TEXTANALYSEMETHODEN FÜR BAUDOKUMENTE MIT LEXIKALISCH ORGANISIERTEM KONTEXTWISSEN AUS PRODUKTMODELLEN

St. Scheler, R. J. Scherer
TU Dresden, Computeranwendung im Bauwesen

Kurzfassung: *Die in den verschiedensten Formen auftretenden Baudokumente enthalten Informationen, die mit einer einfachen Suchanfrage nicht extrahiert werden können. Zur erfolgreichen Dokumentenklassifikation wird neben einfachen Textanalysemethoden zusätzlich projektspezifisches Kontextwissen benötigt. Mittels einer vollständig objektorientiert strukturierten Datenbank, die das zur Suche notwendige Kontextwissen repräsentiert, werden Suchanfragen generiert und in interaktiver Kopplung zum Nutzer dahingehend strukturiert, dass komplexe Suchanfragen, ausgehend bei einfachen Schlüsselwörtern bishin zu projektspezifisch relevanten Suchwortkonglomeraten, möglich werden. Der Fokus des Forschungsprojektes ist auf die Einbindung von Produktdatenmodellen in die sonst unscharfen Informationstrukturen von linguistischen Textanalysemethoden gerichtet. Es werden Methoden und Verfahren für die Extraktion und Strukturierung des Bauprojektwissens unter Einsatz dieses aus Produktdatenmodellen eingebundenen Kontextwissens untersucht.*

1 Problemstellung

Baudokumente sind sehr heterogen, vom Handzettel bis zur Bauzeichnung entstehen im Laufe eines Projektes jede nur mögliche Dokumentform. Mit den modernen Methoden des Dokumentmanagements kommt man bei der Auswertung von Baudokumenten nicht zum Ziel, da eine DTD (Dokumenttypdefinition) für diese unstrukturierten Dokumente in der Praxis in absehbarer Zeit nicht definiert werden. Im Gegensatz zu SGML (*Standard Generalized Markup Language*) strukturierten Dokumenten, was sich für ein durchgängiges Dokumentmanagement empfiehlt, können daher auch keine vordefinierten Suchmechanismen zu Analysen genutzt werden. Um das Bauprojektwissen langzeitlich zu sichern, müssen die in den Baudokumenten gespeicherten Informationen mittels Suchalgorithmen, die in Such*clustern* gebündelt

sind, extrahiert werden und unter Verwendung von Kontextwissen in einer verständlichen Form verfügbar gemacht werden. Die verständliche Form definiert die jeweils projektspezifische Sicht eines Nutzerkreises. So verbinden sich z.B. bei der Suche nach einem Statikproblem aus der Sicht eines Fachplaners für Personenaufzüge völlig andere Aspekte zur Auffindung relevanter Dokumente als beispielsweise aus der Sicht des Statikers selbst. Es müssen daher aus der Produktmodellwelt Informationsstrukturen dahingehend nutzbar gemacht werden, dass im Sinne des Wissensmanagements die in Produktdatenmodellen definierten Klassenbibliotheken als Ontologien interpretiert werden. Es werden so umfangreiche Begriffdefinitionen vorgenommen, die gerade für den angesprochenen Nutzerkreis Relevanz besitzen. So müssen, ausgehend vom Beispiel „Aufzugsschacht für Personenaufzüge“, durch die Integration eines Standard-Produktmodells wie STEP (*Standard for Exchange of Product Model Data*) die Assoziationen mit weiteren Begriffen (Gebäudekern, Fahrschacht oder Rauchabzugsöffnung) aus den Klassendefinitionen der Produktmodellkategorien und weitergehend den direkten Beziehungen zu einzelnen Regelwerken wie die DIN 276 vorgenommen werden. Es entsteht so ein Begriffsnetzwerk um das gesuchte Schlüsselwort, das so in den zu durchsuchenden Dokumenten der Bauplanung und des Bauablaufes für den eingegrenzten Nutzerbereich instantiiert wiedergefunden werden muss. Die dadurch aufgefundenen Dokumente sind als hinreichend zum eingegrenzten Themenbereich (im Beispiel Aufzugsschacht) interpretierbar.

Die im ESPRIT Projekt ToCEE (*Towards a Concurrent Engineering Environment*) [1] erarbeiteten Ergebnisse im *Document Modelling* sind direkt übertragbar, da gerade hinsichtlich der tatsächlich in der Baupraxis vorliegenden Dokumenttypen und der Einführung einer speziellen Klassifizierung nach den Beteiligten (actors) im jeweiligen Bauvorhaben eine in den allgemeinen Produktmodellen der Planung nicht vorhandene Unterscheidung gemacht wurde.

Untersucht werden neben den Assoziationen, die sich aus der Verknüpfung mit dem Normenwerk für Produktmodellierung STEP - ISO/FDIS 10303 (hier speziell das Applikationsprotokoll AP225 [2] für das Bauwesen) ergeben, ebenfalls solche aus dem mehr von der Industrie bevorzugten IFC Modell [3]. Die Anwendung der Klassifikationen aus den verschiedenen existierenden Produktdatenmodellen und den sich daraus ergebenden Querverbindungen für die Kontextspezifizierung des Schlüsselwortes eröffnen letztlich mit reinen Textanalyseverfahren bisher nicht mögliche Schlussfolgerungen über das zu interpretierende Dokument.

2 Wissen in Standards

Bei der Einbindung von Produktdatenmodellen wird hauptsächlich auf die Normenreihe STEP zurückgegriffen, da im Versuchsstadium die stabile und bezüglich Konsistenz

verifizierte Modellstruktur ein wichtiger Ausgangspunkt für die Verfeinerung und Stabilisierung des angestrebten Relationsnetzes der Schlüsselwortanalyse ist [Abb. 1]. STEP (ISO/FDIS 10303) beschreibt Produkte vollständig und in allen Lebensphasen - von der Konstruktion bis zur Fertigung. Dabei wird jeder Technologiebereich von der Elektrotechnik, Mechanik bis hin zum Bauwesen unterstützt. Wichtig ist hier die durch den Austausch bedingte Konzeption einer einheitlichen, durchgängigen Methodik für verschiedene Anwendungen wie es im Bauwesen der Fall ist. STEP ereicht auch eine hohe Konsistenz bei der Beschreibung der Semantik der Daten, es werden Anwendungsdatenmodelle unter Verwendung von Ressourcenmodellen (z.B. Geometrie) herangezogen. Die Ansatzpunkte zur Implementierung von STEP-konformen Modellbegriffen bedingen weiterhin, dass das Bibliotheks- und Variantenkonzept vollständig mit integriert wird. Das in den Begriffen und deren Verknüpfungen enthaltene Wissen wird einschließlich der konsistenten und logischen Klassenstruktur eingebunden.

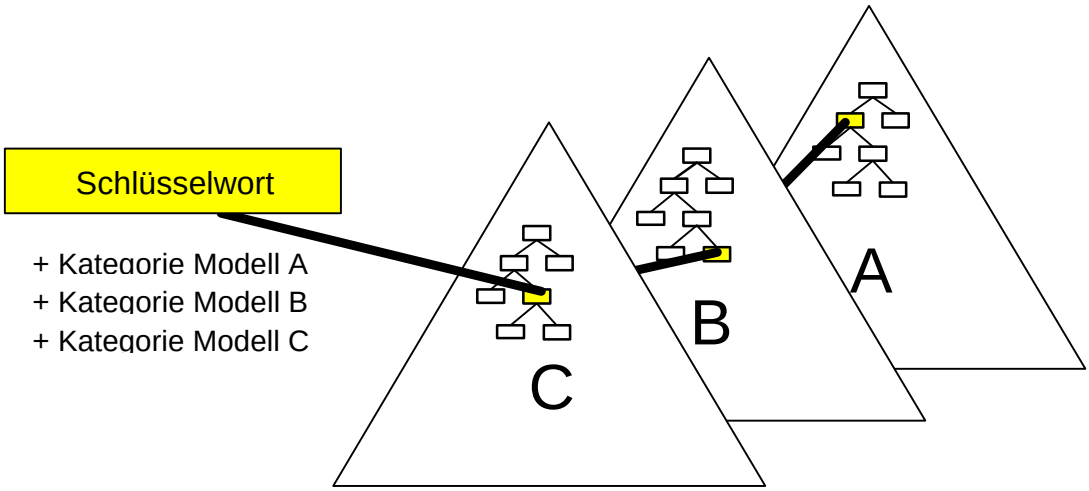

Abbildung1: Relationen des Schlüsselwortes zu den angesprochenen Modellkategorien

Nach der Extraktion der gesuchten Dokumente mittels der strukturierten Schlüsselwortanalyse muss eine sichere Langzeitarchivierung der erhaltenen Daten erfolgen. Dazu ist das Dokument an sich zu klassifizieren. Die Klassifikation erfolgt dynamisch mit dem System LOOM [4], das mit dem Produktinformationssystem PINS [5] um eine Wissensbasis für das Bauwesen erweitert wurde. Hierbei wurden elektronische Produktkataloge und Produktkomponenten nicht alleinstehend genutzt, sondern über ein einheitliches Produktdatenmodell in ein Suchsystem für beliebige Anwendungen eingebettet. Eine derartig integrierte Lösung hat vor allem im Hinblick auf die Konsistenz und die Aktualität der Produktpräsentationen Bedeutung und ermöglicht darüber hinaus eine redundanzfreie Datenerfassung und -verwaltung.

Die gilt es nun in einem spezifisch für einen angesprochenen Nutzerkreis (Architekt, Tragwerksplaner) als Dokumentdatenmodell zu konsolidieren, da es in STEP gerade für den Bereich Bauwesen an Basismodellen mangelt, die Aspekte der o.g. Anwendungen zu berücksichtigen. So sind in STEP die Dokumenteigenschaften allgemein unter

Information requirements ➔ Application objects ➔ Building_document_reference ➔

document_type + *identifier* + *item_in_document*

beschrieben. Zum Vergleich beschränkt sich die Dokumentbeschreibung im IFC-Standard hingegen auf

IfcDocumentType + *IfcDocumentReference.*

Hier sind die Dokumente nur als *Resource layer* zur Festlegung von *Properties* erfaßt.

Für eine eindeutige Klassifizierung im dynamischen Sinne werden neben dem Dokumentnamen und -typ eine Reihe von inhaltlich beschreibenden Eigenschaften erwartet, die über einen *identifier*, wie dies in STEP beschrieben wird, hinausgehen. Das hier vorgeschlagene Dokumentmodell nutzt diese Inhaltseigenschaften zur Beschreibung der einzelnen Klassen, in welche das gefundene Dokument zugeordnet wird, ohne dass vorher das Dokument Strukturen für eine Identifikation entsprechend SGML aufweisen muss. Mit der Auffindung eines Dokuments über die strukturierte Schlüsselwortanalyse einerseits und die damit verbundene Möglichkeit der nachträglichen inhaltlichen Klassifizierung des Dokuments andererseits zeichnet sich eine weiterreichende Nutzung für die Langzeitsicherung der Dokumente ab.

3 Produktmodelldaten lexikalisch strukturiert

Das aus den Produktmodellen gewonnene Kontextwissen wird in einer vollständig objektorientiert strukturierten Datenbank repräsentiert. Interaktionen mit dieser Datenbak werden mit JAVA-Applikationen ausgeführt, was hinsichtlich einer späteren Nutzung von Internetressourcen in Form einer Client-Server-Applikation von Vorteil ist [Abb. 2].

Als erster Schritt ist ein geeignetes wissensbasiertes Lexikon mit differierenden Strukturen [6] aufgestellt worden, das in Anlehnung an bereits vorhandene Produktmodelle einen hierarchischen Aufbau besitzt. So werden entsprechend des Bauablaufes verschiedene Planungsphasen gebildet:

Konzeption ➔ Konstruktion ➔ Bauvorbereitung ➔ Ausführung

Ein weiterer Teil des Wissens wird anhand gesetzlicher Vorschriften, wie die Einteilung des Ablaufes in Phase 1-9 nach HOAI [7] und aus Regelwerken, wie DIN/ISO, die hauptsächlich die Kontextbeschreibungen liefern, erarbeitet.

Aus dem Projekt ToCEE [1] werden die Methoden zur Unterscheidung nach Beteiligten (*actors*) übernommen. Zur Erkennung eines spezifischen Problems ist es von vordringlichem Interesse, welchen Kreis der Projektbeteiligten diese Information überhaupt zugänglich ist. Die Unterscheidung ist differenziert vom Bauherrn bzw. Auftraggeber mit dem leitenden Projektmanager oder Architekten über den ausführenden Bauleiter und Fachplaner bis hin zur Behörde und dem Gutachter.

Im zweiten Schritt werden Dokumentmodelle anhand der vorliegenden Produktmodelle erstellt, die der Datenbankstrukturierung zugrunde liegen und in welche die lexikalisch geordneten Begriffe eingelesen werden. Die objektorientierte Datenbank wird anschliessend mit dem an die einzelnen Schlüsselwortbegriffe gebundenen Kontextwissen instantiiert. Die mittels Relationen verknüpften Begriffe bilden ein Netzwerk von verwandten Worten und Wortgruppen, das dem jeweils gesuchten Zielbegriff zugeordnet ist. In diesen Netzwerken wird nun eine Strukturanalyse hinsichtlich der modellierten Umgebung durchgeführt. Jedes Dokument besitzt dann differenzierte Zugehörigkeit zu den Modellen aus Planung, Ausführung und Dokumentation [Abb. 1] und bekommt aus dem Kontext Referenzen wie Plannummer, Arbeitsschutzverordnungen oder Baurechtsparagraphen zugewiesen [Abb. 3].

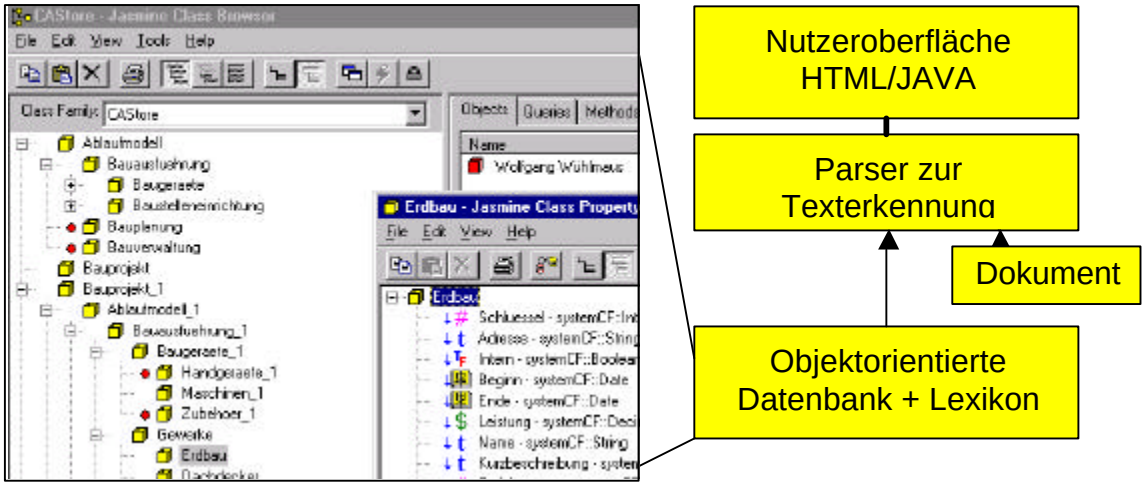

Abbildung 2: Anbindung der objektorientierten Datenbank zur Bereitstellung des Kontextwissens

Eine besondere Berücksichtigung gilt dann im dritten Schritt dem Zeitablauf. Die eingebundenen Produktmodelle werden durch Prozessmodelle erweitert. Mit der Integration der Zeitachse nun können auch Objekte differenziert erfasst werden, die mehrmals mit jeweils unterschiedlichen Eigenschaften entlang dieser Zeitachse auftreten. So kann am Beispiel des zu betonierenden Aufzugsschachtes eines Personenaufzugs die Wortgruppe „Schacht betonieren“ mehrmals parallel auftreten, da in den verschiedenen Bauabschnitten - Erdgeschoss bis Dachgeschoss - dieser

Prozess jeweils erneut initiiert wird. Grundlegend handelt es sich aber um einen identischen Begriff, der mit speziellen Eigenschaftsparametern wie Bauabschnitt zeitlich unterschieden wird. Der Begriff wird also in der Datenbank nur einmal instantiiert und nicht aufgrund der verschiedenen Zeitebenen mehrfach.

4 Datensuche und Anfragekonsistenz

In Hinblick auf die Datensuche in einem virtuellen, vernetzten Dokumentenarchiv wurde im zweiten Schritt ein auf Java basierender einfacher Parser zur Extraktion von Begriff*strings* aus den heterogenen Dokumentbestand entwickelt. Die Basis bildet eine Beschreibung in natürlicher Sprache [8]. Die Parserapplikation setzt direkt auf die objektorientiert strukturierte Datenbank auf. Die in der Datenbank lexikalisch angeordneten Begriffe sind vorerst nach dem dynamischen Klassifikationssystem vorstrukturiert. Dem Nutzer werden die Klassenzugehörigkeit und somit die lokale Verknüpfungsumgebung und die zusätzlich mit eingebundenen Relationen angezeigt. Hinter der Strukturierung nach Superklassen und Subklassen verbirgt sich das jeweils integrierte Produktmodell. Zur Vorstrukturierung und Aggregation der einem *Belief*netzwerk ähnlichen Schlüsselwortumgebung wird entlang des ausgewählten Analysebaumes vorgegangen.

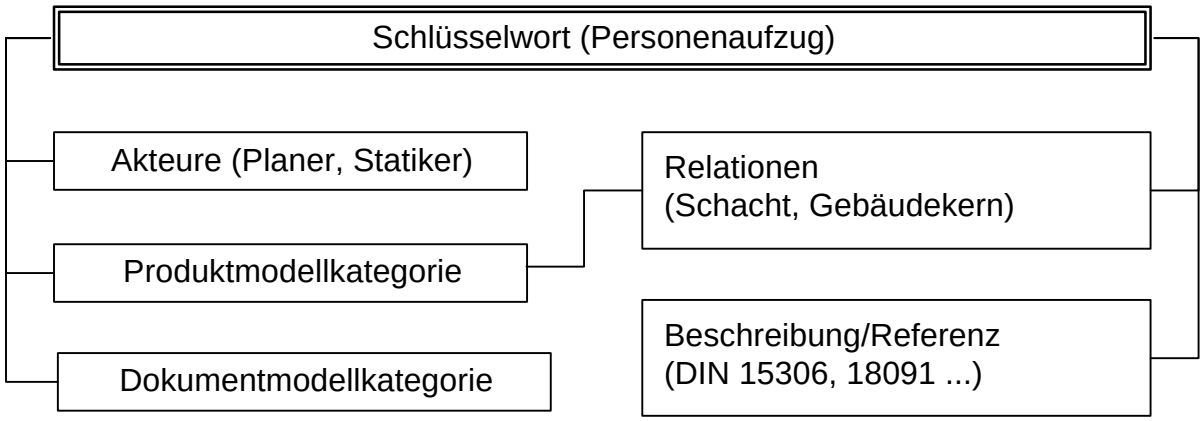

Abbildung 3: Strukturierung und Instantiierung der Schlüsselwortzuordnung

Es gibt beispielsweise ein Problem beim Einbau eines Aufzuges [Abb. 3]. Der Schacht entspricht nicht den angeforderten Maßen. Dem Problem gehen zwei Nutzer mit unterschiedlichen Sichtweisen und demnach differierenden Anforderungen nach. Der Planer sucht Dokumente, die dieses Problem schon im Vorfeld der Bauausführung tangieren, wogegen der Statiker seinen Focus auf das Gebäudemodell und die darin enthaltenen mathematisch nachweisbaren Berechnungen setzt. Es ergeben sich zwei unterschiedliche Suchanfragen, die das gleiche Schlüsselwort enthalten, aber eine unterschiedliche Wichtung in der Beurteilung der Zieldokumente ergeben. Der Nutzer

muss die beteiligten Akteure und die zusätzlich aufgezeigten Relationen verifizieren. Eine Hilfe hierbei ist die jeweils mit propagierte Beschreibung oder Referenz des Begriffes.

Mit Hilfe des Parsers wird dann die syntaktische Analyse des Zieldokumentes durchgeführt und in einer Strukturbeschreibung abgelegt. Anhand dieser Beschreibung wird das Dokument im Anschluss klassifiziert und kann nach Abschluss der verschiedenen Suchdurchläufe verifiziert werden. Der Vorteil einer dynamischen Klassifikation kommt bei unterschiedlichen Bewertungen der einzelnen Suchdurchläufe zum Tragen.

5 Ausblick

Die lokale Umgebung der Schlüsselwörter mit ihrer Schlüsselwortstruktur unter Einbindung von Produktmodelldaten durch unmittelbare Verweise der betrachteten Instanzen bildet ein Such*cluster*. Ziel ist es nun nach der Modellbildung von probabilistischen Netzwerken eine umfassende Strukturanalyse in diesen Netzwerken durchzuführen. Die Bewertung von Übereinstimmungen zwischen den einzelnen Such*clustern* innerhalb eines Dokumentes führt schließlich dazu, den anfänglich umfangreichen Kontext in unmittelbarer Umgebung der Schlüsselwörter sinnvoll einzuschränken. Durch Einbindung einer interaktiven Kopplung zwischen der zugrundeliegenden Datenbank und der Nutzeroberfläche wird es ermöglicht, dem jeweiligen Anwender ein Werkzeug zur Verfügung zu stellen, das eine individuelle Strukturierung der Verknüpfungen zwischen den Schlüsselwörtern, den Produktmodellinstanzen und weiterer kontextspezifischer Relationen vornimmt.

Literatur

[1] SCHERER R.J.: *A Framework for the Concurrent Engineering Environment,* Proc. of the 2nd European Conference on Product and Process Modelling in the Building Industry, Amor R., Scherer R.J. (ed.), Building Research Establishment, Watford, Great Britain, October 1998.

[2] ISO/FDIS 10303-225: *Product Data Representation and Exchange – Application protocol: Building elements using explicit shape representation,* Supersedes ISO TC184/SC4/WG3 N719, ISO Central Secretariat, Geneva, January 1998.

[3] IAI : *An Introduction to the International Alliance for Interoperability and the Industry Foundation Classe,*, IFC Release 2.0, IAI Publ., Oakton, VA., 1999.

[4] BRILL, D.: *LOOM - Reference Manual,* University of Southern California-School of Engineering, Information Sciences Institute, 1993.

[5] SCHERER R.J., NOLLAU C., BUCHWALTER J., SCHELER S.: *Produktinformationssysteme unterstützt durch dynamische Klassifikation und ähnlichkeitsbasierte Suche,* in: Hartmann D. (ed.) Veröffentlichung zum Abschlußkolloquium des DFG-Schwerpunktprogramms 694 "Objektorientierte Modellierung in Planung und Konstruktion", Springer-Verlag, 1998.

[6] LEMAIRE, F.: *Knowledge Bases, Texts and Lexicon,* Towards Very Large Knowledge Bases pp 281-287, N.J.I.Mars ed. Knowledge Building & Knowledge Sharing University of Twente, Enschede, IOS Press, 1995.

[7] HOAI: *Verordnung über die Honorare für Leistungen der Architekten und Ingenieure,* Kohlhammer-Verlag, Stuttgart, 1996.

[8] TAYLOR, A.: *Extracting Knowledge from Biological Descriptions,* Towards Very Large Knowledge Bases pp 114-119, N.J.I.Mars ed. Knowledge Building & Knowledge Sharing University of Twente, Enschede, IOS Press, 1995.