Exploiting Contextual Independencies in Web Search and User Profiling

C.J. Butz
Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: butz@cs.uregina.ca

Abstract - Several researchers have suggested that Bayesian networks be used in web search and user profiling. One advantage of this approach is that Bayesian networks are more general than the probabilistic models previously used in information retrieval. In practice, experimental results demonstrate the effectiveness the modern Bayesian network approach. On the other hand, since Bayesian networks are defined solely upon the notion of probabilistic conditional independence, these encouraging results do not take advantage of the more general probabilistic independencies recently proposed.

In this paper, we show how to exploit contextual independencies in both web search and user profiling. Whereas a conditional independence must hold over all contexts, a contextual independence need only hold for one particular context. For web search applications, it is shown how contextual independencies can be modeled using multiple Bayesian networks. We also point to a more general learning approach for user profiling applications.

I. INTRODUCTION

In practice, probabilistic inference would not be feasible without making independency assumptions. Directly specifying a joint probability distribution is not always possible as one would have to specify 2^n entries for a distribution over n binary variables. However, *Bayesian networks* [8], [16], [17] have become a basis for designing probabilistic expert systems as the *conditional independence* (CI) assumptions encoded in a Bayesian network allow for a joint distribution to be *indirectly* specified as a product of *conditional probability tables* (CPTs). More importantly, perhaps, this factorization can lead to computationally feasible inference in some applications. Thereby, Bayesian networks provide an elegant framework for the formal treatment of uncertainty.

One problem domain that involves reasoning under uncertainty is *information retrieval* [1], [5], [10], [13]. Several researchers have naturally suggested that Bayesian networks be applied in traditional information retrieval [7], [9], [12], [18], web search [11] and user profiling [15]. Although encouraging results have been reported, these works are based on the strict notion of probabilistic conditional independence. To the best of our knowledge, no study has tried to incorporate the more general *contextual independencies* [2], [14] for these purposes.

It is well-known that the notion of conditional independence is too restrictive to capture independencies that only hold in certain contexts. This kind of contextual independency was formalized as *context-specific independence* (CSI) by Boutilier et al. [2]. The important point is that Zhang and Poole [19] have empirically demonstrated that CSI can significantly speed up inference. At the same time, Wong and Butz [14] emphasized that CSI is a special case of a more general contextual independency called *contextual weak independence* (CWI).

In this paper, we show how to exploit contextual independencies in both web search and user profiling. It has already been shown how a single Bayesian network can be used to model information retrieval concepts [9] and web search concepts [11]. Unfortunately, contextual independencies cannot be captured by a single Bayesian network. Thus, we first show how contextual independencies can be *modeled* using multiple Bayesian networks. An alternative approach is to apply the method in [2] to *detect* the contextual independencies that hold in a given Bayesian network. Similarly, our earlier work on user profiling [15] was based on learning conditional independencies from data. Although some work has been done on learning contextual independencies [4], [6], these results have never been applied in an information retrieval setting. Moreover, these learning techniques only refer to the notion of CSI. Here we suggest that these contextual learning algorithms should be extended to the more general notion of CWI [14] for user profiling.

This paper is organized as follows. Section II contains a review of the pertinent notions of Bayesian networks and contextual independencies. In Section III, we show how to exploit contextual independencies in web search.

Similarly, we suggest ways to utilize contextual independencies in user profiling in Section IV. The conclusion is given in Section V.

II. BACKGROUND KNOWLEDGE

A. Bayesian Networks

Consider a finite set $U = \{A_1, A_2, \ldots, A_n\}$ of discrete random variables, where each variable $A \in U$ takes on values from a finite domain V_A. We may use capital letters, such as A, B, C, for variable names and lowercase letters a, b, c to denote specific values taken by those variables. Sets of variables will be denoted by capital letters such as X, Y, Z, and assignments of values to the variables in these sets (called configurations or tuples) will be denoted by lowercase letters x, y, z. We use V_X in the obvious way. We shall also use the short notation $p(a)$ for the probabilities $p(A = a)$, $a \in V_A$, and $p(z)$ for the set of variables $Z = \{A, B\} = AB$ meaning $p(Z = z) = p(A = a, B = b) = p(a, b)$, where $a \in V_A, b \in V_B$.

Let p be a *joint probability distribution* (jpd) [8] over the variables in U and X, Y, Z be subsets of U. We say Y and Z are *conditionally independent* given X, if given any $x \in V_X$, $y \in V_Y$, then for all $z \in V_Z$,

$$p(y \mid x, z) \quad = \quad p(y \mid x), \qquad \text{whenever } p(x, z) > 0. \quad (1)$$

We write Equation (1) as $p(Y \mid X, Z) = p(Y \mid X)$ for convenience.

Based on the *conditional independence* (CI) assumptions encoded in the Bayesian network in Figure 1, the jpd $p(A, B, C, D, E)$ can be factorized as

$$\begin{aligned} & p(A, B, C, D, E) \\ = \quad & p(A) \cdot p(B) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D). \quad (2) \end{aligned}$$

Using the CPTs $p(D|A, B)$ and $p(E|A, C, D)$ shown in Figure 2, we conclude this section with an example of probabilistic inference.

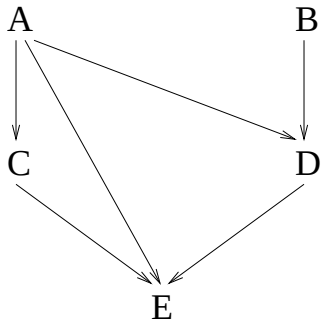

Fig. 1. A Bayesian network.

A	B	D	$p(D\|A, B)$
0	0	0	0.3
0	0	1	0.7
0	1	0	0.3
0	1	1	0.7
1	0	0	0.6
1	0	1	0.4
1	1	0	0.8
1	1	1	0.2

A	C	D	E	$p(E\|A, C, D)$
0	0	0	0	0.1
0	0	0	1	0.9
0	0	1	0	0.1
0	0	1	1	0.9
0	1	0	0	0.8
0	1	0	1	0.2
0	1	1	0	0.8
0	1	1	1	0.2
1	0	0	0	0.6
1	0	0	1	0.4
1	0	1	0	0.3
1	0	1	1	0.7
1	1	0	0	0.6
1	1	0	1	0.4
1	1	1	0	0.3
1	1	1	1	0.7

Fig. 2. The conditional probability tables $p(D|A, B)$ and $p(E|A, C, D)$ in Equation (2).

The distribution $p(A, B, C, E)$ can be computed from Equation (2) as

$$\begin{aligned} & p(A, B, C, E) \quad (3) \\ = \quad & \sum_D p(A, B, C, D, E) \\ = \quad & \sum_D p(A) \cdot p(B) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D) \\ = \quad & p(A) \cdot p(B) \cdot p(C|A) \cdot \sum_D p(D|A, B) \cdot p(E|A, C, D). \end{aligned}$$

Computing the product $p(D|A, B) \cdot p(E|A, C, D)$ of the two distributions in Figure 2 requires 32 multiplications. Marginalizing out variable D from this product requires 16 additions. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to obtain our desired distribution $p(A, B, C, E)$.

A	B	D	$p(D\|A,B)$
0	0	0	0.3
0	0	1	0.7
0	1	0	0.3
0	1	1	0.7
1	0	0	0.6
1	0	1	0.4
1	1	0	0.8
1	1	1	0.2

$\rightarrow$

A	D	$p(D\|A=0)$
0	0	0.3
0	1	0.7

Fig. 3. Variables D and B are conditionally independent in context $A = 0$.

B. Contextual Independencies

The Bayesian network factorization of $p(A, B, C, D, E)$ in Equation (2) only reflects conditional independencies $p(y|x, z) = p(y|x)$ which hold for *all* $x \in V_X$. In some situations, however, the conditional independence may only hold for certain *specific* values in V_X.

Consider again the CPT $p(D|A, B)$ in Figure 2. Although variables D and B are *not* conditionally independent given A, it can be seen in Figure 3 that D and B are independent in context $A = 0$, that is,

$$p(D = d|A = 0, B = b) \quad = \quad p(D = d|A = 0).$$

Similarly, for the CPT $p(E|A, C, D)$ in Figure 2, it can be seen in Figure 4 that variables E and D are independent given C in context $A = 0$, while variables E and C are independent given D in context $A = 1$, i.e.,

$$\begin{aligned} & p(E = e|A = 0, C = c, D = d) \\ = \quad & p(E = e|A = 0, C = c) \end{aligned}$$

and

$$\begin{aligned} & p(E = e|A = 1, C = c, D = d) \\ = \quad & p(E = e|A = 1, D = d). \end{aligned}$$

This kind of contextual independency was formalized as *context-specific independence* (CSI) by Boutilier et al. [2] as follows. Let X, Y, Z, C be pairwise disjoint subsets of U and $c \in V_C$. We say Y and Z are *conditionally independent* given X in *context* $C = c$, if

$$p(y \mid x, z, c) = p(y \mid x, c), \qquad \text{whenever } p(x, z, c) > 0.$$

III. WEB SEARCH

It has already been shown how a single Bayesian network can be used to model information retrieval concepts [9] and web search concepts [11]. Unfortunately, contextual independencies cannot be captured by a single Bayesian network. Thus, we begin our discussion on how to model contextual independencies using multiple Bayesian networks.

$p(E|A, C, D)$ $\nearrow$

A	C	E	$p(E\|A=0,C)$
0	0	0	0.1
0	0	1	0.9
0	1	0	0.8
0	1	1	0.2

$\searrow$

A	D	E	$p(E\|A=1,D)$
1	0	0	0.6
1	0	1	0.4
1	1	0	0.3
1	1	1	0.7

Fig. 4. Variables E and D are conditionally independent given C in context $A = 0$, while E and C are conditionally independent given D in context $A = 1$.

Recall the conditional probability table $p(E|A, C, D)$ in Figure 2. The parents of node E in the Bayesian network in Figure 1 are $\{A, C, D\}$. As already mentioned in the last section, variables E and D are independent given C in $p(E|A = 0, C, D)$, while E and C are independent given D in $p(E|A = 1, C, D)$. These contextual independencies can be represented by *two* Bayesian networks. One Bayesian network is constructed for $A = 0$ in which the parents of E are $\{A, C\}$. A second Bayesian network is constructed for $A = 1$ in which the parents of E are $\{A, D\}$. (See [14] for a more detailed discussion on modeling contextual independencies.) The other given CPTs can be examined for contextual independencies in a similar fashion using the detection algorithm in [2]. We now turn our attention from modeling to inference.

In order to utilize the above three context-specific independencies for more efficient probabilistic inference, Zhang and Poole [19] generalized the standard product operator $\cdot$ as the *union product* operator $\odot$. (It should be mentioned that Zhang and Poole [19] also pointed out that the notion of CSI can also be applied in the problem of constructing a Bayesian network [16].) The *union product* $p(Y, X) \odot q(X, Z)$ of functions $p(Y, X)$ and $q(X, Z)$ is the function $p(y, x) \odot q(x, z)$ on YXZ defined as

$$\begin{cases} p(y,x) \cdot q(x,z) & \text{if both } p(y,x) \text{ and } q(x,z) \text{ are def.} \\ p(y,x) & \text{if } p(y,x) \text{ is def. and } q(x,z) \text{ is undef.} \\ q(x,z) & \text{if } p(y,x) \text{ is undef. and } q(x,z) \text{ is def.} \\ \text{undefined} & \text{if both } p(y,x) \text{ and } q(x,z) \text{ are undef.} \end{cases}$$

Note that $\odot$ is commutative and associative [19].

The union product operator allows for a single CPT

to be horizontally partitioned into more than one CPT, which, in turn, exposes the contextual independencies. Returning to the factorization in Equation (2), the CPT $p(D|A,B)$ can be rewritten as

$$\begin{aligned} & p(D|A,B) \\ = \; & p(D|A=0,B) \odot p(D|A=1,B) \\ = \; & p(D|A=0) \odot p(D|A=1,B) \end{aligned} \tag{4}$$

while $p(E|A,C,D)$ is equivalently stated as

$$\begin{aligned} & p(E|A,C,D) \\ = \; & p(E|A=0,C,D) \odot p(E|A=1,C,D) \\ = \; & p(E|A=0,C) \odot p(E|A=1,D). \end{aligned} \tag{5}$$

By substituting Equations (4) and (5) into Equation (2), the factorization of the jpd $p(A,B,C,D,E)$ using CSI is

$$\begin{aligned} & p(A,B,C,D,E) \\ = \; & p(A) \cdot p(B) \cdot p(C|A) \odot p(D|A=0) \odot p(D|A=1,B) \\ & \odot \; p(E|A=0,C) \odot p(E|A=1,D). \end{aligned} \tag{6}$$

The use of CSI leads to more efficient probabilistic inference.

Computing $p(A,B,C,E)$ from Equation (6) involves

$$\begin{aligned} & p(A,B,C,E) \\ = \; & \sum_D p(A) \cdot p(B) \cdot p(C|A) \\ & \odot p(D|A=0) \odot p(D|A=1,B) \\ & \odot p(E|A=0,C) \odot p(E|A=1,D) \\ = \; & p(A) \cdot p(B) \cdot p(C|A) \odot p(E|A=0,C) \\ & \odot \sum_D p(D|A=0) \odot p(D|A=1,B) \odot p(E|A=1,D). \end{aligned} \tag{7}$$

Computing the union product $p(D|A=0) \odot p(D|A=1,B) \odot p(E|A=1,D)$ requires 8 multiplications. Next, 8 additions are required to marginalize out variable D. Eight more multiplications are required to compute the union product of the resulting distribution with $p(E|A=0,C)$. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to give $p(A,B,C,E)$.

The important point in this section is that computing $p(A,B,C,E)$ from the CSI factorization in Equation (7) required 16 fewer multiplications and 8 fewer additions compared to the respective number of computations needed to compute $p(A,B,C,E)$ from the CI factorization in Equation (3).

For simplicity, the discussion in this section focused on the contextual independence CSI. A more general contextual independence called *contextual weak independence* (CWI) was proposed in [14]. More importantly, we show in [3] how more efficient probabilistic inference can be achieved in a CWI approach using independencies that would go unnoticed in a CSI approach.

IV. USER PROFILING

In [15], we proposed a method for constructing a user profile using a Bayesian network. The input to our approach is a sample of documents that the user has marked as either relevant or nonrelevant. We can then learn a probabilistic network which encodes the user's preferences. Such a network provides a formal foundation for probabilistic inference. Documents can then be ranked according to the conditional probability defined by the network. Our approach has several advantages. The class of probability distributions represented in traditional approaches is a proper subset of the distributions represented in our approach. Our method can take full advantage of the established techniques already employed for uncertainty management in artificial intelligence. For instance, many algorithms exist for learning a Bayesian network from sample data. Moreover, efficient inference techniques exist for query processing in Bayesian networks.

For example [15], consider a user who receives numerous *electronic mail* (email) messages each day. This particular user is too busy to read all of the new email messages received each day. Thereby, she would prefer to read the most relevant email messages of the unread messages. For simplicity, let us assume in this example that every email message is represented by the same fixed set $\{A_1, A_2, \ldots, A_n\}$ of terms. Suppose further that there is available an auto-indexing program which will assign values $A_1 = a_1$, $A_2 = a_2$, ..., $A_n = a_n$ to each newly arrived email message. Given a sample of email messages that the user has marked as either relevant or nonrelevant, we can apply one of many algorithms to learn a Bayesian network. Whenever a new email message arrives, we rank it according to the following conditional probability defined by the Bayesian network:

$$p(Rel = relevant \mid A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n).$$

where *Rel* stands for *Relevance*. Based on the original probabilistic network, let $e_1, e_2, e_3, e_4, e_5, e_6$ be (the vector representations of) six new email messages with conditional probabilities:

$$\begin{aligned} p(Relevance = relevant \mid e_1) &= 0.5, \\ p(Relevance = relevant \mid e_2) &= 0.1, \\ p(Relevance = relevant \mid e_3) &= 0.7, \\ p(Relevance = relevant \mid e_4) &= 0.0, \end{aligned}$$

$$
\begin{aligned}
p(Relevance = relevant \mid e_5) &= 0.9, \\
p(Relevance = relevant \mid e_6) &= 0.3.
\end{aligned}
$$

Thereby, these new email messages would be ranked as $e_5, e_3, e_1, e_6, e_2, e_4$. Following this ranking, let us assume that the user has time to read e_5, e_3, e_1, which she ranks as *relevant*, *nonrelevant*, and *relevant*, respectively. Using these three new samples, the original probabilistic network can be *refined* in an offline mode. Suppose that the ranking, specified by the refined network, of the previously unread messages e_6, e_2, e_4 and the newly arrived messages e_7, e_8 is:

$$
\begin{aligned}
p(Relevance = relevant \mid e_8) &= 0.9, \\
p(Relevance = relevant \mid e_6) &= 0.7, \\
p(Relevance = relevant \mid e_2) &= 0.2, \\
p(Relevance = relevant \mid e_7) &= 0.1, \\
p(Relevance = relevant \mid e_4) &= 0.0.
\end{aligned}
$$

Notice that the refined network defines a different conditional probability for the messages e_6 and e_2.

We now suggest ways in which contextual independencies can be utilized for user profiling. The many conditional independence learning algorithms reported in [15] can simply be replaced by the contextual independence learning algorithms [4], [6]. On the other hand, these two learning algorithms only refer to the notion of CSI. As already mentioned, we have suggested CWI [14] as a more general contextual independence than CSI. Thus, there are two approaches to incorporating CWI for user profiling applications. Either extend, if possible, the learning methods [4], [6] to the notion of CWI or develop an original CWI learning algorithm. We are currently investigating these two options.

V. CONCLUSION

Bayesian networks provide an elegant framework for the formal treatment of uncertainty. Several researchers have suggested that Bayesian networks be applied in information retrieval [7], [9], [12], [18], web search [11] and user profiling [15]. Although encouraging results have been reported, these works are based on the strict notion of probabilistic conditional independence. To the best of our knowledge, no study has tried to incorporate the more general *contextual independencies* [2], [14] for these purposes.

In this paper, we suggested that contextual independencies should be incorporated into these retrieval problems. It was shown that contextual independencies can either be modeled using multiple Bayesian networks or detected in given conditional probability tables. More importantly, perhaps, we demonstrated how contextual independencies can be utilized for more efficient probabilistic inference than in a conditional independence approach. Our earlier work on user profiling [15] was based on learning conditional independencies from data. Although some work has been done on learning contextual independencies [4], [6], these results have never been applied in an information retrieval setting. Moreover, these learning techniques only refer to the notion of CSI. We suggested here that these contextual learning algorithms should be extended to the more general notion of CWI [14] for user profiling.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, New York, 1999.

[2] C. Boutilier, N. Friedman, M. Goldszmidt and D. Koller, "Context-specific independence in Bayesian networks," Twelfth Conference on Uncertainty in Artificial Intelligence, 115–123, 1996.

[3] C.J. Butz and M.J. Sanscartier, "On the Role of Contextual Weak Independence in Probabilistic Inference", submitted to the Fifteenth Canadian Conference on Artificial Intelligence, 2002.

[4] D.M. Chickering, D. Heckerman and C. Meek, "A Bayesian approach to learning Bayesian networks with local structure," Thirteenth Conference on Uncertainty in Artificial Intelligence, 80–89, 1997.

[5] S. Dominich, *Mathematical Foundations of Information Retrieval*, Kluwer Academic Publishers, Dordrecht, 2001.

[6] N. Friedman and M. Goldszmidt, "Learning Bayesian networks with local structure," Twelfth Conference on Uncertainty in Artificial Intelligence, 252–262, 1996.

[7] R. Fung and B. Del Favero, Applying Bayesian networks to information retrieval. *Communication of ACM*, **38**(3), 42-48,57, 1995.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Francisco, 1988.

[9] B. Ribeiro-Neto, I. Silva and R. Muntz, "Bayesian network models for information retrieval", Soft Computing in Information Retrieval: Techniques and Applications. F. Crestani and G. Pasi (Eds.), Springer Verlag, 259–291, 2000.

[10] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, New York, 1983.

[11] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura and N. Ziviani, Link-based content-based evidential information in a belief network model. *Twenty-third International Conference on Research and Development in Information Retrieval*, 96–103, 2000.

[12] H.R. Turtle and W.B. Croft, Inference networks for document retrieval. *Thirteenth International Conference on Research and Development in Information Retrieval*, 1-24, 1990.

[13] C.J. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.

[14] S.K.M. Wong and C.J. Butz, "Contextual weak independence in Bayesian networks," Fifteenth Conference on Uncertainty in Artificial Intelligence, 670–679, 1999.

[15] S.K.M. Wong and C.J. Butz, A Bayesian Approach to User Profiling in Information Retrieval. *Technology Letters*, **4**(1), 50-56, 2000.

[16] S.K.M. Wong and C.J. Butz, Constructing the dependency structure of a multi-agent probabilistic network. *IEEE Transactions on Knowledge and Data Engineering*, **13**(3), 395–415, 2001.

[17] S.K.M. Wong, C.J. Butz and D. Wu, On the implication problem for probabilistic conditional independency, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-A, **30**(6), 785–805, 2000.

[18] S.K.M. Wong and Y.Y. Yao, On modeling information retrieval with probabilistic inference, *ACM Transactions on Information Systems*, **13**, 38-68, 1995.

[19] N. Zhang and D. Poole, "On the role of context-specific independence in probabilistic inference," Sixteenth International Joint Conference on Artificial Intelligence, 1288–1293, 1999.