Antal P., De Moor B., Timmerman D., Meszaros T., Dobrowiecki T., "Domain knowledge based information retrieval language : an application of annotated bayesian networks in ovarian cancer", in *Proc. of Fifteenth IEEE Symposium on Computer-Based Medical Systems (CBMS2002)*, Marobor, Slovenia, Jun. 2002, pp. 213-218., Lirias number: 182113.

DOMAIN KNOWLEDGE BASED INFORMATION RETRIEVAL LANGUAGE: AN APPLICATION OF ANNOTATED BAYESIAN NETWORKS IN OVARIAN CANCER DOMAIN

P. Antal, B. De Moor
Dept. of Electrical Eng.
Katholieke Univ. Leuven
{antal, demoor}@esat.kuleuven.ac.be

D. Timmerman
Dept. of Obst.&Gyn.
University Hospitals Leuven
Dirk.Timmerman@uz.kuleuven.ac.be

T. Mészáros, T. Dobrowiecki
Dept. of M.&Inf. Syst.,
Budapest U. of Tech& Eco.
{meszaros, tade}@mit.bme.hu

Abstract

The increasing amount and variety of domain knowledge and the availability of increasingly large quantities of electronic literature requires new types of support for the development of complex knowledge models. In previous publications we proposed the application of so called Annotated Bayesian Networks (ABN), textually enriched probabilistic domain models, which help knowledge engineers and medical experts to find and organize the information necessary in model building. In this paper we describe an information retrieval language in which the formalized domain knowledge and the attached textual information can be accessed in an integrated fashion and can be used to define various retrieval schemes and relevance measures. This language, on one hand, provides maximum flexibility for knowledge engineers to exploit the available annotated domain model as contextual information. On the other hand, it allows the definition of complex, high-level queries, in which the contextual use of the annotated domain model can be optimized for clinical situations. We compare the performance of the standard and the proposed query language in the ovarian cancer domain.

1. Introduction

Bayesian networks have become a prevalent method to build probabilistic medical knowledge bases [1]. This modeling technique is capable of integrating both available background knowledge and observations. Furthermore it can serve as a universal probabilistic predictive model for decision support. We developed a Bayesian network model to assist clinicians discriminating between benign and malignant ovarian tumors [2]. To track the route of the incorporated prior background information (studies, papers, experts, etc.) to the Bayesian network model we investigated the use of annotation in the Bayesian network model (ABN) [3]. The annotations consist of natural language comments and references to various levels of the Bayesian network. Based on this concept, we implemented an integrated knowledge engineering and information retrieval system in which the knowledge model with the attached textual information can provide various contexts for the queries. During the evaluation of this system it became clear that traditional query languages used in information retrieval provide poor performance and no means of using domain knowledge and annotations already present in the ABN.

In this paper we introduce a conceptually new query language that allows the efficient inclusion of the domain knowledge and personal annotations to define relevance measures. In this way, the ABN model can provide flexible contextual information and help to establish better relevance measures for clinicians and knowledge modelers to specialize the queries.

We thank Karen Jermy (Dept. of Obstet&Gyn, St George's Hospital Med. School, Cranmer Terrace, London SW17 ORE) and Frand De Smet (Katholieke Univ. Leuven, ESAT/SCD) for their help.

2. Annotated Bayesian networks in the ovarian cancer domain

A Bayesian network represents a joint probability distribution over a set of variables [1]. The model consists of a directed graph and dependency models. For a particular domain, the vertices of the graph represent the domain variables and there is a dependency model for every vertex (i.e. for the corresponding variable) to describe its probabilistic dependency on the parent vertices (i.e. on the corresponding variables).

2.1. Annotated Bayesian networks

An extension of the Bayesian network representation of probability distributions, the Annotated Bayesian Networks (ABN), allows the definition of hierarchy of classes over the domain variables and the attachment of free text to the objects in the representation, such as values, variables, edges, dependency models and classes [3]. This ABN representation and a corresponding implemented system provided the formal framework and the experimental environment for our investigations.

2.2. The ovarian cancer domain

We perform the investigations in the ovarian cancer domain in the context of the International Ovarian Tumour Analysis (IOTA), which is a multicentre study on the pre-operative characterisation of ovarian tumours based on artificial intelligence models. This study includes the multicentre collection of patient data. The corresponding data collection protocols and study descriptions provide a consistent source of domain literature for our experiments. For example, the IOTA data collection protocol and a domain dictionary/glossary contain (1) general information about the domain and goal, (2) inclusion and exclusion criteria for patient records, (3) a description of each variable with its format, value list, mandatory/optional constraints, and possible inter-variable dependency rules, (4) the grouping of variables into classes and (5) diagnostic methods for all tumor variables together with figures [4,5].

2.3. Annotations of the domain variables

To ensure objective annotations, we used the IOTA protocols as primary sources. A corresponding thesis [6], the Merck Manual[1], the On-line Medical Dictionary[2] and the CancerNet Dictionary[3] provided an extension for the IOTA descriptions. Together, these compose the free text annotation, on average a hundred-word description for each of the domain variables.

In the experiments we used twenty-six domain variables, such as parity, drug treatment for infertility, use of oral contraceptives, family history of breast and ovarian cancer, age, bilaterality of the tumor, pain, descriptors of the morphology, echogenicity and vascularization of the mass, level of CA125 tumor marker, etc. [4,5].

3. An information retrieval language based on ABNs

Information retrieval (IR) deals with methods for indexing, searching, and recalling data, particularly text and other unstructured data forms [7,8]. Two major trends leading to an increased efficiency of the IR process are the utilization of user specific

[1] http://www.merck.com/pubs/mmanual
[2] http://www.graylab.ac.uk/omd/index.html
[3] http://thymoma.de/meddict.htm

information and domain specific information. The ABN can serve both as the user profile and as the domain ontology. To increase the efficiency of information retrieval, particularly in knowledge modeling and model building, a special language is proposed for the detailed characterization of necessary documents. It allows the definition of complex expressions based on the ABN (i.e. based on the domain model and the personal annotations), which are interpreted and evaluated as a relevance measure to select the matching documents by the information retrieval system.

3.1. Representations of annotations

According to our assumption for each domain variables a free text annotation exists, possibly containing references to related document collections. The algebraic representation called the vector space model encodes a document in a D-dimensional space where each component represents a corresponding word, neglecting the grammatical structure of the text. We used the Porter stemmer to canonize the words [9], and the essential domain specific phrases and synonyms are processed appropriately. Based on these, a domain vocabulary was constructed containing more than one million words (D=1135017), which is used to generate a controlled index. The weights in the vector model were computed using the tf-idf (term frequency – inverse document frequency) term weighting scheme [7].

3.2. An ABN-based IR language

The proposed language uses the following constructs:

1. *ABN-(entity)-set*: composed by the union, intersection, and difference of domain variables and classes. The properties of the Bayesian network and the ontology can be used in the set definitions (such as parents or Markov-blanket of a variable [1], descendants/ascendants of a class).
2. *Annotation-set:* composed of *Queries* and the annotations of *ABN-(entity)-sets.*
3. *Index-set*: the corresponding set of vector representations of the members of an *Annotation-set.* It may be a certain combination of the vector representations of members, such as the average, minimum or maximum value of the index weights. The result can be restricted to a *word-set.*
4. *Word-set*: the keywords of an *index-set* with average weights above a specified threshold or the union, intersection, and difference of *word-sets.*
5. *Relevance-measure*: a logical and arithmetic expression using operators such as '+ - * / ^ max min' and '< > = & |', and operates on the distances of the vector representation of a (matched) document from *index-sets* (various types of distance measures are possible, such as the minimum or average distance from the members).

An outline of a formal definition can be given as follows:

ABN-stochastic -variable:= [A_0|..| A_L]
ABN-ontology-class:= [C_0|..| C_L]
Query:= [Q_0|...| Q_K]
ABN-set:= [*ABN-stochastic -variable* | *ABN-ontology-class* |
($ABN\text{-}set_i \cup ABN\text{-}set_j$) | ($ABN\text{-}set_i \cap ABN\text{-}set_j$)|($ABN\text{-}set_i$ / $ABN\text{-}set_j$)
Parents(ABN-set)| *Children(ABN-set)*| *Markov_blanket(ABN-set)*|
Classes(ABN-set)| *ClassMembers(ABN-set)*|...]
Annotation-set:=[*A(ABN-set)* | *Query* | $Annotation\text{-}set_i \cup Annotation\text{-}set_j$]
Index-set:= I(*Annotation-set*$_i$ [*min*|*max*|*average*][/*Word-set*])

Word-set:= [Ω(*Index-set*$_i$, *scalar*)| (*Word-set*$_i \cup$ *Word-set*$_j$) |
(*Word-set*$_i \cup$ *Word-set*$_j$) | (*Word-set*$_i$ / *Word-set*$_j$)]
Relevance:= [Δ(*Index-set*$_i$ [*min*|*max*|*average*])| *scalar*| (*Relevance*$_i$+*Relevance*$_j$)|
(*Relevance*$_i$**Relevance*$_j$)| (*Relevance*$_i$^*Relevance*$_j$)|...]

3.4. A query in the ABN-based IR language

For example, to retrieve documents about a (1) variable "CA125 serum test" and (2) its influencing factors (3) in the ovarian cancer context (4) with emphasis on "Meigs' syndrome" the following expression can be formulated:

Min(Δ(I(A(CA125$\cup$ *Markov_b*(CA125)))), Δ(I(A(IOTA)))+Δ(I(Meigs' syndrome))

This relevance measure induce the selection of documents that are relevant both to the description of the 'IOTA class' and on average to the descriptions of related variables, additionally the presence of the keyword 'Meigs' syndrome' is preferred.

4. Results

Our ABN-based system, which integrates a knowledge engineering and an information retrieval subsystem was extended with the described IR language. To evaluate certain aspects of the ABN-based information retrieval language, two medical experts selected independently electronically available medical references most relevant in the IOTA context for the domain variables. They selected sixty-one separate references, which are attached in a non-exclusive way to the domain variables, on average three to five references for the eighteen variables that are covered in the model. This selection of the relevant documents for the domain variables and the annotation of the domain variables were performed by different experts and separated in time.

Additionally we asked the experts to select the most relevant (2), highly relevant (5) and relevant journals for the domain (93). We constructed three collections of abstracts containing 5367, 71845 and 378082 documents selected from the US National Library of Medicine's MEDLINE[4] corpus dated between January 1982 and November 2000. The existence of document collections and the selected relevant references make it possible to evaluate quantitatively the performance of the ABN based information retrieval.

We selected a typical task, the overall checking of the latest publications in a certain domain, context. In our case, it means the selection of relevant publications for each of the IOTA domain variables (i.e. for the variables in the correspondent ABN model).

A standard solution would be the checking of publications by a search engine like PubMed[4] about each of the variables with a query, which is partly variable specific and partly describe the IOTA context. This would include the main domain keywords and keywords about domain variables. Because of the required time and the difficulty of formulation of the queries, the potential of the automated use of domain knowledge and personal information is obvious.

The ABN model provides many options to perform this task with a single query, in which the relevance for the domain and the relevance for the domain variables are expressed. We used the following ABN-based relevance measure Δ^{ABN}=Δ(I(A(*Descendants*(IOTA))) in the experiments, which defines an average distance from all the domain variables and classes (the *'IOTA'* ABN class includes all the other classes and domain variables).

---

[4] http://www.ncbi.nlm.nih.gov/PubMed/

We compared the performance of the following queries: restricted keyword, standard keyword, domain description, domain keywords, an ABN-based and their combination. Fig. 1. shows the correspondent Area Under the ROC curve (AUC) values [10] for the three document collections and some selected ROC curves computed on the 'most relevant' document collection.

Restricted keywords ($\Delta^{R..keywords}$)= Δ(I(Ovarian, cancer))

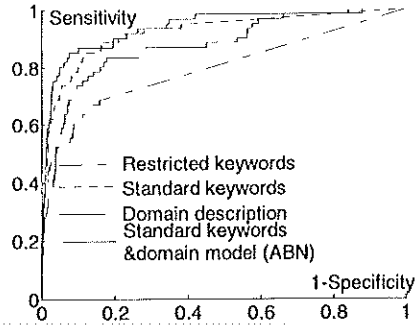
Standard keywords ($\Delta^{St.keywords}$)= Δ(I(Ovarian_cancer, preoperative, discrimination, benign, malignant, adnexal,mass))

Domain keywords ($\Delta^{D.keywords}$)= Δ(I(classification, score, factor, feature, combination, ultrasonography, ultrasound, morphology, vascularization tumor_marker, ca125, unilocular,multilocular,cyst, solid, papillary_projections, septum,locular, wall, echogenicity, age, menopausal, breast_cancer, family_history, personal_history, hormonal_therapy, oral_contraceptives, color_Doppler,color_score, blood_flow, pulsitality_index, resistance_index))

Domain description ($\Delta^{Description}$)=Δ(I(A(*IOTA*)))

ABN-based (Δ^{ABN})=Δ(I(A(*Descendants*(IOTA)))),

In the last two cases we used the ABN class '*IOTA*'. In the case of $\Delta^{Description}$, only its annotation is used, while in the definition of Δ^{ABN} it is indirectly used to refer to all the other classes and domain variables.



	AUC^{MR}	AUC^{HR}	AUC^{R}
$\Delta^{R..keywords}$	0.722	0.883	0.903
$\Delta^{St.keywords}$	0.921	0.964	0.966
$\Delta^{Description}$	0.874	0.972	0.973
Δ^{ABN}	0.913	0.979	0.980
$\Delta^{St.keywords}+\Delta^{D.keywords}$	0.903	0.974	0.975
$\Delta^{St.keywords}+\Delta^{Description}$	0.919	0.982	0.983
$\Delta^{St.keywords}+\Delta^{ABN}$	0.938	0.986	0.988

Figure 1. Left: ROC curves for some standard and a domain model (ABN) based information retrieval methods on the 'most relevant' document collections. Right: Area under the ROC curve (AUC) values for various IR methods on the collections of MEDLINE abstracts from the 'most relevant' (MR), 'highly relevant' (HR) and 'relevant' (R) journals.

5. Discussion

The primary purpose of the development of the local information retrieval system and databases to help knowledge engineers and medical experts in building complex statistical and knowledge models. The proposed ABN-based IR language allows the definition of complex, high-level queries by providing flexible access to the domain knowledge through ABNs, to the already organized domain literature in the annotations and to personal comments. Additionally, the information retrieval system is specialized for the ovarian cancer domain by using a domain specific vocabulary – incorporating also results from the IOTA terminology definitions [4] - and it applies an advanced indexing scheme (compared to the usual Boolean text representations used by most of the public search engines). However, the functionalities of this local, integrated knowledge engineering and information retrieval system could be accessed in general

through a currently planned web-based interface, providing possibly advanced access to appropriately linked public document collections. This raises the question of special queries for typical medical situations. The previously introduced ABN-based queries demonstrate the easy incorporation of a patient description or keywords from patient description in a 'patient-based' ABN query.

The quantitative evaluation has indicated that the ABN-based queries with multiple reference points offers new ways to express the interest, i.e. the region of relevance in the document space [7]. Additionally, the multiple reference points can be accessed and combined according to the domain.

6. Conclusion

The ABN-based information retrieval is only one application of linking a probabilistic, computational domain model with textual resources. The ABN representation can be useful for documenting the incorporated heterogeneous information in the network and it provides a framework for the investigations of how the textual knowledge can be used in the model building and identification. In the manual case investigated in this paper, the ABN serves as a modeler's context for information retrieval while building the model. In the automatic case under evaluation, the incorporated textual information supports the learning of the structure of the Bayesian network based on the data. The preliminary positive results here also suggest that the development of integrated text and data analysis methodologies, closely coupled representations and models for domain literature and statistical data are a promising challenge.

7. Acknowledgments

Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Peter Antal is a Research Assistants with the K.U.Leuven. Supported by grants from: Research Council KUL: Concerted Research Action GOA-Mefisto 666 (Math.Eng.), IDO (IOTA Oncology, Genetic networks), Flemish Gvt.: Fund for Scientific Research Flanders (projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02 research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary BIL2000/19), IWT (Soft4s), STWW-Genprom,GBOU-McKnow Eureka-Impact, Eureka-FLiTE,); Belgian Fed. Gvt: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006)), Program Sustainable Development PODO-II (CP-TR-18); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS; from National Fund for Scientific Research (OTKA) under contract number T030586; from National Fund for Scientific Research (OTKA) under contract number F-030763;

8. References

[1] J. Pearl: Probabilistic reasoning in intelligent systems, MIT Press, 1992

[2] P. Antal, D Timmerman, H. Verrelst, B. De Moor, S. Van Huffel: Bayesian networks in ovarian cancer diagnosis: potentials and limitations, in Proc. of the 13th IEEE Symp. on Comp.-Based Med. Syst., Houston, Texas, 2000, pp. 103-108

[3] P. Antal, T.Meszaros, B. De Moor, T.Dobrowiecki: Annotated Bayesian Networks: A tool to integrate textual and probabilistic medical knowledge", in Proc. of the 14th IEEE Symp. on Comp.-Based Med. Syst., Bethesda, MD, 2001, pp. 177-182

[4] D. Timmerman et al., Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol 2000; 16(5), pp. 500-505.

[5] IOTA Group., Official website of the IOTA project, http://www.iota-group.org

[6] D. Timmerman, Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology, Ph.D. dissertation, Leuven University Press, D/1997/1869/70

[7] R. Baeza-Yates, B.Ribeiro-Neto: Modern Information Retrieval, ACM Press, 1999

[8] R. Korfhage, Information Storage and Retrieval, New York: Wiley Computer Pub., 1997

[9] W.B.Frakes, Stemming algorithms, in W.B.Frakes et al., Information retrieval, Prentice Hall, 1992

[10] J. A. Hanley et al., The Meaning and Use of the Area under Receiver Operating Characteristic (ROC) curve, Radiology, 143, April 1982, pp. 29-36