Using Business Process Models to Retrieve Information from Governing Documents

Tarjei Lægreid, Paul Christian Sandal, Jon Espen Ingvaldsen,
and Jon Atle Gulla
Norwegian University of Science and Technology,
Department of Computer and Information Science,
Sem Saelands vei 7-9,
NO-7491 Trondheim, Norway
{tarjeil, paulchri, jonespi,
jag}@idi.ntnu.no

Abstract

Governing documents are long textual documents containing information about how different operations are to be carried out in a business. The information spans from overall policies and standards to detailed guidelines, and the amount of such documents in larger corporations tend to be substantial. In business process models the aim is to model the same domain from a process perspective. In this project work we investigate the potential of utilizing text mining technologies together with a standard information retrieval system to link these two sources of information in a dynamic way. To test out the concepts we employ business process models and governing documents from Statoil ASA, a Norwegian oil and gas company.

1. Introduction

In large enterprise architecture solutions corporate policies, operations, and standards are commonly defined using graphical business process descriptions and textual governing documents. These documents are usually lengthy, and not specific to particular tasks or processes, and the user is left to read through a substantial amount of irrelevant text to find the fragments that are relevant for the execution of a specific activity. Since the users tend to use the process model as a guide in their daily work, it is often desirable to start with the activity in the process model and automatically retrieve the parts of the governing documents that pertain to this activity.

Even though they document the same domain, the existence of both governing documents and graphical business process models are necessary. The expressiveness of business process models can not eliminate the importance of governing documents. Similarly, the importance of business process models can

not be eliminated by giving governing documents a process oriented structure. The challenge is to find methods that enable

1. Content consistency between the two information sources
2. Retrieval of information across both representation formats.

In this work, three different text mining approaches (Latent Semantic Indexing, Association Rules and document expansion using WordNet) are applied to establish links between business process model elements and relevant parts of governing documents.

To reduce the burdensome browsing, it is desirable that the links do not only point to the document where relevant information is found, but also to the precise location of the information within the document. Further, dynamic linking reduces the manual effort of maintaining such links.

The techniques are thoroughly studied and implemented in a simple prototype. The approaches are evaluated based on available documents and accompanying model fragments covering the Procurement and Logistics (P&L) area in Statoil ASA.

The work is part of the KUDOS project, a collaborative effort including Statoil ASA and the Information Systems Group at the Norwegian University of Science and Technology. The main objective of KUDOS is to improve the retrieval and management of corporate information by use of anthologies and text mining techniques. The evaluation of our results indicates that the information retrieval approach to integration has potential. Of the text mining techniques investigated, Latent Semantic Indexing (LSI) gives the most promising results, but both this and the other techniques should be more thoroughly evaluated in a continuation of this work.

2. Statoil ASA

Statoil is an integrated oil and gas company with more than 25 000 employees and activities in 31 countries. The group is operator for 60 percent of all Norwegian oil and gas production. Statoil is also a major supplier of natural gas in the European market and has substantial industrial operations. The size and complexity of such a company introduce substantial challenges regarding coordination and management, both within each functional unit and at the higher cross-functional level.

2.1. Business Process Model

As a tool to help managing the complexity Statoil has done an initiative to document the different business processes of their enterprise. The main purpose of this initiative is to change their established functional view of the enterprise areas into business processes, as a part of building an enterprise architecture for the corporation. The Statoil Business Process Model (BPM) is used to document relations between business processes, information and IT systems. The main

driver of the effort was that Statoils IT portfolio contained, and still contains redundant and overlapping systems.

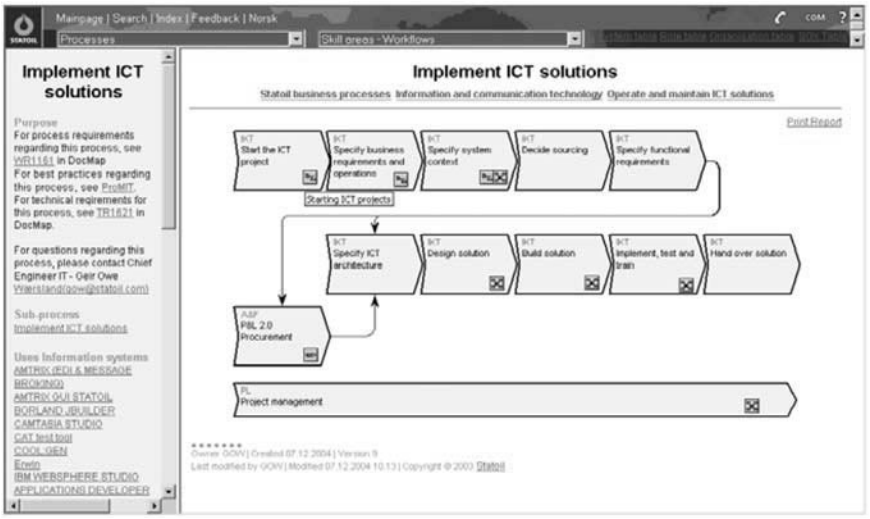

Figure 1. A screenshot of the BPM system that shows a decomposed view of the "Implement ICT solutions" process, which is part of the main process "Information and communication technology". A description, and references to three governing documents for the process is displayed on the left side of the figure.

BPM is a top-down hierarchical model of the Statoil enterprise. At each level business processes are described by graphical business process models and related governing documents. A screenshot of the BPM showing both a graphical business process model and links to related governing documents is shown in figure 1. The graphical business process models visualize subsequent and aggregated processes, involving resources (like documents), events and decision points.

The governing documents are related to the graphical business process models through hyperlinks in the BPM. While the graphical models give a process-oriented overview of the business processes in the BPM, the governing documents contain all the information (guidelines, procedures, descriptions, etc.) that are necessary for a successful execution.

Today, elements in the graphical business process models are manually related to relevant governing documents. For each process, a list of relevant governing documents is maintained. As governing documents are lengthy, formally structured bodies of text and describe issues that are not directly relevant for the execution of specific business activities, it is in many cases bothersome to locate the fractions of text that are of importance. The links point

from the graphical models to whole documents and no indication of where in the document the relevant part(s) occur is given.

2.2. Case Study

In our work we were granted access to ten governing documents with related graphical business process models, all dealing with the P&L area. Independent of hierarchy level, the content of the governing documents is structured into sections, subsections, and paragraphs. All sections and subsections start with a heading. All sections and subsections contain a heading and one or more paragraphs (overall statistical data of the governing documents is listed in table 1). In addition a few paragraphs have their own heading as well. The lengths of the different sections are varying. The content is mainly pure text and point-lists, in some cases supplied with figures. All of the governing documents are written in English.

Table 1. Statistical data of the governing documents.

Average number of words	3792
Average number of sections/subsections	32
Average number of paragraphs	160

3. Implementation

To be able to evaluate text mining techniques for dynamic linking, a test framework has been implemented. The framework is based on Lucene, which is an open source search engine Application Programming Interface (API) released by the Apache Software Foundation.

The graphical business process models in the BPM are encoded in HTML and to be able to extract information from these models both parsing and preprocessing were necessary. The processes elements in the models consist of three text key elements which we make use of in the indexing and result ranking. That is:

1. *Title* is the name of the process / activity.
2. *Description* is a short natural language written explanation of the process.
3. *Links to super and sub processes.* A process may be decomposed into several levels of sub processes (The title and description of super and sub processes are included as an integrated part of a process query).

To find relations between processes and activities in the graphical models and sections in the governing documents, we treat the key text elements of the processes and activities in the graphical models as queries and search for related fractions in the governing document index.

It is reasonable to assume that if a title in an element in a graphical model matches a title of a governing document, there is a relation between the two. Further, a match between titles is probably more relevant than matches in the description contents. This motivates for keeping and treating the different elements of the processes and activities separately. With such a separation it is possible to boost sub parts individually during the retrieval process. In our implementation, process title matches are boosted by 2, while matches in sub process titles are boosted by 1.5.

To be able to search for relevant fractions of the documents they must be segmented prior to indexing. The governing documents related to Statoils BPM system have a common overall structure and it is reasonable to expect that the structure reflects the semantic content to some extent, and thus the explicit structure can be used as a basis for fragmentation. Each fraction of the original documents is considered as a new self-contained document with title and text. Finally, these documents are indexed by removing all stop words, stemming the remaining words, transforming the documents into inverted files, and storing the inverted files in the index. The inverted files are weighted by the tf-idf weighting scheme. Then the title terms are boosted by 1.5. The actual integration of the graphical model elements and the fractions of the governing documents rely on a cosine similarity search.

3.1. Latent Semantic Indexing

The vector space model is based on the assumption that the same terms occur in both the query and the relevant document [Baeza-Yates and Ribeiro Neto, 1999]. However, in natural language texts it is common to use different terms for describing the same concepts. In an attempt to reduce this problem we have implemented a variation of LSI.

The Lucene API offers a simple way to extract the term-document frequency matrices for each indexed field. We extract the title and text fields from all documents and create one term document matrix. The term-document matrix is decomposed into three matrices by Singular Value Decomposition (SVD). The central idea is to let a matrix A represent the noisy signal, compute the SVD, and then discard small singular values of A. It can be shown that the small singular values mainly represent the noise, and thus the rank-k matrix A_k represents a filtered signal with less noise. The SVD has a variety of applications in scientific computing, signal processing, automatic control, and many other areas [Hansen, 1987; Hansen and Jensen, 1998]. By first applying SVD, the dimensionality of our term-document matrix was reduced to from 349 to 50.

Optimally the resulting LSI-matrix should replace the original index in Lucene. This has proven to be difficult, since present and official versions of the Lucene API lack functionality for creating own indices. As an attempt to approximate the effects imposed by LSI, we instead create new pseudo documents. This is done by multiplying cells in the LSI-matrix by a constant factor to simulate term frequencies. Each term in the LSI-matrix is then added to

the pseudo document the number of times it is simulated to occur. This is done for all positive values in the matrix.

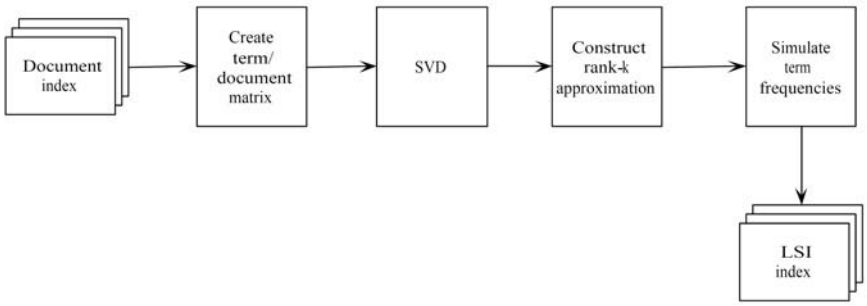

Figure 2. Set up of the term/document matrix. Each document is a document segment from a governing document. In the created matrix, each cell corresponds to the frequency of the term in the corresponding segment. When the matrix is created, the frequencies of the title and its corresponding text are combined, e.g. a document containing the title "my title" and text "my text", then the corresponding frequencies are "my 2","title 1", "text 1".

These new pseudo documents, with term frequencies reflecting the LSI-matrix, are then finally indexed by Lucene. The LSI process is visualized in figure 2.

3.2. Association Rules

Our implementation of association rules mining is aimed at utilizing rules for query expansion. Since process descriptions in the graphical business process models tend to be short, a direct consequence is sparse query-vectors. Fewer words to specify the process, means fewer words to match for relevance in the document index.

If one could augment the query vector based on effective association rules, i.e. rules reflecting semantic associations in the domain of discourse, the probability of a more correct ranking could be increased and one could even get matches in documents lacking the words in the original query.

A general problem when mining text is the high number and varying quality of the extracted rules [Holt and Chung, 1999]. Even though the amount of rules can be limited by adjusting thresholds, the element of "random rules" is not reduced. As an attempt to limit these side effects, two main adjustments from straightforward mining in the original documents are done. First we investigate the use of sections/paragraphs (segments) as units (transactions) instead of mining whole documents. This approach employs the idea of proximity in texts to a higher extent than if the mining is done in larger blocks of text.

The underlying assumption is that words occurring close to each other in a document are more likely to be somewhat related than words co-occuring at

longer distance. Second we perform part-of-speech tagging and execute the association-rule algorithm on reduced text containing only nouns alone, assuming that associations between nouns have a higher probability of reflecting semantic relations than associations between other/different classes of words [Sennelart and Blondel, 2004].

Our implementation is based on the Frequent Pattern Tree algorithm as it is set forth by [Han et al., 2004], and we refer to their article for a detailed description of the algorithm.

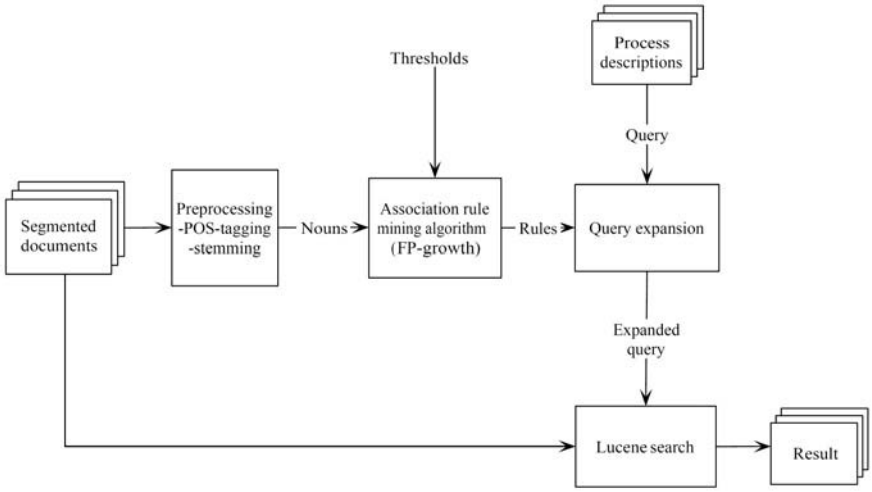

Figure 3. Setup of the association rules implementation.

Our utilization of the extracted association rules is illustrated in figure 3. The text of each section is then run through a part of speech tagger to extract the nouns. The tagger used is a stochastic tagger called QTAG and is implemented by Oliver Mason, University of Birmingham. Before the nouns are fed into the association rule algorithm they are stemmed. The mining algorithm takes as input the segments along with user-defined thresholds. To be regarded a frequent pattern, words must co-occur at a frequency within the range indicated by the thresholds. The output of the algorithm is a set of frequent patterns, or association rules. These are then used for query expansion. That is, during search all the words in the original query are looked up in the set of rules. If a word is found, the query is expanded by the associated word(s). Finally the expanded query is fed into Lucene and the search is executed.

3.3. Synonym and hypernym expansion

Equal concepts may be represented by different terms in the BPM processes and the governing documents. The two latter techniques aim at limit this problem. However, the assumption underlying these methods is that related terms co-occur more often than by chance. This is not always the case.

As an alternative we have implemented an approach that expands the document objects with synonyms and hypernyms from the WordNet lexical database. Synonyms are concepts that have exactly the same meaning, while hypernyms are concepts (nouns) in a "is-a" relation (A is a *hypernym* of B if A is *a* (kind of) B) [Miller, et al., 1993].

As the most expressive terms in the graphical models in the BPM mainly are nouns, we only expand noun terms.

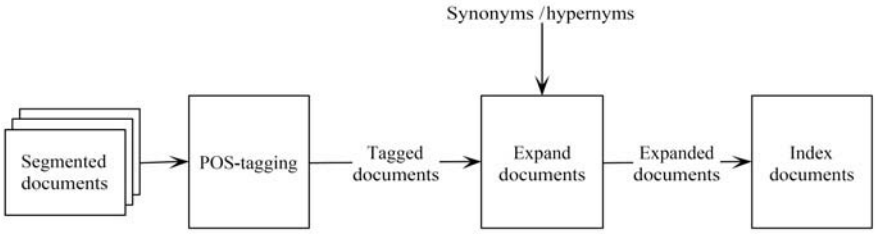

Figure 4. Expansion of noun synonyms and hypernyms from WordNet.

The term expansion is done before the governing documents are indexed. The prototype setup of the expansion is illustrated in figure 4. The following sequence is applied:

1. The documents are part-of-speech tagged to recognize nouns.
2. Each noun is looked up in WordNet, and synonym and hypernym are retrieved.
3. The documents are extended and indexed.

To test the various alternatives we create three separate indices; two for synonyms and hypernyms exclusively and one where both synonyms and hypernyms are included.

3.4. Integrated framework

To ease the task of comparing the different approaches mentioned in this chapter, we have implemented a simple graphical user interface (GUI) for our framework. A screenshot from the GUI is shown in figure 5. At the top of the screen the user can choose which IR technique to apply. When a process in the graphical model is clicked the particular decomposition is visualized and its description used as a search query. The segments found relevant are listed below the model. At the bottom, the user may read any desired segment from the result list.

The GUI is able to display the graphical business process models accompanied with a ranked list of the sections of the governing documents that are found relevant. When selecting a segment in the list, the content is visualized in the bottom of the screen.

Figure 5. A **s**creenshot from the implemented GUI.

4. Evaluation of Results

In our evaluation we used plain Lucene search as a baseline for comparisons. Each technique is evaluated by reviewing the top five ranked items in the result sets. Seven processes from the BPM forms the test data. A Statoil employee familiar with the processes has reviewed the result sets accompanying each process and assigned scores to the result set items according to his opinion of relevance.

The scores $S_{d,expert}$ spans from 1 to 5 where 5 is most relevant. Several segments can be assigned with the same score if they are found equally relevant to the process. To make the comparison explicit we have combined his feedback with the results of the techniques. To do this we use the rank of the five topmost relevant segments of each technique and give them scores $S_{d,technique}$. The most relevant segment gets the score 5, the second gets 4, and so on. We then calculate

a combined score by accumulating the scores of the top five fractions of each technique. To make the comparisons explicit we have combined his feedback with the results of the techniques.

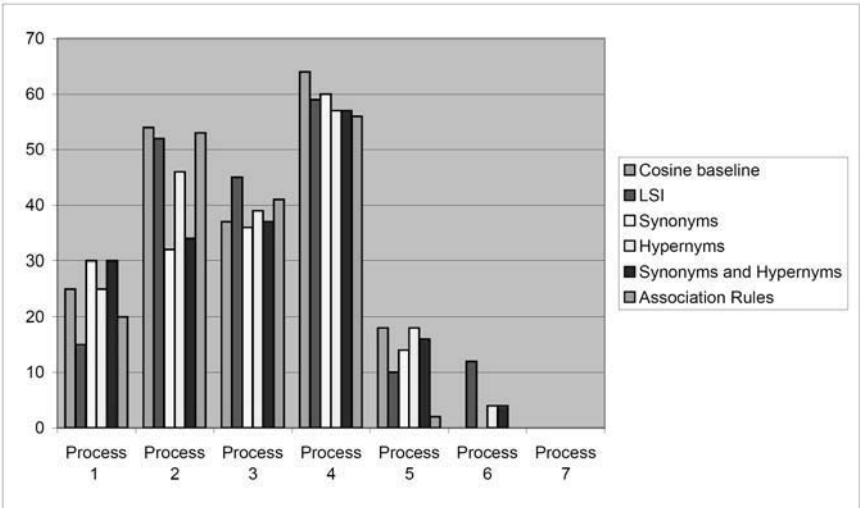

Figure 6. The accumulated evaluation scores.

Figure 6 shows the computed scores of each of the techniques. The first thing that should be noticed is the variance in score between the processes. This reflects the variation in the feedback from the domain expert. In fact, for process 7 none of the fractions were judged as relevant, and naturally all techniques come out with the score 0. Likewise among the fractions of process 6, only two were found relevant limiting the size of the computed score.

On the contrary, for the processes 2-4 more than eight fractions were judged relevant, making possible higher scores. The cosine baseline generally seems to perform well, and when considering the processes as a whole no technique can be said to consistently outperform it. Neither is there any consistent pattern in what is the best technique when disregarding the baseline. If we look at the results on a per process basis, the cosine baseline is outperformed in processes 1, 3 and 6. The difference is most clear for process 3 and 6, both with the LSI implementation as the technique with the highest score.

Precision is defined as the proportion of the retrieved documents which are considered as relevant [Baeza-Yates and Ribeiro Neto, 1999]. Figure 7 shows the precision of the top five fractions of each technique (i.e. percentage of the fractions also judged as relevant by the expert). Again we see that the cosine baseline performs well compared to the other techniques, and that none of the techniques consistently do better than the others. Further, the results conform well to figure 6 in that the processes with high average precision (i.e. processes 2-4) also have high average score.

Even though the extent of these evaluations is not adequate to draw finite conclusions, the following observations can be stated:

- The ability to find relevant information varies from process to process. Some result sets contain highly relevant fractions while others have a total lack of relevant information. Possible explanations are defective techniques, unsuitable process descriptions and lack of documents covering the process.
- None of the implemented techniques seem to consistently outperform the baseline. In defense of the techniques it should be stressed that there are several possibilities of optimization that may lead to better results.
- The main effect of applying document expansion techniques (synonyms, hypernyms and combination) seems to be an alternation of the order and not the contents of the baseline results. This alternation generally seems to reduce the relevance scores compared to the baseline.
- The association rule technique has the most unstable effects of the techniques.
- The LSI implementation is the technique with result sets most different (low overlap) from the baseline. For some processes this gives LSI better results than the other techniques. This may indicate that LSI is capable of identifying relevant information not found by the others in particular cases.

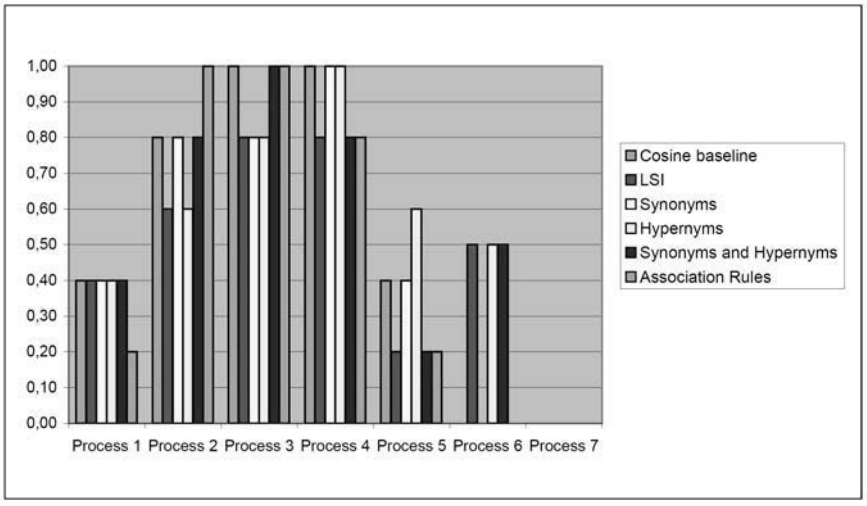

Figure 7. The precision of the top five rank.

5. Discussion

The results described in the previous section can be influenced and improved by several factors. The choice of number of dimensions after the SVD operation and number included association rules are two factors that affect the result. Also other fine tuning attempts can be applied to improve the results and make the retrieval for appropriate for the specific domain.

Another potential for improved retrieval is to make further use of existing business process model information. In the existing implementation, only the hierarchical relationships between aggregated process elements are exploited when the queries are extracted. Here, we could also make extensive use of other flow relationships, like control and resource flow. It is reasonable to believe that two subsequent model elements are somewhat related to the same domain and descriptive governing documents. Such subsequent relationships can be subsequent activities or processes relationships to resources, events and decision points.

In web based search engine solutions the ability to show all relevant documents is neither required nor feasible. Taken into account that the governing documents and the graphical business process models are business critical, it is of major importance that all information that is found to be relevant for a given business process or activity is included in the result set and shown to the user. [Ingvaldsen, et al., 2005] presents one alternative way of displaying relevant information on the governing documents by use of expansive sections and color coding for visualization of relevance estimates. With dynamic linking we are never sure that all relevant information is discovered by the IR algorithms, and the only way to a waterproof solution is by use of manual revision of the result sets.

6. Related work

There exist several attempts to integrate some kind of business process models with underlying knowledge contained in documents, or other sources of information. These efforts show that there are several approaches for such integration. The objective of most attempts is to utilize the models to enable user-support through automatic delivery of relevant information.

EULE [Reimer et al., 2000] is a implemented process-aware retrieval framework. The motivation here is to provide computer-based guidance and support for both novice and experienced office workers performing their day-to-day tasks. The system is integrated with the repository of the corporations' knowledge. Such repositories might be fully or partially computerized. The authors call attention to the importance of making such organizational memories an active system as opposed to a mere passive one, where manual access, e.g. through query interfaces, puts the entire responsibility on the user. Instead the user should be supplied with knowledge, even if he/she does not necessarily know that the information exists. To make this possible the system needs to know what the user is doing. Browsing a workflow diagram or a business process model are examples of interfaces that might provide such input to the system. EULE maintains process-descriptions based on formalized knowledge and data objects encoded in special clauses which make it possible for the system to perform information retrieval, deduction and validity checks just-in-time.

[Abecker et al., 2000] discusses an integration of workflow management systems and agents performing information retrieval (information agents). The KnowMore project is used as an example. The target of this project is, in resemblance with EULE, to support a user performing a specific task with relevant, context-sensitive information without an explicit request from the user. To achieve this, the following features are added to conventional workflow models:

- Extended specifications of complex knowledge intensive tasks, mainly describing the information need of the task, encoded as generic queries.
- Context variables describing the information flow between tasks in the workflow.

These features represent the relevant context of a task and are used to instantiate the generic queries at runtime. Values assigned to the variables must be contained in a domain ontology, making reasoning and thus a more intelligent retrieval possible. When a complex task is reached, the generic query of the specific task is instantiated by the actual context. The query is then shipped to the information agent, which executes the query. Put together the system is claimed to be able to perform ontology-based, situation-specific information retrieval.

A different definition of context is adopted in [Goesmann, 2001]. The article presents an attempt to integrate workflow management systems (WFMS) and organizational memory information systems (OMIS). An implementation, called the KontextNavigator, consists of a process-oriented OMIS, in which the content is organized according to the objects of the workflow system. The context of a process in the WFMS is defined as the set of documents (in the OMIS) containing knowledge relevant to the process. This set is integrated with the WFMS through an event-driven system. When a specific event occurs at a specific process in the WFMS, the context (documents) linked to that event is automatically delivered to the user by the OMIS. How the linking is done is not described in detail. The need for interfaces to browse and store information in different contexts is mentioned by the authors. This indicates that the linking is to be done manually - demanding a considerable human effort.

[Gaizauskas et al., 2004] examine the potential of integrating text mining technology and workflow management systems in the domain of biomedical research. Even though the focus is on a specific domain, the proposed architecture is applicable in a wider range of domains.

The core of their framework is the use of web services, which enables interoperability between applications across the Internet, irrespective of platform and programming language. Three main components constitute the basis of the proposed architecture:

1. A client providing the user interface through which the user may initiate a workflow and browse the results.
2. A workflow server able to execute workflows. A workflow may consist of any number of steps. Each step might involve accessing remote information. Text

mining techniques are exploited to improve the information retrieval, both through query pre-processing and manipulation of the retrieved information.
3. A text database server accessible to external applications through a web services interface. The content might be pre-processed using any suitable techniques.

The workflow server and the text database communicate through a general web services interface. The interface should as a minimum offer basic functionality such as ability to answer queries. The workflow server is responsible of query-creation and response interpretation, thus giving designers of different workflow systems the ability to influence the communication to meet the specific needs of their intended users.

When a workflow-step that demands access to the remote text database is reached, suitable text mining techniques is employed to extract query terms from the outcome of the previous steps along with surrounding context information. The query is then sent to the text database, which retrieves information found relevant to the query and returns the result. The result is then further processed by the workflow system according to the user need. Again text mining techniques are deployed.

Which text mining techniques to be used depends on the structure and amount of the information available to the specific workflow system. It is thus up to the workflow designers to adapt suitable methods to best utilize the information present in their domain.

7. Conclusion and future work

Both governing documents and business process models play important roles in modern enterprises. The standards for running their business operations are laid down in the governing documents by the management, and they expect their staff to follow these policies and guidelines in their daily work. The focus of these documents is on the enterprise's relationships to external entities, like customers or legal frameworks, and they are structured to ensure consistency and completeness with respect to these external aspects. Business process models, on the other hand, are structured by operational people to help the staff carry out their tasks effectively and efficiently. Unfortunately, the dynamic nature of businesses today makes it expensive or even impossible to maintain exact static correspondences between document fragments and models that are subject to continuous changes.

The approach presented here allows us dynamically to relate activities in the business process model to fragments of the governing documents at the time they are needed. This frees the organization from verifying and updating these correspondences whenever a document or business process is modified. It also allows us to find more specific information in the governing documents, as the approach actually retrieves and ranks every fragment or section of the documents

relevant to the activity in the process model. This means that the users can faster check the relevant governing policies when carrying out their activities. Reversely, the management can easier check which processes and activities are affected if they change the content of a (part of a) governing document.

Plain cosine similarity based search and latent Semantic Indexing seems to give the most promising and accurate results for this particular case, but it is evident that further improvements and optimizations are needed before one can assert to what extent the techniques actually improves the integration. Anyhow, the feedback has increased our insight into the techniques performance and future effort will be done to improve existing approaches and to enlarge the scale of our studies.

8. References

[Abecker et al., 2000] Abecker, A., Bernardi, A., Maus, H., Sintek, M., and Wenzel, C. Information supply for business processes: coupling workflow with document analysis and information retrieval. *Knowledge-Based Systems*, 13(5), pp. 271–284.

[Baeza-Yates and Ribeiro Neto, 1999] Baeza-Yates, R. and Ribeiro Neto, B. Modern information retrieval. ACM Press Books.

[Gaizauskas et al., 2004] Gaizauskas, R., Davis, N., Demetriou, G., Guo, Y., and Roberts, I. Integrating biomedical text mining services into a distributed workflow environment. Proceedings of the third UK e-Science Programme All Hands Meeting (AHM 2004).

[Goesmann, 2001] Goesmann, T. KontextNavigator: A workflow-integrated organizational memory information system to support knowledge-intensive processes. INAP 2001, pp. 393-403

[Han et al., 2004] Han, J., Pei, J., Yin, Y., and Mao, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery (8), pp. 53–87

[Hansen, 1987] Hansen P.C., The truncated SVD as a method for regularization, BIT, 27, pp. 534-553.

[Hansen and Jensen, 1998] Hansen, P.C., and Jensen, S.H., FIR filter representation of reduced-rank noise reduction, IEEE Trans. Signal Proc., 46 (1998), pp. 1737-1741.

[Holt and Chung, 1999] Holt, J. D. and Chung, S. M. Efficient mining of association rules in text databases. Proceedings of the eighth international conference on Information and knowledge management, pp. 234–242.

[Ingvaldsen et al., 2005] Ingvaldsen, J. E., Gulla, J. A., Su, X., and Rønneberg, H. "A text mining approach to integrating business process models and governing documents". OTM Workshops 2005, pp. 473–484.

[Miller, et al., 1993] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. Introduction to WordNet: An On-Line Lexical Database, Accessible from: www.cogsci.princeton.edu/~wn/5papers.pdf

[Reimer et al., 2000] Reimer, U., Margelisch, A., and Staudt, M. Eule: A knowledge-based system to support business processes. Knowledge-Based Systems, 13(5), pp. 261–269.

[Sennelart and Blondel, 2004] Sennelart, P. P. and Blondel, V. D. Automatic discovery of similar words. Survey of Text Mining: Clustering, Classification and Retrieval, pp. 25–43.