Some Reflections on the Use of Structural Equation Modeling for Investigating the Causal Relationships that Affect Search Engine Results

Massimo Melucci
massimo@unipd.it

University of Padua

Abstract. Search engines and recommender systems pervade everyday life and continuously make decisions regarding what information should be retrieved and how it should be ranked in order to meet the user's information needs on the user's behalf. Unfortunately, bias affects automated decision systems and as a consequence fairness cannot be taken for granted. Understanding whether and how bias affects search results can be a necessary and useful condition to every user and designer who aims to investigate the reasons that the systems fail or succeed. In this paper, we discuss whether Structural Equation Modeling (SEM) can be a useful methodology to investigate the causal relationships between the variables describing the content representation and retrieval processes of search engines and recommender systems. Understanding how and why a retrieval system retrieves certain documents can help understand when the system provides biased results. To this end, we provide a general illustration of the issues and the potential of SEM for causal discovery in Information Retrieval.

1 Introduction

The evidence of the widespread support provided by search engines and other web applications for human activities should cause all of us to feel frightened by the possible bias occurring in the search engine result pages which provide information relevant to the end user's information needs [1].

The possible bias on the web calls for theories, methods, data structures and algorithms for supporting the end users to recognize unfair results and find alternative sources of information. Recent scientific initiatives such as research workshops [6, 14] and legislative initiatives [7] signal the importance of fair and transparent search and recommendation systems.

The search for the reasons that a search or recommendation system and in general an Information Retrieval (IR) system provide a certain result page to the end user suggested to us as well as other researchers that we should frame the

 BIRDS 2020, 30 July 2020, Xi'an, China (online).

problem within a causality scheme. The internal mechanism of a search engine that ranks a document can be viewed as a cause and the actual document ranking as an effect.

In this paper, we suggest that Structural Equation Modeling (SEM) can be considered as a possible framework for searching for the causes which can produce an observed effect, that is, the internal mechanisms of a search engine producing an observed document ranking. A structural equation model may be an appropriate conceptual instrument for investigating the causal relationships between the mechanics of a search engine and the search engine result pages because these pages can be represented as matrices of unit-feature pairs, i.e. data matrices and the features that are observed in the pages can correspond to the manifest variables of the structural equation model. The regression coefficients and the inter-variable covariances can represent possible causal relationships which will be tested for a given sample of data, i.e. the observed covariance matrix. In [13] we illustrated how SEM can be utilized to describe the mechanisms underlying retrieval. Instead, some aspects regarding causality and SEM in the context of IR are discussed in this paper.

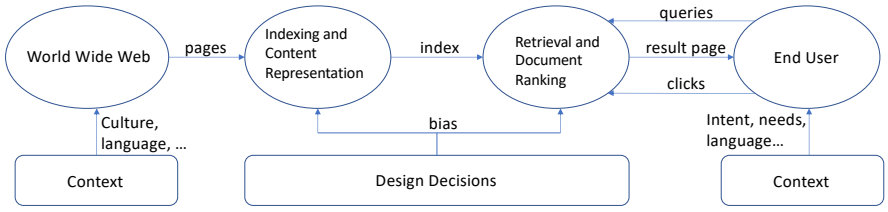

Fig. 1. The architecture of a retrieval system and how design and context may affect web page retrieval and ranking and ultimately the end user's experience.

2 How Context can Affect an IR

An IR system is a computer system designed and implemented to perform IR activities, i.e. those activities aiming to deliver all and only relevant information to meet a user's information need. A search engine is the most popular implementation of such a system yet IR technology pervades any device such as desktop computers and smartphones. Figure 1 depicts the general architecture of an IR system and how it relates to contextual factors and design decisions. The web pages that are crawled from the web are indexed in order to implement indexes and a representation of the content. The web page content is then retrieved and ranked to answer the end user's queries and respond to the clicks and eventually the end user's information needs. Because of the size of the index, a system has got to decide which tiny subset of pages should be retrieved and how these pages should be ranked and displayed on the user's device screen. Of course, the system

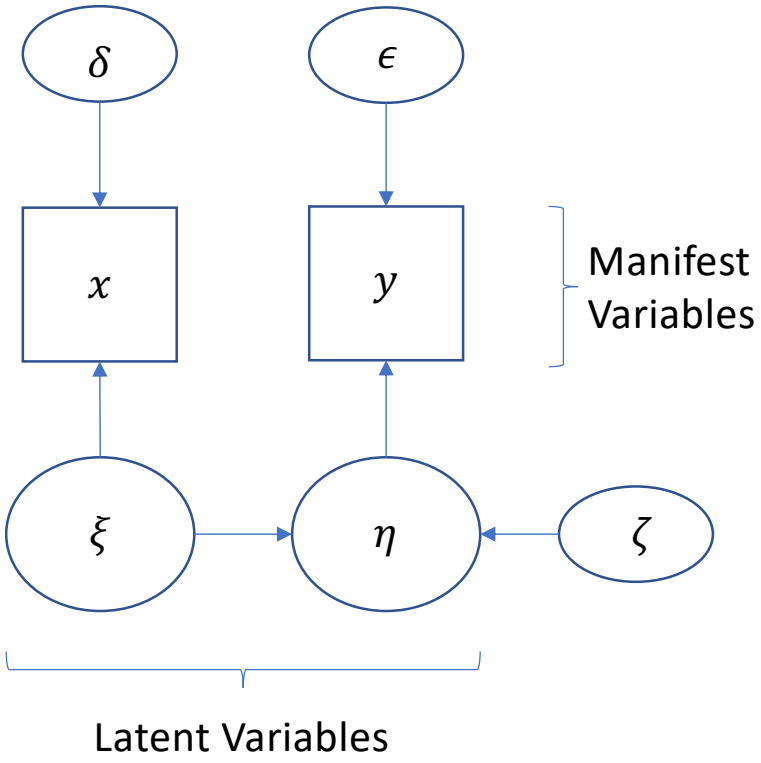

Fig. 2. A pictorial representation of a structural equation model

cannot decide on its own, since it is just an implementation of data structures and algorithms designed by programmers, engineers and scientists. The theoretical models, such as deep neural networks, that are utilized as the basis of the system implicitly decide what to retrieve and how to rank the indexed web pages. These models might be so complex that even their designers may not be aware of all the internal mechanisms driving the system toward a certain ranking. Such ignorance may be a source of bias since the designers of a system may bring some hidden selection mechanisms into operation.

3 What is SEM

SEM refers to the complex of multivariate statistical methods aiming to specify, estimate and fit a system of linear equations to a dataset [3]. The variables of the linear equations can be either exogenous or endogenous and in parallel they can be either manifest or latent, thus yielding four types of variable (latent endogenous, latent exogenous, manifest endogenous, manifest exogenous). A

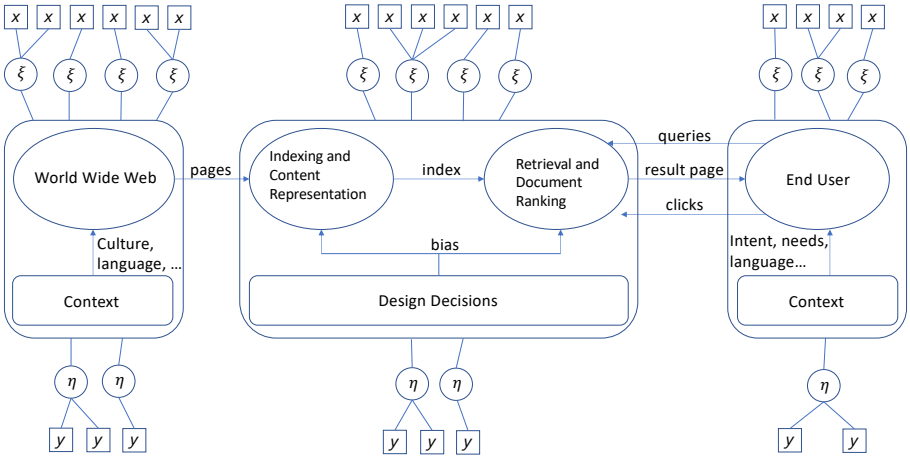

Fig. 3. How a structural equation model relates with a retrieval system

structural equation model can be specified in general terms as follows (Figure 2):

$$\eta = \mathbf{B}\,\eta + \mathbf{\Gamma}\,\xi + \zeta \tag{1}$$

$$\left.\begin{aligned} y &= \mathbf{\Lambda}_y\,\eta + \epsilon \\ x &= \mathbf{\Lambda}_x\,\xi + \delta \end{aligned}\right\} \tag{2}$$

where Eq. 1 is called "latent model" and Eq. 2 is called "measurement model". In particular, η is a vector of endogenous latent variables, ξ is a vector of exogenous latent variables, x is a vector of exogenous manifest variables, and y is a vector of endogenous manifest variables. $\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Lambda}_y, \mathbf{\Lambda}_x$ are coefficient matrices, whereas ζ, ϵ, δ are vectors of error uncorrelated with the variables. It can easily be seen how to define a certain linear model by imposing some constraints on the coefficient matrices. The constraints imposed on a structural equation model correspond to the "causal" relationships; for example, a null coefficient means that no causal relationship can be assumed between two variables.

4 How to use SEM to Understand IR Systems

The first step of the procedure to understand how a retrieval system decides about retrieval and ranking is the collection of the data of the manifest variables. The manifest variables that should be collected at this step are related to two main conceptual entities of the search process, i.e. documents and users; Figure 3 provides a generic illustration of the relationships between the variables of a structural equation model and the components of a retrieval system under scrutiny.

The documents such as web pages are the container of information delivered to the end user by the retrieval system. The documents are mainly a source of exogenous variables, since they are input data for the retrieval system, which is not allowed nor is it requested to update the document content. The data that is observed from documents may be

- structured data such as time and location, e.g. the Uniform Resource Locator (URL) and the embedded metadata,
- semi-structured data such as logical organization using titles, sections and paragraphs, and
- unstructured data such as keywords, which are measured in terms of Term Frequency $\times$ Inverse Document Frequency (TFIDF) and Best Match No 25 (BM25) weights.

The amount and the quality of the document features depend on the degree to which the applications performing causal analysis are allowed to access the index(es). In the event the applications cannot access the indexes, all the document features can only be extracted from the search engine result pages as illustrated in [13]. Moreover, the effectiveness of document parsing may be crucial; for example, the document author's gender may be inferred, thus providing the data necessary to check whether the retrieval system is biasing the results according to gender. Document quality can be considered one of the latent variables that is associated to the documents and is a source of retrieval bias; see for example [2,18].

The users can be viewed as sources of streams of data rather than containers of data. They express their own information needs mainly as (streams of) queries, clicks and display or dwelling time intervals. The amount and the quality of the data that can be gathered from the users depend on the type of experimental setting prepared for the investigation of the cause-effect relationships; in the event of controlled experiments, the user can be selected and trained by the scientists and the data can be gathered in a laboratory environment; otherwise, the search engine query logs are the main source of data. The design of a user study can be a complex task depending on the aims and the available resources [9, 10]; for example, user profiles can be built and utilized for the purposes of the causal analysis to understand whether some user's features, such as gender, affect how he or she formulates queries and then how document retrieval can be affected. User intent can be considered one of the latent variables that can be associated to the users and that can be a source of retrieval bias. In this respect, some noticeable research in user simulation was carried out in Information Science [5].

The definition of a structural equation model is perhaps the most crucial step because it is the step when the analyst can add constraints to the structural equation model and, in this way, express the possible causes and effects under investigation. However, the use of SEM might be complicated. The highly supervised nature of SEM should be regarded as both a strength and a weakness. When specifying a structural equation model, an analyst imposes her own viewpoint on the mechanics of a retrieval system; the addition of one constraint or the removal of another constraint is definitely a subjective decision. The mechanical procedure of model fitting is nothing but a computational procedure providing

a measure of fit and the significance level thereof. As discussed in [3] and [11], the lack of rejection of a structural equation model for a given sample of data or sample covariance matrix cannot be regarded as the sign that the model *is* the true and only one – there might be other, equally acceptable structural equation models which might significantly be different from the tested model. Nevertheless, the supervision exercised by the analyst guarantees that the discovery of causal relationships is not completely entrusted to an automated system, which might in turn be affected by the bias which is supposed to affect the scrutinized retrieval system.

5 About Causality, SEM and IR

In this section, we discuss the relationships between interpretability, explainability and causality within the domain of utilization of SEM in IR. The main aim of the discussion of the relationships between interpretability, explainability and causality is to understand the way structural equation models may explain the principles that govern the mechanics of a retrieval system and eventually the reasons behind the production of a certain search engine result page. Interpretability, explainability and causality are three broad concepts which appear to be interrelated and, in some cases, largely overlapping; furthermore, there are many other related concepts yet they were already addressed in [12], for example, and we will not further address them in this paper. Despite being overlapped, we consider interpretability, explainability and causality as three distinct yet related notions.

In particular, cause-effect relationships cannot be explained without interpretation. In the context of IR, an interpreter of the mechanics of a retrieval system is necessary to explain the reasons behind the production of a certain search engine result page. Our argument that cause-effect relationships cannot be explained without interpretation rests on an meaning basis. Causality[1] is the relationship between two things where one thing makes the other happen as if there were a sort of physical action between these two things. On the other hand, interpretability[3] refers to a broker which is able to find an agreement between the two parties who are trading goods and services. The implicit assumption is that (1) each party is using its own language that cannot be understood by the other party and (2) the interpreter has the ability to perform a sort of translation. In the context of automated decision systems, the broker i.e. the interpreter is thus the agent which translates the model's language to the user's language. Therefore, we consider interpretation in the sense that the model might not at all be directly understandable and an interpreter is needed to make the model understandable for the user. Instead, explainability[4] refers to the ability to remove the folds in

[1] The root of "causality" is from the Latin *causa*, i.e. a thing[2] which should be regarded as fact or event.

[3] The root of "interpretability" is from the Latin *inter*, i.e. between, *pretĭum*, i.e. price, and *habĭle*, i.e. led by hand.

[4] The root of "explainability" is from the Latin *ex*, i.e. out of and *planus*, i.e. plain.

order to make the internal meaning and content explicit. The action of removing the folds, i.e. explaining the causes and the effects, can only be performed by an interpreter which is able to understand the languages of both parties. Indeed, an automated decision system such as a retrieval system cannot by assumption explain what caused it to produce a certain search result page; furthermore, the end user of a retrieval system cannot be asked to understand the causes because of the high complexity of the retrieval system.

We argue that a structural equation model may play the role of an interpreter between a retrieval system and the end user, thus making an explanation of the internal system's mechanics possible and explicit. A structural equation model may play the role of an interpreter because of the following reasons.

1. First, the model can organize latent and manifest variables as well as exogenous and endogenous variables within a network of paths, also known as path diagrams, making possible "causes and effects" easily readable. In a path diagram, each symbol has a well-defined meaning; ovals represent latent variables, boxes represent observed variables, and oriented edges represent the "causal" relationship between the variable at the base of the edge and the variable at the head of the edge [15–17]. It is the graphical and visual feature of path diagrams that make a structural equation model an effective means of explanation to human users.
2. Second, the graphical representation of a structural equation model has a dual representation consisting of a system of linear equations where
   (a) each node of the path diagram corresponds to a variable,
   (b) each edge between two nodes corresponds to the co-occurrence of the corresponding variables in one equation at least,
   (c) the weight of an edge corresponds to one coefficient of an equation, and
   (d) the direction of an edge varies due to the fact that the coefficient of a left-hand variable contributing to the right-hand variable of an equation would differ from the the coefficient of a left-hand variable contributing to the right-hand variable if the two variables were swapped.
3. Moreover, whether a variable is endogenous or exogenous can be induced and represented by the linear system since the dependent variables are endogenous whereas the independent variables are exogenous.
4. Finally, latent variables can be expressed with a structural equation model because each variable in a system of linear equations do not necessarily correspond to observable quantities; from a mathematical point of view, it is only required that they be defined over the real field.

Some misconceptions surrounding SEM might negatively affect the utilization of structural equation models to explain the mechanics of a retrieval system to the end user. In [4] the authors explained that SEM had often been underestimated or even misunderstood and tried to clarify the false beliefs and uncover the potential of structural equation models. The belief that structural equation models aim to establish causal relations from associations alone is one significant misconception with respect to the use of structural equation models to explain the mechanics of a retrieval system. In contrast, a structural equation model has a different

objective, since it provides the methods to test the hypothesis that a sample fits a structural equation model, the latter being the hypothetical explanation provided by the expert user *a priori*. As a consequence, SEM can never establish causal relations from associations alone because those relations are already encoded into the structural equation model postulated by the user.

The demystification of the role played by SEM to establish causal relations is important within the context of a retrieval system, since a structural equation model would be a means to explain how a retrieval system works and as a consequence it can be a means to understand if and the degree to which the system follows some fairness guidelines in delivering the search results to a certain query. If a structural equation model were the outcome of an automated causal discovery algorithm, the assessment of the degree to which the system follows some fairness guidelines in delivering the search results to a certain query would be moved from the level of retrieval and search to the level of causal discovery and thus recreated. The fact that SEM can only assess whether the observed data fit a certain structural equation model allows the end user to get control of the most delicate step of the process, which is understanding the causes of the production of a search result. Thus, only the mere fitting is left to the computational methods.

The notion of causality or causation can be another source of issues when using SEM in general and within IR, in particular. It is well known that correlation does not imply causation; for example, increase in weight does not cause increase in height although they are two correlated variables when measured in a population. Moreover, it is impossible for causality to be framed only within the situations in which manipulation alone can be considered as the sole source of cause, i.e. the statement "no causation without manipulation" [8] can hardly be taken for granted [4]. However, manipulation can play a role in the context of retrieval systems in more than one way:

- First, the end user can manipulate the structural equation model, and as a result, the possible causes and effects; indeed, the addition or removal of nodes or edges correspond to the process of imposing constraints on the system of linear equations and more importantly to stating that a non-zero coefficient means a possible cause-effect relation between two variables.
- Second, the manipulation is physically possible in case of a retrieval system. If a structural equation model fits a certain sample of data, it is possible to investigate the effects on the endogenous variables, such as the effects of the rank of a retrieved document upon the variations of the exogenous variables such as the frequency of a query term. In other words, if the end user observed that the rank of relevant documents improves because of the increase in the frequency of a query term, the retrieval model could be tailored to this observation and the retrieval system could change the retrieval score and as a result the rank of the documents matching the query term.

6 Future Directions

The size of the data processed to fit a structural equation model can be a significant issue for the future research. In the context of IR, the primary source of data consists of the search engine result page. Such a page implements a non-random sample, i.e. the sample cannot uniformly be drawn from the collection of the page crawled by the search engine. The sample cannot be random because the focus of the investigation of the cause-effect mechanisms underlying the performance of a retrieval system and its impact on the user's information needs in terms of bias can only be observed from the top-ranked retrieved pages, since the top-ranked hits will be the ones accessed by the users. The size of the data processed to fit a structural equation model is not an issue for computational problems; it is rather an issue for statistical reasons, since sample size affects model estimation and significance testing. The issue of size coupled with the fact that the top-ranked hits should be considered implies that the sampled pages are not equal; in particular, the top ten ranked pages which usually correspond to the displayed "blue links" are the most frequently accessed by the end user. How to consider these top-ten or top-twenty hits should be addressed in future work.

References

1. Baeza-Yates, R.: Bias on the web. Communication of the ACM **61**(6), 54–61 (2018)
2. Bendersky, M., Croft, W.B., Diao, Y.: Quality-biased ranking of web documents. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. pp. 95–104. WSDM '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/1935826.1935849, `http://doi.acm.org/10.1145/1935826.1935849`
3. Bollen, K.A.: Structural Equations with Latent Variables. Wiley (1989)
4. Bollen, K.A., Pearl, J.: Eight myths about causality and structural equation models. In: Morgan, S.L. (ed.) Handbook of causal analysis for social research, pp. 301–328. Springer (2013)
5. Borlund, P.: A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. Journal of Documentation **72**(3), 394–413 (2016). https://doi.org/10.1108/JD-06-2015-0068, `https://dblp.org/rec/journals/jd/Borlund16`
6. Cuzzocrea, A., Bonchi, F., Gunopulos, D.: CIKM 2018 co-located workshops summary. In: Proceedings of CIKM. pp. 2309–2311. CIKM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3269206.3274267, `http://doi.acm.org/10.1145/3269206.3274267`
7. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj` (2016)
8. Holland, P.W.: Statistics and causal inference. Journal of the American Statistical Association **81**(396), 945–960 (1986), `http://www.jstor.org/stable/2289064`

9. Kelly, D.: Measuring online information seeking context. Part 1: Background and method. Journal of the American Society in Information Science and Technology **57**(13), 1729–1739 (2006). https://doi.org/http://dx.doi.org/10.1002/asi.v57:13
10. Kelly, D.: Measuring online information seeking context. Part 2: Findings and discussion. Journal of the American Society in Information Science and Technology **57**(13), 1862–1874 (2006). https://doi.org/http://dx.doi.org/10.1002/asi.v57:14
11. Kline, R.B.: Principles and Practice of Structural Equation Modeling. The Guilford Press, fourth edn. (2015)
12. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30:31–30:57 (Jun 2018). https://doi.org/10.1145/3236386.3241340, http://doi.acm.org/10.1145/3236386.3241340
13. Melucci, M., Paggiaro, A.: Evaluation of information retrieval systems using structural equation modeling. Computer Science Review **31**, 1–98 (2019)
14. Olteanu, A., Garcia-Gathright, J., de Rijke, M., Ekstrand, M.D.: Workshop on fairness, accountability, confidentiality, transparency, and safety in information retrieval (facts-ir). In: Proceedings of SIGIR. pp. 1423–1425. SIGIR'19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331644, http://doi.acm.org/10.1145/3331184.3331644
15. Wright, S.: On the nature of size factors. Genetics pp. 367–374 (1918)
16. Wright, S.: Correlation and causation. Journal of Agricultural Research **20**, 557–585 (1921)
17. Wright, S.: The method of path coefficients. Annals of Mathematical Statistics **5**, 161–215 (1934)
18. Zhou, Y., Croft, W.B.: Document quality models for web ad hoc retrieval. In: Proceedings of CIKM. pp. 331–332. CIKM '05, ACM, New York, NY, USA (2005). https://doi.org/10.1145/1099554.1099652, http://doi.acm.org/10.1145/1099554.1099652