Can Structural Equation Models Interpret Search Systems?

Massimo Melucci
massimo.melucci@unipd.it
University of Padova
Padova, Italy

ABSTRACT

Interpretability refers to the model's ability to explain the decision to relate a exogenous variable to an endogenous variable. Current approaches to interpretability utilize some conceptual structures such as decision trees and clusters that may need to provide further explanation to the end user who cannot intervene in making these structures interpretable. In order to make a model interpretable, the expert user should be involved in the design of an interpreter and provide her/his expertise in terms of variables and relations thereof. In this paper, we propose Structural Equation Modeling (SEM) as an approach to defining models that are simpler and can provide an interpretation of the learning models. A structural equation model can organize latent and manifest variables as well as exogenous and endogenous variables within a network of paths making possible "causes and effects" explicit. In particular, we focussed on search systems and on the interpretability of the reasons leading such a system to retrieve and display a certain list of documents to the end user. We provided some examples of structural equation models that can be used to interpret search results.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Learning to rank**; *Query intent*; • **Computing methodologies → Machine learning**.

KEYWORDS

Interpretability; Fairness; Explainability

ACM Reference Format:
Massimo Melucci. 2019. Can Structural Equation Models Interpret Search Systems?. In *Proceedings of SIGIR 2019 Workshop on ExplainAble Recommendation and Search (EARS'19).* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

In Machine Learning, a model is designed to instruct a machine to autonomously and effectively assign exogenous or independent variables to endogenous or dependent variables. The assignment of exogenous variables to endogenous variables must fulfil some constraints and minimize a cost function. The learning model would result very complex since the end user is not expected to understand it.

Recently, one important problem that has required the attention of the designers of complex learning models is the difficulty in understanding the reasons that caused a model to decide in favor of a certain value of an endogenous variable on the basis of the observed exogenous variables. The learning model can combine many components to the extent that a human user can hardly understand the reasons that lead the model to a certain decision. The problem and the solutions thereof have become urgent since the advent of "black-box" learning models that hide the reasons underlying a decision from the end users and even from expert users; the publication of regulations in the US and in the EU has made the need even more urgent.

Interpretability refers to the model's ability to explain the decision to relate an exogenous variable to an endogenous variable. We use "interpretability" according to the meaning provided by its root, "interpret", which is from the Latin *inter*, i.e. between, and *pretĭum*, i.e. price, and refers to an agent that explains and translates the model's language to the user's language in order to reach an agreement between the two parties in the same way a trade agent aims to reach an agreement between seller and buyer. Therefore, we consider interpretation in the sense that the model might not at all be understandable and an interpreter is needed to make the model understandable for the user.

Current approaches to interpretability utilize some conceptual structures such as decision trees and clusters to provide some explanations about a learning model in terms of decision rules and membership assertions; for example, a decision tree can provide an explanation in terms of "if-then-else" rules and leverage the use of these rules in programming languages and to some extent in the natural language in the hopes of making the decisions taken by a learning model explicit and clearly understandable. One limitation of the current approaches may be due to the need of providing further explanation of the reasons that lead a tree, a classifier or a regressor to decide about an assignment between variables, thus only moving the problem from one model to another model. Another limitation may be caused by the poor expressivity of generic patterns such as trees and clusters since they are often mined by unsupervised algorithms without any interventions made by an expert of the domain in which the data are interpreted and the patterns are utilized.

Our approach to interpretability differs from the current approaches in that it is our opinion that the expert user should be involved in the design of an interpreter and provide her/his expertise in terms of variables and relations thereof. To the aim of providing an intelligible scheme that describes how a machine makes decisions, we leverage a large field of linear models which by definition consists of numerical variables related by simple weighted sums. In Statistics and in particular within Statistical Learning it is customary to adopt models that act as interpreters between a

EARS'19, July 25, 2019, Paris, France

natural, social or economic phenomenon and the human users of the model who are interested in understanding the phenomenon. Another reason that makes linear models attractive is the potential to display the mathematical expressions in terms of expressive path diagrams.

Our research interest is thus focussed on the utilization of highly expressive statistical models as interpreters that are simpler the learning models utilized within the "black-box" systems. In this paper, we propose Structural Equation Modeling (SEM) as an approach to defining models that are simpler and can provide an interpretation of the learning models. A structural equation model is a linear statistical model that is more general than traditional linear models because it incorporates both latent and manifest variables within complex and highly expressive path diagrams. By using SEM, the expert user is involved in interpretation since s/he is requested to provide a structural equation model which is then evaluated against a dataset observed from a phenomenon in order to test the appropriateness of the structural equation model's explanation of the phenomenon. Thus, two key elements of the utilization of SEM to intepret learning models are (1) the direct participation of the expert user and (2) the ability to hypothesize a structural equation model that explain the decisions made by a machine.

In this paper, we focussed on search systems and on the interpretability of the reasons leading such a system to retrieve and display a certain list of documents to the end user. A search system can be queried by the end user by means of natural language phrases or more simply "bag of words". The reasons that lead the system to retrieve a certain list of documents might not be clear to the user. A structural equation model may provide such an interpretation since it can relate variables in a way that these relationships make some "causes" of the main effect, i.e. ranking explicit. Therefore, the structural equation model is a sort of interpreter which may show how the interactions among variables can be an explanation of search results. The structural equation model may also be useful for the expert user such as a designer of the search system to analyze the failures and the successes of the search system, thus finding the way towards an improvement of the search system's effectiveness.

2 RELATED WORK

Interpretability was recently addressed in the survey of Guidotti et al [10], in two introductory articles published in the Communications of the ACM [17], [21] and in a paper describing an approach to Data Science [14]. The rise of interest in the topic was also fueled by the European Union regulations on algorithmic decision-making and on the "right to explanation" [9], especially in the light of the emergence of deep neural network-based learning which made automated learners "black-box" machines. We surveyed SEM in the context of Information Retrieval (IR) system evaluation in [20]; in this paper, our focus is on interpretability of search results when the system internals (e.g. document index, term statistics) are unavailable.

Interpretability was addressed within some data mining and machine learning tasks such as recommendation, classification, prediction and topic modeling. Despite the variety of approaches, our proposal differs from them since it suggests utilizing SEM.

Interpretability was investigated in the design of recommender systems to address different research questions: how to provide accurate recommendations while preserving interpretability [1], [16], [28]; whether visual interface may help interpret recommendations [7], [27]; which is the "best" matrix decomposition in terms of interpretability [13]; and whether reviews can improve interpretability of recommendations [26].

As regards classification some research proposals were made to provide interpretable classifiers, for example an evolutionary classifier based on a small set of interpretable "if-then-else" rules [4], [29] and an algorithm that can faithfully explain a classifier or a regressor [24].

Regarding prediction, some empirical studies that aimed to obtain interpretable and accurate predictors were reported in [5] whereas estimates of uncertainty regarding interpretable early prediction were addressed in [8]. How well predefined patterns such as decision tables and trees can interpret classifiers and predictors was reported in [12] and [22].

Interpretability has also been an issue addressed in topic modeling since topics are usually modeled as vectors without any labels; the assignment of a label to a topic is thus crucial in order to make it interpretable [18], [23]. The problem with topic modeling is that accuracy is sometimes unrelated to interpretability in the sense that "models which achieve better predictive perplexity often have less interpretable latent spaces" [6].

The use of graphical data structures to interpret the causal effects in experimental or observational studies was addressed in [2] who suggested a methodology based on the idea of cross-validation to measure the differences in treatment effects across subpopulations.

A research work reported in [11] is somehow related to our work since it proposed non-additive interactions within any set of variables. The utilization of hidden factors to boost interpretability was suggested in [30]. However, both papers do not suggest to use SEM and for this reason our proposal differs from them.

3 STRUCTURAL EQUATION MODELING

In this section, we provide some basic terms and notions about SEM; two compendia can be found in [3] and [15].

SEM refers to the complex of multivariate statistical methods aiming to specify, estimate and fit a system of linear equations to a dataset observed from a phenomenon. SEM consists of two main conceptual pillars:

- the data observed from a phenomenon are encoded as variables, and
- the variables are inter-related by linear equations.

In particular, variables can be either exogenous or endogenous and in parallel they can be either manifest or latent, thus yielding four types of variable. An exogenous variable takes values from outside the model, i.e. it cannot be determined by other variables of the equations; in contrast, an endogenous variable can be determined within the model. A manifest variable can directly be observed whereas a latent variable cannot. For example, intelligence can be an endogenous latent variable whereas the number of questions asked to measure text comprehensibility and the number of mathematical problems solved to measure numerical skill can be exogenous manifest variables. Moreover, income can be an endogenous manifest variable whereas attitude to entrepreneurship can be an exogenous latent variable.

A structural equation model can be specified in general terms as follows:

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

$$\left.\begin{aligned} y &= \Lambda_y\,\eta + \epsilon \\ x &= \Lambda_x\,\xi + \delta \end{aligned}\right\} \tag{2}$$

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. $\mathbf{B}, \mathbf{\Gamma}, \Lambda_y, \Lambda_x$ 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. One of the simplest instances of a structural equation model is a linear regression model like $y = a + bx + \epsilon$ where x, y are two variables, a, b are two real coefficients and ϵ is the error. However, a structural equation model may comprise many equations and variables of different kinds.

SEM has two main steps: estimation and specification. Estimation, which is also known as identification, consists of computing the entries of the coefficient matrices in order to reproduce the covariances between the manifest variables; therefore, the m numbers placed on the diagonal and in the upper triangle of the covariance matrix are the sufficient statistics used to estimate the coefficients by means of methods such as Maximum Likelihood Estimation (MLE). The differences between actual correlation and estimated correlations are called "residuals". Specification consists of deciding on the "shape" of the coefficient matrices, thus deciding how complex the structural equation model is to estimate. Specification is a crucial step because the number of coefficients determines the identifiability of a structural equation model; a necessary condition to make a structural equation model identifiable is that the number of coefficients must not be greater than m.

4 AN INTERPRETABILITY FRAMEWORK

In this section, we introduce an interpretability framework based on SEM. We called the framework "Two Levels / Two Steps (2L2S)" because it is based on the two main types of variables that can be found in a structural equation model and on the two main steps of SEM. In particular, on the one hand, "2L" refers to the utilization of two types of variable: latent variables and manifest variables; it is within 2L that we have both the latent model Eq. 1 and the measurement model Eq. 2 of a structural equation model. SEM is quite a complex methodology comprising a number of interrelated steps. On the other hand, "2S" refers to the two steps of the structural equation model employed in this interpretability framework: the first step consists of specification, which aims to specify a structural equation model within 2L, whereas the second step consists of estimation, which aims to evaluate the structural equation model.

The main idea underlying the 2L2S interpretability framework is that a structural equation model is an interpreter providing an explanation of the internal mechanics of a system that yields the data under observation. The 2L2S interpretability framework is thus designed to provide the expert user with the methods which are sufficient to define and validate the structural equation model used as an interpreter. The 2L2S interpretability framework is depicted in Figure 1.

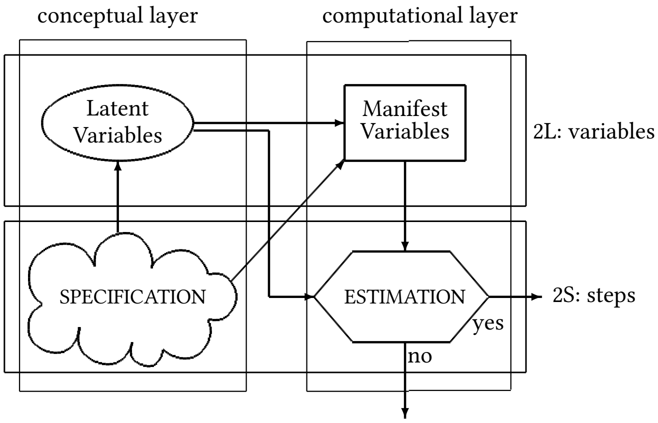

Figure 1: A pictorial representation of the 2L2S interpretability framework

At the conceptual level, latent variables and specification have been placed. The conceptual level refers to the intellectual activity performed by the expert user to define the latent variables that can explain the manifest variables and thus provide a means for interpreting what was observed altogether[1].

At the conceptual level, the issues of validity and reliability exist. A manifest variable is a valid indicator of the concept which the measured latent variable refers to when the manifest variable measures what it is supposed to measure. For example, a query term frequency is a valid indicator of relevance which aboutness refers to when query term frequency measures the degree to which a document is about the user's information need. The support for the utilization of query term frequency was developed by means of a long series of experiments which, on the one hand, proved a correlation between aboutness and frequency; on the other hand, the link between aboutness and relevance was highly debated [25] and it was more recently enriched by further dimensions of relevance [19], which would suggest the existence of additional latent variables such as user intent and document quality.

A manifest variable is a reliable indicator of the concept which the measured latent variable refers to when the repeated observation consistently produces the same values of the manifest variable. For example, a query term frequency can be considered as a reliable indicator because the number of occurrences of the query term in a document yields the same value whenever it is performed. In contrast, the relevance assessment provided by a user in terms of numerical score or ordinal labels cannot be considered as a reliable indicator because the user may be subject to a number of both exogenous and endogenous factors, which might induce the user to change his/her own mind regarding the actual relevance of the informative content of a document to her/his information need.

[1] "Conceptual" comes from the Latin word *concept*, meaning "conceived" which comes from *concipere*, which in turn comes from *com* ("together") and from *capere* ("take").

At the computational level, manifest variables and estimation have been placed. The computational level refers to the computations performed by an automated system to fit the structural equation model designed by the expert user at the conceptual level to the manifest variables; at this level, there is no expert user's interventions. The expert user is expected to receive an outcome from the estimation step in terms of error measures, fit indexes and probability of significance upon which s/he can decide whether her/his interpretation can be accepted with respect to the data.

The arrow between the latent variables and the manifest variables represents the relationship between the two main types of variable which is defined within a structural equation model, that is, the manifest variables result from the latent variables. The arrows between specification and the variables mean that this step aims to define a structural equation model and as a consequence select the variables and the relationships between the variables which may provide an interpretation to the observed data. The arrows between the variables and estimation mean that this step tries to fit the observed data to the structural equation model defined for interpreting these data. The estimation step ends the process with a binary outcome, i.e. either the observed data fits to the structural equation model or it does not. In the former case, the structural equation model may be viewed as an interpretation of the internal mechanics of the system that generated the observed data.

Consider the general structural equation model defined in terms of the latent model Eq. 1 and the measurement model Eq. 2. A constraint on the coefficient matrix would allow the expert user to design a different type of structural equation model and therefore represent a different hypothesis about the reasons that affect search results. Suppose there are three latent variables referring to relevance, user's intent and document quality. The expert who is trying to understand the reasons of the search results may want to hypothesize that intent and quality are determining relevance. Under this hypothesis, relevance is an endogenous variable whereas intent and quality are exogenous variables since the intent can only be formulated by the user and the quality can only be due to the document's author. In terms of Eq. 1, we have that η refers to relevance whereas ξ_1, ξ_2 refer to intent and quality, and

$$\eta = \begin{pmatrix} \gamma_1 & \gamma_2 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \zeta$$

where $\mathbf{B} = 0$ since relevance cannot be self-related. However, if another endogenous variable such as aboutness were hypothesized by the structural equation model, one further relation could be added to obtain the following model:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & b_{1,2} \\ b_{2,1} & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

where η_1, η_2 refer to relevance and aboutness, respectively. As latent variables are unobservable by definition, some manifest variables are necessarily defined and linked to the latent variables. Suppose a two-term query has been sent to the system by a user; the query term frequencies can be computed for each retrieved document to obtain two exogenous variables x_1, x_2. The rank y_1 assigned to each document by the system can be considered as an endogenous variable which depends on query term frequency whereas the relevance assessment y_2 assigned to each document by the user can be considered as another endogenous variable which depends on relevance, intent and quality. As a consequence, the measurement model of the structural equation model can become as follows:

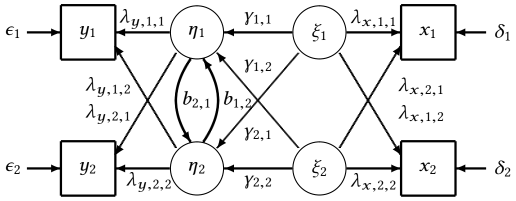

Figure 2: The path diagram of Eq. 3

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \lambda_{y,1,1} & \lambda_{y,1,2} \\ \lambda_{y,2,1} & \lambda_{y,2,2} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \lambda_{x,1,1} & \lambda_{x,1,2} \\ \lambda_{x,2,1} & \lambda_{x,2,2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \tag{3}$$

Interpretability can be ensured by the matrix coefficients, the value and the sign thereof, as well as by the direction of the relation between the variables due to the fact that the matrices may be asymmetric. A matrix coefficient of row i and column j measures the influence of variable j on variable i. When the coefficient c is negative, for each unit shift in variable j an exact c shift in variable i occurs. One strength of SEM is the possibility to depict a structural equation model by using path diagrams; for example, the structural equation model Eq. 3 is depicted as in Figure 2.

The specification of a structural equation model in the terms required by 2L2S may be limited by the constraints due to the necessity of selecting the variables and the relations thereof in advance. In light of such a limitation, an analysis performed within the 2L2S interpretability framework may be complemented by factor analysis where a simple structural equation model is constructed in advance by only setting the number of latent variables while the relationships with the manifest variables are automatically computed in an unsupervised way. A model resulting from factor analysis with r hidden factors and k manifest variables can be described as follows:

$$\xi_1 \longrightarrow x_1 + \cdots + x_k \qquad \cdots \qquad \xi_r \longrightarrow x_1 + \cdots + x_k$$

where the arrows denote a "directed" relation between the variables. The main limitation of factor analysis is that all the factors might be linked to every manifest variable although the weights assigned to the link between a factor and a manifest variable may result almost zero. Therefore, a less rigid approach to finding a structural equation model that explains the observed data may start with factor analysis, which avoids specifying the model that relates the latent to the manifest variables. Indeed, on the basis of the outcome of a factor analysis, the expert may select some manifest variables that are believed to be influenced by one latent variable before adding further manifest variables or latent variables. As an alternative, the expert may test the goodness of a structural equation model and then adopt an exploratory analysis in order to change the structural equation model and improve the fit. The main advantage of factor analysis is thereby unsupervision since the analysis computes some hidden factors without a predefined structural equation model. On the other hand, the factors suggested by factor analysis might not indicate a structural equation model which is effective in terms of goodness of fit; moreover, one factor computed by factor analysis consists of vectors of real weights measuring the degree of participation of each manifest variable in the factor, thus not providing the expert user with a binary answer to question as to whether to include the variable in a structural equation model.

5 EXAMPLES

This paper focusses on search systems and, therefore, the interest of our research was on the interpretability of search results. To the aim of the interpretability of search results, some examples were made as described in this section. The main aim of the examples described in this section is to provide some ideas about how to utilize structural equation models to interpret search results in practice. In order to explain the pratical utilization of structural equation models, a search scenario was chosen; in particular, in this section we considered the interaction between the user and the Search Engine Search Pages (SERPs), which can be obtained by means of a search engine that answers the user's queries. The examples reported in this section cannot be considered as an exhaustive evaluation as this is a task which will be addressed in the future work.

The search scenario that was chosen for the aims of this paper consists of, on the one hand, the index of a document collection and, on the other hand, a user who sends queries to the search system. The search system accesses the index and returns the SERP as a response to a certain query. A scenario is imagined where the user receives the SERP and tries to understand the reasons that made the system send the SERP. The user may want to assess the relevance of the returned documents; however, relevance assessments are relatively little of importance in these examples because the main aim of the examples was to ease the interpretability of search results rather than the effectiveness of the search system. It is important to recall that the user is assumed to be an expert who can formulate SEMs.

The index that was utilized in the examples reported in this paper was created from the GOV2 document corpus consisting of about 25 million documents. The full text of each document was indexed without special filters or schemas aiming to emphasize the importance of some document fields. The Text Retrieval Conference (TREC) 13, 14 and 15 topic sets were utilized as sources of queries; however, we could only refer to very few example queries, which were generated from the topic title fields. Stopwords were removed and no stemming was performed. As for SEM, we used the lavaan R library (lavaan 0.6-3 with MLE).

Consider an input query. The search system will retrieve and rank at most, say, $n = 300$ documents, where n is the sample size. Suppose the search system can display the most important snippets for each retrieved document to the user who may as a consequence be able to extract some features from each document by using an automated feature extraction tool. In the examples of this section, we considered the following document features: rank, document length[2], frequency of each query term within the document, average distance between two query term occurrences, and number of documents including two query terms. As fitting may be sensitive to large variances, a logarithm was applied to both rank and document length in order to reduce their variances; note that logarithm is a monotonic function which keeps order and therefore does not influence the association measures, thus making this transformation a common practice in statistics. The following structural equation model was then defined:

$$\text{rank} \leftarrow x_1 + \cdots + x_k \tag{4}$$

where x_i is the frequency of the i-th query term; the structural equation model Eq. 4 is basically a linear regression model. Suppose the user entered the query generated from topic number 701 ($k = 4$). Fitting the data with structural equation model Eq. 5 provides unsatisfactory results because R^2 is low and the regression coefficients are statistically not significant.

As our interest is in the latent model, the following structural equation model is instead utilized:

$$\xi_1 \rightarrow x_1 + \cdots + x_k \qquad \xi_1 \rightarrow \text{rank} \tag{5}$$

where ξ_1 is a latent variable who is supposed to be the common source of any manifest variable. We considered the model chi-square with its degrees of freedom and p-value; the Root Mean Squared Error (RMSE) and its 90% confidence interval; the Standardized Root Mean Square Residual (SRMR); and the correlation residuals. The analysis of the unexplained residuals of the correlations between the manifest variables indicated that the tested structural equation model under-predicts the correlation between x_1 and x_2; in this case, the hypothesis of no direct effect between x_1 and x_2 may be revised. As a consequence, we tested the following structural equation model:

$$\begin{array}{lll} \xi_1 \rightarrow x_1 + x_2 + x_3 + x_4 & x_2 \leftrightarrow x_3 & x_1 \leftrightarrow x_3 \\ x_3 \leftrightarrow x_4 & \xi_1 \rightarrow \text{rank} & \end{array} \tag{6}$$

The resulting residuals and, as a consequence, SRMR, all were almost zero. Moreover, the chi-square statistic values were less than their expected values and its p-value was about 0.19 although the sample was quite large[3], thus signalling that rejecting the hypothesis of good fit is a rather costly decision. Indeed, it should be noted that within SEM the researcher wants to not reject the null hypothesis of fit while within a traditional experimental setting the researcher aims to reject the null hypothesis of equality between two treatments with low probability of error. The RMSE calculated for Eq. 6 was 0.0 and its 90% confidence interval was $[0.00 - 0.07]$. Therefore, the chi-square statistic of this structural equation model will not reject the good fit hypothesis. Moreover, the p-value of the hypothesis that RMSE is not greater than 0.05 was 0.34, thus confirming that the hypothesis of good fit should not be rejected. Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were 1.0 and 0.96, respectively. These indexes aim to measure the departure from the baseline structural equation model; they assume covariances of zero between the endogenous variables which is a

[2] The document length may be calculated after downloading the page from the given URL.

[3] In hypothesis testing, the larger the sample is, the more frequently the null hypothesis is rejected.

rather unrealistic assumption. However, their utilization may further support the other evidence in favor of non-rejection of the null hypothesis.

The coefficients between ξ_1 and the manifest variables suggested that the former was positively correlated with rank and x_4 and was negatively correlated with x_3, thus suggesting the role of x_3 in promoting the documents to the top ranks and the role of x_4 in demoting the documents from the top ranks, as confirmed by the negative correlation between x_3 and x_4.

In general, a structural equation model might not fit all samples. For example, the data observed from the result document returned to the query of topic 707 could not be explained by Eq. 6 and another structural equation model was defined after analyzing the residuals; specifically, we considered

$$\begin{array}{lll} \xi_1 \longrightarrow x_1 + x_2 + x_3 + x_4 & x_1 \leftrightarrow x_2 & x_2 \leftrightarrow x_3 \\ \xi_1 \longrightarrow \text{rank} & \text{rank} \leftrightarrow x_2 & \text{rank} \leftrightarrow x_3 \end{array} \tag{7}$$

which exhibited a very good fit since all the residuals and SRMR were zero up to the third decimal digit. Moreover, the chi-square statistic values were much less than their expected value and its p-value was about 0.9 although the sample was quite large. Moreover, the RMSE calculated for Eq. 7 was 0.0 and its 90% confidence interval was $[0.00 - 0.08]$. The p-value of the hypothesis that RMSE is not greater than 0.05 was 0.92, thus confirming that the hypothesis of good fit should not be rejected. Both CFI and TLI were 1.0.

The event of query 708 was another example where the structural equation model fitting the data returned was inadequate. In this case, the solver could not find any solution for the structural equation models Eq. 5 and Eq. 7 and an incremental approach to finding the appropriate structural equation model was defined. Therefore, we considered

$$\xi_1 \longrightarrow x_1 + x_2 + x_3 \qquad \xi_1 \longrightarrow \text{rank} \tag{8}$$

and found that $x_1 \leftrightarrow x_3$ should be added, thus significantly improving the fit. As above, the chi-square statistic, its p-value and the evidence provided by RMSE and SRMR confirmed the decision not to reject the hypothesis of good fit. Nevertheless, the regression coefficients between the variables cast doubt on the reliability of the structural equation model. Indeed, for each unit shift of x_1, there would be a -3.5 shift of x_2, a -8.2 shift of x_3, and 9.5 shift of rank, thus suggesting that the improvement of the rank of a document was promoted by x_2 and x_3. However, these coefficients were not statistically significant since their p-values were relatively large, thus making a wide range of alternative coefficient values possible; for example, the probability that the coefficient between ξ_1 and rank can be greater than 9.5 or less than 9.5 is about 40%.

As another example, consider topic 708. The following structural equation model could well fit the top 100, 300 and 500 ranked documents since the p-value of the chi-square test statistic was about 0.34, the p-value of the hypothesis that RMSE was less than 0.05 was 0.43 and the SRMR was almost zero:

$$\xi_1 \longrightarrow \text{rank} + x_1 + x_3 + x_4 \qquad x_1 \leftrightarrow x_4$$

The coefficient of the model suggested that x_3 was the most important factor in determining rank where the other query terms were little significant and also inversely correlated as if they were occurring in different subsets of retrieved documents. We performed the same analysis yet only the relevant retrieved documents were added to the sample; the analysis was then repeated with only the non-relevant documents. We found that the aforementioned structural equation model exhibited a good fit with the relevant documents, but it did not with the non-relevant documents. This example suggests that the best structural equation model may be detected with no knowledge about the relevant documents, i.e. without prior knowledge other than the knowledge used to specify the model.

We also considered a very short query. Topic 731 consists of two terms. The following structural equation model could well fit the top 500 retrieved documents:

$$\xi_1 \longrightarrow \text{rank} + x_1 + x_2 \quad \xi_2 \longrightarrow d_{1,2} \quad x_1 \leftrightarrow d_{1,2}$$

The regression coefficient between rank and x_1 is about 1.3, i.e. the rank will increase by 1 unit for each 1.3 variation of x_1. Similarly the regression coefficient between rank and x_2 is about 30.3, i.e. the rank will increase by 1 unit for each 30 variation of x_2. These coefficients suggest that (1) the variation of rank for each unit shift of x_2 is much slower than the variation for each unit shift of x_1 and (2) both query terms demote documents and the documents that include few occurrences of the query terms are ranked above the documents that include many occurrences.

Further structural equation models can be defined to investigate the relationships among manifest variables and latent variables; for example, one may want to investigate whether the latent variables referring to the user's assessment of relevance and the system's assessment of relevance are inter-related and connected to the latent variables referring to document and query content. One model that can represent these connections may be expressed as follows:

$$\begin{array}{lll} \xi_1 \longrightarrow \eta_1 + \eta_2 & \xi_2 \longrightarrow \eta_1 + \eta_2 & \eta_1 \leftrightarrow \eta_2 \\ \xi_1 \longrightarrow x_1 + \cdots + x_k & \xi_2 \longrightarrow c_{1,2} + \cdots + c_{k-1,k} & \\ \eta_1 \longrightarrow \text{rank} & \eta_2 \longrightarrow \text{qrel} & \end{array} \tag{9}$$

where η_1 and η_2 refer to the user's assessment of relevance and the system's assessment of relevance, respectively. This example shows that the specification and the estimation of a structural equation model can be a difficult task because the model might be identified or estimated due to a number of methodological issues, which should be addressed when an interpretation of search results is made by using structural equation models. These issues are discussed in the next section.

6 DISCUSSION

The 2L2S framework for interpretability is based on two main steps (i.e. specification and estimation or identification) and two types of variable (i.e. manifest and latent) which may be either exogenous or endogenous. Therefore, the framework has the strengths and the weaknesses which are typical of SEM. These strengths and weaknesses are thus crucial to the exploitation of the framework for defining structural equation models in the role of interpreters.

The power of the use of SEM in interpretability is the tight connection with the task of the system that produced the results under scrutiny. Indeed, differently from general patterns such as decision trees, which are automatically discovered, a structural equation model is specific to the type of system – a structural equation model designed for one query will likely be inappropriate for another query. It is this peculiarity of SEM that makes specification a crucial step in an approach to interpretability based on SEM. As a consequence, the particular domain (e.g. IR) in which interpretability is investigated should be carefully considered before suggesting a structural equation model as an interpreter of a search result. In other words, the expert user should invest her/his expertise in the domain when defining structural equation models.

As stated above, the use of latent variables makes SEM different from other statistical methods such as regression, which only employs manifest variables. The use of latent variables have good potential which may turn into a problem if they are used without giving them an appropriate meaning. Indeed, the manifest variables can exhibit a clear meaning since they correspond to real concepts; for example, the values of term frequency is a clear measure of the occurrence of a term. In contrast, a latent variable such as relevance can become ambiguous and as a consequence useless if its meaning cannot be made clear and explicit; in the aforementioned situation, relevance should be specified according to the particular search task and should, for example, be redefined if the task refers to homepage finding or patent search to name just a few contexts of search. Besides, the redefinition of a latent variable such as relevance would help the expert user to link the variable to the most appropriate manifest variables.

Directionality refers to the distinction between exogenous variables and endogenous variables and to the different roles played by them – an exogenous variable directly or indirectly determines an endogenous variable and the viceversa cannot hold. For example, the system's assessment of relevance, which may be indicated by document rank, can determine the user's assessment of relevance, which may be indicated by a binary variable; the viceversa cannot hold. In other cases, the distinction might be less clear; for example, the latent variable indicated by term frequency may determine the latent variable indicated by the distance between term occurrences, but the viceversa can also hold.

Regarding the examples in Section 5, the specification of the best structural equation model was perhaps the most difficult step. Two reasons that made specification difficult are (i) the necessity that a structural equation model is identifiable and (ii) the tendency to add more latent variables than what can be "tolerated" by the available observations. Fortunately, identification can be checked by some conditions. We already mentioned the necessary condition, that is, the number of variances and covariances of the manifest variables (i.e. observations) must be greater than the number of "arcs incidents on" the variables (i.e. parameters). The so-called rank condition[4] is also sufficient, thus providing the expert with an easy-to-check condition that can be implemented by a straightforward algorithm before estimating the model [15].

Obtaining samples is relatively easy when interpreting search results since a search system can in general be queried to obtain any number of retrieved documents. The ease of obtaining samples is a significant advantage of the 2L2S interpretability framework. However, some attention should be paid to the selection of the sample units in the event of a list of retrieved documents since the sample will not be random, i.e. sampling cannot be repeated to obtain a representative of the document collection since the top ranked documents will always be retrieved. Furthermore, the list of retrieved documents should not be a random sample because the aim of 2L2S is to improve the interpretability of the search results and, in particular, the top ranked documents which will be delivered to the end user. From a methodological point of view, the lack of randomness and repeatability of sampling demonstrates the need for another concept of sampling, and the need to consider a list of retrieved documents as an exemplar of the universe of retrieved document lists that can be retrieved by the search systems answering a certain query yet within different contexts. As a consequence, the number of retrieved documents may be a parameter of a structural equation model in the sense that there might be different models for different retrieved document list sizes.

It is worth mentioning one word regarding feature extraction. From the examples in Section 5, it seems that a structural equation model can only handle ranking models that are built with manually crafted features, while the state-of-the-art search systems often use embedding features learned with neural networks. In principle, it would be very promising if we could develop an explanation model that can handle these latent features. However, it is the "black-box" nature of neural networks that makes structural equation models a useful approach to interpreting search results and makes the extraction of features from these results necessary and the only viable source of evidence.

A structural equation model appears to be only a linear equation system. However, in theory, it is possible to extend SEM to nonlinear models [3], but the question is whether the interpretability of the search systems can still be provided. In the relevant literature, the degree of interpretability contrasts with model complexity and it is believed that linear models are the best instruments in light of interpretability; one reason is that a linear equation can be accompanied by sentences like "one unit shift of x determines one unit shift of y" whereas non-linear models would require "non-linear" sentences.

As for the evaluation of the use of structural equation models to the aims of interpretability, further research is necessary to provide evaluation frameworks, metrics, search scenarios, and public datasets. Absolute fit indices or incremental fit indices are only a measure of fit of a structural equation model, but they little information about the user's overall satisfaction, the availability of user experts, the ability to scale up the model without requiring considerable human effort and how to interpret the model when the model gets complicated with more complex variables.

7 CONCLUSIONS

We reported on a preliminary investigation aiming to answer the following research question: "Can structural equation models interpret search results?". Scientific research requires a critical attitude especially when the answer to a research question is complex and cannot be reduced to a binary outcome. In this section we will make an effort to report all the major issues that deserve further research.

The utilization of SEM for interpretability purposes might seem like an oxymoron. On the one hand, the specification of a structural

[4] The rank condition (1) requires that the matrix of the graph connecting all the variables to the endogenous variables be reduced according to a given algorithm and (2) states that the rank of the reduced matrix must be greater than or equal to the number of endogenous variables minus 1.

equation model requires some knowledge about the mechanics of a search system or at least about the relationships between the manifest variables. On the other hand, the structural equation model should serve as an interpreter and provide knowledge about the aforementioned mechanics. Nevertheless, the apparent contradiction can be overcome because the structural equation model can, at least initially, be quite simple and can be further made complex according to the residuals; moreover, factor analysis can suggest possible latent variables, calculate the loadings between latent variables and manifest variables, and support the expert to find the right structural equation model.

Any structural equation model requires some interpretation and a great deal of attention should be paid to statistical significance of fitting. Indeed, when a structural equation model has a good fit with the observed data, the p-value will be large enough to let the expert state that the model should not be rejected. However, the good fit does not imply that the model is the sole interpretation of the search results; there might be other structural equation models exhibiting a comparable fit. To select the "right" model, power analysis[5] should be performed to compare different models. The presence of multiple structural equation models may make interpretability more difficult than in the event of one structural equation model; however, the degree to which these structural equation models overlap is still unclear and may, on the contrary, be a strength in that they may provide additional information to the interpretation as they provide different angles on the same subject.

The search for the structural equation model that can interpret search results requires the expert user's intellectual activity because of the presence of latent variables and the need to specify the relations between the variables – it is our opinion that the role played by the expert user is necessary. However, such an activity cannot be the only one since it may turn out to be tedious and prone to error, especially when the structural equation model is complex and there are many queries whose search results necessitate an interpretation. In this case, an algorithm that make the entire process automatic or at least requiring minimal interaction would be very useful. Future research will address the quest for such an algorithm that can go beyond the current software tools, which are primarily used to perform calculations such as estimations and optimizations.

In summary, this paper presents a preliminary study of the utilization of SEM to the purposes of interpretability of search systems. Further research is required to address some methodological issues. Despite being preliminary, structural equation models may provide an effective way to approach interpretability thanks to their intrinsic expressibility as shown in other research fields such as Economics and Psychology.

REFERENCES

[1] B. Abdollahi and O. Nasraoui. Using explainability for constrained matrix factorization. In *Proceedings of RecSys*, pages 79–83, New York, NY, USA, 2017. ACM.

[2] S. Athey and G. W. Imbens. Machine learning methods for estimating heterogeneous causal effects. https://arxiv.org/abs/1504.01132v1, 2015.

[3] K. A. Bollen. *Structural Equations with Latent Variables.* Wiley, 1989.

[4] A. Cano, A. Zafra, and S. Ventura. An interpretable classification rule mining algorithm. *Information Sciences*, 240:1 – 20, 2013.

[5] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of SIGKDD*, pages 1721–1730, New York, NY, USA, 2015. ACM.

[6] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of NIPS*, pages 288–296, USA, 2009. Curran Associates Inc.

[7] S. Cleger-Tamayo, J. M. Fernandez-Luna, and J. F. Huete. Explaining neighborhood-based recommendations. In *Proceedings of SIGIR*, pages 1063–1064, New York, NY, USA, 2012. ACM.

[8] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Utilizing temporal patterns for estimating uncertainty in interpretable early decision making. In *Proceedings of SIGKDD*, pages 402–411, New York, NY, USA, 2014. ACM.

[9] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". https://arxiv.org/abs/1606.08813v3, 2016.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, Aug. 2018.

[11] G. Hooker. Discovering additive structure in black box functions. In *Proceedings of SIGKDD*, pages 575–580, New York, NY, USA, 2004. ACM.

[12] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141 – 154, 2011.

[13] S. Hyvönen, P. Miettinen, and E. Terzi. Interpretable nonnegative matrix decompositions. In *Proceedings of SIGKDD*, KDD '08, pages 345–353, New York, NY, USA, 2008. ACM.

[14] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, Oct 2017.

[15] R. B. Kline. *Principles and Practice of Structural Equation Modeling.* The Guilford Press, fourth edition, 2015.

[16] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of SIGKDD*, pages 1675–1684, New York, NY, USA, 2016. ACM.

[17] Z. C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018.

[18] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of SIGKDD*, pages 490–499, New York, NY, USA, 2007. ACM.

[19] M. Melucci. *Contextual Search: A Computational Framework.* Foundations and Trends in Information Retrieval. Now Publishers, 2012.

[20] M. Melucci and A. Paggiaro. Evaluation of information retrieval systems using structural equation modeling. *Computer Science Review*, 31:1–98, 2019.

[21] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 2019.

[22] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić. What makes classification trees comprehensible? *Expert Systems with Applications*, 62:333 – 346, 2016.

[23] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of SIGKDD*, pages 457–465, New York, NY, USA, 2011. ACM.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of SIGKDD*, pages 1135–1144, New York, NY, USA, 2016. ACM.

[25] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[26] S. Seo, J. Huang, H. Yang, and Y. Liu. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of RecSys*, pages 297–305, New York, NY, USA, 2017. ACM.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. https://arxiv.org/abs/1312.6034, 2014.

[28] N. Wang, H. Wang, Y. Jia, and Y. Yin. Explainable recommendation via multi-task learning in opinionated text data. In *Proceedings of SIGIR*, pages 165–174, New York, NY, USA, 2018. ACM.

[29] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. A bayesian framework for learning rule sets for interpretable classification. *J. Mach. Learn. Res.*, 18(1):2357–2393, Jan. 2017.

[30] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of SIGIR*, pages 83–92, New York, NY, USA, 2014. ACM.

[5] Power is the probability of observing a sample under the hypothesis of non-fitting.