INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED'07
28 - 31 AUGUST 2007, CITE DES SCIENCES ET DE L'INDUSTRIE, PARIS, FRANCE

ITERATIVE SEMI-AUTOMATIC MODULARIZATION OF DOCUMENTS AND TOPIC MAP CREATION FOR PRODUCT DEVELOPMENT KNOWLEDGE

Harald Weber, Marko Lenhart and Herbert Birkhofer
Darmstadt University of Technology

ABSTRACT

The aim of the current pinngate-project [1] at the product development and machine elements (pmd) department at the Darmstadt University of Technology is a flexible knowledge system for the learning, teaching and application of product development knowledge. Close interaction between the cognitive and the information technical part is essential for its success. The first specifies the requirements for the contents of documents, while the second deals with the modularization of the documents. Both are taken into account by the creation of a semantic net for advanced information retrieval. The problem is finding an appropriate methodology for this task. While manual methods are very work intensive and time intensive, fully automatic methods do not normally provide the required quality. A compromise is a semi-automatic procedure, which supports the user in preparing a document for modularization, textual analyses and integrating it in the system by means of several tools. Further analyses also help the user to identify weak points in the structure of the original documents. The information gained can then be used to improve the documents and their modularization.

Keywords: knowledge management, modularization, semantic nets, text mining

1 INTRODUCTION

Knowledge management systems support users by retrieving the required documents and information in the shortest time possible. To supplement the query possibilities of full text searches, the knowledge units are provided with metadata for classifying and filtering. However, full text and metadata searches are mostly only successful if the user vaguely knows in advance what he or she is actually looking for. E.g. the name of a method must be known in order to request a description of it, but this is often not the case [2]. If the user is aware of related terms or the context of the unknown objective term (e.g. the expected result of the sought-after method), a semantic net is a useful extension of the system. Especially in a complex and ill-structured knowledge domain like product development, this helps to build up a systematic order by presenting the relevant topics and their relations to each other. By means of a graphical user interface, the user is supported in navigating to the prospected subject. Then, he or she is supplied with the documents describing the subject, such as definitions, method descriptions or theoretical background information.

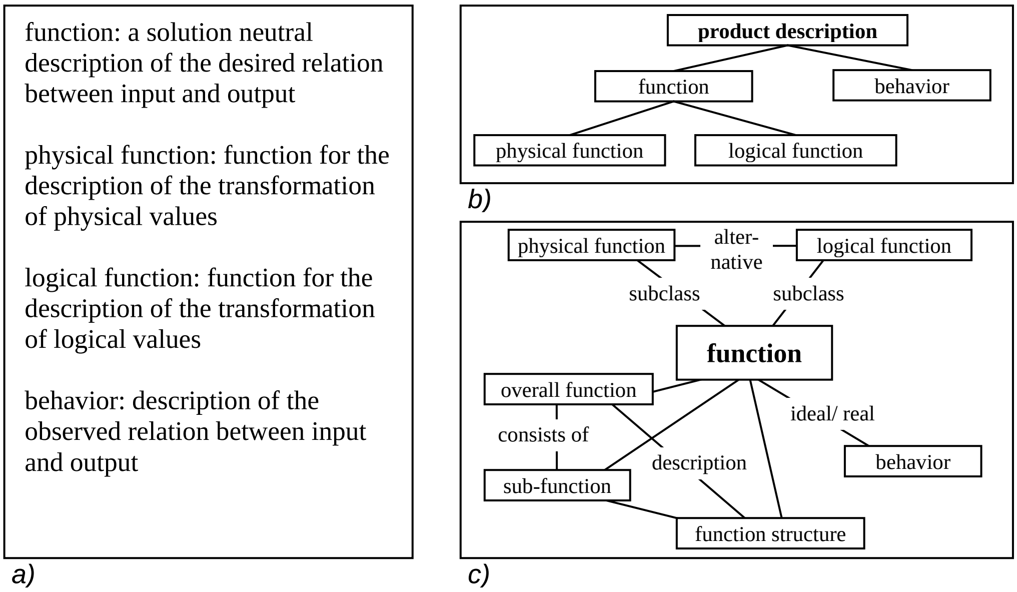
For this reason, the pinngate-system is now to be extended by such a semantic net, based on the international topic map standard (ISO 13250) [3]. The arising problem is the creation of the semantic net, especially on the basis of a high number of subjects/ terms and documents. On the one hand, manual creation is not feasible due to the time expenditure. On the other hand, fully automatic techniques lead to compromised quality. This is why we propose a semi-automatic method. The computer analyzes the available documents and creates the semantic net with the aid of human input in an alternating proposal and correction process. For this, a modularization of the documents is used to gain extra information. But as a consequence, the quality of the modularization influences the value of the automatically generated results, and thus determines the scale of the necessary manual effort.

Now, it happens that an already available topic map can support the modularization and analysis of the documents. Therefore, our approach integrates the document modularization, textual analyses and topic map creation for a mutual positive influence. In the first step, a document is broken down roughly manually, e.g. into paragraphs and chapters. From this, the topic map is created in several steps. Finally, the modularization of the original document is revised. As a result, the modularization of the document and the topic map are adapted and linked to each other.

The following sections of this paper first introduce the basics of semantic nets and topic maps, modularization and text mining methods as stand-alone techniques. Afterwards, the integrative approach is presented.

2 SEMANTIC NETS, TOPIC MAPS

Semantic nets are used to store and visualize terms and their relations to each other. This distinguishes semantic nets from glossaries and controlled vocabularies (vocabularies with eliminated synonyms), which only include terms and partial explanations without relations (Figure 1a). If the terms are arranged in a strict hierarchical order, the result is a taxonomy [4] (Figure 1b). No other information, e.g. about the differences between the levels or relations between two terms at the same stage, is added. This is the benefit of a thesaurus. Here, a standardized set of relation types can be used to label terms as (preferred) synonyms, broader/ narrower or otherwise related ("see also") [5]. The highest level of a semantic net is reached by an ontology where any desired relation type can be used (Figure 1c).



function: a solution neutral description of the desired relation between input and output

physical function: function for the description of the transformation of physical values

logical function: function for the description of the transformation of logical values

behavior: description of the observed relation between input and output

Figure 1. Sample a) glossary with explanations, b) taxonomy and c) ontology.

The most common languages for the formal description of semantic nets are those of the RDF (Resource Description Framework) family (RDF-Schema, DAML+OIL, OWL) and the topic maps which are standardized by ISO 13250 [3]. Each has its own, yet partially similar, data model and different syntaxes. The main differences are due to the different intentions of these languages. While RDF is intended to support the vision of the semantic web by adding metadata to resources in the World Wide Web, topic maps were created to support the information retrieval in a specified set of resources [6]. Because of this and their easy handling, topic maps are used in this approach.

Topic maps consist mainly of three elements [7]:

- Topics
- Associations
- Occurrences

A topic is a symbol for anything the author of the topic map wants to make statements about (rectangles in Figure 1c). Topics can be real objects, like machines or machine parts, as well as imaginary things, like physical effects or design methods. It is only important that each topic is unique. If two topics with different notations ultimately describe the same thing, they have to be merged. This means that synonyms have to be detected. If two topics are not exactly about the same thing, the author has to decide if the differences between them are relevant for the map and justify their separation. Each topic is then provided with several elements, like a unique identifier, one or more names (e.g. for different languages or synonyms), to be displayed to the user, and a type of topic. The last one allows a first classification of the topics in user-defined classes.

Associations describe the relations between the topics (lines in Figure 1c). Like topics, each association has a type predefined by the author ("description", "alternative", "consists of" …). The connections of topics to the associations are implemented by roles. In asymmetrical associations (e.g. "subclass-superclass"), it is important to define which topic has which role, in contrast to symmetrical associations, like "alternative."

As the third main element of a topic map, occurrences build the bridge from the semantic net to any document, dealing with the particular topics. The documents themselves are not part of the topic map. With the help of a mechanism called "reification," associations can be treated like topics and consequently have occurrences. This makes it possible to assign a description of the difference between "function" and "behavior" exactly to the association between these topics.

3 MODULARIZATION OF DOCUMENTS

The aim of a modular document basis is to have small pieces of text, tables and pictures provided with metadata in order to reuse these units in different documents. In this way, memory can be saved and consistency will increase. A change in one text will automatically change all documents in which it is contained.

In previous research, the so-called EMC-model has been developed [8]. In this model, mainly already existing documents are divided into small, formally defined syntactical units as basic parts (elements). From them, the larger semantic units (modules and containers) are composed. Figure 2a) shows the original EMC-model, where modules may only consist of elements and containers are built out of elements and modules. In the current approach, a relaxation of the EMC-model is used by only distinguishing between elements and modules. Elements are plain texts and pictures. Modules are represented by lists of elements and – this is the main difference – other modules. Hence, there is no limitation to a specific number of levels (Figure 2b).

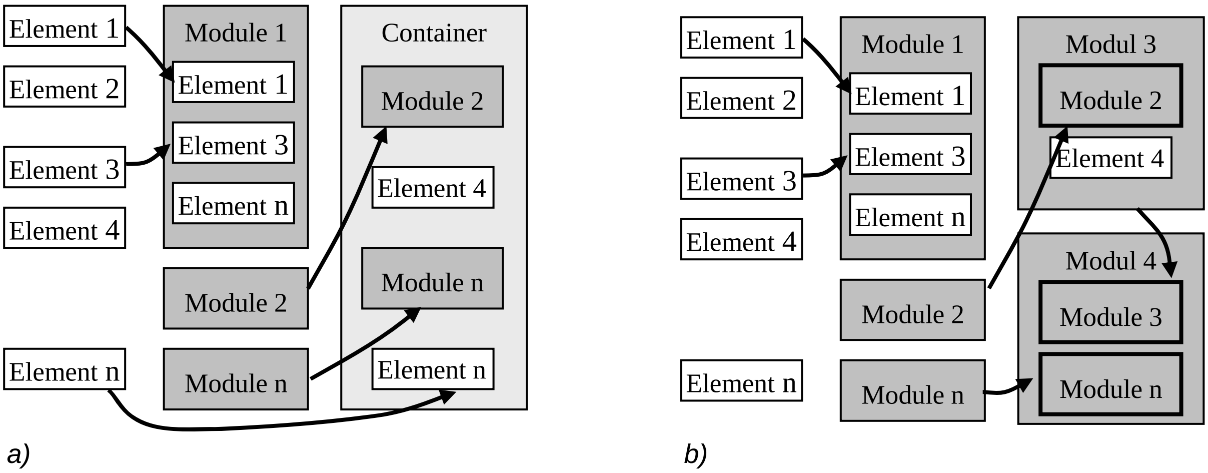


Figure 2. a) Original EMC-model and b) relaxed modularization.

The modularizing of an existing linear document processes in two steps:

- Dividing the document into elements and modules.
- Assigning metadata to the elements and modules.

In the first step, coherent units have to be identified. This does not mean that each paragraph is automatically an element and a chapter a module. Depending on the document and its structure, some paragraphs can be senseless without their neighbors. Only if a part of a document is coherent and individually reusable, does it makes sense to save it as a separate element or module. For example, a caption must not be separated from its picture or table.

Metadata are mainly used to describe the particular type of each element and module (definition, theory, method description, example, etc.) and which concepts are dealt with (keywords). Other metadata, like author, date, version, language and format, may be added.

Issues

Modularizing documents is relatively time intensive without adequate user support. In this case, each element has to be edited manually (possibly with the aid of drag and drop) and the modules have to be composed. That is why a tool to facilitate the modularizing of an existing document has been developed (see section 5.1).

4 TEXT MINING

4.1 Overview

A simple definition for text mining is “data mining from text” [9]. This may suggest only a one-way information flow from texts to data. Terms and relations are extracted from a text and, with this information, a topic map can be built. But this is not the only aim of text mining. Conclusions about the original text are also important. For example, analyses of the most significant terms, and thus topics, dealt with in the text are used for assigning metadata and links between documents. Considering this view, text mining should be understood as the acquisition of information *from and about* a text.
Methods of text mining can be divided into:

- Statistical methods, which analyze term frequencies.
- Linguistic method, e.g. part-of-speech tagging, which analyze the grammatical structure of sentences with the help of vocabularies and syntactical rules.
- Natural language processing, where a semantic understanding of the content by means of a computer is aspired to.

This approach makes use of statistical methods only, because they provide good enough results for our purpose with little effort. Although much research is being done in the field of natural language processing, we are still far from a complete understanding of a text through a computer. Increasing these methods is not the aim of our research. Additionally another problem occurs with domain-specific texts, if no corresponding lexical database containing the relevant technical terms is available.
The statistical text analyses used in this approach can be divided into three steps:

- Topic extraction
- Association extraction
- Text clustering

4.2 Topic extraction

The aim of topic extraction is to ascertain the most important topics of a text or a part of it. Because topics are symbols denoted by (technical) terms and with the assumption that the notations are non-ambiguous, topic extraction is reduced to term extraction here. One of the best known term weighting methods for evaluating the importance of a term is the term frequency - inverse document frequency approach (TF-IDF) [10]. This calculates the ratio of the term frequency in the analyzed document and the total number of documents in the collection which contain this term. It is assumed, that a term is more important for the current document, the more frequently it is used in it and the fewer other documents contain this term. This is consistent with Luhn’s theory, that the most important terms have an overall middle frequency [11]. Hence high frequency words like “the” or “and” (so-called stop words), which have a high rate of occurrence in most documents, are eliminated.
If no extensive document collection for calculating the inverse document frequency is available, an alternative is to use a corpus. This is a large collection of text pieces that provide natural language in a machine readable form [12]. Well-known corpora of the English language are, among others, the British National Corpora (BNC) and the Brown corpus. Because of the current limitation to German documents, the freely accessible text corpus provided by the Leipzig University in the Projekt Deutscher Wortschatz [13] is used.

Issues

One problem is the size of the documents and the distribution of the terms within the document. Let there be a large document, e.g. the current lecture notes for product development used at the

Darmstadt University of Technology. The term “Systemgrenze” (German for “system boundary”) has a relative high frequency in the chapter about product function and it is an important term there. But regarding the whole document, the term has a low frequency, because it is not often mentioned outside the chapter “product function.” If now each chapter is regarded as a separate document, the TF-IDF weighting for “Systemgrenze” will increase. The modularization can partly remedy this deficiency by dividing documents of different sizes into more equally sized elements and modules.

Another problem is the identification of technical terms, if they are widely used in common language or other domains, perhaps with different meanings. For example, consider the different meanings of the term “function” in the domains product development, organization, mathematics and informatics.

4.3 Association extraction

After identifying the most important terms, the collocations, i.e. the joint appearances, of these terms can be examined. Topics whose corresponding terms often occur together are prospective candidates for being parts in an association. Out of the number of common occurrences and the individual term frequencies, a significant measure is calculated e.g. on the basis of a Log-Likelihood Test [14]. This test makes a statement about the independence of the regarded terms. Let p_1 and p_2 be the probabilities, with which two terms appear in one specific piece of text (e.g. sentence or paragraph). Then the expected number of pieces of text, which contain both terms, in a collection of n pieces is $n\ p_1\ p_2$. If the number of observed common occurrences exceeds this value by a significant rate, p_1 and p_2 are assumed to not be independent. Thus the terms, and consequently the topics, are likely to be related to each other. Advanced approaches like Latent Semantic Analysis [15] do not only regard common occurrences but also common absences of terms.

Issues

It is necessary to specify the range in which common occurrences should be counted. This can be a fixed number of words, whole sentences, paragraphs or chapters. An unambiguous proposal for the optimal range size cannot be given. The integrated approach solves this, by taking advantage of the modularization of the document.

Another problem is that purely statistical methods provide no statement about the type of associations found, but linguistic methods fail again because of the domain specificity.

4.4 Text clustering

Clustering in general is the partitioning of a data set into subsets (clusters), so that the similarity within the clusters is maximal and the similarity between the clusters is minimal. In text clustering, similar documents are to be found on the assumption that documents with similar content use the same terms.

One possible method for this task is the document vector model [10]. Hence, each document d_j is represented by a t-dimensional vector $\vec{d}_j = (w_{1,j}, w_{2,j},\ldots,w_{t,j})$, where t is the number of different terms and $w_{i,j}$ is the term weight of term i in document j. The similarity of two documents is then calculated by the cosine of the angle between the corresponding vectors:

$$sim(d_1, d_2) = \frac{\vec{d}_1 \bullet \vec{d}_2}{\left|\vec{d}_1\right| \times \left|\vec{d}_2\right|}$$

If Latent Semantic Analysis is used for association extraction, it can also be used for evaluating the similarities of the documents. After calculating the distance between all documents, a clustering algorithm (e.g. k-means or the EM-algorithm [16]) divides the documents into a predefined number of groups.

Issues

The common clustering algorithms give no answer to the question about the optimal number of groups, which often has to be specified in advance. The more clusters are generated, the greater is the similarity of the documents within one cluster. In the extreme case, each cluster contains one document. Thus, a limit for the minimum increase of the similarity in/ dissimilarity between the clusters when raising the number of clusters has to be set.

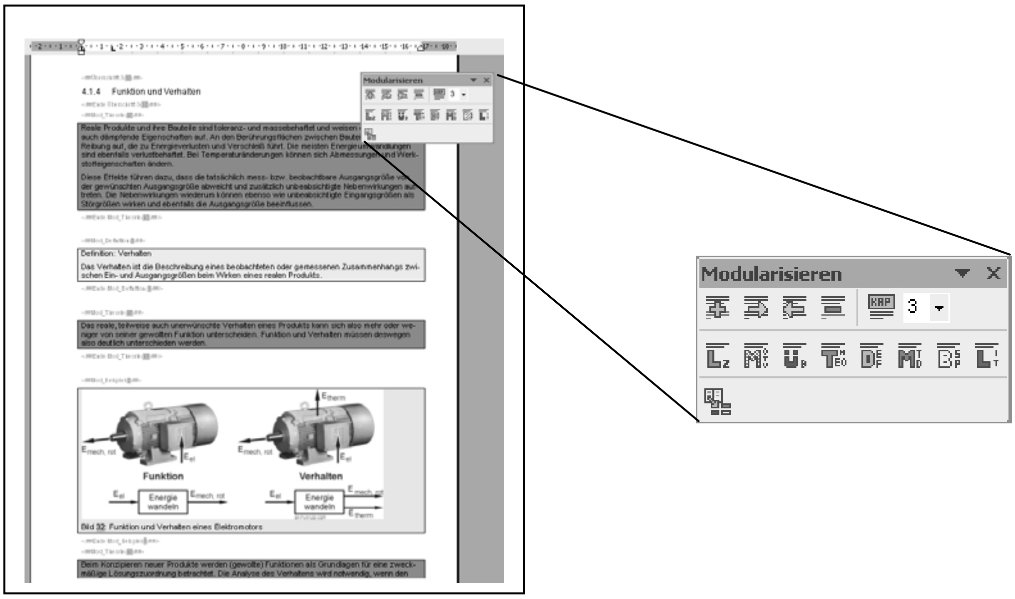
5 THE INTEGRATED SYSTEM

The preceding considerations have detected several issues of the described methods, especially if used as stand-alone techniques. In an integrated approach and with computer-aided assistance, more information generated by one method can be used in the further steps.

The fact that, in the current state of the project, only a single document model is regarded has an essential influence on the analyses. That means only one (more or less extensive) document, which may cover several topics, is examined. As a consequence, the linear structure of the given text is used as additional information.

5.1 Modularization tool

For preparing a document for modularization, a tool has been developed on the basis of Microsoft Word. Any desired text can be copied to a prepared template. With the help of an additional command bar, the user marks the chapters and functions of the paragraphs (Figure 3). Currently, we take the needs of imparting product development knowledge into consideration by distinguishing between educational objectives, motivation, overview, theoretical background, definition, method description, example and literature.



4.1.4 Funktion und Verhalten

Reale Produkte und ihre Bauteile sind toleranz- und massebehaftet und weisen auch dämpfende Eigenschaften auf. An den Berührungsflächen zwischen Bauteil Reibung auf, die zu Energieverlusten und Verschleiß führt. Die meisten Energieumwandlungen sind ebenfalls verlustbehaftet. Bei Temperaturänderungen können sich Abmessungen und Werkstoffeigenschaften ändern.

Diese Effekte führen dazu, dass die tatsächlich mess- bzw. beobachtbare Ausgangsgröße von der gewünschten Ausgangsgröße abweicht und zusätzlich unbeabsichtigte Nebenwirkungen auftreten. Die Nebenwirkungen wiederum können ebenso wie unbeabsichtigte Eingangsgrößen als Störgrößen wirken und ebenfalls die Ausgangsgröße beeinflussen.

Definition: Verhalten

Das Verhalten ist die Beschreibung eines beobachteten oder gemessenen Zusammenhangs zwischen Ein- und Ausgangsgrößen beim Wirken eines realen Produkts.

Das reale, teilweise auch unerwünschte Verhalten eines Produkts kann sich also mehr oder weniger von seiner gewollten Funktion unterscheiden. Funktion und Verhalten müssen deswegen also deutlich unterschieden werden.

Bild 32: Funktion und Verhalten eines Elektromotors

Beim Konzipieren neuer Produkte werden (gewollte) Funktionen als Grundlagen für eine zweckmäßige Lösungszuordnung betrachtet. Die Analyse des Verhaltens wird notwendig, wenn den

Figure 3. Marked paragraphs in the modularization tool.

Three kinds of information are gained during the marking process:

- Boundaries of the elements: An element may consist of one or more paragraphs. Although the tool automatically highlights a single paragraph for marking, the area can be extended to include the following paragraphs or completely manually selected.
- Function of the elements: The defined function of an element is stored as metadata and determines how the element is later treated (e.g. terms in a definition have high priority).
- Structure of the modules: Each marked chapter and sub-chapter results in a module.

After the marking process, each element is saved to a single text file or image. While saving the elements, the modules are stored as so-called “constructor codes.”

The described procedure was applied to 167 pages of the current lecture notes for product development about the concept process. The marking of the paragraphs was done in approx. 90 minutes, which implies approx. two pages per minute. This rate may vary depending on the complexity of the document and how familiar the user is with its content. The following modularization resulted in 830 elements (632 texts and 198 pictures/ tables) and 179 modules.

5.2 Term extraction

The modularized document is the input for the statistical analyses of the term distribution. The complete document is transcribed to a database, so that each word can be accessed individually. In the next step, the significant terms are extracted. But before the frequency of the words in the examined document can be compared with the frequency in the corpus, the different word forms (singular, plural, cases) have to be reduced, in order to sum up their occurrences. One possibility is to use a large lexical database, which contains the stems for different word forms. Here again difficulties arise, if this database does not cover the domain-specific technical terms. This problem is avoided by the use of an automatic stemming algorithm, which contains rules for the removal of suffixes from the words. Most popular is the Porter stemming algorithm [17], which has been adapted to several languages. The German version, according to [18], is implemented in the system. Although an automatic algorithm does often not produce the linguistically correct stems, only minor errors are generated by overstemming (too many letters are deleted from the end of the word) and understemming (too few letters are removed).

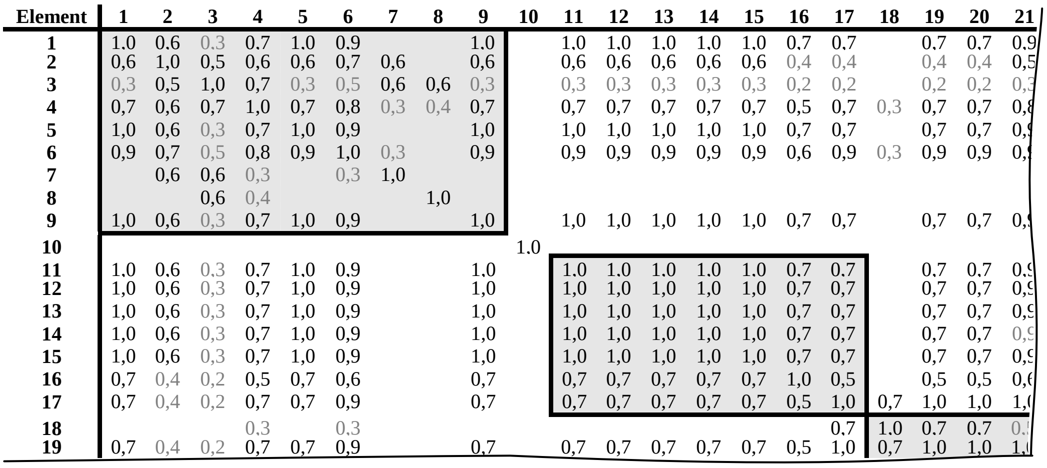
Because of the absence of multiple documents for calculating an inverse document frequency, the downloadable offline version of the German Wortschatz corpus, which consists of about 48 million running words in 3 million sentences, is used. It contains about 1.6 million different words which have been reduced to 1.2 million stems. Each distinct stem in the document is now rated by the ratio of the frequency in the document (total frequency or number of elements/ modules containing the stem) and the relative frequency in the corpus. One positive consequence of using a corpus is that domain-specific terms mostly receive a high weight, since they are seldom used in common language.

The result of the analysis is a ranked list of terms as candidates for becoming relevant topics. This list has to be reviewed by an expert. Because it may be very large (in our example 3,892 entries), some automatic reduction has to be done beforehand. Terms with a low frequency in the document and a low relative rating are eliminated and only the remaining terms are presented to the user. With a second combination of thresholds, a pre-selection is done which reduces the time necessary for the review process.

One risk is overlooking significant terms, which can happen if a text deals with a topic without mentioning the corresponding term very often (e.g. if many pronouns or periphrases are used). In this case, an existing topic map is helpful. If some terms of the topic map can be identified in the document, a closer look at the associated terms is advised. In addition, multiple terms can be assigned to a topic in the map, which can be helpful by treating the problem of synonyms.

5.3 Association extraction, clustering and topic map creation

Clustering, association extraction and topic map creation are closely related in the integrated system. Text clustering is originally used for grouping several documents. Here it is adapted to the modularization approach. First, each element e_i is regarded as a separate document and their similarities are calculated. Second, it is assumed that the (linear) structure of the original document is reasonable. Then the original clustering is reduced to find thematic borders, i.e. changes of topics/ combination of topics in the document. Figure 4 shows an extract of a sample similarity table.



Element	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1.0	0.6	0.3	0.7	1.0	0.9			1.0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
2	0,6	1,0	0,5	0,6	0,6	0,7	0,6		0,6		0,6	0,6	0,6	0,6	0,6	0,4	0,4		0,4	0,4	0,5
3	0,3	0,5	1,0	0,7	0,3	0,5	0,6	0,6	0,3		0,3	0,3	0,3	0,3	0,3	0,2	0,2		0,2	0,2	0,3
4	0,7	0,6	0,7	1,0	0,7	0,8	0,3	0,4	0,7		0,7	0,7	0,7	0,7	0,7	0,5	0,7	0,3	0,7	0,7	0,8
5	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
6	0,9	0,7	0,5	0,8	0,9	1,0	0,3		0,9		0,9	0,9	0,9	0,9	0,9	0,6	0,9	0,3	0,9	0,9	0,9
7		0,6	0,6	0,3		0,3	1,0														
8			0,6	0,4				1,0													
9	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
10										1.0											
11	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1.0	1.0	1.0	1.0	1.0	0.7	0.7		0,7	0,7	0,9
12	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
13	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
14	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
15	1,0	0,6	0,3	0,7	1,0	0,9			1,0		1,0	1,0	1,0	1,0	1,0	0,7	0,7		0,7	0,7	0,9
16	0,7	0,4	0,2	0,5	0,7	0,6			0,7		0,7	0,7	0,7	0,7	0,7	1,0	0,5		0,5	0,5	0,6
17	0,7	0,4	0,2	0,7	0,7	0,9			0,7		0,7	0,7	0,7	0,7	0,7	0,5	1,0	0,7	1,0	1,0	1,0
18				0,3		0,3											0,7	1.0	0.7	0.7	0.5
19	0,7	0,4	0,2	0,7	0,7	0,9			0,7		0,7	0,7	0,7	0,7	0,7	0,5	1,0	0,7	1,0	1,0	1,0

Figure 4. Similarity table with marked high similarity ranges

The following algorithm is used to find text ranges $r(i,j) = \{e_i, e_{i+1},\ldots,e_{j-1},e_j\}$ with a high inner similarity.

1. Rate each possible coherent text range *r(i,j)* by the sum of the similarity of each element to all others in the range, minus the similarities of these elements to the elements next to the range.
2. Mark the range with the maximum rating.
3. Mark the range with the highest rating in the remaining part of the document.
4. Repeat step 3 until each element of the document is marked or the remaining ranges have only non-positive ratings. The latter means that the inner similarity of the elements is not greater than the similarity to the neighbor elements.

After all ranges are marked, the contained terms are analyzed. The term combinations which are mostly responsible for the similarity are used as candidates for association. These are shown to the user for the final selection and the assignment of association types.

The last step is the creation of the topic map from the extracted terms and associations with the identified text ranges as occurrences. If a topic map already exists from the analysis of another document, the contained associations can be used in the clustering process. During the calculation of the similarity measures, term combinations that are part of associations can be favored by a higher rating.

The association extraction and clustering process is, at the same time, a revision of the modularization. The original borders of the modules are replaced by the found clusters, which are finally used in the topic map.

5.4 Didactic rules

The results of the preceding analyses can also be used to increase the quality of the original document. For example, repetitions of the same topic in different chapters can be detected and shown to the author. Then it is his task to decide if the text should be reduced or not, since repetitions are sometimes explicitly desired. Another rule concerns the order in which the terms are used. In most cases, it is sensible to first define terms before they are used to describe methods or used in other definitions. Exceptions can be made in overviews. Then the main terms are mentioned without defining them all, in order to attract and direct the attention of the reader in the succeeding parts of the text.

6 CONCLUSION AND FURTHER WORK

The semi-automatic approach leads to satisfactory results in an adequate time. The significant terms can be extracted from a document and the integrated association extraction and clustering simplifies the topic map construction. This improves the navigation and retrieval possibilities through the complex domain of product development knowledge.

One condition for successful document analyses is that they have to be supervised by an expert in the particular domain, who is familiar with the documents. The statistical analyses are not able to increase the understanding of a document, but they significantly reduce the manual effort. Also, the analyses in connection with didactic rules give valuable information for increasing the quality of a document, and as a consequence, the whole system. Thereby the specialties of product development knowledge with the principal aim of imparting design methods have to be taken into consideration.

Currently, only a single document is regarded at once. In the future, a multiple document model will be developed to analyze the cross connections between documents. One topic will be dealt with in several documents with various main focuses at different levels. This has to be mapped in a consolidated topic map. Although an existing topic map can be used for the analysis of a new document, there is the task of identifying new topics. In addition, if the topic map is changed because of new documents, this can have effects on the older documents (clustering, assignment to topics/ associations), which leads to a continuous, iterative adjustment.

Besides this, an advantage arises when the system contains a high number of documents. In this case, this collection can be used as a domain-specific corpus in addition to the general language corpus.

REFERENCES

[1] Birkhofer, H., Weiß, S., Berger, B., Modularized Learning Documents for Product Development in Education at the Darmstadt University of Technology. In *International Design Conference - DESIGN 2004*, Dubrovnik, Croatia, May 2004, Vol. 1, pp. 599-604.

[2] Ahmed, S., Encouraging Reuse of Design Knowledge: A Method to Index Knowledge. *Design Studies,* 2005, 26(6), pp. 565-592.

[3] ISO/IEC 13250-2:2006, Information technology – Topic Maps – Part 2: Data model, International Organization for Standardization, 2006 (Geneva, Switzerland).

[4] Garshol, L.M., Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. *Journal of Information Science*, 2004, 30(4), pp. 378-391.

[5] ISO 2788:1986, Documentation – Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization, 1986 (Geneva, Switzerland).

[6] Garshol, L.M., Living with topic maps and RDF. In *XML Europe 2003,* IDEAlliance, London, UK, 2003; available online: http://www.ontopia.net/topicmaps/materials/tmrdf.html.

[7] Pepper, S., The TAO of Topic Maps – Finding the Way in the Age of Infoglut. In *XML Europe 2000,* Paris, France, GCA, 2000; available online: http://www.ontopia.net/topicmaps/materials/tao.html.

[8] Berger, B., Jänsch, J., Weiss, S., Birkhofer, H., Modularisation of Product Development Contents as a Basis for a Flexible and Adaptive Use in Learning, Teaching and Practice. In *International Conference on Engineering Design 2003, ICED '03,* Stockholm, Sweden, August 2003, on CD-ROM.

[9] Alonso, O. and Ford, R., *Text Mining with Oracle Text,* Oracle Corporation, Redwood Shores, CA, USA, 2005; available online: http://www.oracle.com/technology/products/text/pdf/10gR2text_mining.pdf.

[10] Baeza-Yates, R., Ribeiro-Neto, B., *Modern Information Retrieval*, 1999 (ACM Press, New York).

[11] Luhn, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 1958, 2(2), pp. 159-165.

[12] McEnery, T., Wilson, W., *Corpus Linguistics*, 2nd edition, 2001 (Edinburgh University Press, Edinburgh).

[13] Quasthoff, U. Richter, M. and Biemann, C., Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genoa, 2006, pp. 1799-1802; available online: http://corpora.informatik.uni-leipzig.de/ download/CorpusPortal.pdf.

[14] Dunning, T, Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics,* March 1993, 19(1), pp. 61-74.

[15] Landauer, T. K., Foltz, P. W. and Laham, D., An Introduction to Latent Semantic Analysis. In *Discourse Processes*, 1998, 25, pp. 259-284.

[16] Weiss, S., Indurkhya, N., Zhang, T. and Damerau, F., *Text mining: Predictive Methods for Analyzing Unstructured Information,* 2005 (Springer-Verlag, New York).

[17] Porter, M.F., An algorithm for suffix stripping. *Program*, 1980, 14(3), pp. 130-137.

[18] Snowball Main Page: *German stemming algorithm*; available online: http://www.snowball.tartarus.org/algorithms/german/stemmer.html.

Contact: Harald Weber
Darmstadt University of Technology
Department of product development and machine elements
Magdalenenstraße 4
64289 Darmstadt
Germany
Phone: +49 6151 16-3378
Fax: +49 6151 16-3355
e-mail: weber@pmd.tu-darmstadt.de
URL: http://www.pmd.tu-darmstadt.de