A Four Layer Bayesian Network for Product Model Based Information Mining

S.-E. Schapke, Institute for Construction Informatics, Technical University of Dresden,
(Sven.Schapke@cib.bau.tu-dresden.de)

R. J. Scherer, Institute for Construction Informatics, Technical University of Dresden,
(Raimar.J.Scherer@cib.bau.tu-dresden.de)

Summary

Business and engineering knowledge in AEC/FM is captured mainly implicitly in project and corporate document repositories. Even with the increasing integration of model-based systems with project information spaces, a large percentage of the information exchange will further on rely on isolated and rather poorly structured text documents. In this paper we propose an approach enabling the use of product model data as a primary source of engineering knowledge to support information externalisation from relevant construction documents, to provide for domain-specific information retrieval, and to help in re-organising and re-contextualising documents in accordance to the user's discipline-specific tasks and information needs. Suggested is a retrieval and mining framework combining methods for analysing text documents, filtering product models and reasoning on Bayesian networks to explicitly represent the content of text repositories in personalisable semantic content networks. We describe the proposed basic network that can be realised on short-term using minimal product model information as well as various extensions towards a full-fledged added value integration of document-based and model-based information.

1 Introduction

In the construction industry, project and corporate document repositories provide the most comprehensive collections of business and engineering knowledge. Unfortunately, most of this knowledge is retained only implicitly in isolated and poorly structured text documents. Even with an increasing use of model-based systems to share design and construction information on a semantic level, a large percentage of the information exchange will rely on isolated and rather poorly structured documents. Current product model standards such as the IFCs provide classes to reference documents, but the efficient interlinking among the numerous documents and related modelling objects remains a challenging task. There is a need to explore methods to automatically externalise information from common text documents and to flexibly integrate this information with operational model based information systems.

During the last years, various new technologies for information retrieval, mining text corpora and generating corresponding information maps have been explored. These approaches have proven to work well by harmonised repositories or by domain-specific analysis tasks employing a great amount of background knowledge. However, for retrieving and reusing information from heterogeneous document repositories of highly multi-disciplinary domains such as AEC/FM, it is necessary to enable more flexible utilisation of appropriate background knowledge. The different mental models and current information needs of users from various disciplines need to be considered in the processes of information identification, analysis and evaluation.

Indeed, due to the diversity of operations in AEC/FM, model-based systems remain limited to selected business and engineering functions. To ensure vertical and horizontal communication they need to be complemented with basic file and document based management systems. A first step to enhance the management of traditionally textual encoded information of e.g. specifications, contracts, protocols, etc. has often been the introduction of forms and standardisation of document structures. In parallel to human evaluation and editing, the resulting semi-

structured and partially labelled documents to a certain extend provide for exchanging information among selected users and applications. In this context, standardisation initiatives, software developers and manufacturers of building systems have introduced several refined classification systems and schemata to structure the content of bidding, tendering and controlling documents (cf. http://www.gaeb.de, http://www.cite.org.uk). However, the integration of this information with related engineering and management applications is still limited due to the content's granularity and the degree to which the concepts of the rather problem specific schemata can be directly translated to new application areas.

In addition to the operational information exchange, different types of information requests need to be supported on document repositories of both current as well as previously finished projects for operational information and knowledge management. Furthermore, to effectively identify and access document information of a domain not directly related to the source document, available context information and various ontological views of AEC/FM have to be considered. To meet these requirements, several studies have extended the content schemata and exploring AEC/FM specific ontologies to provide for more flexible and detailed tagging based on common semantics, as e.g. in the EU-projects eCognos (Lima et al. 2002) and eConstruct (Tolman et al. 2001). On the long run, we expect these approaches to quite successfully allow for exchanging and versioning information among so called "intelligent" or "smart" documents and corresponding model-based systems within a single domain. However, developing expressive and consistent ontologies spanning multiple disciplines, providing for efficient information tagging as well as monitoring the numerous relations among documents and information models remains a challenging (if at all possible) task.

Because of these limitations of structure-based approaches, we see the need to complement the available technologies with context-based methods allowing to analyse a document's content in the absence of suitable structure and semantics and by rather fuzzy information requests. For that purpose, a flexible framework is required that supports both numerical and statistical methods of e.g. data mining and language processing as well as logic-based methods recognising the content structures and providing for integration with the semantic product models. We suggest a probabilistic information mining framework combining methods analysing text documents, filtering product models and reasoning on Bayesian networks to explicitly represent the content of text corpora in personalisable semantic content networks. The developed model extends and adapts Bayesian network retrieval models developed for classical information retrieval tasks during the 90s.

2 Approaches to Information Retrieval and Text Mining

In this section we shortly examine methods of Information Retrieval, Text Mining and Information Extraction to be supported by the mining framework. The discussed context-based approaches recognise text units as natural language elements to consider knowledge of lexicology, syntax and semantic in their analyses. Graphical content elements such as charts, figures and drawings are not in the scope of discussion.

Traditionally the discipline of Information Retrieval deals with the representation, storage, organisation and access to information items in text documents. The methods of Information Retrieval are usually based on a lexical view dealing with natural language sets, that are not always well structured and often semantically ambiguous (Baeza-Yates and Ribeiro-Neto 1999). They are optimised to robustly perform on large repositories of domain-independent, often multilingual documents. In today's document management systems the retrieval of relevant documents based on user queries is an important basic function. However, the methods rely on rather simple term-based formulations to represent document content and information needs, that are not sufficient for retrieving detailed or even distinctive information elements.

In recent years methods from Data Mining and Machine Learning have increasingly been applied to problems involving textual data to achieve better retrieval results and a higher degree of automatic adaptation to new application domains and knowledge areas (Hearst 1999, Nahm and Mooney 2002). Today, a large number of automatic document management tasks such as text categorisation, clustering, segmentation, summarisation and topic detection can be subsumed under the topic of *Text Mining* (cf. Grobelnik and Mladenic 2001). In contrast to Information Retrieval returning documents explicitly having been assigned index terms or keywords, these approaches focus on the discovery of implicit, previously unknown, and potentially useful information in unstructured and semi-structured texts. Depending on the mining task the texts need to be represented in simple vector space models or complete phrase structures of each sentence. The effective application of Text Mining algorithms in construction have been proven e.g. in (Kosovac et al. 2000), for the extraction of key phrases to construct a thesaurus for the roofing domain, as well as in (Caldas et al. 2002), providing for a hierarchical classification of weekly progress reports. Albeit limited to rather homogenous collections and only a few topics of interest, the results of these studies encourage to further explore mining methods and try to apply them to complete heterogeneous project repositories in practice.

Using similar text representations and methods, *Information Extraction* is often considered a part of Text Mining. However, while Text Mining is mainly concerned with the discovery of new knowledge, the goal of Information Extraction is to find and link relevant information in the text (while ignoring others) and transform it into detailed, structured databases (Neumann 2001, Nahm and Mooney 2002). Thus, Information Extraction mostly requires a more detailed representation of the text relying on additional lexical, morphological and syntactical information on terms, phrases and text compounds. Important tasks of Information Extraction are e.g. term extraction, entity recognition and ontology acquisition. Grimme (2003) has demonstrated the use of Information Extraction in construction by automatically extracting structured work specifications indicating e.g. objects, amounts, and regulations from unstructured construction contracts. Despite the relatively high computational cost and the current restrictions in application domains and document types, we expect Information Extraction to be useful in retrieving detailed information required for the direct integration with product models.

Common to all approaches is a rather flexible combination of linguistic pre-processing methods from Computational Linguistics and Natural Language Processing to build representation models of the texts, that are not necessarily theoretically sound but provide for optimal retrieval and mining results. On a document level a large variety of the retrieval and mining tasks can already be supported by a comprehensive vector based representation of the corpora. Thus, we expect to only build more detailed models when the type of a paragraph can be explicitly specified for certain Information Extraction tasks e.g. through its content structures. In these cases, the mining framework should allow for considering the extracted information in the respective vector space model (Nahm and Mooney 2002).

3 Bayesian Networks for Information Retrieval

In Information Retrieval probabilistic models from Salton's vector space model to conditional logic models by Rijsbergen have been used to evaluate a document's probability of relevance given a user's information need (Crestani et al. 1998, Baeza-Yates and Ribeiro-Neto 1999). The variables, their evaluation and methods to estimate a final probability of relevance in these models have been chosen based on different simplification assumptions to first of all achieve computational efficiency and satisfactory retrieval results but also to provide for a theoretically sound framework of information retrieval. Especially for the latter aspect Bayesian network models have been proposed that extend earlier probabilistic models by conditional dependencies in a real environment. These generalised probabilistic network models provide for a solid yet very flexible framework allowing to integrate several representation schemes and different sources of evidence.

Bayesian networks are *directed acyclic graphs* the nodes being random variables of a problem to be solved (here: finding relevant information) and the arcs the causal relationships among them (Pearl 1988). The graphs represent knowledge (a) qualitatively, showing the (in)dependencies among variables, and (b) quantitatively, expressing the strength of these dependencies by means of conditional probability distributions. For each node, the parents of that node reflect all its direct causes given by a respective set of conditional probability distributions. From these conditional distributions we can recover a downsized decomposition of the joint probability distribution over all nodes, due to the independencies graphically encoded in the graph.

Bayesian network based information retrieval models associate index terms, documents, user queries and/or user's information needs with random variables to model the retrieval task as an evidential reasoning process. In this context the observation of a document, index or query term is considered to be the cause for an increased belief in the respective variable that is further propagated throughout the network. The model most often consists of a document subnetwork representing the static knowledge on the overall text corpora and a query subnetwork that is build according to a user's query. However, according to their underlying causal model and simplification assumptions the networks differ significantly in their structure, causal ordering and probability distributions (Figure 1).

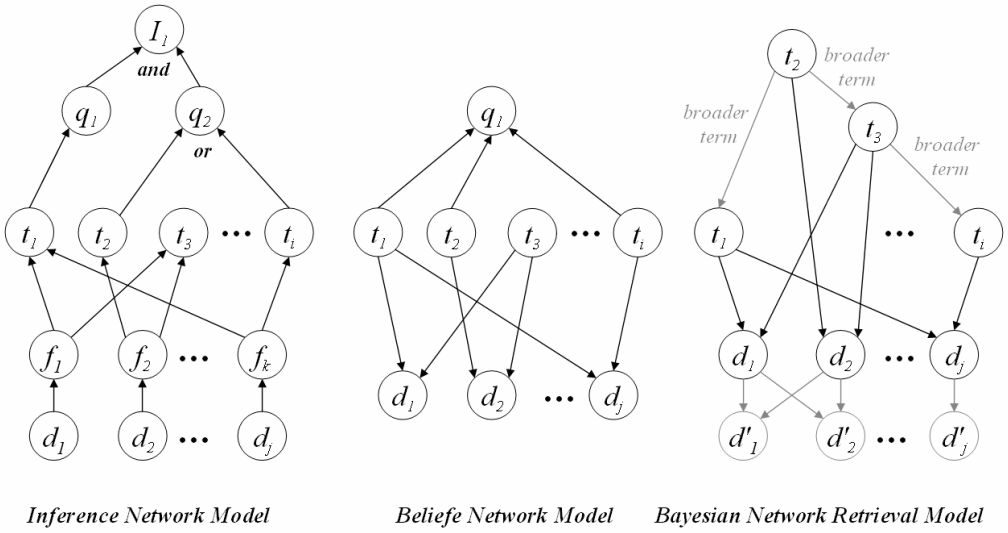

Figure 1: Models of Bayesian network based Information Retrieval

The first Bayesian network based information retrieval model called **Inference Network Model** depicted on the left side in Figure 1 was suggested by Turtle and Croft (Turtle and Croft 1991). Their document network is composed of document nodes d_j, text representation nodes f_k and concept representation nodes t_i corresponding to the events of having observed respective documents, text fragments/content elements contained therein and associated descriptive concept representations such as index terms or keywords. For simplification, the fragmentation of documents into elements such as text paragraphs, figures or multimedia objects is neglected in the following, so that document and text nodes correspond one-to-one (j=k). In this simplified network the documents can be indexed connecting the document nodes with related *concept representation nodes* whenever respective terms are observed in the text. While mostly the index terms themselves are used for representing the concepts, these nodes can also depict other representations such as manually assigned keywords, document type features or weighted index terms (for conciseness, in the following we refer to concept representation nodes as concept or term nodes, as appropriate). Conceptually this is a major advantage to the traditional information retrieval models in that it allows for considering multiple document representation schemes in parallel. The query network is repeatedly build for each query and consists of a

single node representing the user's information need and one or more query nodes to express this need. Besides standard keyword queries, it allows for alternative Boolean-like query formulations so that also multiple sources of evidence can be considered in parallel.

Turtle and Croft show that the probability distributions in the network can be quite gradually extended to represent **ranking strategies** from Boolean to probabilistic models. While the Inference Network Model gives no preference to the documents assigning a prior probability of 1/(collection size) to the document nodes, the conditional probabilities of the term nodes given its set of parent nodes can incorporate effects of indexing weights and/or term weights. For term and query nodes a canonical representation of the probability distributions is used. Corresponding closed-form expressions implementing Boolean operators (AND, OR, NOT) as well as weighted-sums for probabilistic retrieval are used to more efficiently compute the probability $p(I|d_j)$ that the user's information need is met given the document d_j as a function of static term weight probabilities $p(t_i|d_j)$. The retrieval is carried out on one-by-one basis, instantiating each document node d_j and then ranking the nodes according to the probabilities $p(I|d_j)$ obtained.

While the Inference Network Model ranks documents based on posterior probabilities obtained at query/information nodes, the following approaches assume the opposite causal ordering. We agree with de Campos saying that it is more intuitive to speak about the probability that a document d_j is relevant given a query q_k than the opposite (de Campos et al. 2002). Furthermore, only one propagation step is required to propagate the probabilities triggered by a new query throughout the whole network.

The foundation of the **Belief Network Model** proposed by Ribeiro and Mutz (Baeza-Yates and Ribeiro-Neto 1999) is a single sample or concept space given by the index terms of the corpora, viewed as elementary concepts defining the universe of discourse. Documents and queries are defined by concepts k_l given by a subset of this concept space, i.e. essentially a set of term nodes t_i. These assumptions yield to the network topology depicted in the middle network of figure 1, where the arcs are directed from the concept nodes to both the query and the document nodes. Instantiating the index term variables in this network the document and the query part of the network are logically separated, so that a document ranking can be obtained by:

$$p(d_j \mid Q) \sim \sum_{T_i} p(d_j \mid t_i)\, p(Q \mid t_i)\, p(t_i) \qquad (1)$$

De Campos, Fernández-Luna and Huete (de Campos et al. 2001, de Campos et al. 2002) have confined their basic **Bayesian Network Retrieval Model** to the document network of the Belief Network Model illustrated on the right in Figure 1. In contrast to the preceding models it does not consider query nodes at all. A query is simply simulated setting the status of the respective term nodes with a prior probability of 1/(number of index terms) as "relevant". The ranking strategy could again be encoded in the conditional probability distributions of the document nodes using exact propagation. However, due to the typically great number of parent term nodes t_i a probability function is called during the propagation process. Considering indexing and/or term weights within this weighing function w_{ij} the propagation process can be substituted by a single evaluation of equation (2) obtaining equal results for each document node. Equation (3) provides an example for a weighing function that implements the tf.idf model, where tf_{ji} is the frequency of the term t_i in the document d_i , idf_i is its inverse document frequency and α_j a normalising constant.

$$p(d_j \mid Q) = \sum_{T_i \in D_j} w_{ij}\; p(t_i \mid Q) \qquad (2)$$

$$p(d_j \mid Q) = \alpha_j \sum_{T_i \in D_j} tf_{ji}\; idf_i^2\; p(t_i \mid Q) \quad with: \alpha_j = \frac{1}{\sqrt{\sum_{T_i \in D_j} tf_{ji}\; idf_i^2}} \qquad (3)$$

The authors of the Bayesian Network Retrieval Model have proposed several **extensions** to the basic network to account for user's relevance feedback (de Campos et al. 2001) and inter-dependencies among terms or documents (de Campos et al. 2002). To overcome the restrictions of the most common assumptions of term and document independency, additional arcs are introduced within the term or the document sub-network. Directed dependencies such as citations and hyperlinks or broader/narrower term relations can be easily included in the subnetwork (s. term network in Figure 1). However, to consider undirected dependencies (e.g. synonyms, same document type) while ensuring a directed acyclic graph, the respective subnetwork needs to be mirrored, interconnecting the duplicates one-to-one and inserting two corresponding directed links as depicted in the document sub-network in Figure 1. The dependencies within the respective document network can represent document similarities that are used for clustering documents and generating corresponding information maps.

The above brief discussion of the networks shows that all models allow to flexibly store and interrelate various information and evidences used in retrieval processes. The explicit visualisation of the diverse variables and their influence on the retrieval results supports the understanding of the relations between text documents and user domains as well as constructing corresponding retrieval and mining approaches.

We argue that extending the query side of the networks to knowledge models of AEC/FM the illustration of respective aspects and concepts will support the domain specific configuration of queries and information needs as well as the identification of new approaches to automatically interrelate the respective model and text domains.

Taking advantage of the described network capacities, in the following section we propose a novel Bayesian network to support a set of common retrieval and text mining techniques and at the same time to consider general and populated product models as a first source of AEC/FM background knowledge. In further iterations, the topology of the network can be complemented to integrate additional features and interdependencies. Furthermore, the conditional probability distributions can be optimised to more effectively support selected reasoning and mining approaches. Despite the computational costs of exact propagation we adopt Bayesian inference to most flexibly explore different mining methods on a small test collection, replacing them with computationally more efficient algorithms when the mining methods are sufficiently tested.

4 Suggested Layered Bayesian Network Connecting Document and Model-based Information

We propose a mining framework extending the retrieval networks discussed in the previous section. Central to this framework is the **four-layered Bayesian inference network** illustrated in Figures 3 and 4. On each of these layers a respective subnetwork represents a distinctive set of variables thereby describing the knowledge on

- the product domain (product model layer),
- the engineering and mining context (concept layer),
- the documents' content (descriptor layer), and
- the overall document collection (document layer).

The combined network is used to re-configure the collected domain and context information in an evidential reasoning process to most effectively support the given retrieval or mining tasks.

In the following we describe our approach to generate and configure the layers of the mining model discussing first the modelling process, then a first basic configuration of the network, and finally possible extensions bringing added functionality to the network model.

4.1 The *dokmosis* mining network

To explore the possibilities of the suggested novel mining network a document and knowledge modelling suite (*dokmosis*) is being implemented. It provides a set of modules integrating several applications and services to analyse documents and product model data. Based on these modules, the four layers of the mining network are generated in the three analysis steps shown in Figure 2.

A 2-step analysis is required to configure the information on the collected documents for the repository subnetwork comprising the document and the descriptor layer. Firstly, importing documents from external document management systems the underlying document collection and corresponding network layer is built. Afterwards the sub-network for the descriptor layer is created based on the results of selected text analyses providing for pre-processing, indexing and weighing the documents' contents.

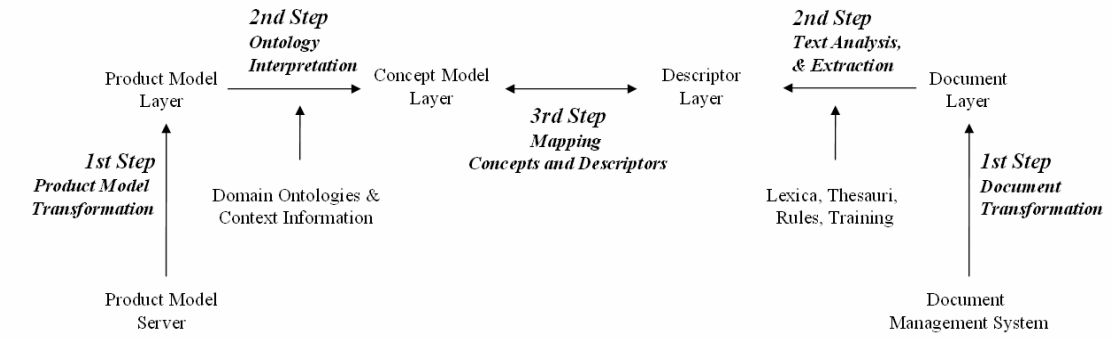

Figure 2: The *dokmosis* mining network

In parallel to the repository network, the two model layers of the network are generated in two consecutive analysis steps. In the first step, the product model data to be considered is imported from an external product model server, transformed and represented on the product model layer. Ontological and additional context information is then used in a second step to derive a discipline or problem specific concept model from the available product model information and represent this on the concept model layer.

In the final third step the knowledge model network and the repository network are inter-connected analysing the similarities among the variables on the concept model and the descriptor layer.

4.2 Basic Mining Network

The basic mining network (Figure 3) is designed to rapidly evaluate the overall concept of the network approach and enable achievement of short-term results. It mainly adopts the Bayesian Network Retrieval Model to represent the knowledge on the document collection using document nodes d_j and (in extension to the original model) fragment nodes f_k on the document layer as well as concept representation (or descriptor) nodes t_i on the descriptor layer. Quite similar to the one-time query networks in section 3, a configurable knowledge model sub-network is build on the remaining two layers. The knowledge model represents product model classes denoted p_m on the product layer and engineering concepts c_x on the concept layer, derived by *interpretation* of the model information. In the basic network, these concepts are essentially the names of the corresponding product model classes, with only little additional attributive information.

For the causal ordering we adopt a continuous direction of the arcs from the product model to the document layer. Thus, based on an Information Retrieval like concept formulation, *direct* propagation determines the relevance of fragments or documents, given the knowledge on the concepts in the domain specified by the selected product models and ontologies.

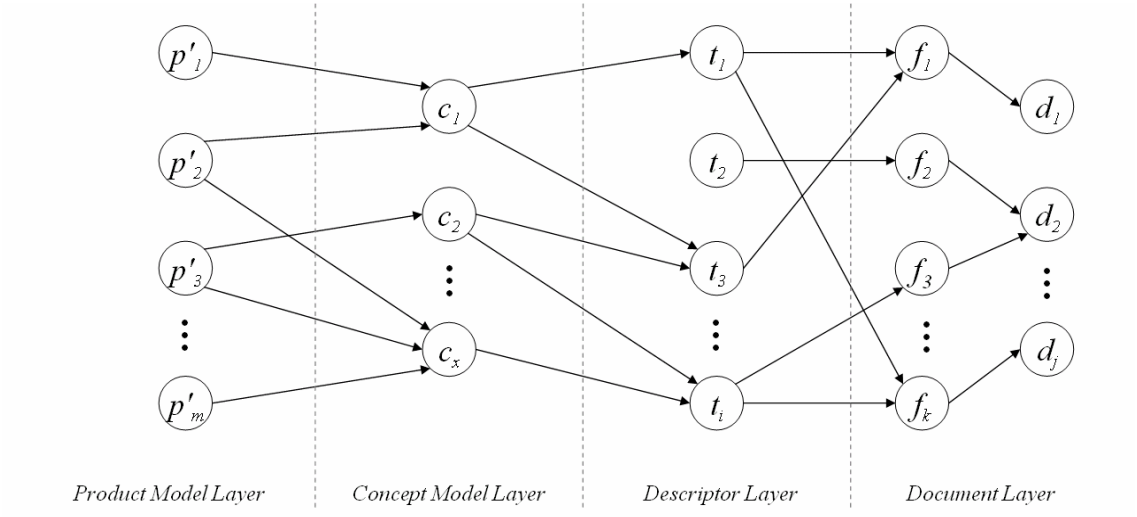

Figure 3: Four layer basic mining network with concepts of two mental models

4.2.1 Building the Repository Network

The repository network is build using three text related modules of the *dokmosis* suite. Firstly, the document collection module provides for importing the documents and converting them to a common format based on a small subset of the *DocBook* specifications (see http://www.oasis-open.org). The intention is to preserve basic structural information on e.g. sections, headers and figures for later analyses. For an additional information extraction module based on the *SpecEx* Extractor (Grimme 2003), it is planed to complement the schema so that specific AEC texts e.g. in GAEB-based specification and work descriptions can be recognised (see http://www.gaeb.de). Furthermore, to adequately allow for more focused information access a heuristic fragmentation algorithm compiles sections and text paragraphs in 'self-contained text units' of similar size depicted as fragments f_k in the network. Interdependencies among the fragments and documents are all considered as *part-of* relations.

The pre-processing of the fragments in the text analysis modules includes tokenisation, entity recognition, and selected lexical analyses. However, for the basic mining network only tokenisation, morphological analysis and stop-word removal are performed. Based on these text analyses the current version of the *dokmosis* suite provides for building a vector space model considering raw term frequencies. In a further extension of *dokmosis* we expect to provide for building two types of Repository Networks: one based on a Boolean and one based on the tf.idf weighing strategy.

4.2.2 Building the Knowledge Model Network

As mentioned above, the knowledge model network is built by importing product models or schemata and performing corresponding ontological transformations. 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 iCSS product model server (cf. iCSS 2002), complemented with methods to identify both the classes defined in an EXPRESS schema as well as those used in a corresponding instantiated product model. A transformation of the model is achieved by simply adding an independent node to the product model subnetwork for each class (name) in the returned set.

As concepts can hardly be a one-to-one reflection of the classes in the product model layer, additional background knowledge is required to build a more expressive concept model network. For this purpose the *dokmosis* suite uses an adapted version of the *ontology interpreter* developed in the EU ISTforCE project that enables translating model information based on respective engineering ontologies (Katranuschkov et al. 2003). To achieve an easy to process discipline-specific flattened network, the ontological mapping specifications are confined to mappings that filter or aggregate classes from the original set. Thus, 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. Furthermore, the specifications have been complemented with lexical descriptors to label each engineering concept with suitable terms.

Finally, the connection of the knowledge model with the repository network is achieved by interlinking concepts and descriptor nodes that represent the same term or respective label. Aspects of comprehensive labelling and weighing of the labels is omitted in the basic model, again considering Boolean dependencies.

4.3 Extending the Mining Network

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

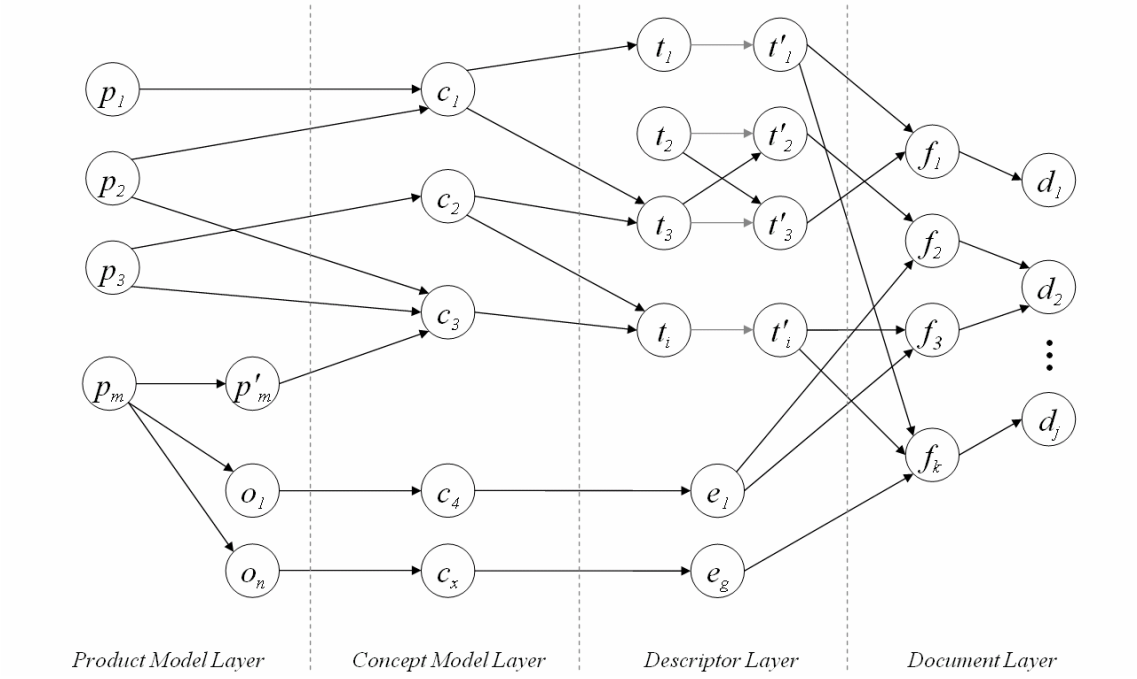

Figure 4: Extended Mining Network

Especially the concept layer allows for a distinctive, personalised configuration of the applied background knowledge, without changes to the original model-based information. Context and additional user information can be utilised to re-label classes, filter certain levels of abstraction or aggregate classes into more meaningful concepts for a user's discipline and retrieval task. However, to explore further retrieval options we first of all focus on extensions to the product model and the descriptor layer.

4.3.1 Considering Term Interdependencies

A first useful extension to the mining network is to account for term interdependencies in natural language text. The main goal is to increase the influence of the comparably few distinctively labelled concepts nodes and, respectively, e.g. the recall of network-based information retrieval. The interdependencies among the index terms can be determined by evaluating *thesauri* or performing a term similarity analysis in the document collection. For the extended mining network shown on Figure 4 the approach to duplicating the representation nodes (cf. section 3) has been chosen to consider term similarities determined in the most frequent domain-specific document collection of interest. The network in Figure 4 illustrates that the importance of comprehensive labelling and label term weighing can be reduced by introducing synonyms and related terms in the network.

4.3.2 Considering an Entity Representation Scheme for the Descriptors

To be able to account for named entities identified in the text in preliminary analyses, we propose to consider a second representation scheme illustrated by the nodes e_k on the descriptor layer. Currently the *dokmosis* suite provides for regular expressions based entity recognition of persons, organisations, addresses, codes and regulations, scales, formulas etc. Furthermore, the text analysis module includes sentence identification and part of speech tagging to allow for building and training information extractors for explicitly classified text elements of e.g. specifications or punch lists. To simplify the network we assume the two representation schemes to be independent. Thus, the network based information retrieval using instantiated product model classes in the basic mining network is separated from the propagation of beliefs on respective product model objects described in the following.

4.3.3 Considering Instantiated Product Model Objects

With a second representation scheme on the product model layer corresponding to the entity representation scheme, a separated set of retrieval paths is provided by the extended mining network. By more accurately representing the product model information we now distinguish between classes and instantiated model objects (the *p* and *o* nodes on Figure 4 respectively). Whilst currently the object relations defined in the product model are not yet considered, the product model layer already accounts for *instance-of* dependencies among objects and classes.

The explicit representation of modelling objects and text entities demonstrates the possibilities but also the challenges of directly interlinking product model and text information. Via concept nodes objects can be connected with corresponding entities on the descriptor layer. However, according to the various types of possible objects and entities, a set of similarity measures needs to be evaluated to determine the probability that *c*-*e* node pairs really represent the same aspect. Furthermore, the concept layer is required to first group, abstract, generalise the product model objects to provide for a mapping to the identified entities.

5 Conclusions

Bayesian inference networks have been identified as a very flexible technology that allows to identify and externalise information from respective AEC text documents and to represent and interrelate various information resources.

By combining a classical Bayesian network approach with product model data management we have defined a framework that enables to better and more efficiently use available project information. Due to the possibilities to encode the knowledge on a variable in terms of causal relations and weights the network can be configured to support logic based mapping as well as numerical mining techniques in addition to the complete Bayesian inference. This is an essential capacity of the networks enabling to interlink the “rigid” world of model based systems with the rather fuzzy world of text and language processing.

Some of the opportunities that can be achieved through the suggested network are:

- By pre-selecting a product model schema or emphasising certain aspects on the concept layer, the user can intuitively reconfigure the content structure, search for concepts and highlight certain relationships in the text corpora.
- Based on the inference network semantic content networks of the repository can be generated and document maps can be re-calculated and further optimised
- The established links between documents and product model data can enable navigating a text corpus through related product or concept models

Further possibilities are expected to emerge when the implementation of the approach is extended and tested on a larger set of use cases.

6 References

Baeza-Yates R., Ribeiro-Neto B. (1999): *Modern Information Retrieval*. Addison-Wesley, Wokingham, UK, 1999

Caldas C.H., Soibelman L., and Han J. (2002): *Automated classification of construction project documents*. In: Journal of Computing in Civil Engineering, Vol. 16, No. 4, pp.234-243, 2002

Crestani F., Lalmas M., van Rijsbergen C., Campbell I. (1998): *"Is This Document Relevant? ... Probably": A Survey of Probabilistic Models in Information Retrieval*. In: ACM Computing Surveys, Vol. 30, No. 4, December 1998

De Campos L.M., Huete J.F., Fernández-Luna J.M. (2001): *Document Instantiation for Relevance Feedback in Bayesian Network Retrieval Model*. In: ACM SIGIR MF/IR Workshop on Mathematical/Formal Methods in Information Retrieval. Nueva Orleans, USA 2001

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

Grimm S. (2003): *Untersuchung des Einsatzes von Methoden zur Informationsextraktion im Bauwesen am Beispiel der Angebotskalkulation*. Diploma Thesis at the Institute for Construction Informatics, Prof. R.J. Scherer, Technical University of Dresden, February 2003

Grobelnik M., Mladenic D. (Eds) (2001): *Workshop on Text Mining*. Proceedings of IEEE International Conference on Data Mining, San Jose, California, USA, November 2001

Hearst, M. (1999): *Untangling Text Data Mining.* In: Proceedings of 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 1999

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

Katranuschkov P., Gehre A. and 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, 2003

Kosovac B., Vanier D.J., Froese T.M. (2000): *Use of Keyphrase Extraction Software for Creation of an AEC/FM Thesaurus*, In: Electronic Journal of Information Technology in Construction, Vol. 5, pp. 25-36, 2000

Lima C., Fies B., Zarli A., Bourdeau M., Wetherill M., Rezgui Y. (2002): *Towards an IFC-enabled ontology for the Building and Construction Industry: the e-COGNOS approach*, In Proc. of eSM@RT 2002, Salford, UK, pp. 254-264, November 2002

Neumann G. (2001): *Informationsextraktion.* In Klabunde et al. (Eds): Computerlinguistik und Sprachtechnologie - Eine Einführung. Spektrum Akademischer Verlag, Heidelberg 2001

Nahm U.Y., Mooney R.J. (2002): Text Mining with Information Extraction. In: Proc. *of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pp. 60-67, Stanford, CA, March 2002

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

Tolman F., Böhms M., Lima C., van Rees R., Fleuren J., Stephens J. (2001): *eConstruct: expectations, solutions and results*, In: Electronic Journal of Information Technology in Construction, Vol. 6, Special Issue ICT Advances in the European Construction Industry, pp. 175-197, 2001, http://www.itcon.org/2001/13

Turtle H.R., Croft W.B. (1991): *Evaluation of an Inference Network-Based Retrieval Model*. In: ACM Transactions on Information Systems, Vol. 9, No. 3, pp. 187-222, July 1991