K-106

A File Selector Design Based on Bayesian network

W. Xie[1], H.Yoshida[1,2] and Y. Fujiwara [1,2]

1 Satellite Venture Business Laboratory, Kitami Institute of Technology,

2 Dept. of Computer Sciences, Kitami Institute of Technology,

165 Kouenchou, Kitami-shi, Hokkaido, 090-8507, Japan

Email: xiewei@mail.kitami-it.ac.jp

Abstract: In this paper, a file selector is constructed based on Bayesian network. In order to help users easily estimating use possibility of each file in the file directory, it offers users with the information about selected probability or popularity degree for each file and directory. Moreover, the order of files and directory is sorted in terms of this selected probability value.

Keywords: File selector, Bayesian network.

1. Introduction

Up to now, a file selector is only a directory viewer, allowing you to open files from a displayed directory in the current frame. It has been used widely in a variety of computer programming applications with respect to file operating. Always it offers information of file such as the name, the size and the time that file was created.

When we deal with a lot of files or search a requested file from a lot of files in a directory and the information of frequency that a file was used is necessary for users in order to presume use possibility for the future, the functions of traditional file selector are not satisfied for this aim. In this paper a file selector based on Bayesian network is proposed in order to realize above aim. This file selector is designed to supply users with information of the file's popularity degree. In other words, the selected probability that each file or directory will be used in the next selection is estimated. Moreover, the order of files and subdirectory are sorted in terms of this probability in the file selector.

As is well known that Bayesian network [1] is a graphical representation of uncertain knowledge that most people find easy to construct and interpret; it is also a network with the added property that the parents of each node are its direct causes. Furthermore, the representation has formal probabilistic semantics, which makes it suitable for statistical manipulation [2, 3]. Over the last decade, the Bayesian network has become a popular representation for encoding uncertain expert knowledge and data in expert system [4]. The techniques have been shown to be remarkably effective in some domains [5]-[12].

In this paper, we deal with relationship of directory and files represented as a Bayesian network. It has two important advantages. One, we can easily encode knowledge of directory and files in a Bayesian network and use this knowledge to increase efficiency and accuracy of interpretation and learning. Two, nodes and arcs in learned Bayesian networks often correspond to causal relationships. Consequently, we can easily interpret and understand the knowledge encoded in the representation to construct such a file selector that supplies the users with information of the file's popularity degree. When a certain event (a file is opened or used) occurs in the file selector, the occurring probability of other events (other files are selected) will be shown directly for users. Using this statistic information concerning file and file selector, users can easily estimate which one of files in a directory is the most popular. Based on this representation a simple algorithm is proposed to compute the selected probability for each file and file selector.

This paper is organized as follows, first Baye's Theorem is overviewed, and then factor analysis about the relationship of the file selector and files is represented with Bayesian network. In terms of the factor analysis, a simple algorithm is proposed. Finally, an example using Java language programming demonstrates construction of a two-layer file selector.

2. Bayesian network of File Selector

In this section, before we discuss the Bayesian network among directory and files, it is helpful to review the Bayesian interpretation of probability.

Baye's Theorem: Let (Ω, F, p) be a probability space, where Ω be a finite set of sample points and F is a set of events relative to Ω. $\{E_1, E_2, \cdots, E_n\}$ is a set of mutually exclusive and exhaustive events in F such that for $i = 1, \cdots, n$, $p(E_i) > 0$,where $p(E_i)$ is the probability of E_i according to the classical definition of probability. Then for any events B such that $p(B) > 0$, we have that for $i = 1, \cdots, n$

$$p(E_i / B) = \frac{p(B/E_i)p(E_i)}{\sum_{j=1}^{n} p(B/E_j)p(E_j)}$$

Now, we represent the relationship of file selector and files with a Bayesian network as the following Fig.1.

$T\ (t,\bar{t})$ is the event of main directory with states selected and unselected.

$B_{mj}\ (b_{mj},\bar{b}_{mj})$ is event of the jth directory with states selected and unselected on the mth layer.

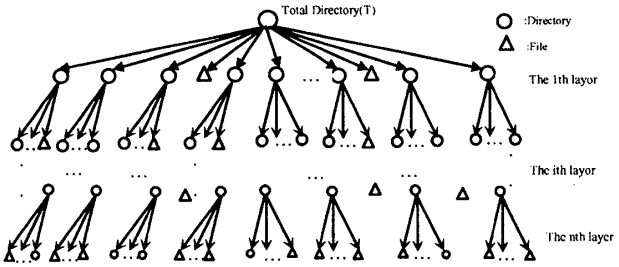

Fig.1. Factor analysis represented as a Bayesian network

$E_{mi}\ (e_{mi},\bar{e}_{mi})$ is event of the ith file with states selected and unselected on the mth layer.

In order to reflect a fact that the more a file was used, the selected probability is bigger; meanwhile the files in a directory are used more, the used probability of this directory will be higher. Thus the selected number for a file is only decided by the event that the file in a directory is used or not, and the selected number for a directory is dependent on sum of selected number of all files in this directory. First, we define the selected probability for the ith file on the mth layer as

$$p(e_{mi} = file\ i\ is\ selected / m) = \frac{num_{mi}}{sum_T}, i = 1,\cdots,l\ .$$

,where l is the total number of files on the mth layor,

$num_{mi} = selected\ number\ of\ the\ ith\ file$,

$sum_T = total\ selected\ number\ for\ all\ files$.

In addition, we also get

$$\sum_{m=1}^{n}\sum_{i=1}^{l} num_{mi} = sum_T \quad \text{for main directory } T\ .$$

Based on the definition of selected probability for file, the selected probability for the jth directory on the mth layor is discussed as follows: there are two cases to be considered,

One, this directory includes no sub-directory but files, since all of files E_{mjk} belong to the file selector B_{mj}, when a file in this directory is selected, it means that file selector B_{mj} is also selected. Namely, we have

$$p(b_{mj} / e_{mjk}, m) = 1, \qquad k = 1,\cdots,mje\ .$$

Using Baye's Theorem,

$$p(b_{mj} = jth\ directory\ is\ selected / m)$$
$$= \sum_{k=1}^{mje} p(b_{mj} / e_{mjk}, m)\cdot p(e_{mik} / m+1)$$
$$= \sum_{k=1}^{mje} p_e(kth\ file\ is\ selected\ \ in\ the\ jth\ directory / m+1)$$

,where mje denotes the total number of files in the jth directory on the (m+1)th layor.

The other, this directory includes not only sub-directory but also files. the selected probability for the jth directory on the mth layor can be derived as follows:

$$p(b_{mj} = jth\ directory\ is\ selected / m)$$
$$= \sum_{k=1}^{mje} p_e(kth\ file\ is\ selected\ \ in\ the\ jth\ directory / m+1)$$
$$+ \sum_{q=1}^{mjd} p_d(qth\ \ subdirectory\ is\ selected\ \ in\ the\ jth\ directory / m+1)$$
$$, j = 1,\cdots,g.$$

Here g is the total number of file selectors on the mth layor, mjd denotes the total number of subdirectory in the jth directory on the (m+1)th layer. In this case the computation of selected probability for subdirectory can be referred to case one.

Here it is an n-layer network with directed graphical models. This directed graphical model compactly represents the probability distribution related with file selectors and files. When a certain event (a file is opened or used) occurs, the occurring probability of other events (other files are selected) will be represented directly for users. Using this statistic information concerning file and file selector, users can easily estimate which one of files in a directory is the most popular. That is, the use frequency for each file and directory is shown for users. Then according to above Bayesian network analysis, a simple algorithm is proposed to compute the probability for both file and directory in the main directory.

Procedure of algorithm:

Initialization

Set each file in the main directory T with the identical value, that is, $p(e_{mi} / t) = 1 / sum_T,\quad i = 1,\cdots,l$.

Based on selected probability of each file, according to selected probability definition of file selector or directory, Computer selected probability value for each file selector; the order of file selector is sorted in terms of this value.

Updating

Update the selected numbers for each file and directory in main directory T, and compute the selected probability for all files and file selectors as

$$p(e_{mi} = file\ i\ is\ selected / m) = \frac{num_{mi}}{sum_T}, i = 1,\cdots,l\ .$$

$$p(b_{mj} = jth\ directory\ is\ selected / m)$$
$$= \sum_{k=1}^{mje} p_e(kth\ file\ is\ selected\ \ in\ the\ jth\ directory / m+1)$$
$$+ \sum_{q=1}^{mjd} p_d(qth\ \ subdirectory\ is\ selected\ \ in\ the\ jth\ directory / m+1)$$
$$j = 1,\cdots,g.$$

Sort the order of files in this directory and file selector in terms of the probability value. It is convenient for users to obtain the on-line information of frequency for files in a directory.

3. Example of Two-Layer File Selector

In this section, based on the Bayesian network of file selector, construction of file selector is shown. An example using Java language programming demonstrates construction of a simple two-layer file selector. This file selector is characteristic of file's name, created time, size, selected time and selected probability for the next selection. Thus it is very convenient for uses to master the knowledge about use frequency and probability for the next selection (or the popularity degree for each file) in the file selector.

Initialization

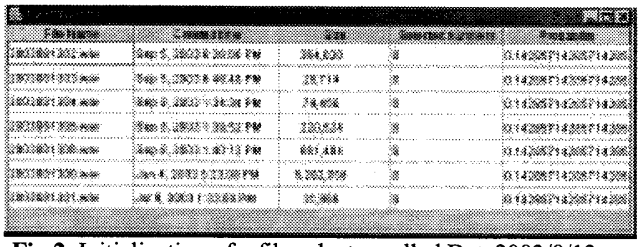

Fig.2. Initialization of a file selector called Data2003/8/13

From Fig.2, the file selector called Data2003/8/13 is made up of seven files. Then at the beginning, the selected numbers for each file in the directory is zero, however, the probability of each file is identical, that is, $p(e_i) = 0.143$.

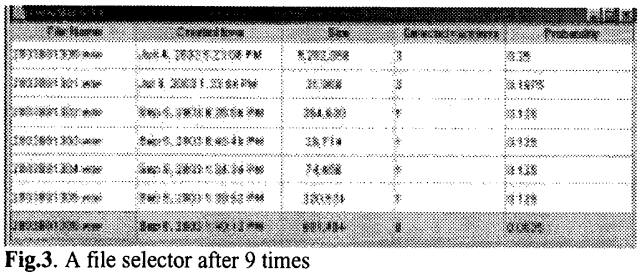

Fig.3. A file selector after 9 times

From Fig.3, it shows the change of selected numbers and probability of each file in the file selector. When a certain event (a file is opened or used) occurs, the selected probability of other files will be also represented on-line directly for users after nine times. Using this statistic information concerning file and file selector, users can easily estimate which one of files in a directory is the most popular. We also can find that the order of files is sorted according to the probability of each file.

The advantages that a file selector has this construction are as followed, with the increase of sum selected numbers in this file selector, it can inflect the popularity degree for each file more correctly, and meanwhile, it provides the knowledge about the selected probability for each file more clearly. Furthermore, in practice, it also provides users with an effective way of selecting a file under the condition of no other knowledge about these files.

4. Conclusions

In this paper, as an effective tool to represent uncertain knowledge, Bayesian network is used to represent the probability distribution related with file selectors and files. There are two important advantages. One, we can easily encode knowledge of directory and files in a Bayesian network and use this knowledge to increase efficiency and accuracy of interpretation and learning. Two, nodes and arcs in learned Bayesian networks correspond to causal relationships. In order to help users easily estimating use possibility of each file in the file directory, a type of file selector is modeled based on this Bayesian network. Furthermore, an example demonstrates construction of a two-layer file selector by using Java language programming.

References

[1] Pearl,J: Causality: Models, Reasoning, and Inference, NewYork: Cambridge University Press(2000)

[2] Howard,R. and Matheson,J: "Reading on the Principles and Applications of Desion Analysis", volume II, pp.721-761. Strategic Decisions Group, Menlo Park,CA(1981).

[3] Pearl,J: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo,CA(1988).

[4] Heckerman,D., Mamdani,A., and Wellman,M: "Real world applications of Baysian networks", Communications of the ACM,38.

[5] Cooper,G. and Herskovits,E: "A Bayesian method for the induction of probabilistic networks from data", *Machine Learning*,9,pp.309-347(1990).

[6] Alifers,C. and Cooper,G. "An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets", In proceedings of Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, pp.8-14.Morgan Kaufmann(1994).

[7] Fujiwara, Y., Matsuzawa, B. and Okada, S.: "Learning Assistance Expert System Based on Java Production System with a Self-Adaptive Function", International conference on Computers in Education, Volume I, pp158-159, Auckland, New Zealand (2002).

[8] Dean T., and Kanazawa, K., A model for reasoning about persistence and causation, Computational Intelligent, 5:142-150.

[9] Friedman, N., The Bayesian structural EM algorithm, in Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference, (G.F.Cooper and S.Moral,Eds.),san Mateo: Morgan Kaufmann, pp.129-138.

[10] Goldszmidt, M., and Pearl, J., 1996, Qualitative probabilities for default reasoning, belief revision, and causal modeling, Artificial Intelligence, 84:57-112.

[11] Neapolitan, R.E., Probabilistic Reasoning in Expert System Theory and Algorithm. John Wiley & Sons, 1990.

[12] Lauritzen, S.L., and Spiegelhalter, D.J., 1988, Local computations with probabilities on graphical structures and their application to expert systems(with discussion), Journal of the Royal Statistical Society, Series B, 50:157-224.