Constructing Building Information Networks from Proprietary Documents and Product Model Data

R.J. Scherer & S.-E. Schapke
Institute for Construction Informatics, TU Dresden, Germany

ABSTRACT: The paper presents a novel *Building Information Mining Framework* (BIMF) that allows utilising building information captured in product model data as a valuable source of background knowledge in information retrieval and mining. Central to the framework is a *four layered Bayesian Network* adapted from probabilistic Information Retrieval models developed in the 90s. Capturing, combining and visualising the results of various text and model analyses as well as representing aspects of the current mining context, the network allows for explicitly representing content of the repository in personalisable information networks. These networks enable not only the retrieval of information from the text documents but also the explicit interlinking of the document and the product model domain to also support the understanding of the available interrelations and the exploration of new mining and integration strategies. The paper introduces the principal approach, explains the components of the basic network and suggests several further extensions that are currently still under development.

1 INTRODUCTION

The standardisation of product and process information has been a major focus for overcoming interoperability problems and enabling integrated networked collaboration. However, in the practice of project-centred, highly fragmented sectors such as the Architecture Engineering and Construction (AEC) industries information exchange still heavily relies on isolated text documents. Even with an increasing integration of model-based systems with project communication platforms, a large amount of the business and engineering knowledge will remain captured in large document repositories (Froese 2004). Hence, in addition to the innovative planning and modelling techniques, possibilities to retrieve project knowledge from traditional text documents and integrate it with operational model-based information systems need to be further explored.

The knowledge in today's project repositories is difficult to access due to the project-specific organisation of the documents and the complexity of their mostly unstructured text content. Document management systems and project extranets provide standardised metadata sets for labelling document files, but detailed document schemata or even ontologies for a more comprehensive classification of the information items are still missing. Furthermore, due to the highly interdisciplinary and often ad hoc work organisation in AEC, the comprehensive annotation of document information is very complex and has remained an unsolved task so far.

Furthermore, with the increasing use of model-based planning system there is a need for integrating document information with the related information models to achieve consistent information bases. Most business and product information models provide classes to reference documents, but the efficient interlinking among the numerous documents and related modelling objects remains a challenge. In order to integrate document with operational model-based information, methods are needed to automatically identify, externalise and track information from common text documents in relationship with available product data classes and instances.

We consider both the knowledge retrieval as well as the information integration a *context-specific information mining task*. This means that, first of all, a more detailed analysis of the document content is required to compensate the absence of suitable structure and semantics. Methods of computer linguistic, information retrieval and text mining provide for a first identification of text content e.g. in the course of full-text search, entity recognition or text-clustering. However, for flexible and sustainable information sharing among the involved disciplines, projects and business functions the working context in which the information was generated (and the one it is currently needed in) needs to be considered as well. Corresponding background knowledge is re-

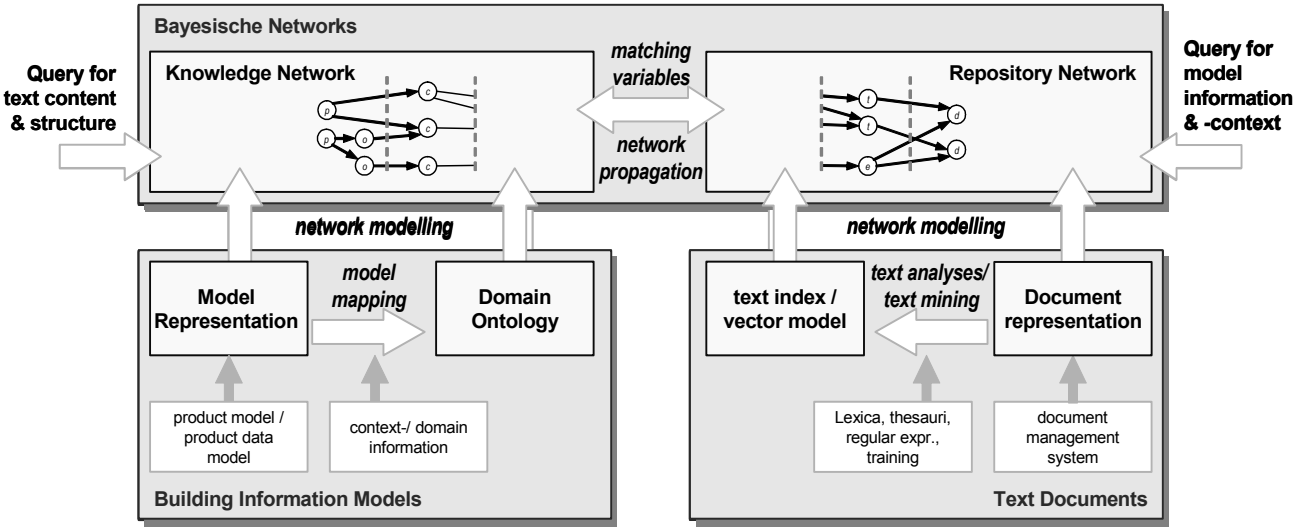

Figure 1: Analyses and Integration of Results in the Building Information Mining Framework (BIMF)

quired to take into account the varying structures, semantics, and granularity of the different disciplines' information models. This cannot be achieved solely by the currently available text analysis and mining techniques.

The goals of our research are to explore possibilities for accessing the AEC domain knowledge and project information stored in building information models (BIM) and utilise it in the processes of common text retrieval and mining. Standardised product data models such as IFC or ISO 10303-225 to a certain degree already enable the exchange of technical and functional descriptions of buildings and engineering structures, as well as the integration of heterogeneous software applications.

We argue that both the topology of the data models and the corresponding instantiated product models comprise general and specific AEC knowledge that can enable more focused, contextualised identification and reconfiguration of the document information. In contrast to specialised linguistic resources, building information models (1) do already represent existing, standardised models that are visualisable and shared among several AEC disciplines, (2) exhibit a notion of the AEC domain that correspond to the employed model-based systems, and (3) maintain continuously up-to-date context information on the project.

In our research, we explore the use of Bayesian Networks to represent both the results of statistical language processing and knowledge discovery as well as the deterministic model information. In the following sections we present the basic ideas, the architecture and the major components of the developed *Building Information Mining Framework* (BIMF) integrating the different components. More information on the developed framework is available in (Schapke & Scherer 2004) which provides also a detailed state-of-the-art analysis and further references.

2 RESEARCH APPROACH

The basic idea of the research is in the integration of existing analysis and modelling techniques into an overall *Building Information Mining Framework* that allows for bringing together different representations and methodologies from the separated document and model domain. This is achieved by means of a modular approach organising the necessary resources and the respective processing methods and tasks into three distinct, yet inter-related information spaces as depicted on figure 1. The integrating network enabling efficient utilisation of the individual analyses results is provided by the superordinate Bayesian Network space.

Various methods were reviewed and respectively selected for adoption to (1) externalise the content of text documents, (2) externalise the domain information from building information models, and (3) represent the knowledge on the text corpus and the application domain in a Bayesian Network. They are shortly reviewed below.

2.1 *Analyses of Document Repositories*

For the analysis of text information a variety of information retrieval and text mining technologies can be adopted. Most commonly a vector space model is used to represent content features and perform further analyses. However, while the different vector space models are usually limited to single representation schemes and an optimised number of document features, we pursue a more comprehensive representation to allow for a flexible interlinking with domain BIMs. Accordingly, our text analysis comprises:

– Normalisation, fragmentation and pre-processing of project documents to enable access to relevant text sections and more focused mapping to the model objects. Ready available text analysis tools such as converters and tokenisers can be adopted

for specific document domains (cf. Cunningham et al. 2002). An increasing utilisation of standardised document schemata will be important to allow for effective filtering, segmentation and content classification within the industry.

– Identification of specialised terms and phrases to enable a more concise representation of domain-specific content. Methods for entity recognition are available in several text applications but current lexica, thesauruses and expression bases are limited to general constructs such as names and addresses. More comprehensive domain-specific linguistic resources such as the LexiCon (Woestenenk 2002) need to be developed. Information extraction technologies may also provide for identifying complex information units when the document type can be determined through prior classification or schema information.
– Further text mining approaches such as text classification and clustering are increasingly explored in AEC (cf. Caldas et al. 2002, Froese 2004). We intend to implement some of these techniques within the Bayesian Network Model, when a connection between the model and the text information has been established and the available domain knowledge can be considered.

2.2 *Analyses of Building Information Models*

The goal of the model analyses is to automatically create discipline-specific domain representations from building information models that can be used to describe user and task-specific 'search contexts'. Extraction of the domain knowledge can be pursued by translating existing BIM data into discipline-specific ontologies. This provides a method for selecting only descriptive concepts and relations, as well as for reconfiguring the model to fulfil additional modelling constraints of the Bayesian network space. Furthermore, the model-based information can be supplemented with additional data about typical appearances of terms or phrases that can be associated with each concept. However, the success of this approach is strongly dependent upon related developments in the areas of model translation, model mapping, multi databases and ontologies for information platforms in AEC (Hyvärinen et al. 2004). Model transformations can adopt and extend existing specifications (e.g. EXPRESS-X, VML, CSML) and related tools, interfacing respective ontology construction software to provide the final 'mapped' results.

2.3 *Creating Bayesian Network Representations*

The goal of the investigations for Bayesian network modelling is the exact and comprehensive representation, interlinking and weighting of the different analysis results and further context information into an integrated Bayesian network (Pearl 1988, Baeza-Yates & Ribeiro-Neto 1999). The advantage of that network is that it can be used (to a certain extend) to represent both the results of numerical and statistical methods for language processing and knowledge discovery, and the results of deterministic model information and reasoning. To limit the complexity of the network the context information is modelled with binary variables. In this context, the probabilities of individual variables can be regarded a measure for the relevance (or importance) of the model objects, concepts or document feature represented by these variables. By manual instantiation of an individual variable the respective search context and need for information can be described and the relevance of further concept and document variables can be determined through network propagation.

Here, different types of Bayesian networks and network configurations can be explored. Taking into account the specific characteristics of the analysed information domains we suggest to first distinguish between a knowledge and a document network. The separate presentation of different sources allows for successive, straight-forward modelling of the individual knowledge domains, and the various influences on the overall information mining process, respectively. A matching analysis is then performed to interlink the two networks.

3 THE FOUR LAYERED BAYESIAN MINING NETWORK

The most critical issue in establishing the *Building Information Mining Framework* is the combination of the document and the model world in a superordinate information space. As stated above, we explore Bayesian networks to integrate the statistical text analysis and the deterministic model analysis. Our *Bayesian Mining Network,* adapting and extending probabilistic Information Retrieval models developed during the 90s (Baeza-Yates & Ribeiro-Neto 1999, de Campos et al. 2002, Schapke & Scherer 2004), combines the knowledge and the document network as depicted in figure 2.

On four separate layers the developed network represents (1) the knowledge on the building information models (product model layer), (2) the discipline-specific domain ontology representation (concept layer), (3) the contents of the text documents (descriptor layer), and (4) the overall document collection (document layer). The combined network can be used, in an evidential reasoning process, to reconfigure the collected information to most effectively support various retrieval or mining approaches, i.e. the available structure and context information is canonized and weighted for a subsequent combined analysis.

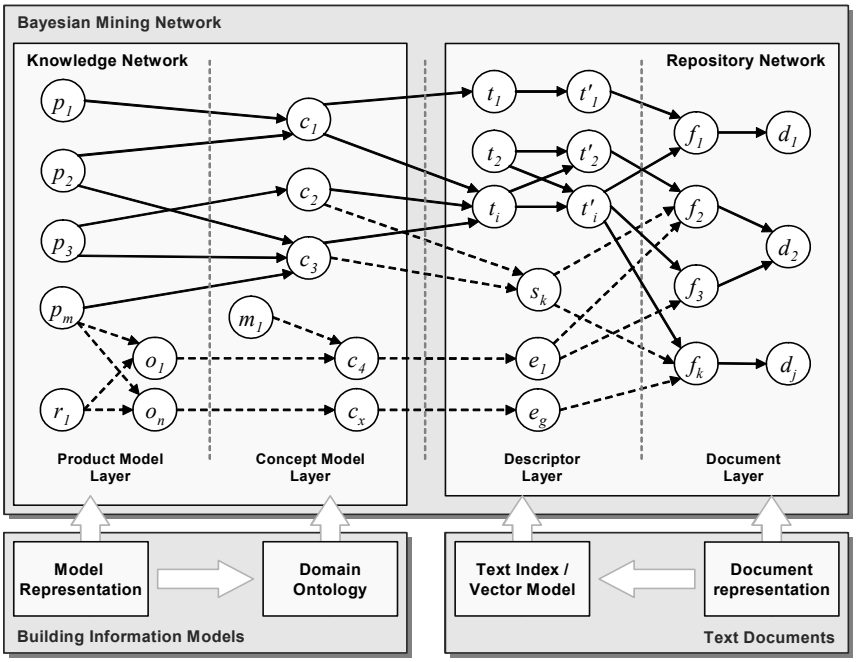

Figure 2: Analyses and Integration of Results in the Building Information Mining Framework (BIMF)

To validate the overall approach we have implemented a software suite named *dokmosis (Document and Knowledge Modelling Services)*. In the following subsections a first basic network configuration (depicted by the variables connected by solid line arcs in figure 2) that utilises selected text and model analyses is described in more detail.

3.1 *Constructing a Basic Repository Network*

The repository network is comprised of the document and the descriptor layer. In the basic version of the network it represents the knowledge on the document collection using document nodes d_j and fragment nodes f_k on the document layer, as well as descriptor nodes t_i on the descriptor layer.

The two layers are built using four text analysis modules of the *dokmosis* suite. Firstly, a collection module provides for importing documents and converting them to a common format based on the *DocBook* specification (see http://www.oasis-open.org). Secondly, a heuristic fragmentation algorithm is used to compile text paragraphs into equally large, self-contained text fragments; the resulting *part-of* relations are represented by arcs connecting the corresponding fragment and document variables. Thirdly, the fragment's text content is pre-processed, performing tokenisation, morphological analysis and stop-word removal. At last, by indexing the fragments a vector space model considering different term weighing can be built.

Based on the term weights the conditional probability distributions of the fragment variables are configured for exact Bayesian inference. Considering all index terms to be equally important we can assume a marginal probability distribution of $p(t_i=true)=1/M$ and $p(t_i=false)=1-1/M$, with M being the number of index terms for each concept node. For the possible value combinations of all $t_i \in d_j$ the conditional probabilities $p(f_k|t_1,\ldots,t_i)$ are computed by the sum of the respective normalized term weights (cf. Schapke & Scherer 2004). Thus, simulating a query Q by manually instantiating certain term variables, we obtain a posterior probability ranking of the variables that is equivalent to the fragments' ranking in classical information retrieval.

Furthermore, to increase the influence of the comparably few labelled concept nodes and, respectively, the recall of network-based information retrieval, term interdependencies such as term similarities and synonyms can be considered on the descriptor layer. To ensure a directed acyclic graph, the descriptor variables are duplicated and interconnected one-to-one. By inserting additional arcs and probability distributions the term similarities obtained from thesauri or by a term similarity analysis of the collection can be modelled.

3.2 *Constructing a Basic Knowledge Model Network*

The product model layer and the concept model layer together represent the configurable knowledge model used to trigger and control information retrieval and mining processes. In the basic version, the knowledge model network represents the user's knowledge via product model classes, denoted p_m, and engineering concepts c_x. The two layers are generated in the following consecutive analysis steps.

Firstly, the product model data to be considered is imported, transformed and represented on the product model layer. In the basic version the underlying knowledge model is restricted to a set of classes obtained from a product model server. For this purpose

the *dokmosis* suite integrates a client to the *VooDaMaS* product model server developed in the iCSS project (cf. iCSS 2002), which has been complemented with methods to identify both the classes defined in an EXPRESS schema and those used in corresponding instantiated product models. For each class in the returned result set, an independent variable is added to the product model layer.

Secondly, ontological information is used to derive a discipline specific concept model from the available product model information. For this purpose the *dokmosis* suite uses an adapted, somewhat simplified version of the *ontology interpreter* developed in the EU ISTforCE project which enables mapping of the model information to an engineering ontology (cf. Katranuschkov et al. 2003). To achieve an easy to process 'flattened' discipline-specific network, the ontological mapping specifications are confined to 1:1, 1:C, and C:1 mappings that allow for filtering or aggregating classes from the original result set. Hence, for the time being, only Boolean relations interlinking corresponding model and concept variables are represented in the knowledge model network, while independence is assumed among *all* nodes on the same layer.

Finally, the mapping specifications are supplemented by lexical descriptors to label each engineering concept with suitable terms. Thus, in the basic network, the concept nodes essentially represent the names of selected product model classes.

3.3 *Combining and Querying the Basic Repository and the Knowledge Model Networks*

To enable adequate reasoning on the overall mining network, the knowledge model and the repository network are interconnected, matching the concept labels with the descriptor nodes. Based on the concept mapping logical operation can be modelled with the conditional probability distributions.

Whilst the presented basic mining network is relatively easy to establish and process, it already provides for some new possibilities to formulate queries and express information needs compared to pure text mining approaches. In parallel to initiating a full text search on the descriptor layer, discipline specific concepts or model classes can be instantiated. Exemplary data models as well as common engineering classification schemes can be used to visualise the indices on the three top layers.

4 EXTENDING THE BASIC MINING NETWORK

The basic mining network can be beneficially extended in several ways. On each layer enhanced representation schemes and interdependencies among the variables can be identified to increase the expressiveness of the mining network.

4.1 *Extensions to the Repository Network*

To account for annotations and other structural information on the document's content, we propose to consider a second representation scheme depicted by the variables s_k on the descriptor layer. We expect the content meta information to provide for evidence on the characteristic syntax and semantics e.g. of domain-specific documents such as specifications, punch lists or protocols, as well as the user's respective domain of interest (cf. Caldas et al. 2002).

A third representation schema denoted e_g is used to represent named text entities and content objects embedded within the fragments. The text entities can be identified through previously assigned annotations as well as further content analysis.

Currently, the *dokmosis* suite integrates two entity recognition modules. First, the text analysis module provides for direct, regular expression based entity recognition of e.g. persons, organisations, addresses, codes and regulations, scales, formulas. Secondly, an information extraction module based on the *SpecEx* Extractor (Grimme 2003), that identifies information elements such as actors, tasks and responsibilities within functional, full-text work specifications. Manually labelled work specifications are used to train instance-based classifiers for automatic annotation of respective tokens and phrases. The on-going prototype implementation will provide a first indication of the potentials of respective extractors.

4.2 *Extensions to the Knowledge Network*

Corresponding to the *entity representation,* an *object representation scheme* is introduced on the product model layer. Differentiating between model classes and instantiated model objects, denoted by the nodes p_m and o_n, respectively, more detailed background knowledge on a given project can be provided. Distinctive concept nodes are added to the concept model layer for every class and object node. Their interrelations are recognised via the corresponding product model root node. It seems reasonable to first limit the extension to *instance-of* relations. Typical model relations can be represented on the model layer as illustrated by the variable r_1 in figure 2. However, more comprehensive analyses and information deductions are required to obtain meaningful relations from the product model information to truly enhance information retrieval and mining.

The main idea of the enhanced concept layer is to allow for personalised configurations of the applied background knowledge, without having to change the original model-based information. Additional context and user information can also be utilised to

re-label classes or alter discipline-specific concept views. To consider the influence of discipline-specific aspects without having to rebuild the concept model network, we introduce mental model nodes, denoted m_l, to allow for conditioning the relevance of each concept variable on distinctive mental models representing e.g. architectural, managerial or engineering domain views.

4.3 *Querying the Extended Mining Network*

The described additional representation schemes provide for numerous new ways to interconnect the variables of adjacent network layers. However, to limit the complexity of the network topology, we focus on two separated retrieval paths as illustrated in figure 2. The network based information retrieval of the basic mining network using product model classes is separated from the propagation of beliefs on respective instantiated product model objects. Furthermore, we assume the variables of different representation schemes on a layer to be independent among each other, even though possibly interrelated via variables of the preceding layers.

The explicit representation of modelling objects and text entities demonstrates very well the possibilities but also the challenges of directly interlinking product model and text information. Via the concept nodes product information can be explicitly connected with corresponding text elements to be automatically retrieved from the repository. The interlinking of the document and the model domain already supports collecting information elements such as punch list items or errors/omissions corresponding to certain building elements for subsequent information analysis. However, according to the various types of possible objects and entities, a set of similarity measures needs to be established to determine the probability that *'c-e'* node pairs really represent the same aspect. The previously applied ontology-based transformation to group, abstract or generalize model objects to meaningful concepts can greatly affect the possibilities to identify the 'best matches' among the concept and descriptor nodes.

5 CONCLUSIONS

Bayesian Network based information retrieval models have been identified as a very flexible technology that allows for representing various information resources and evidences to retrieve relevant information from document repositories. We argue that the presented approach provides a good basis to utilise appropriate background knowledge and additional context information in the processes of externalising information from respective AEC documents.

Due to the possibilities to encode the knowledge on the variables in terms of both causal relations and conditional probabilities, networks can be configured to support simultaneously logic operations and numerical mining techniques. This is an essential capacity allowing to interrelate the rather deterministic world of model-based systems with the rather fuzzy world of text and language processing.

By explicitly interlinking product model and document information we expect the mining network to support the understanding of available interrelations among the two domains, thereby revealing new retrieval, mining and integration strategies for more efficient and reliable information management.

REFERENCES

Baeza-Yates R. & Ribeiro-Neto B. 1999. Modern Information Retrieval. Wokingham, Addison-Wesley, UK.

Caldas C.H., Soibelman L., Han J. 2002. Automated classification of construction project documents. In: Journal of Computing in Civil Engineering, Vol. 16, No. 4, 2002.

Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.

DeCampos L.M., Fernández-Luna J.M., Huete J.F. 2002. A layered Bayesian Network Model for Document Retrieval. In: Proceedings of 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval, Glasgow, UK: pp.169-182.

Froese T. M. 2004: Integration of Product Models with Document-based Information. Proceedings of the European Conference on Product and Process Modeling, Istanbul, Turkey, September, 2004.

Grimme S. 2003. Untersuchung des Einsatzes von Methoden zur Informationsextraktion im Bauwesen am Beispiel der Angebotskalkulation. Diploma Thesis at the Institute for Construction Informatics, Technical University of Dresden, Germany.

Hyvärinen J. et al. 2004. InteliGrid – State of the Art and Market Watch Report, Deliverable D11, (c) InteliGrid Consortium c/o University of Ljubljana, www.InteliGrid.com.

iCSS 2002. Integrated Client-Server System for a Virtual Enterprise in the Building Industry - Project Description. available under: http://cib.bau.tu-dresden.de/icss/factsheet-en.html.

Katranuschkov P., Gehre A., Scherer R. J. 2003. An Ontology Framework to Access IFC Model Data, In: Electronic Journal of Information Technology in Construction, Vol. 8, Special Issue "eWork and eBusiness": pp. 413-437.

Pearl J. 1988. Probabilistic Reasoning in intelligent Systems: Networks of Plausible Inference. Morgan and Kaufmann, San Mateo, USA.

Schapke S., Scherer R. J. 2004: Interlinking Unstructured Text Information with Model-Based Project Data: An Approach to Product Model Based Information Mining. Proceedings of the European Conference on Product and Process Modeling, Istanbul, Turkey, September, 2004.

Woestenenk K. 2002. The LexiCon: Structuring Semantics, Proceedings of CIB W78 conference on Distributing Knowledge in Building, Aarhus, Denmark, June, 2002