies [3]. The G-lattice provides a tool to represent all the possible developmental phases to reach a given total body of knowledge via different partial knowledge structures. Galois lattices are based on the triple (A, E, I) defined by two-mode social network data. A and E are finite nonempty sets and I is in a binary relation with A × E. As an example, consider the hypothetical data matrix shown in <Figure 3(a)>. There are six actors, labeled 1 through 6, and there are four social events, labeled A through D. An entry of 1 in a cell indicates that the actor designated by the row was involved in the event designated by the column.

To automatically extract a set of rules from text, first of all, a chart-keyword map contain-



〈Figure 3〉 Generated Galois Lattice

〈Table 5〉 A Galois Lattice (see 〈Figure 3〉) as a Rule Set

Rule #	IF	THEN
E1	5, 6, 7, 9, 11, 12, 13, 17, 18, 19	F
E2	2, 3, 7, 9, 10, 12, 13, 15, 16, 18	R
E3	1, 4, 6, 8, 10, 13, 15, 16, 18	M
	⋮	
E7	4, 10, 17	A, C
E8	4, 8, 10, 13	A, M
	⋮	
E21	4, 10	A, C, M
E22	8, 13	A, B, M
E23	9, 12, 13	B, F, R
E24	13, 18	F, M, R
E25	10, 13	A, M, R
E26	12, 17	B, C
	⋮	
E33	13	B, F, M
E34	17	A, C, F
	⋮	

where 1: clustered, 2: correlations, 3: cumulative, 4: layered, 5: multi, 6: categorical, 7: connected, 8: variation, 9: vertical, 10: xy, 11: zoom, 12: axes, 13: linear, 14: plot, 15: rectangle, 16: scatter, 17: summaries, 18: trend, 19: bar, 20: stacked, A: area chart, B: bar chart, C: bubble chart, F: column chart, M: line chart, R: pie chart.

ing the entire chart and all keywords is created. The problem in this study, however, is that there are so many keywords that the value is likely to be a sparse matrix. Therefore, the following procedure is proposed to solve this problem:

1. Determine the h-value, which is a threshold based on the TF-IDF value of keywords.
2. The h-value receives the query word set suggested by the user.
3. Only charts with h-value or higher thresholds are selected for all words in the query word set. If there is only one chart, select it and exit. If there is no chart, lower the h-value by a percentage, then go back to step 3. If there is more than one chart, proceed to step 4.
4. Create a chart-sub map of keyword map and a Galois lattice consisting of selected charts only.
5. Create a rule set with the generated Galois lattice and generate an additional query.
6. Select a chart according to the user's answer as a result of querying according to the query.

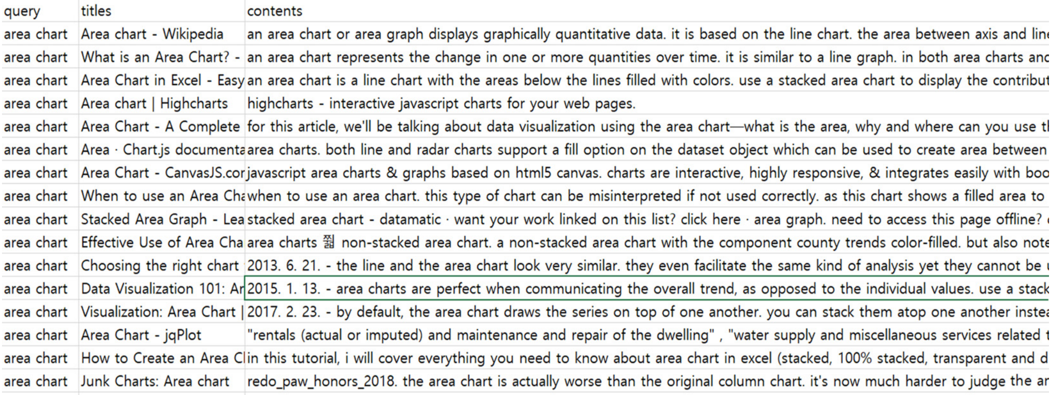
For example, if we have a Galois lattice as shown in <Figure 3>, a rule setis derived from the Galois lattice (see <Table 5>).

4. Experiment

4.1 Procedure

We used R, an open source programming language, to implement the proposed method. First, we crawled the web to build a corpus in which to identify the key words of each business chart. In the Google search box, each chart name was entered in English and Chinese, and the search results showed collected titles and content from all pages. In total, 17 business charts were considered. All the crawled data were converted into a data frame and saved in the form of a spreadsheet, as shown in <Figure 4>.

Based on the results of the crawling process, a corpus was constructed by selecting sentences in which a specific pattern appears. All patterns were converted to regular expressions to ensure selection of sentences that contained a specific pattern. Next, preprocessing was per-



query	titles	contents
area chart	Area chart - Wikipedia	an area chart or area graph displays graphically quantitative data. it is based on the line chart. the area between axis and lin
area chart	What is an Area Chart? -	an area chart represents the change in one or more quantities over time. it is similar to a line graph. in both area charts anc
area chart	Area Chart in Excel - Easy	an area chart is a line chart with the areas below the lines filled with colors. use a stacked area chart to display the contribut
area chart	Area chart | Highcharts	highcharts - interactive javascript charts for your web pages.
area chart	Area Chart - A Complete	for this article, we'll be talking about data visualization using the area chart—what is the area, why and where can you use t
area chart	Area · Chart.js documenta	area charts. both line and radar charts support a fill option on the dataset object which can be used to create area between
area chart	Area Chart - CanvasJS.cor	javascript area charts & graphs based on html5 canvas. charts are interactive, highly responsive, & integrates easily with boo
area chart	When to use an Area Cha	when to use an area chart. this type of chart can be misinterpreted if not used correctly. as this chart shows a filled area to
area chart	Stacked Area Graph - Lea	stacked area chart - datamatic · want your work linked on this list? click here · area graph. need to access this page offline?
area chart	Effective Use of Area Cha	area charts 쮦 non-stacked area chart. a non-stacked area chart with the component county trends color-filled. but also note
area chart	Choosing the right chart	2013. 6. 21. - the line and the area chart look very similar. they even facilitate the same kind of analysis yet they cannot be
area chart	Data Visualization 101: Ar	2015. 1. 13. - area charts are perfect when communicating the overall trend, as opposed to the individual values. use a stack
area chart	Visualization: Area Chart |	2017. 2. 23. - by default, the area chart draws the series on top of one another. you can stack them atop one another inste
area chart	Area Chart - jqPlot	"rentals (actual or imputed) and maintenance and repair of the dwelling" , "water supply and miscellaneous services related
area chart	How to Create an Area Cl	in this tutorial, i will cover everything you need to know about area chart in excel (stacked, 100% stacked, transparent and d
area chart	Junk Charts: Area chart	redo_paw_honors_2018. the area chart is actually worse than the original column chart. it's now much harder to judge the ar

〈Figure 4〉 Sample Raw Data Obtained by Web Crawling

〈Table 6〉 Sample TF-IDF for Each Chart-Keyword

Business Chart	Keyword	n	TF-IDF
candlestick chart	bullish	102	0.159709
candlestick chart	bearish	101	0.098174
candlestick chart	dozens	96	0.083749
candlestick chart	patterns	107	0.079742
tree map chart	tree	45	0.076502
candlestick chart	live	100	0.074516
histogram chart	visualizes	42	0.068795
calendar chart	calendar	82	0.044454
area chart	swimlane	3	0.005228

〈Table 7〉 Group Description

Group	List of business charts	Number of keywords (TF-IDF) provided	Chart(s) recommended (Galois lattice)
Group 1	O	0	X
Group 2	O	10	X
Group 3	O	5	X
Group 4	O	0	O

formed so that the computer could recognize the sentences in natural language. After that, the TF-IDF value was derived to determine the importance of the main word or the related word in each chart. <Table 6> shows examples of the frequency and TF-IDF values of keywords derived for each business chart.

The experiment was conducted with business students majoring in business administration. The participants were divided into four groups (see also <Table 7>) as follows.

Group 1: Control Group - These participants were first presented with examples, a list of sentences, and business charts and asked to identify the business charts each sentence described. After a certain amount of time, the same problem was reattempted.

Group 2: Experimental Group (Full) - These participants were first presented with an example, a list of sentences, and business charts and asked to identify the business chart each sentence described. After a certain period of time, the top 10 keyword sets with high TF-IDF values ere presented for each business chart.

Group 3: Partial - These participants were first presented with an example, a list of sentences, and business charts and asked to identify the business chart each sentence described. After a certain period of time, the top five keyword sets with high TF-IDF values ere presented for each business chart.

Group 4: Experimental Group (Galois) - These participants were first presented with an example, a list of sentences, and business charts and asked to identify the business chart each sentence described. After a certain amount of time, one or more business charts were recommended for each problem by applying the business chart recommendation rule generated by the Galois lattice generation method based on the problem context.

〈Table 8〉 Keywords Provided to the Partial Group

Business Chart	Keywords
waterfall chart	sequential, cumulative, bridge, positive, negative
tree map chart	tree, hierarchical, patterns, space
scatter chart	rectangle, dimensional, xy, zoom
radar chart	multivariate, dimensional, polar, quantitative, spider
pie chart	slices, sectors, divided, circle, aggregation
organization chart	organization, structure, department, corporate, directors
line chart	straight, line, dashed, connected
histogram chart	continuous, distribution, interval, period, frequency
gauge chart	needle, gauge, frequently, variability
Gantt chart	project, management, schedule, timeline
donut chart	pie, hole, geom_label, donut, center
column chart	stacked, layer, column, comparative
candlestick chart	dozens, patterns, stock, future
calendar chart	daily, calendar, bold, time
bubble chart	dimensions, geographic, dots, scatter, size_data
bar chart	bars, rectangular, categorical, horizontal, silent
area chart	area, filled, stacked, bar_chart, series

The sample keywords provided to the partial group are shown in <Table 8>.

The questionnaire used in the experiment is shown in <Appendix B>.

4.2 Subjects

The groups who participated in the experiment may be described as follows. Except for insincere responses, there were 55 participants in the control group, 30 in the full group, 20 in the partial, and 30 Galois. The genders of the participants were 57 males and 78 females there were100 business majors and 35 IT majors. There were also a few participants who majored in both business and IT. In such cases, they were classified as business majors. Differences between groups were determined by ANOVA test.

4.3 Results

The results of the experiment are shown in <Tables 9> and <Table 10>. In terms of accuracy, the group with the Galois lattice results improved the accuracy by 10.26 points (from 5.24 points on average to 15.50 points on average out of 17 points), which was more

〈Table 9〉 Results

		N	Mean	Std. Dev.
Improvement of accuracy	partial	20	6.00	3.784
	full	30	5.37	3.987
	no	55	0.07	1.913
	Galois	30	10.26	1.654
	total	135	4.39	2.593
Confidence	partial	20	4.85	1.725
	full	30	4.53	1.756
	no	55	2.40	1.382
	Galois	30	5.12	1.687
	total	135	3.84	1.583
Satisfaction with recommendation system	partial	20	4.95	1.538
	full	30	4.67	1.807
	no	–	–	–
	Galois	30	5.12	1.486
	total	80	4.91	1.619
Intention to use recommendation system	partial	20	5.40	1.231
	full	30	4.67	1.749
	no	–	–	–
	Galois	30	5.92	1.32
	Total	80	5.32	1.459

〈Table 10〉 Performance Comparison

		SS	d.f.	MS	F	p
Improvement of accuracy	Between group	810.753	3	405.376	44.428	0.000
	Within group	930.676	132	9.124		
	Total	1741.429	134			
Confidence	Between group	134.974	3	67.487	27.621	0.000
	Within group	249.217	132	2.443		
	Total	384.190	134			
Satisfaction with recommendation system	Between group	49.795	2	24.897	12.067	0.000
	Within group	210.453	83	2.063		
	Total	260.248	134			
Intention to use recommendation system	Between group	38.881	2	19.441	9.956	0.000
	Within group	199.176	83	1.953		
	Total	238.057	85			

than the partial group or the full group. The increase in accuracy of the partial group was higher than that of the full group or control group. On the other hand, when it comes to confidence about their answers, the Galois lattice group had the highest level of confidence in their answers compared to the other groups. In addition, the Galois group had the highest level of satisfaction with business recommendations and intention to use the recommendation system. This means that the rules created by the Galois lattice method are useful for selecting business charts, participants were more satisfied with them, and they were more likely to use them later. The control group, on the other hand, had lower values han the experimental group in terms of accuracy and confidence. As the control group did not use the recommendation system, neither satisfaction nor intention to use the recommendation system were investigated in this group.

5. Discussion

5.1 Implications

This study proposed a method for recommending appropriate business charts according to the task and situation. The results of the experiment showed that the proposed method outperforms other methods typically used for this purpose. This study has the following academic implications. First, this study is the first to recommend business charts based on textual information related to management, which is unstructured data. In the past, the selection of visual symbols or business charts has been entirely based on the user's prior knowledge and solicited expert help although existing visualization tools support business charts with widget-like features, recommending business charts for the task at hand has always been an ad hoc process [2, 28, 30, 38]. However, in this study, we describe a recommendation system that analyzes text, which is unstructured data, based on the theoretical background and classification system of business charts and recommends visualization techniques specific to the business areain which the user operates.

Second, the proposed text mining method is superior in terms of efficiency and ability to build recommendation knowledge. In general, existing recommendation systems must accumulate vast amounts of data to secure knowledge on which to base their recommendations [23]. This is a very knowledge-intensive process, and hence not effective in terms of time and cost [17]. Our proposed method efficiently constructs high-quality rules using pattern extraction and the Galois lattice rule generation method. The basic assumption underlying our approach to discovering rules for data visualization through online documentation is that visualization insights are generic, reusable, and likely to be crowd-oriented rather than individual because they are similar and do not vary significantly across business problem domains. In addition, visualization insights are inherent in the representation of text and can be classified into several finite types. The distinction in this study is to support data visualization by finding associations (data characteristics, business charts) from sentences in unstructured text.

Third, this study presents a new approach to recommend visualization methods by constructing a keyword-image matrix using a text mining technique. There have been proposals for visualization frameworks [29, 44], data visualization methods [10, 16, 21, 22, 25], and visualization recommendation methods using recommendation systems technology [44, 46]. However, these approaches remain conceptual and have the limitation of impracticality due to the cold start problem described in traditional studies [47]. In contrast, this study established a chart-keyword matrix based on management reports using actual business charts, demonstrating that it is possible to recommend accurate business charts without sufficient information about users, unlike with conventional visualization recommendation methodologies.

Finally, this study empirically verifies users' intention to use the proposed business chart recommendation system. The results indicate thatthe business chart selection rules extracted from the text significantly improved accuracy and reduced the time required for selection. While most existing studies suggest recommending by simulationof hypothetical situations [17, 47] or focus on accuracy and measurement performance [36], we conducted a rigorous experiment involving actual users to measurereal performance improvement. In addition, confidence [41, 50] and satisfaction with the recommendation outcome [48, 49], which are both important in evaluating recommendation systems, were significantly higher than for those who did not experience our system.

Also, the recommendation method which is represented in this study has the following practical implications. First, the results suggest that the proposed method enables users to visualize their data according to their intended use. In other words, users can choose appropriate charts to show their data despite the lack of relevant knowledge on chart selection through the recommendation system proposed in this study. A visual tool with this capability will attract users to experience it more easily than other data visualization tools. Therefore, this business chart recommendation system has a certain contribution in practice.

Second, the developers of data visualization tools can be encouraged to let the business chart recommendation system proposed in this study be plugged into their tools. Although there were some legacy visualization tools to assist chart selection before, the previous systems are too cumbersome to use or just show users the concept of the relevant chart, but did not really recommend the business chart.

Hence, we believe that the proposed system may create the benefit for the visualization tool developing companies.

5.2 Limitation and Future Research Directions

This study has several research issues, as follows. First, the business chart recommendation method proposed in this study has not yet been fully implemented. In future, we plan to complete the implementation, developing a graphic user interface and including it in the form of plug-ins to existing data visualization tools. In addition, the Galois lattice generation algorithm was developed using R, but for more sophisticated implementations, it should be implemented in a language such as C++ or Java in the future.

Second, since participants in the experiment were students who measured the outcome of the experiment, we must be very careful in generalizing the results. However, the business chart selection support method proposed in this study may be used for educational purposes, and the main user group will be students thus, the results of the experiment involving students should not be underestimated. In addition, since all participants were majoring in business administration or IT, the bias caused by a lack of understanding of business charts was minimized. Nevertheless, it would be meaningful to experiment with subjects other than data analysts in the future. Finally, intention to use or purchase is arguably limited in a cohortof students with relatively weak purchasing power.

Third, since the Galois lattice generation method finds all existing rules, scalability is more difficult to guarantee as the numbers of business charts and keywords increase. In our experiment, we selected less than 100 keywords using the TF-IDF criteria for 17 business charts, implemented them on 64-bit computers with16G RAM, and adjusted the size of the chart-keyword map by adjusting the h-value. No problems resulted from this procedure, but in future, consideration should be given to optimization of the TF-IDF criteria in the selection of keywords for optimal performance of the recommendation system and minimizing the amount of time required to generate rules.

6. Conclusion

In this paper, we discuss a new, cost-effective way to generate knowledge for selecting business charts from open online text. We hope that the methodology proposed in this study contributes to improving the quality of data visualization tools in the future. We also hope that this study will be a comprehensive resource for mining of existing and future knowledge and data analysis. Its insights may be useful to data analysts, engineering researchers, business practitioners, or educators. Many opportunities and challenges remain in this relatively new research area, which provide potential directions for future research in this field.

References

[1] Al-Kassab, J., Ouertani, Z. M., Schiuma, G., and Neely, A., "Information visualization to support management decisions," International Journal of Information Technology & Decision Making, Vol. 13, No. 2, pp. 407-428, 2014.

[2] Alvarado-Uribe, J., García, A. B., Gonzalez-Mendoza, M., Espinosa, R. L., Martín, J., and Espinosa, M., "Semantic approach for discovery and visualization of academic information structured with OAI-PMH," Acta Polytechnica Hungarica, Vol. 14, No. 3, pp. 129-148, 2017.

[3] Andor, C., Joó, A., and Mérö, L., "Galois-lattices: A possible representation of knowledge structures," Evaluation in Education, Vol. 9, No. 2, pp. 207-215, 1985.

[4] Anwar, A., Nagel, T., and Ratti, C., "Traffic origins: A simple visualization technique to support traffic incident analysis," in IEEE Pac, Vis. Symp., pp. 316-319, 2014.

[5] Blasco, J., Aleixos, N., Cubero, S., Gómez-Sanchís, J., and Moltó, E., "Automatic sorting of satsuma (Citrus unshiu) segments using computer vision and morphological features," Computers and Electronics in Agriculture, Vol. 66, No. 1, pp. 1-8, 2009.

[6] Bowman, R. L., Wang, Q., Carro, A., Verhaak, R. G., and Squatrito, M., "GlioVis data portal for visualization and analysis of brain tumor expression datasets," Neuro-oncology, Vol. 19, No. 1, pp. 139-141, 2016.

[7] Brown, L. D., Hua, H., and Gao, C., "A widget framework for augmented interaction in SCAPE," In Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 1-10, 2003.

[8] Cassol, I. and Arévalo, G., "A methodology to infer and refactor an object-oriented model from C applications," Software: Practice and Experience, Vol. 48, No. 3, pp. 550-577, 2018.

[9] Chang, T. W., "A literature review on information visualization of manufacturing industry sector," The Journal of Society for e-Business Studies, Vol. 21, No. 1, pp. 91-104, 2017.

[10] Chen, W., Guo, F., and Wang, F. Y., "A survey of traffic data visualization," IEEE Transactions on Intelligent Transportation Systems, Vol. 16, No. 6, pp. 2970-2984, 2015.

[11] Choe, E. K. and Lee, B., "Characterizing visualization insights from quantified selfers' personal data presentations," IEEE Computer Graphics and Applications, Vol. 35, No. 4, pp. 28-37, 2015.

[12] Enzenhofer, M., Bludau, H. B., Komm, N., Wild, B., Mueller, K., Herzog, W., and Hochlehnert, A., "Improvement of the educational process by computer-based visualization of procedures: Randomized controlled trial," Journal of Medical Internet Research, Vol. 6, No. 2, p. e16, 2004.

[13] Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T., "Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips," IEEE Trans. Vis. Comput. Graphics, Vol. 19, No. 12, pp. 2149-2158, 2013.

[14] Gmati, H. and Mouakher, A., "Fast and compact cover extraction from big formal contexts," In 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 209-212, 2018.

[15] Hansen, C. D. and Johnson, C. R., The Visualization Handbook, USA: Academic, San Diego, CA, 2004.

[16] Herman, I., Melançon, G., and Marshall, M. S., "Graph visualization and navigation in information visualization: A survey," IEEE Transactions on Visualization and Computer Graphics, Vol. 6, No. 1, pp. 24-43, 2000.

[17] Hwangbo, H., Kim, Y. S., and Cha, K. J., "Recommendation system development for fashion retail e-commerce," Electronic Commerce Research and Applications, Vol. 28, pp. 94-101, 2018.

[18] Ifenthaler, D. and Pirnay-Dummer, P., "Model-based tools for knowledge assessment," In J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop (Eds.), Handbook of research on educational communications and technology (4th ed.), New York, NY: Springer, pp. 289-301, 2014.

[19] Ifenthaler, D., "Learning analytics," In J. M. Spector (Ed.), The SAGE encyclopedia of educational technology, Vol. 2, pp. 447-451, Thousand Oaks, CA: Sage, 2015.

[20] Ifenthaler, D., "Toward automated computer-based visualization and assessment of team-based performance," Journal of Educational Psychology, Vol. 106, No. 3, pp. 651-665, 2014.

[21] Keim, D. A., "Information visualization and visual data mining," IEEE transactions on Visualization and Computer Graphics, Vol. 8, No. 1, pp. 1-8, 2002.

[22] Key, A., Howe, B., Perry, D., and Aragon, C., "Vizdeck: Self-organizing dashboards for visual analytics," In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 681-684, 2012.

[23] Kim, J. Y. and Kim, D. "A Study on the method for extracting the purpose-specific customized information from online product reviews based on text mining," Journal of Society for e-Business Studies, Vol. 21, No. 2, pp. 151-161, 2017.

[24] Knaflic, C. N., Storytelling with data: A data visualization guide for business professionals, John Wiley & Sons, 2015.

[25] Kreuseler, M., Lopez, N., and Schumann, H., "A scalable framework for information visualization," In IEEE Symposium on Information Visualization 2000, pp. 27-36, 2000.

[26] Lee, B., Brehmer, M., Isenberg, P., Choe, E. K., Langner, R., and Dachselt, R., "Data visualization on mobile devices," In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 7, 2018.

[27] Lee, H. Y. and Ong, K. L., "Visualization support for data mining," IEEE Expert, Vol. 11, No. 5, pp. 69-75, 1996.

[28] Mehta, A., Makkar, P., Palande, S., and Wankhede, S. B., "Semantic web search engine," International Journal of Engineering Research and Technology, Vol. 4, No. 4, pp. 687-691, 2015.

[29] Morton, K., Balazinska, M., Grossman, D., and Mackinlay, J., "Support the data enthusiast: Challenges for next-generation data-analysis systems," Proceedings of the VLDB Endowment, Vol. 7, No. 6, pp. 453-456, 2014.

[30] Mouromtsev, D., Pavlov, D., Emelyanov, Y., Morozov, A., Razdyakonov, D., and Galkin, M., "The simple web-based tool for visualization and sharing of semantic data and ontologies," In International Semantic Web Conference (Posters & Demos), 2015.

[31] Myers, B. A., Goldstein, J., and Goldberg, M. A., "Creating charts by demonstration," In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 106-111, 1994.

[32] Oh, J., Kim, J., Kim, J., and Kim, D., "Analysis of web traffic change using change ratio visualization," Proceedings of the Korea IT Service 2014, pp. 89-92.

[33] Perkel, J. M., "Data visualization tools drive interactivity and reproducibility in online publishing," Nature, Vol. 554, No. 7690, pp. 133-134, 2018.

[34] Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., and Andrienko, G., "Visually driven analysis of movement data by progressive clustering," Information Visualization, Vol. 7, No. 3/4, pp. 225-239, 2008.

[35] Scheepens, R., Willems, N., van de Wetering, H., and Van Wijk, J. J., "Interactive visualization of multivariate trajectory data with density maps," In 2011 IEEE Pacific Visualization Symposium, pp. 147-154, 2011.

[36] Shi, C., Hu, B., Zhao, W. X., and Philip, S. Y., "Heterogeneous information network embedding for recommendation," IEEE Transactions on Knowledge and Data Engineering, Vol. 31, No. 2, pp. 357-370, 2018.

[37] Shin, D. H., "Analysis of online social networks: A cross-national study," Online Information Review, Vol. 34, No. 3, pp. 473-495, 2010.

[38] Singhal, A., "Introducing the knowledge

graph: Things, not strings," Available: https://googleblog.blogspot.mx/2012/05/introducing-knowledge-graph-things-not.html, 2016.

[39] Streit, A., Pham, B., and Brown, R., "Visualization support for managing large business process specifications," In International Conference on Business Process Management, pp. 205-219, Springer, Berlin, Heidelberg, 2005.

[40] Symons, D., Konczewski, A., Johnston, L. D., Frensko, B., and Kraemer, K., "Enriching student learning with data visualization," 2017.

[41] Tang, M., Dai, X., Cao, B., and Liu, J., "Wswalker: A random walk method for QoS-Aware Web service recommendation," In 2015 IEEE International Conference on Web Services, pp. 591-598, 2015.

[42] Tegarden, D. P., "Business information visualization," Communications of the Association for Information Systems, Vol. 1, No. 4, pp. 1-38, 1999.

[43] Tummarello, G., Delbru, R., and Oren, E., "Sindice. com: Weaving the open linked data," In The Semantic Web Springer, Berlin, Heidelberg, pp. 552-565, 2007.

[44] Vartak, M., Madden, S., Parameswaran, A., and Polyzotis, N., "SeeDB: Supporting visual analytics with data-driven recommendations," Proceedings of the VLDB Endowment, Vol. 8, No. 13, 2015.

[45] Vessey, I., "Cognitive fit: A theory-based analysis of graphs versus tables literature," Decision Sciences, Vol. 22, pp. 219-240, 1991.

[46] Voigt, M., Pietschmann, S., Grammel, L., and Meißner, K., "Context-aware recommendation of visualization components," In The Fourth International Conference on Information, Process, and Knowledge Management (eKNOW), pp. 101-109, 2012.

[47] Wei, J., He, J., Chen, K., Zhou, Y., and Tang, Z. "Collaborative filtering and deep learning based recommendation system for cold start items," Expert Systems with Applications, Vol 69, pp. 29-39, 2017.

[48] Xu, C., Peak, D., and Prybutok, V., "A customer value, satisfaction, and loyalty perspective of mobile application recommendations," Decision Support Systems, Vol. 79, pp. 171-183, 2015.

[49] Zhang, L., Yan, Q., Lu, J., Chen, Y., and Liu, Y., "Empirical research on the impact of personalized recommendation diversity," In Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[50] Zhu, D. H., Wang, Y. W., and Chang, Y. P., "The influence of online cross-recommendation on consumers' instant cross-buying intention: The moderating role of decision-making difficulty," Internet Research, Vol. 28, No. 3, pp. 604-622, 2018.

〈Appendix A〉 Generated Concept Patterns

(a/an/the) * allow(s) *
(a/an/the) * comprise(s) *
(a/an/the) * consist(s) of *
(a/an/the) * describe(s) *
(a/an/the) * enable(s) *
(a/an/the) * include(s) *
(a/an/the) * is (an) (generic) term/mean of *
(a/an/the) * is a (specific) class of *
(a/an/the) * is a chart/method that/for *
(a/an/the) * is a concept (in) *
(a/an/the) * is a reference to *
(a/an/the) * is a set/kind/feature/type/sort/piece/portion/form/collection/family/representation/member of *
(a/an/the) * is a specification for *
(a/an/the) * is an effort to *
(a/an/the) * is/are ((most) widely) used to *
(a/an/the) * is/are a *
(a/an/the) * is/are defined *
(a/an/the) * is/are made up of *
(a/an/the) * is/are useful *
(a/an/the) * is/was/are/were/has been/have been used *
(a/an/the) * provides (a/an/the) *
(a/an/the) * refer(s) to *
(a/an/the) * require(s) *

〈Appendix B〉 Questionnaire Used in the Experiment

- Refer to the example given and provide an answer about the business chart as described.

Problem	Answer
Commonly one compares two or more quantities with an *** chart. *** charts are used to represent cumulated totals using numbers or percentages over time. Use the area chart for showing trends over time among related attributes. The *** chart is like the plot chart except that the area below the plotted line is filled in with color to indicate volume.	
A *** chart is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. *** graphs/charts provide a visual presentation of categorical data.	
A *** chart is a type of chart that displays three dimensions of data. You can use a *** chart instead of a scatter chart if your data has three data series that each contain a set of values. The sizes of the *** are determined by the values in the third data series.	
A *** chart is a visualization used to show activity over the course of a long span of time, such as months or years.	
A *** chart is a style of financial chart used to describe price movements of a security, derivative, or currency.	
*** charts are a good way to show change over time because it's easy to compare column lengths.	
Displaying values or percentages in data labels is very useful in a *** chart. You have one or more data series that you want to plot. *** charts are more space-efficient than pie charts because the blank space inside a *** chart can be used to display information inside it.	
A *** chart is a type of bar chart that illustrates a project schedule.	
On a *** chart, the value for each needle is read against the colored data range or chart axis. This chart type is often used in executive dashboard reports to show key business indicators.	
A *** is an estimate of the probability distribution of a continuous variable.	
A *** chart is often used to visualize a trend in data over intervals of time - a time series - thus the line is often drawn chronologically.	
An *** chart shows the structure of an organization and the relationships and relative ranks of its parts and positions/jobs.	
In a *** chart, the arc length of each slice is proportional to the quantity it represents.	
A *** chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point.	
A *** can be used either when one continuous variable is under the control of the experimenter and the other depends on it or when both continuous variables are independent.	
The *** chart is created based on this technique of data visualization. The *** chart is used for representing hierarchical data in a tree-like structure.	
A *** chart helps in understanding the cumulative effect of sequentially introduced positive or negative values. These intermediate values can either be time-based or category-based.	
Business Charts	1 scatter chart, 2 Gantt chart, 3 tree map chart, 4 bubble chart, 5 donut chart, 6 waterfall chart, 7 candlestick chart, 8 line chart, 9 gauge chart, 10 calendar chart, 11 pie chart, 12 radar chart, 13 organization chart,14 column chart, 15 bar chart, 16 histogram chart, 17 area chart

Hints	
waterfall chart	sequential, cumulative, bridge, positive, negative
tree map chart	tree, hierarchical, patterns, space
scatter chart	rectangle, dimensional, xy, zoom
radar chart	multivariate, dimensional, polar, quantitative, spider
pie chart	slices, sectors, divided, circle, aggregation
organization chart	organization, structure, department, corporate, directors
line chart	straight, line, dashed, connected
histogram chart	continuous, distribution, interval, period, frequency
gauge chart	needle, gauge, frequently, variability
Gantt chart	project, management, schedule, timeline
donut chart	pie, hole, geom_label, donut, center
column chart	stacked, layer, column, comparative
candlestick chart	dozens, patterns, stock, future
calendar chart	daily, calendar, bold, time
bubble chart	dimensions, geographic, dots, scatter, size_data
bar chart	bars, rectangular, categorical, horizontal, silent
area chart	area, filled, stacked, bar_chart, series

저자소개

Qingxuan Zhang (E-mail: bbandft@khu.ac.kr)
2017년 Guilin University of Technology, China 경영학 (학사)
2018년~현재 경희대학교 경영학과 석사과정
관심분야 Big Data, Data Analysis, Text Manning,

권오병 (E-mail: obkwon@khu.ac.kr)
1995년 한국과학기술원 경영과학과 (박사)
1996년~2004년 한동대학교 경영경제학부 부교수
2002년~2003년 Carnegie Mellon University, ISRI, School of Computer Science (방문과학자)
2004년~현재 경희대학교 경영대학 교수
관심분야 빅데이터, 인공지능 경영, IoT