# CAUSATION-DRIVEN VISUALIZATIONS FOR INSURANCE RECOMMENDATION

*Zhixiu Liu*[1], *Chengxi Zang*[2], *Kun Kuang*[1], *Hao Zou*[1], *Hu Zheng*[3], *Peng Cui*[1]

[1]Tsinghua University, [2]Cornell University, [3]Datebao Insurance Ltd.

## ABSTRACT

Recommender systems provide insurance enterprises with insights on customer behaviors and demands, thereby serving as an important source of revenue generation. Key success factors for a recommender include incorporation of multi-dimensional data from diverse media, clear and interpretable presentation of analytical results, and accurate recommendation of insurance products. However, most existing methods in the industry produce merely a list of recommendations without effectively visualizing them, and meanwhile are based on correlation-driven algorithms that can omit significant causal patterns in both customer and product data. In this paper, we propose a causation-driven visualization system that fundamentally transforms cross-media insurance data into network diagrams and performs recommendation reasoning. The system leverages data retrieved from both local insurance service providers and resources across the Web, and utilizes a multi-valued confounder balancing algorithm to compute the causal effect between customer and products, aiming to explore the reasons behind customers' purchasing behaviors and make appropriate recommendations accordingly. Furthermore, we conduct domain expert studies to evaluate the effectiveness of knowledge discovery by our causation-driven visualization system and correlation-driven tools currently adopted in the insurance industry. Results show that our system considerably outperforms others in terms of all evaluation metrics, and can potentially help insurers devise practical decision-making strategies.

***Index Terms***— Causation-driven, recommendation reasoning, confounder balancing, insurance recommendation

## 1. INTRODUCTION

Abundant cross-media data bring huge opportunities for the data owners, such as insurance companies, to discover knowledge and then support their business analysis and decision-making. Insurers are seeking for a more efficient recommendation system to help them deliver the most suitable products to their customers, and to develop comprehensive customer profiles for business intelligence. Visualizing those insurance recommendation results is one of the most important methods to business analysts, to reveal customers' preferences and their underlying reasons of behaviors, facilitating the decision-making process.

Data visualization methods are widely used for recommendation systems in many fields, such as finance [1], education [2], also fostering research in academia [3, 4]. Visualization tools such as Excel, Microsoft Power BI, Tableau, BOARD, *etc.* offer tools such as spreadsheets, bar/pie/line graphs, which are suitable for gaining statistical results, especially correlation-based statistics from their recommendation systems. However, existing visualizations for the recommendation systems mainly focus on the customer profile space [5], which largely ignores the intrinsic relationships between customers and products. Furthermore, the majority of those existing methods are correlation-based approaches, leading to the lack of interpretability for their recommendations, which makes them less attractive in many applications, especially those requiring decision making. How to visualize the reasons behind customers' purchasing behaviors and why a certain product is recommended by a causation-driven approach is still an open problem.

Causal inference is an efficient statistical modeling tool for explanatory analysis, widely adopted in various industrial realms, such as social marketing [6] and advertising [7]. Many methods have been proposed to analyze causation in observational data, such as propensity score based methods [8, 9, 10] and directly confounder balancing methods [11]. These methods are gaining ground in applied work, but most of these methods focus on causal inference on binary variables, leading to their limit on mining the causation between multi-value variables.

In this paper, by marrying the causal analysis technique and the visualization methods, we propose a causation-driven visualization for insurance recommendation, aiming to acquire better interpretable insights for recommendation reasoning. Specifically, we adapt a recent causal algorithm, Differentiated Confounder Balancing (*DCB*) [11], for causal inference on multi-valued variables, and propose the causation-driven Sankey diagram [1], a network visualization approach to present both the correlational and causal relationships between customers and the products they purchased. In order to extract features from the product space as comprehensive and accurate as possible, we not only use the provided data from

[1] The Sankey diagram is a type of network flow diagrams with the width of the arrows shown proportionally to the flow quantity, and has been widely adopted in taxonomy and uncertainty visualization [12] previously.

our cooperating insurance enterprise, but also consider the relevant media resources by retrieving insurance articles from the Web via search engine. Finally, we extract the features of insurance products by aggregating resources from cross-media platforms into an overall knowledge base.

By analyzing the data from both the insurance enterprise and cross-media resources from the Web, our causation-driven visualization approach, *DCB* algorithm driven Sankey diagram, helps to reveal interpretable insights on the actual predilections of customers when purchasing insurance products. Furthermore, we conduct several task-oriented expert studies to evaluate how our causation-driven visualization methods could improve the efficiency in knowledge discovery and better facilitate decision-making process.

We summarize our contributions as follows:

1. We propose a novel causation-driven visualization approach to reveal both the correlations and the causal relationships between customers and products for insurance recommendation.

2. To the best of our knowledge, we are the first to leverage both a causal inference model, that is, the *DCB* algorithm [11] and network visualization tools, that is, Sankey diagrams to interpret the reasons of insurance purchases and recommendations, with data from cross-media formats and platforms.

3. We conduct extensive task-oriented expert studies to evaluate our causation-driven visualization system, and our system indeed improves the efficiency in knowledge discovery and reasoning for the business analysts.

A video showcase of our work can be found at https://vimeo.com/310414360.

## 2. METHODS

### 2.1. Data and Problem

In current insurance recommendation systems, most of the visualizations are based on correlations between customers and products, ignoring the attempts hidden in customers' purchasing behaviors. In order to bridge this gap, we need a causation-driven visualization approach, to uncover the reasons why customers choose a specific type of insurance product. Specifically, we need to visualize that, in a specific category of the insurance products, how much weight each value takes up for a specific attribute (*e.g.,* age) of the customer profile. For example, of the products categorized "travel", we would like to see which age group is the most likely willing to purchase, assuming the effects of all other confounders are balanced.

To achieve this, we need an appropriate type of network, which is able to: (1) present the different values of a specific attribute of the customer profile space; (2) present the different categories of the product profile space; (3) reveal the link of the nodes from the two spaces, as well as the weights (calculated based on causal model) of each link. Sankey diagrams here (API supported by Echarts [13]) can effectively reveal the weights of the causal relationships between customers and the products they purchased. Different colors can represent different dimensions, meanwhile the flows can effectively reveal the proportions taken up by various categories, allowing different stakeholders to easily get perceptions at a first glance, as well as enabling more detailed comparisons.

**Data specifics.** We cooperate with datebao.com (*DTB*), one of the leading online insurance platforms in China. *DTB* is currently diverting its promotion strategies from posting promotional articles on social platforms such as *TopFeed* and *Baidu*, excavating royal customers based on its top-paid ranking list, due to the low conversion rate of its former approach.

The dataset covers 50,000 examples, each with a customer profile of 23 attributes as well as the payment information during a 4-year time period from Dec 14th, 2014 to Nov 18th, 2018. A fraction of the attributes are discarded either because the indicators are based on predictions by *DTB*, (for instance, "whether the customer has children or not" is based on his/her corresponding purchasing records) which cannot effectively indicate the real situations of those customers; or because the indicators are irrelevant in this context. Final attributes include: *age, gender, geographical region, purchasing platform, day-of-week, hour-of-day.*

The products purchased by users in our dataset are among the 52 specific types, currently on sale by our data provider *DTB*. Corresponding information of these products is mainly described in text, including some basic features such as insurance liability, insurance rate, premium delivery method, insurance period, insurance claims, and insurance payment method. Specifically, we utilize LDA topic model to extract the topics of the products descriptions, based on insurance text data provided by *DTB*.

In addition, cross-media resources on the Web that are related to these products also contain complementary information targeting at customers, as well as the information to the advantages of *DTB*'s products against existing products by competing institutions. In order to extract the features of these insurance products more accurately, we retrieve those relevant media resources from the web and extract corresponding features by the following two steps:

1. Web querying: The *Bing* searching API is deployed to retrieve insurance-related online articles. Although results can be confined to specific querying types such as the entire web, social media specifically, and geographic locations, in this work we focus on Chinese articles without further restrictions.

2. Features extracting: We conduct label mining of the corresponding insurance products in these articles, by

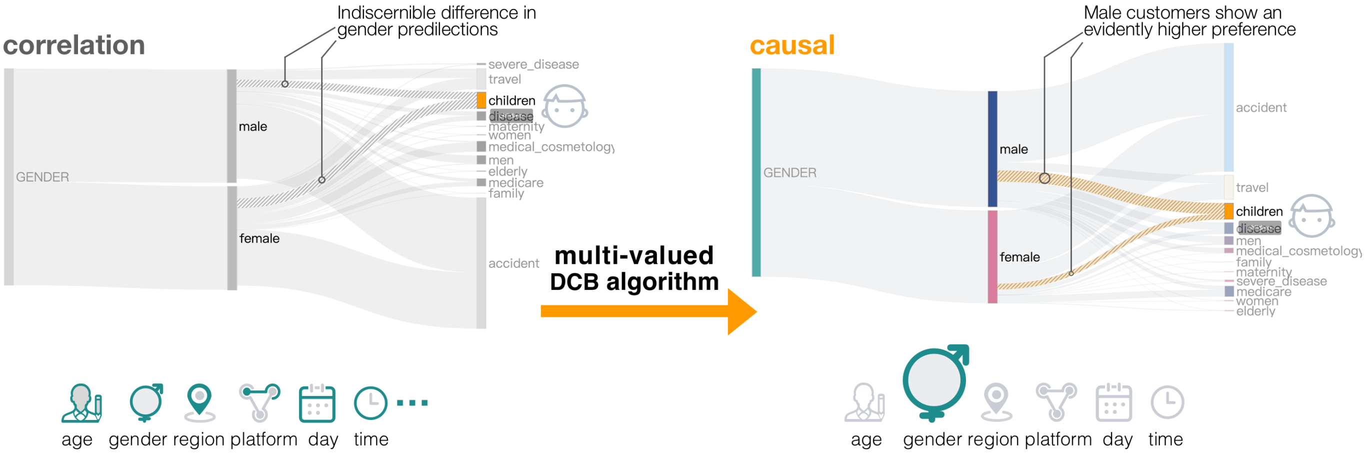

**Fig. 1**. Causation-driven Sankey diagram for insurance product recommendations. Graphs on the left and right are based on correlational and causal analysis respectively; for instance, for products categorized ”children”, the correlation graph reveals indiscernible difference in gender predilections; while by balancing the effects of other confounders, the causal Sankey graph reveals that male customers show an evidently higher preference.

implementing Bi-LSTM and CRF models. For example, from the articles introducing and promoting the products, we can extract specific features (in verbs or adjectives). By conducting correlation analysis with other web resources, we can acquire information in other aspects. This way, we can extract plentiful and multi-dimensional information of the insurance products, covering their inherent attributes, their social attributes, economic attributes, and more. An example result of extraction is as follows: travelling, overseas, family, Europe, low-risk.

Finally, by aggregating the topics of insurance products' description and the features extracted from the cross media sources, we can obtain integrated labels for the insurance products, ready for our following causal analysis.

## 2.2. Causal Model

In calculating the weights of each link thus manifesting the causal relationship between customers and products, we define an causal problem that estimating the causal effect [14, 15] of a particular treatment variable $T$ (*e.g.*, users' gender) on outcome variable $Y$ (*e.g.*, purchasing on "TRAVEL"). The casual effect (CE) of treatment variable $T$ with value of $t$ can be defined as:

$$CE(t)=E\big[Y(t)|T=t\big]-E\big[Y(0)|T=t\big],\ \ t=1,2,\cdots,k \tag{1}$$

where $Y(t)$ and $Y(0)$ represent the potential outcome of units with treatment status as treated $T=t$ and control $T=0$, respectively.

Unfortunately, we cannot directly estimate the $E(Y(0)|T=t)$, since we cannot observe the potential outcome $Y(0)$ for the units with $T=t$. Under unconfounderness [15] assumption, $E(Y(0)|T=t)$ is usually estimated by re-weighting observed units with sample weights $W$ to make the distribution of confounders $\mathbf{X}$ on control units mimic the distribuion on treated units. The recently proposed algorithm, Differentiated Confounder Balancing algorithm [11], achieved great performance on treatment effect estimation by jointly optimize sample weights and confounder weights. However, these methods are designed for causal effect estimation on binary variables, thus cannot be directly applied to our problem, where many variables that we are interested of are multi-valued.

To address this problem, we extend the *DCB* algorithm to our multi-valued scenario, and learn sample weights $W$ by balancing confounder between each treated group with $T=t$ (where $t=1,2,\cdots,k$) and the control group with $T=0$ as following:

$$W=\arg\min_{W}\sum_{t=1}^{k}\|\overline{\mathbf{X}}_0-\sum_{j:T_j=t}W_j\cdot X_j\|_2^2, \tag{2}$$

where $\overline{\mathbf{X}}_0$ represents the mean value of confounders $\mathbf{X}$ on units with $T=0$.

With the learned sample weights $W$ by our multi-valued *DCB* algorithm, we can estimate the causal effect of treatment variable $T$ on outcome variable $Y$ as follows:

$$\hat{CE}(t)=E\big[WY(T=t)\big]-E\big[WY(T=0)\big],\ \ t=1,2,\cdots,k \tag{3}$$

Finally, visualizations and conclusions could be drawn based on the causal results discovered by our algorithm. *e.g.*,

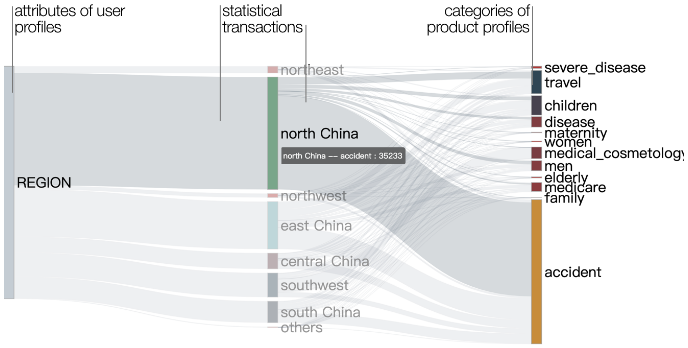

**Fig. 2**. Visualization based on correlations. From left to right, nodes represent customer attributes, clustered customer segments, and categories of products, respectively. Apparently, customers from ”north China” contributed most to the products among all ”REGION” segments.

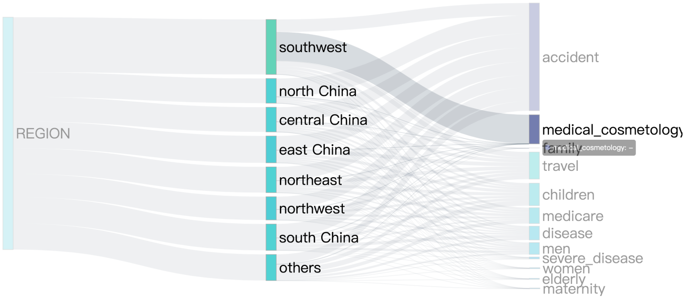

**Fig. 3**. When hovering on a specific node, corresponding links will be highlighted, allowing for detailed inspections; for instance, among links to ”medical cosmetology” on causation-based graph, ”southwest” presents the most weight.

for insurance type “medical cosmetology”, we can analyze customers in which particular geological region present the most predilections.

### 2.3. User Interface

The user interface consists of interactive Sankey diagrams that depict the information flow from one end to the other. Since the arrow width is drawn in proportion to the flow quantity, dominant transfers are given visual emphases and hence made easily discernible. Here in our context, we construct the two ends of Sankey diagrams from customer features and insurance labels, with the flows representing the causal relationship between them. By separating the two spaces, the diagrams enable a focus on the linkage, elaborating the causal relationship we would like to demonstrate.

**Demonstration.** Fig. 1 shows the user interface in the form of visual explorer. By switching between six customer profiles (age, gender, regions, purchasing platforms, day of week, and time of day), users can examine the corresponding Sankey diagram that encompasses: nodes and links of the clustered customer profile, nodes of the categorized product profile, and links between the two profile spaces. Link weights are calculated based on statistical correlation or causal analysis. A navigation bar at the top right facilitates toggling between the two analyses. Color of the nodes is calibrated according to the weights, and weights of approximating values share similar colors.

### 2.4. Use Cases

Traditional visualization approaches focus mainly on statistical correlation either between the customer profile space and the product profile space, or of one space only. We embed statistical correlation analysis in our visualization approach as a baseline, in order to make comparison and illustrate the superiority of our causation-driven method. The link weights here are based on statistical information. In addition, for each customer feature, we use k-means clustering to group customers into segments based on the transaction data. Then we calculate the transactions of each segment as weights of the presented links. Fig. 2 displays the correlation Sankey diagram for the ”REGION” feature as an example. The weight from node “REGION” to node “north China” is 35233, which means that there have been 35233 transactions from customers from ”north China” taken place.

**Comparison between the two analytical approaches.** Considerably biased interpretations could be drawn by exploring the customer profile nodes from the correlation visualization. For example, as presented in the “REGION” diagram, apparently customers resided in “north China” and “east China” make the top two most contribution to the overall insurance purchase, while the causation-driven diagram reveals that “southwest” takes up the largest weight. This discrepancy can be due to confounders (such as the distribution of customers) overlooked by the correlation analysis. With such confounders being balanced, our causation-driven result reveals a more likely speculation that customers in southwest China have the strongest purchasing power (Fig. 3). In this way, our approach helps stakeholders acquire true insights on valuable patterns behind customers’ purchasing behaviors.

## 3. RESULTS

We conducted several task-oriented user studies based on cognitive walkthrough (CW) method [16] to evaluate our causation-driven visualization approach in helping insurance enterprises to discover knowledge. We recruited several experts from *DTB* with more than three years of industry experience. Their career backgrounds included marketing and web-based product development with a focus on insurance.

The user study was organized as follows: we designed a series of tasks to evaluate the efficiency of knowledge discovery by our visualization system for business analysts in

**Table 1**. Results of the expert evaluations

| Indicators[3] | Existing Tools[4] | Our Approach |
|---|---|---|
| Time on specific tasks | 1hour | **less than 1min** |
| Ability on discovering hidden knowledge | 4 | **7.5** |
| Aesthetic integrity | 5 | **6.8** |
| Overall rating | 5 | **6.5** |

*DTB*. Our system was compared with widely-adopted data analytical tools and *DTB*'s current tools in terms of quantitative ratings (1-10 points). In addition, we noted down experts' feedback for further improvements. With regard to the current tools as the control group, the questions were designed as follows:

1. Please specify the current tools which you are using for data analysis.
2. Are you experiencing the following problems using above tools when analyzing data in assisting marketing strategies: (options included lack of technical support, lack of trained employees or in time of training employees, lack of confidence in investment return)
3. How much time on average do you currently spend on analyzing data to support decision-making?
4. Where else are you encountering obstacles when using the current tools for data analysis?

With regard to our visualizations as the experimental group, task-oriented questions included but were not limited to:

1. Please specify which regions contribute the most transactions to the overall flow, and the time you spent on finding the answers.
2. Please specify which of the age groups present the most purchasing power regarding medical cosmetology products, and the time you spent on finding the answers.

The results are shown in Table 1. In general, the domain experts preferred our causation-driven visualizations than current tools in decision making and knowledge discovery. 75% said that they observed novel phenomenon which are in contrast with their former expectations from the correlation-driven tools, and that they would consider adjusting their marketing strategies based on our causation-driven system.

## 4. FUTURE WORK

*Discovering how current approach could better support enterprises*

As feedback gathered from the experts revealed, it is not convenient for decision-makers to investigate visualizations on customized types of data. Therefore, we plan to carry out further user experience studies with the stakeholders, define what specific types of data could be analyzed and fed into our visualization approach, then design and implement the corresponding API. This is also in consistency with our ultimate goal, which is to provide a lightweight data visual analytics platform, whose supporting algorithms can help SMEs draw novel and insightful conclusions from their data. We will also conduct research on what additional algorithms could be incorporated into visualizations, to explore various possibilities of algorithm-empowered business intelligence.

*Enabling stakeholders from various fields to gain insights*

Recommender systems have been proven to be efficient in customer-oriented products and tackling cold-start problems. However, due to exponential growth of data and increasing complexity of algorithms, mechanisms behind successful recommendations seem a black-box to customers, especially those with little knowledge of insurance in our context. Our approach could be designed to better fit their needs and diminish information asymmetry. Using our visualizations, they could easily obtain a clear idea on predilections of other customers who share the same attributes with them.

## 5. ACKNOWLEDGEMENT

Thanks for the support of Datebao Insurance Ltd., for providing the data. The authors would also like to thank the reviewers and editors for their insightful comments and valuable suggestions which have led to substantial improvements of the paper.

## 6. REFERENCES

[1] M. L. Huang, J. Liang, and Q. V. Nguyen, "A visualization approach for frauds detection in financial market," in *2009 13th International Conference Information Visualisation*, July 2009, pp. 197–202.

[2] Antonio R. Anaya, Manuel Luque, and Manuel Peinado, "A visual recommender tool in a collaborative learning experience," *Expert Systems with Applications*, vol. 45, 10 2015.

[3] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory analysis via faceted browsing of visualization recommendations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 649–658, Jan 2016.

[4] David Gotz and Zhen Wen, "Behavior-driven visualization recommendation," in *Proceedings of the 14th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2009, IUI '09, pp. 315–324, ACM.

[3] For time: the shorter the better; for ratings: the greater the better.

[4] Spreadsheets/bar/pie/line graphs as mentioned earlier.

[5] Akrivi Katifori, Maria Golemati, Costas Vassilakis, George Lepouras, and Constantin Halatsis, "Creating an ontology for the user profile: Method and applications.," 01 2007, pp. 407–412.

[6] Kun Kuang, Meng Jiang, Peng Cui, and Shiqiang Yang, "Steering social media promotions with effective strategies," 12 2016, pp. 985–990.

[7] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang, "Causal inference via sparse additive models with application to online advertising," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, AAAI'15, pp. 297–303, AAAI Press.

[8] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert, "Evaluating online ad campaigns in a pipeline: Causal models at scale," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2010, KDD '10, pp. 7–16, ACM.

[9] Heejung Bang and James M Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, pp. 962–73, 01 2006.

[10] Peter Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, pp. 399–424, 05 2011.

[11] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang, "Estimating treatment effect in the wild via differentiated confounder balancing," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2017, KDD '17, pp. 265–274, ACM.

[12] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex M Kale, and Matthew Kay, "In pursuit of error: A survey of uncertainty visualization evaluation.," *IEEE transactions on visualization and computer graphics*, 2018.

[13] Deqing Li, Honghui Mei, Yi Shen, Shuang Su, Wenli Zhang, Junting Wang, Ming Zu, and Wei Chen, "Echarts: A declarative framework for rapid construction of web-based visualization," *Visual Informatics*, vol. 2, 05 2018.

[14] Guido W. Imbens and Donald B. Rubin, *Assessing Unconfoundedness*, p. 479495, Cambridge University Press, 2015.

[15] Paul R. Rosenbaum and Donald Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, pp. 41–55, 04 1983.

[16] Jakob Nielsen, "Usability inspection methods," in *Conference Companion on Human Factors in Computing Systems*, New York, NY, USA, 1994, CHI '94, pp. 413–414, ACM.