Discover Factors that Affect the use of Wikipedia as a Teaching Resource

Yongheng Zhang YZHAN470@JHU.EDU
Applied Math and Statistics/Data Science
yzhan470

1. Introduction

This project focus on investigating which factors can affect teachers to use Wikipedia as a teaching resource. To explore variety of factors, I used the survey data completed by the faculty members in two Spanish universities from UC Irvine Machine Learning Repository.

The survey focuses on collecting opinions on Wikipedia as a teaching resource. The measurements of the opinions can be categorized into internal opinions (the quality of the Wikipedia) and external opinions (thoughts from others). The scale of the measurements range from 1 (strongly disagree) to 5 (strongly agree). The dataset has 913 instances with missing values, and 53 attributes.

The total number of attributes seem large, but those attributes are categorized into 13 categories including perceived usefulness, perceived ease of use, perceived enjoyment, quality, visibility, social image, shared attitude, use behavior, profile 2.0, job relevance, behavior intension, incentives, and experiences.

This dataset is interesting because it provides a chance to explore what kinds of opinions could lead to the use of Wikipedia as a teaching resource. Rigorous analysis could be beneficial for the staff from Wikipedia because they can develop a better site for providing teaching and research resources. As a result, more teachers and students will use the Wikipedia for teaching/learning reference and the probability of donating to the website will increase. Besides, it could also gives other kinds of website some insights on improvements as well.

The scientific question I want to explore in this analysis is that what are the factors can directly affect the teachers' use behavior in Wikipedia, and how strong are the causual relationships between them. The approach consists four parts. First, I clean the data using Jupyter Notebook, and load the data using Tetrad. Second, I generate a directed acyclic graph using a proper algorithm, a statistical test/score, and parameters from Tetrad. Third, I perform some qualitative and quantitative analysis on the generated graph after causal discovery. Last, I summarize my result and do some discussion.

2. Preliminaries

This analysis report uses Directed Acyclic Graph (DAG) as the graphic model, with realizing the limitations are without considering unmeaured confounders and contagion mechanisms.

A Graph $G = (V, E)$ is directed and acyclic if the following three conditions are satisfied. First, $\mathcal{G}$ must be simple. This means that there will be at most one edge that connects to vertices. Second, the edges in E only contains directed edges. Third, the graph does not contain directed cycles. This means that for any $V_i \in V$, there is no sequence of directed edges in $\mathcal{G}$ such that $V_i \to \cdots \to V_i$.

A distribution $p(V)$ satisfies the factorization property with respect to $\mathcal{G}$ if

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)),$$

where $\text{pa}_{\mathcal{G}}(V_i)$ denotes the parents of V_i in $\mathcal{G}$.

A statistical model of a DAG or Bayesian network is a tuple $(\mathcal{G}, P)$ where $\mathcal{G}$ is a DAG, P is a set of distributions that factorize with respect to $\mathcal{G}$.

A distribution $p(V)$ satisfies the local Markov property with respect to a DAG $\mathcal{G}$ if for every variable V_i we have

$$p(V) = V_i \perp\!\!\!\perp \text{nd}^*_{\mathcal{G}}(V_i) \mid \text{pa}_{\mathcal{G}}(V_i)$$

where $\text{nd}^*_{\mathcal{G}} \equiv \text{nd}_{\mathcal{G}}(V_i) \backslash \text{pa}_{\mathcal{G}}(V_i)$. In other words, every variable V_i is independent of its non-descendant non-parents, given its parents. If a distribution $p(V)$ satisfies the DAG factorization property with respect to a DAG G, then it also satisfies the local Markov property with respect to G. If a distribution $p(V)$ satisfies the local Markov property with respect to a DAG G, then it also satisfies the DAG factorization property with respect to G.

I assume that, for each variable V_i on the DAG $\mathcal{G}$, it can be represented as a function of its parents and a noise term. These noise terms add randomness to the system. Mathematically, the previous description can be written as $V_i \leftarrow f_{V_i}\left(p_{\mathcal{G}}\left(V_i\right), \epsilon_{V_i}\right)$, which also defines the non-parametric structural equation model with independent errors (NPSEM-IE). Each functional $f_{V_i}\left(\text{pa}_{\mathcal{G}}\left(V_i\right), \epsilon_{V_i}\right)$ is non-parametric, and the noise terms are mutually independent. It also induces a probability distribution $p(V)$ that satisfies the factorization property with respect to $\mathcal{G}$. This assumption not only gives a generative story of how statistical model came to be, but also provides a theory of intervention and manipulation via assigning deterministic values to variables.

The casual model of a DAG or Bayesian network is a triple $(\mathcal{G}, P, M)$ where $\mathcal{G}$ is a DAG, P is a set of distributions that factorize with respect to $\mathcal{G}$, and M is the underlying NPSEM-IE equipped with the do-operator. This triple could let $V_i \to V_j$ be interpreted as V_i is a direct cause of V_j, and could let $V_i \to \cdots \to V_j$ be interpreted as V_i is a indirect cause of V_j.

A distribution $p(V)$ satisfies Global Markov Property with respect to a DAG $\mathcal{G}$ if for all disjoint subsets X,Y,Z of V, $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. To facilitate structure learning, I will restrict my analysis to the set of faithful distributions where $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. Thus, d-separation can detect all conditional independences in $p(V)$. In addition, I make the simplifying assumption that the relations between my variables are linear.

3. Methods

3.1 Feature Selection

Since the number attributes is big and attributes can be categorized into 13 categories, I pick one measurement from each category since measurements in each category are similar. The table below lists all 12 variables that will be used for analysis.

Table 1: Variable Explanation

Variable	**Category**	**Detailed Explanation of the Variable**
PU2	Perceived Usefulness	The use of Wikipedia improves students' learning
PEU2	Perceived Ease of Use	It is easy to find in Wikipedia the information you seek
ENJ1	Perceived Enjoyment	The use of Wikipedia stimulates curiosity
QU3	Quality	Articles in Wikipedia are comprehensive
IM1	Social Image	The use of Wikipedia is well considered among colleagues
SA1	Sharing attitude	It is important to share academic content in open platforms
PF3	Profile 2.0	I publish academic content in open platforms
JR1	Job relevance	My university promotes the use of open collaborative environments in the Internet
BI1	Behavioral intention	In the future I will recommend the use of Wikipedia to my colleagues and students
INC1	Incentives	To design educational activities using Wikipedia, it would be helpful: a best practices guide
EXP1	Experience	I consult Wikipedia for issues related to my field of expertise
USE1	Use behaviour	I use Wikipedia to develop my teaching materials.

3.2 Missing Values

The selected dataset done by section 3.1 exists many missing values. All the variables have missing values range from 7 to 35 rows. Considering the fact that teachers can take the survey anonymously and the survey features are answerable, I believe that the type for missing data in this case is *missing completely at random (MCAR)*. This term can be explained as the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random (wik, 2021). To make sure the existed data rows are accurate, I choose to drop all the missing rows and there are still 789 rows left.

3.3 Structural Learning

Since the goal is to identify exact factors that can cause teachers to use Wikipedia as a teaching resource and all variables except JR1,PF3, SA1 could possibly cause the use of Wikipedia, this analysis assumes that there are no latent common causes, and use DAG for graph elicitation.

I choose to use Tetrad to generate the graph since the knowledge box and search box it provides are reliable. Considering the fact that JR1, PF3, SA1 may not directly cause USE1, I set up the knowledge box by putting those three variables into tire1, placing USE1 into tire3, and setting the rest of the variables into tire2. Besides, I add additional restrictions that tire1 can only cause tire2.

In terms of search algorithm, I use Fast Greedy Equivalence Search (FGES) (Ramsey et al., 2017), which is the optimized and parallelized version of Greedy Equivalence Search (GES) (Chickering, 2002). It is a Bayesian algorithm performs heuristically searches and returns the model with the highest Bayesian score it finds. Specifically, the algorithm starts with an empty graph, tries to add edges between nodes to increase the Bayesian score, and stops when no additional edges can increase the Bayesian score if added. In the end, it starts perform edge removals until no removals can increase the Bayesian score. I believe the graphic model with the highest Bayesian score can reflect the relationships well considering the dataset I use for the analysis.

This algorithm requires a decomposable score. The score for the entire DAG model is calculated by the sum of logged scores of each variable given its parents. Since the survey data is discrete, I select the discrete BIC score as the score. It is calculated by $BIC = 2L - K \ln N$, where L is the likelihood, K is the number of parameters, and N is the sample size. The higher the score, the greater dependence it corresponds.

Below lists all the hyperparameter settings for the searching step.

1. FGES

- Maximum degree of the graph = 100 (the number of adjacent edges for each node is 100)

- Max number of neighbors considered in the power set consideration = -1 (the maximum number of T-neighbors in any scoring step is unlimited)
- Assumed faithfulness
- First step of FGES score for both $X \to Y$ and $Y \to X$

2. Discrete BIC Score

- Penalty discount = 1 ($BIC = 2L - cK \ln N$)
- Structure prior coefficient = 1 (use as a structure prior the default number of parents for any conditional probability table)

3. Bootstrapping

- Number of bootstraps = 100
- Percentage of resample size = 90
- Preserved Ensemble method (ensures that an edge that has been found by some portion of the individual sample graphs is preserved in the final graph, even if the majority of sample graphs returned no edge as their answer for that edge)
- Sample with replacement (cases which appear once in the original data set can appear multiple times in the subset)
- Add an original dataset as another bootstrap bootstrapping (an extra run using the original dataset)

4. Results

The graph below is the generated graph by the method described in the previous section.

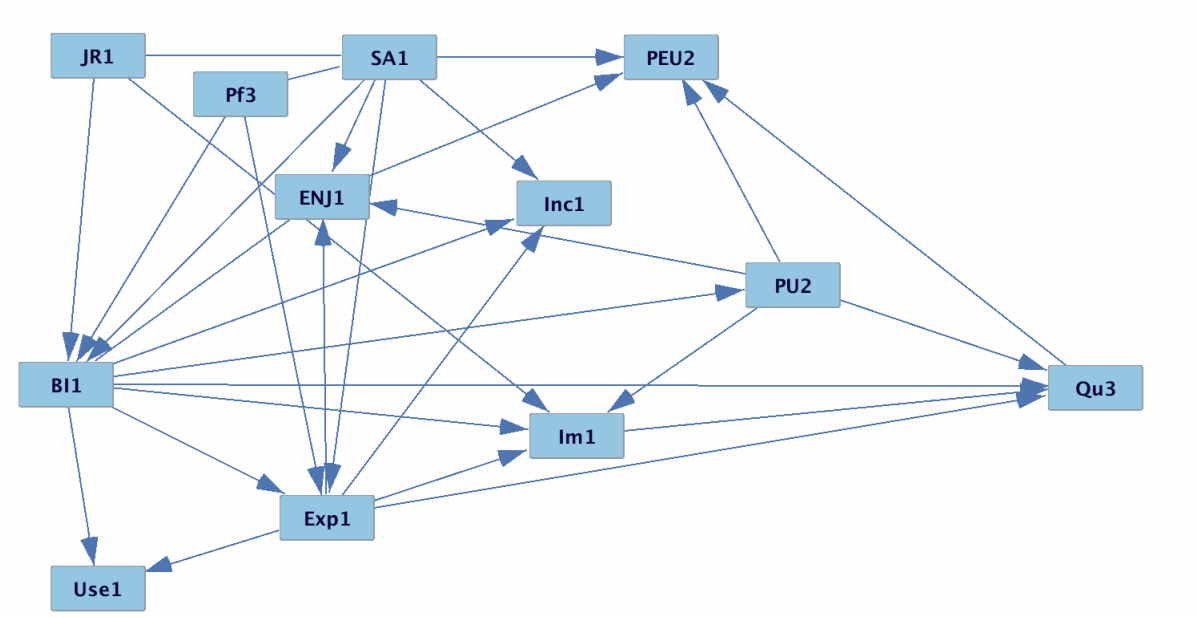

Fig1. Generated Graph

From the graph, one can see that

- BI1 (behavioral intention) and Exp1 (experience) can directly cause the Use1 (Use behaviour).
- BI1 (behavioral intention) is caused by JR1 (job relevance), PF3 (profile 2.0), SA1 (sharing attitude), and may caused by ENJ1 (perceived Enjoyment).
- EXP1 (experience) is caused by BI1(behavioral intention), PF3 (profile 2.0), and SA1 (sharing attitude).
- ENJ1 (perceived enjoyment) is caused by SA1 (sharing attitude), PU2 (perceived ssefulness), EXP1 (experience), and may caused by BI1 (behavioral intention).
- INC1 (incentives) is casued by SA1 (sharing attitude), BI1 (behavioral intention), and EXP1.
- IM1 (social image) is caused by JR1 (job relevance), PU2 (perceived usefulness), BI1 (behavioral intention), EXP1 (experience).
- QU3 (quality) is caused by PU2, BI1 (behavioral intention), EXP1 (experience).
- PU2 (perceived usefulness) is caused by BI1 (behavioral intention).
- PEU2 (perceived ease of use) is caused by ENJ1 (perceived enjoyment), PU2 (perceived usefulness), and QU3 (quality).
- JR1 (job relevance) may cause SA1 (sharing attitude) or vice versa, and PF3 (profile 2.0) may cause SA1 (sharing attitude) or vice versa.

5. Discussion and Conclusion

5.1 Inference Task

It is important to test whether there are edges between BI1 and Use1, and Exp1 and Use1 since those edges determine the causual effect of Use1. In this case, the valid statistic are (a) $\text{OR}(BI1, Use1|Exp1) = 1$ and (b) $\text{OR}(Exp1, Use1|BI1) = 1$. One way to check is by doing the fast Conditional Independence Test (FCIT) (Chalupka et al., 2018). This is a non-parametric conditional independence test that returns the p-value under the null hypothesis that $X \perp\!\!\!\perp Y|Z$. As a result, the test statistic (a) gives the p-value of 0.0174 and (b) gives the p-value of 0.0179. Therefore, we can conclude reject that $BI1 \perp\!\!\!\perp Use1|Exp1$ and $Exp1 \perp\!\!\!\perp Use1|BI1$. Thus, there is an edge exists between BI1 and Use1, and Exp1 and Use1.

One can also evaluate the strengths of the edges by computing the odds ratio. The odds ratio serves as a measure of (conditional) association. Given any pair of reference values x_0 and y_0, the conditional odds ratio of X and Y given Z is defined as:

$$\begin{aligned}\text{OR}(X = x, Y = y \mid Z) &\equiv \frac{p(X = x \mid Y = y, Z)}{p\left(X = x_0 \mid Y = y, Z\right)} \times \frac{p\left(X = x_0 \mid Y = y_0, Z\right)}{p\left(X = x \mid Y = y_0, Z\right)} \\ &\equiv \frac{\text{odds that } X = x \text{ instead of } x_0 \text{ when } Y = y}{\text{odds that } X = x \text{ instead of } x_0 \text{ when } Y = y_0}\end{aligned}$$

The odds ratio is symmetric. which means that $\text{OR}(X, Y \mid Z) = \text{OR}(Y, X \mid Z)$. Besides, $X \perp\!\!\!\perp Y \mid Z$ if and only if $\text{OR}(X, Y \mid Z) = 1$ for all x, y, z

Since values in BI1, Use1, and Exp1 range from 1 to 5, To use the odds ratios to calculate the

strength of the association, I set those values equal to zero if the original values equal to 1,2, or 3. Besides, I set those values equal to one if the original values equal to 4 or 5. As a result, under 200 bootstraps, $\text{OR}(BI1, Use1 \mid Exp1) = 3.865$ with 95% confident interval $(2.921, 5.995)$, and $\text{OR}(Exp1, Use1 \mid BI1) = 5.973$ with 95% confident interval $(3.302, 13.278)$. Therefore, not only one can conclude that BI1 can cause Use1 and Exp1 can cause Use1, but also one can see that the strength of the relationships is strong because the results are far greater than 1.

5.2 Intervention

Since $\text{OR}(Exp1, Use1 \mid BI1)$ has a larger value than $\text{OR}(BI1, Use1 \mid Exp1)$, I would like to further investigate and see how the predicted Use1 values vary by setting fixed Exp1 values and putting observed BI1 into consideration as well. The approach I take is that, firstly, I fit a linear regression model with Use1 as a response variable, BI1 and Exp1 as explanatory variables. Second, I perform intervention by setting all the Exp1 values to the lowest values they can possibly have (which is 1), and to the highest values they can possibly have (which is 5). After setting all the values in the Exp1 column to two different values and generate two different dataset, I take the observed values in BI1 and intervened values in Exp1, put it into the model in step 1, and generate pointwise prediction values for Use1.

I also performed 200 bootstraps by resampling the dataset with the length of the dataset and with replacement. First, I fit a linear regression with resampled data. Second, I take the original cleaned dataset, intervene values in Exp1, and use the model generated in the first step to make predictions. Third, with collected predictions over 200 bootstraps, I sort each row in ascending order and select the 2.5 and 97.5 percentile of each row. Lastly, I combine the variables including BI1, Exp1, predicted Use1, 2.5 percentile data of each row, and 97.5 percentile data of each row. With the combined dataframe, I am able to draw the linear trends between BI1 and predicted Use1 given two extreme Exp1 values, as well as the 95% confidence interval associated with the linear trends.

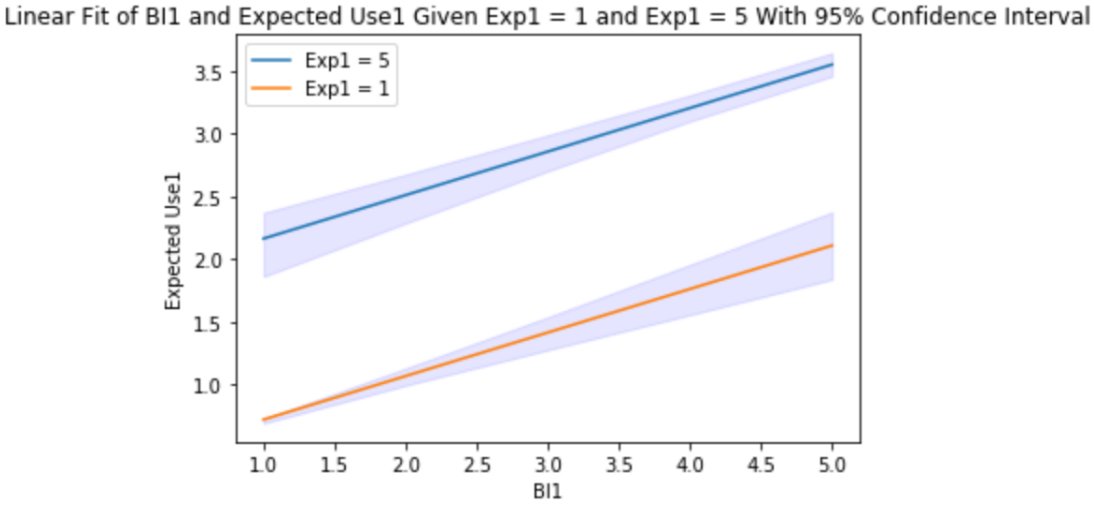

Fig2.

From the graph above, one can see that extreme Exp1 values do affect the relationship between BI1 and Use1 since those two linear trends and their confidence intervals are far apart. Thus, the intervention of Exp1 and observed BI1 values together do lead to significant changes in expected values of Use1.

5.3 Conclusion, Uncertainty, Limitation

From the generated graph and above testing, one can see that behavior intention and experience are significant factors that directly determine whether teacher will use Wikipedia as a teaching resource. Based on the generated graph, the FCIT test, the odds ratio test, and the intervention process in section 5.1 and 5.2, I conclude that those two factors have strong effect in terms of causing the user behavior directly. Those two attributes are also influenced by factors such as perceived usefulness, perceived enjoyment, perceived ease to use, and sharing attitude. This analysis takes a closer look on variables that directly cause the use behavior, and there is still space for analyzing other variables in depth as well.

There are uncertainty in the learning and inference task. In terms of the generated graph, one can notice that there exists undirected edge between JR1 and SA1, PF3 and SA1, as well as BI1 and ENJ3. Those undirected edge indicates that they could possibly cause one another. From Tetrad, the probability that there exists an edge between JR1 and SA1 after ensembling is 0.1188, between BI1 and ENJ1 after ensembling is 0.0099, between BI1 and ENJ1 after ensembling is 0.3762. Therefore, we cannot conclude exactly that there exist an edge between those variables. In terms of the inference task, the odds ratios calculated in 5.1 are uncertain. The differences between upper quantile and lower quantile in 95% confidence interval are still large. Specifically, for $\text{OR}(BI1, Use1 \mid Exp1)$, the difference is around 3; for $\text{OR}(Exp1, Use1 \mid BI1)$, the difference is nearly 10. Although the range for both 95% confidence intervals are significantly greater than 1, there are two problem exist. First, the big difference in the intervals indicates that we are not sure about the quantitative strength of the association. Second, the two intervals have overlap so we cannot conclude exactly which odds ratio is always larger than the other one.

Although the graph generated in this report looks similar to the work in Meseguer-Artola et al. (2015), there are still limitations and improvements need to be addressed. First, there are still space to fine-tune the hyperparameters and choose appropriate algorithms to generate the graph. Second, the dataset I use only considers teachers from two Spanish universities. If more universities from different regions are considered, the results would be more generalizable. Third, although there are already many attributes in the dataset, there could be more features that are useful for determining the user behavior. Last, there is space for me to think about latent causal factors and inference mechanisms such as contagion in this dataset.

References

Missing data — Wikipedia, the free encyclopedia, 2021. URL `https://en.wikipedia.org/wiki/Missing_data#Missing_completely_at_random`.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Antoni Meseguer-Artola, Eduard Aibar, Josep Lladós, Julià Minguillón, and Maura Lerga. Factors that influence the teaching use of wikipedia in higher education. *Journal of the Association for Information Science and Technology*, 67:1224–1232, 2015.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.