3.2 Textual Case-based Adaptation using Semantic Relatedness - A Case Study in the Domain of Security Documents.

Andreas Korger [14] and Joachim Baumeister [15]

Abstract. In previous efforts graph-based and textual knowledge representations were combined for the usage in case-based reasoning. This work proposes first steps for this combination in the domain of security documents and similar document classes. We present an approach pre-processing documents for textual case-based reasoning by adapting methods of natural language processing. We propose a method improving a case-based hierarchical similarity assessment for retrieval by introducing the concept of vector space embeddings and semantic relatedness of words and phrases.

Keywords: Case-based reasoning · Textual similarity · Textual case- based reasoning · Vector space embeddings · Semantic relatedness · Graph- based knowledge

1. Introduction

Security documents for public events represent a special class of documents showing a high percentage of regularity. Similar document classes are for instance house rules, law documents like contracts and public calls for tenders. Their buildup and content follows certain constraints. Thus, the inherent knowledge can be conveniently modeled using graph-based representations. Security documents are a naturally available experience knowledge source. Episodic revision after an event yields incremental improvement. The documents subsequently code tacit knowledge collected in the past. On the one hand, we use this textual information to improve and maintain the graph-based knowledge representation and on the other hand we use the graph-based knowledge representation to evaluate the textual content. The progressive knowledge formalization [Bau11] and the implementation of a textual case-based reasoning framework (TCBR) in this domain of security documents is the later goal [Berg02].

[14] Angesagt GmbH, Dettelbachergasse 2, D-97070 Würzburg

[15] denkbares GmbH, Friedrich-Bergius-Ring 15, D-97076 Würzburg
University of Würzburg, Am Hubland, D-97074 Würzburg

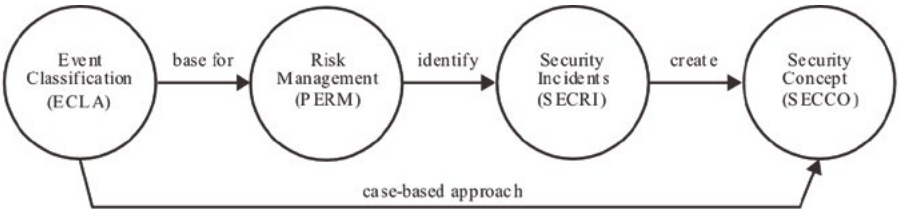

Fig. 1: Classification of events by a classification ontology.

This paper extends previous work [Kor18] that presented a hierarchical classification approach for public events as depicted in Figure 1. Figure 2 shows the manually annotated mapping of a graph-based structure (representing the ontological knowledge) to a security document (representing the textual knowledge). The development and maintenance of this knowledge system required significant efforts of domain experts. Due to limited resources this process has to be supported and facilitated. Closing up to the previous work the first task is to automatically identify the fulfillment of classification characteristics of a public event. For this purpose, we are using the textual content of the related security document. Additionally, we want to save this tacit knowledge for the future support of the generation of new documents. The class of security documents in the domain of public events comes with some distinct characteristics which are described in the following section. Afterwards we show how technologies of information retrieval and extraction can be used to approximate the hierarchical knowledge representation and textual case-based reasoning. We demonstrate our approach by a case study. A similarity assessment for case retrieval by a classification hierarchy is compared to an NLP-approach of entity extraction and cosine similarity of term vectors. Related work will be discussed in the last section.

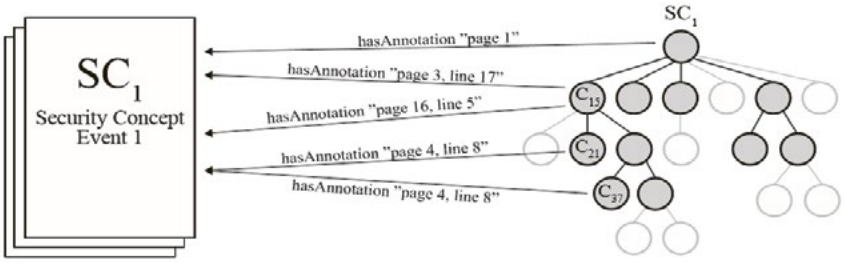

Fig. 2: Annotation of a security document by a graph-based knowledge representation.

2. Characteristics of Security Documents

A security document describes in a certain domain the security incidents that are likely to occur. It shows how to avoid such incidents and gives advice how to react in the case of an occurring incident. An exemplary incident in the domain of public events is a terror threat on a music festival. Additionally, basic information about organizational and environmental parameters is listed: entities that are involved like persons, institutions, and resources, and how these entities interact in certain situations. This information is most commonly presented as a mixture of continuous text, bullet points, tabular data, pictures and figures. It often comes in an arbitrary order not necessarily partitioned into passages headed by a title. Basically, there are few official standards and no strict guidelines for the structure of security documents. In some cases, there is a kind of "ideal" document (exemplary or real world) that holds as a benchmark to measure how "well" a document is written. Those documents are often adapted manually to new scenarios. In this manner, loose standards establish by the common adaptation of the same template document.

The corpus currently available is relatively small summing to about some hundred documents. For reasons of security and data protection security documents are most often published to a limited group of people. This makes the "world of security documents" only partially observable. Thus, techniques that need large amount of (labeled) data cannot be used. We call the set of available and considered documents the *context*.

To a large extend there is no unified domain vocabulary available like, e.g., in the medical domain. There is no domain ontology available, that completely covers the domain vocabulary forming, e.g., a thesaurus. The domain is changing constantly and quickly. New security scenarios arise and have to be mentioned by a document, long before official institutions are capable of giving advice, how to cover it uniformly. For example, christmas markets are exposed to severe terror threats and therefore have a need for different safety measures than some years ago. The vocabulary necessary to adequately describe an event is not closed. Often the vocabulary has to be extended by specific terms of other domains. Unlike for instance in the medical or scientific domain terms often do not have a distinct semantic.

2.1 Recursive Structure of Security Documents

Security documents exploit a recursive character. The underlying structure repeats in the document. The process of the creation of a security document is collaborative and episodic [Bau13]. Many events themselves are episodic and reoccur in certain intervals which reinforces the episodic character of the creational process.

For instance, we consider a classical folk-festival (FF) that takes place every year. The event-site is made up of a big pavilion (BP) and a fairground (FG) surrounding the pavilion with various attractions like fun rides (FR), food sales (FS), and shops (SH). Depending on the organizational importance, influence, and risk-potential of the components, each has to have its own security document. In total they make up the security document

(SD_{FF}) of the actual public event as shown in Figure 3.

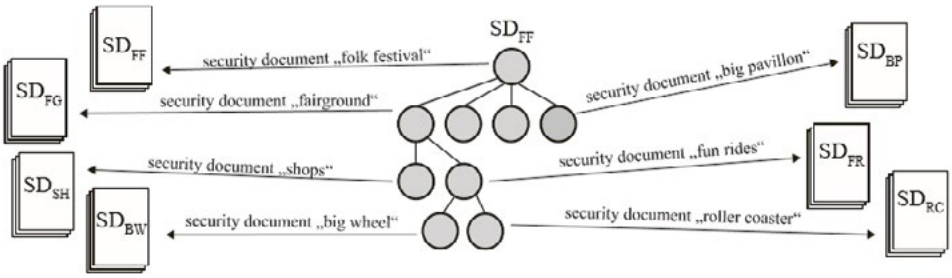

Fig. 3: Knowledge graph showing the hierarchical structure of the security components of an exemplary event.

2.2 Graph-Based Characteristics

In addition to the information coded in the textual corpus there is some structural knowledge available, let that be for instance laws, guidelines, and assessment models. For the facilitation we make the assumption, that any additional knowledge is coded into an ontological structure as a center of multi-modal knowledge representation [Bau14]. All available knowledge resources sum up to a multi-modal knowledge base [Bau11, Mik13]. In the following we introduce formally the scenario for further considerations

Definition 1. *Let SD be the set of all existent security documents. Let $SD_K \subset SD$ be the set of known security documents, $SD_U \subset SD$ the set of unknown security documents, $SD_U \cap SD_K = \emptyset$. Let $SD_C \subseteq SD_K$ be the set of considered security documents also called the context C. Let O_SD be the set of domain ontologies used in SD_C. Let $K_C = SD_C \cup O_{SD}$ be a multi modal knowledge base under the context C. Let $T_K = t_k(O_{SD})$ be the set of ontologically known terms, which are essentially words or phrases of some words (not necessarily contained in SD). Let V_K be the set of known words, the vocabulary contained in SD_K.*

We define a contextual model as a subset of security documents for several reasons. Not all of the known security documents are of the quality to be respected by a knowledge base. Often there is a need for a smaller context, for instance, all security documents of one distinct city. Due to the small corpus there is no need to make any restrictions to the size of the ontology or to introduce a reduced vocabulary set [Mik13]. The experimental corpus that we use for this work consists of 15 security documents. The experimental context of those will be named as C_{15}. In the real world the corpus of security documents is not static. The context is constantly enlarged by new security documents. The context is constantly enlarged by new security documents. Subsequently the vocabulary is enlarged. The set of terms T_C does not cover all elements of the vocabulary contained in SD_C. We therefore need to extend T_C by new elements of SD_C. We also need to maintain O_{SD}. The new terms have to be related to the existing ontological concepts. The scenario of semantic annotation and semantification of the vocabulary is shown in Figure 5. The following definition intro-duces the concept of (partial) term-frequency-vectors for the use in our domain. On this base the semantic relatedness of two elements of the vocabulary is then defined in a first facilitated way as follows. [Kang90, Kor18]

Definition 2. *A document $sd \in SD_K$ is represented by a (full) term vector $x = (x_1,\ldots,x_{|T\_K|})$ of frequencies of elements from T_K in sd. Let $x_p=(x_1,\ldots,x_{|P|})$ be a partial term vector of frequencies of elements of $P \subset T_K$. Let $M_{C\text{-}bow\text{-}5}=T_C \times T_C$ be the co-occurrence matrix of terms in SD_C under the context C, using the bag-of-words-concept and a window of five words before and after an element of T_C. Let $sim_{C\text{-}bow\text{-}5}(t_i,t_j)= M_{ij}$ be an exemplary concept of semantic relatedness of two $t_i,t_j \in T_C$.*

The method "bow" can be substituted by any state-of-the-art co-occurrence-model like e.g. skip-gram. In this manner, the semantic relation of two elements of $_C$ is nudged into a slightly other context with each new security document considered. If all existent security documents where known and considered, a consistent semantic relational model could be calculated using e.g. the *word2vec* methodology [Jones72]. Different context yields different co-occurrences of words. In Figure 4 we see different exemplary context dependent two-dimensional projections of vector space models. Terms are mapped to their counterpart in the fictional global vector space of all existent security documents. The visualization shows, how the semantic relation would change with increasing number of considered security documents.

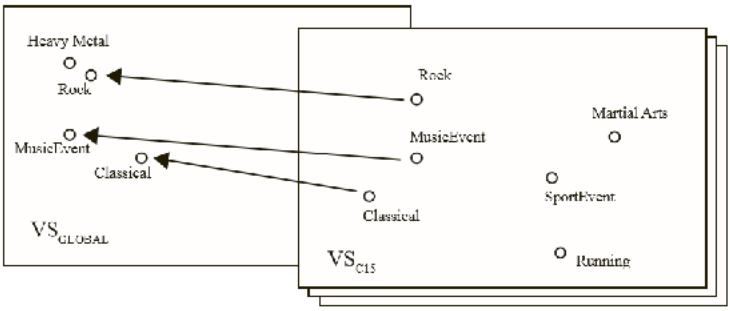

Fig. 4: Mapping of a local specific word-vector space into a partially known global word-vector space.

2.3 Partial Term-Vectors for Adaptation of Case-Based Query

For the safety of an event it is important that the event host has considered the inherent threat of certain scenarios as depicted in Figure 1. For instance, the consumption of alcoholic beverages increases the inherent risk of an event. Subsequently, it is important to know whether a security document considers the topic "consumptions of alcoholic beverages". This stands for the assignation, which characteristics of a document indicate, that a part of an ontological classifier O_{CL} is fulfilled. An exemplary case-based query [Bach12] we used in previous work is q_1 = *PrivateOrganizer* $\wedge$ *Indoor* $\wedge$ *SportEvent*. will most likely not occur in the security document as the text "private organizer". The fulfillment of this classification parameter has to be made accessible by other indicators. In the past, this has been done manually by domain experts as shown by Figure 2. A basic

strategy we propose now is to use partial term-vectors and assign which term-occurrences indicate a fulfillment of an element of O_{CL}. This will surely only hold to a certain grade of detail. The more features [Furth14] are respected by O_{CL}, the more difficult text classification will be. It is easy to extract for instance that a security document covers the topic "alcoholic beverages". What kind of extensiveness of alcohol consumption is expected, is not so easy. In total the structural case-based query has to be translated into a textual case-based query.

Some parts of security documents may be easily generalized and are common for more situations. On the other hand, many parts are very context specific and thus difficult to extract tacit knowledge. Generalization in our model means to enlarge the context under which the information of a document can be used. Figure 5 aggregates the so far made considerations. The process of semantification of new vocabulary elements is shown as well as the partition of the corpus.

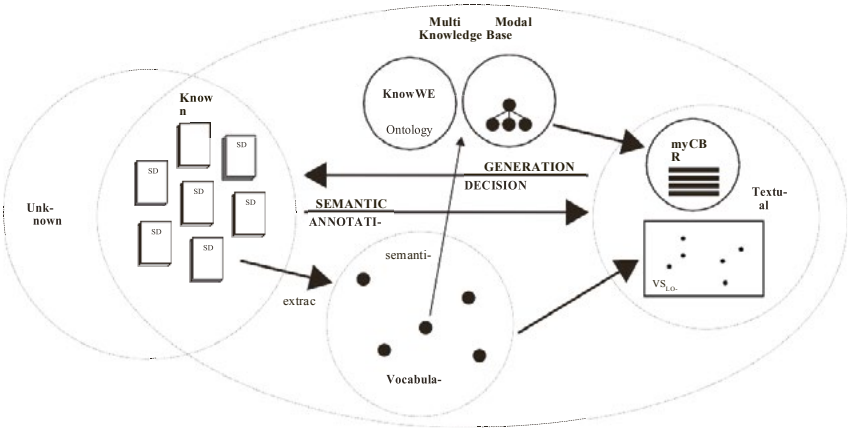

Fig. 5: Integrating textual elements into the semantic annotation workflow.

3. Similarity and Relatedness of Concepts in a Knowledge Organization System

For the ontological representation of security documents we use SKOS [W3C09] as a frame and PROV [W3C13] for the modeling of collaborative and episodic information. Figure 6 shows how SKOS and PROV are merged for the use of structural modeling of security documents.

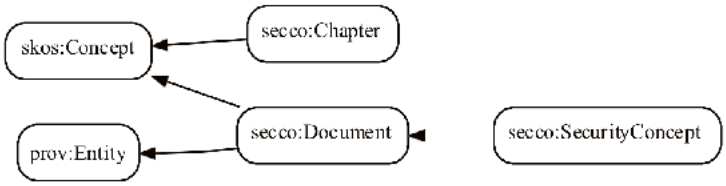

Fig. 6: Integration of SKOS and PROV in the SECCO-Ontology.

The ontological representation of the classification graph can be partially seen in Figure 7. This hierarchical structure is exported to a case-based taxonomy for retrieval purposes.

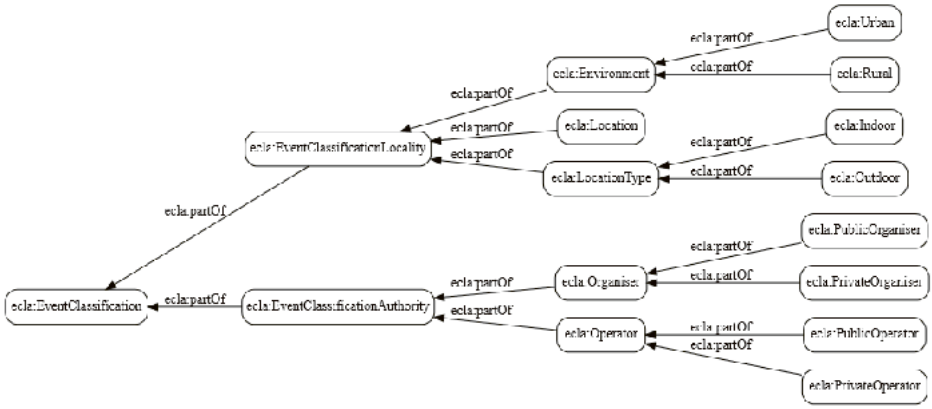

Fig. 7: Excerpt of the event classification hierarchy (16 of 136 concepts).

3.1 Ontology Extension

In a previous work [Kor22] we annotated the security documents with concepts of a manually built knowledge graph. The graph was then exported to different disjunctive case-based classification taxonomies. The weighted and aggregated taxonomies where used for case-based retrieval and adaptation. Figure 8 shows an excerpt of the *EventType* taxonomy and the relation of the classification parameter *running* to (new) elements of the vocabulary. This takes the classification efforts one step further and interlinks the concepts based on information of the corpus. Several issues arise out of the question how textual features can be combined with a knowledge organization system. For instance it has to be considered, whether a new term is represented as an instance of *skos:Concept* or assigned to an existing *skos:Concept* via the property *skos:altLabel*. Actual synonyms can be implemented by the usage of an alternative label. To describe the semantic relation of ontological concepts the schema SKOS provides a variety of predefined properties for instance:

– *skos:semanticRelation*: any relation
– *skos:broader* : upper concept
– *skos:closeMatch*: nearly interchangeable
– *skos:exactMatch*: fully interchangeable
– ...

We assume that "stop-words" are already filtered out of the vocabulary by an appropriate mechanism. We assume that proof for co-occurrence contains natural language processing concepts like "stemming". We propose to implement any new candidates for ontological expansion as concepts named by its "unstemmed" instance. Concepts with at least one co-occurence are linked with *skos:semanticRelation*. *SemanticRelation* is the most unspecific relation and holds because it is approved by the co-occurring in at least one document of the corpus. In a refinement process (left for future work) other measures can be used to change the relation to another type like *skos:broader*, *skos:narrower* or remove the concept and implement it as an alternative label or exact match for interchangeable concepts.

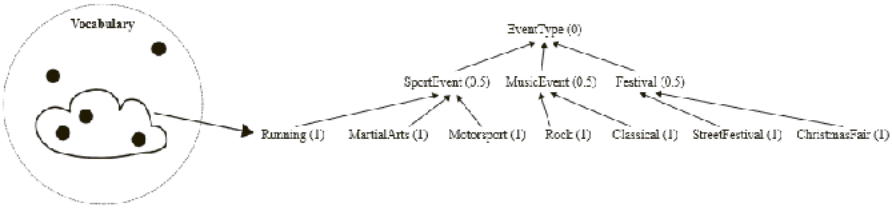

Fig. 8: Mapping vocabulary elements to classification hierarchy.

3.2 Ontological Integration of Textual Elements

We want to save textual elements in the ontology or address textual snippets by their position in the document. Having the application of textual case-based reasoning in mind it is essential to extract and adequately save or point to certain text sections. These informational units may be presented to users for decision support and can be used for the future generation and adaptation of security documents. These text snippets are saved as an instance of *secco:Text*, which is defined as a subclass of *skos:Concept*. This way the textual element becomes an ontological concept itself. *secco:hasTextValue* defined as a sub-property of *skos:semanticRelation*. For explanatory purposes the task is to present adequate text sections to a user if desired. For generation purposes text snippets have to be adapted to a new context [10]. Thereby the convenient substitution of terms by more specific (*skos:narrower*) or more common (*skos:broader*) concepts is a self-suggesting adaptation strategy.

3.3 Retrieval Improvement by Ontological Refinement

The concept of *typicality* can be used to improve the retrieval performance of a case-

based system [Gai15]. This seems very promising for several reasons. The underlying principle is that some subclasses are better representatives for their subsuming class as others. Thus, the reorganization of a classification hierarchy according to typicality improves retrieval and adaptation. For instance, a football or ice-hockey match are typical sport events, a city marathon is a normal sport event and a cross-country walk or boat race are atypical sport events. We assume, that typicality influences the security assessment. Typical events come with typical security measures. Typical security measures lead to typical security documents. We also assume that the security measures of typical events are more easily to adapt than the measures of atypical events. For instance, the security measures for a football match and an ice-hockey match are more similar than those of a cross-country run in the woods. With the use of the textual corpus we expect to have a base for mining of improved typification. A text based approach to measure typicality of security documents is the proof for occurring of certain terms. In a first step these terms are mined manually by domain experts. Figure 9 shows the classification attribute *EventType*. The hierarchy was refined according to the methodology described before.

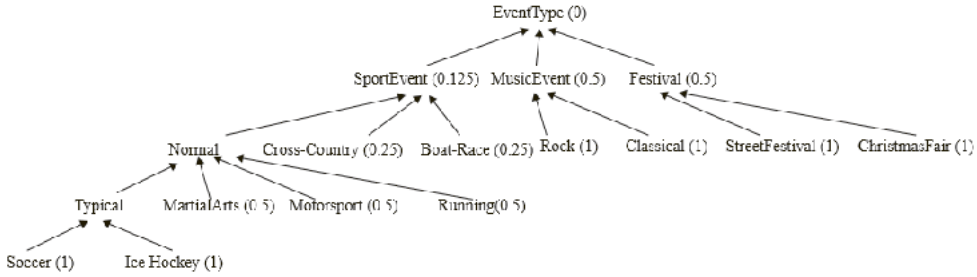

Fig. 9: Classification taxonomy refined according to typicality.

4. Case Study

We demonstrate the presented approach by the continuation of an experiment we did for the evaluation of the case-based event classification structure [Kor18] derived from the classification ontology O_{CL}. In total, 30 real world security documents were collected (all in German language). A corpus was created by manually annotating 15 of the 30 security documents of different events, for instance christmas markets, carnival parades, and city festivals. The process of annotation is very time consuming and has to be supervised by domain experts. In the consequence, due to limited resources, not all of the 30 documents could be annotated. We selected the most meaningful and complete documents covering the domain as good as possible. The coverage of 278 domain specific characteristics were annotated. For the implementation of the ontological representation we used the semantic wiki KnowWE [Bau11b].

	Pages	17	30	31	41	58	72	26	10	64	4	4	9	2	47	16
	Coverage	11,90%	23,70%	23,70%	18,00%	22,30%	14,40%	15,50%	16,20%	30,22%	13,31%	12,23%	13,67%	16,91%	25,18%	23,38%
	Event	christm	wine	wine	folk	city	carne	folk	music	carne	fair	fair	running	camp	arena	campus
Event	Case	ecla0	ecla1	ecla2	ecla3	ecla4	ecla5	ecla6	ecla7	ecla8	ecla9	ecla10	ecla11	ecla12	ecla13	ecla14
christm	ecla0	x	**0.80**	**0.80**	0.65	0.74	0.73	0.71	0.67	0.69	0.69	0.64	0.53	0.67	0.72	0.75
wine	ecla1	0.80	x	**0.95**	0.64	0.76	0.75	0.70	0.71	0.71	0.72	0.68	0.57	0.67	0.73	0.80
wine	ecla2	0.80	**0.95**	x	0.65	0.76	0.75	0.71	0.72	0.72	0.72	0.68	0.58	0.68	0.74	0.77
folk	ecla3	0.65	0.64	0.65	x	0.69	0.78	**0.86**	0.72	0.71	0.73	0.67	0.60	0.69	0.74	0.77
city	ecla4	0.74	0.76	0.76	0.69	x	0.78	0.75	0.76	0.73	0.75	0.69	0.60	0.70	0.77	**0.80**
carne	ecla5	0.73	0.75	0.75	0.78	0.78	x	0.77	0.74	0.75	0.74	0.68	0.59	0.70	0.76	**0.80**
folk	ecla6	0.71	0.70	0.71	**0.86**	0.75	0.77	x	0.72	0.71	0.73	0.67	0.59	0.68	0.74	0.76
music	ecla7	0.67	0.71	0.72	0.72	**0.76**	0.74	0.72	x	0.74	0.68	0.65	0.57	0.64	0.73	0.73
carne	ecla8	0.69	0.71	0.72	0.71	0.73	**0.75**	0.71	0.74	x	0.65	0.64	0.62	0.63	0.72	0.71
fair	ecla9	0.69	0.72	0.72	0.73	**0.75**	0.74	0.73	0.68	0.65	x	0.77	0.55	0.63	0.72	0.72
fair	ecla10	0.64	0.68	0.68	0.67	0.69	0.68	0.67	0.65	0.64	**0.77**	x	0.52	0.60	0.67	0.67
running	ecla11	0.53	0.57	0.58	0.60	0.60	0.59	0.59	0.57	**0.62**	0.55	0.52	x	0.56	0.58	0.57
camp	ecla12	0.67	0.67	0.68	0.69	**0.70**	**0.70**	0.68	0.64	0.63	0.63	0.60	0.56	x	**0.70**	0.69
arena	ecla13	0.72	0.73	0.74	0.74	**0.77**	0.76	0.74	0.73	0.72	0.72	0.67	0.58	0.70	x	0.74
campus	ecla14	0.75	**0.80**	**0.80**	0.77	**0.80**	**0.80**	0.76	0.73	0.71	0.72	0.67	0.57	0.69	0.74	x

Fig. 10: Results of the post mortem analysis of the C_{15} corpus using cosine similarity of term vectors of frequent terms based on a tf-idf-measure.

The corresponding events were ranked by the case-based classification structure using the tool myCBR [Bach14]. Each document respectively public event was considered as a case. For the retrieval we used a similarity function basing on the event classification ontology (ECLA) as depicted in Figure 1. The classification ontology was broken apart into different taxonomies. Those local similarity measures were weighted and combined into a global similarity measure. For the evaluation of this case-based system the technique of "post-mortem analysis" was used. For each case, all other remaining cases in the case-base were ranked by their similarity to the selected case. This process was done for every case in the case-base. The results were compared in a spreadsheet. An analogous ranking of the 15 different events was done by three domain experts. They were asked to inform themselves via e.g. the public events websites and select for each event the three most similar other events concerning security issues and writing the according security documents. The previous scenario is now faced to a text-based similarity measure pairwise comparing the term vectors of the 15 security documents for an analogous post-mortem analysis.

To generate the term-vectors the corpus has to be pre-processed. To bring the documents into a textual format in a uniformed way we used Adobe Acrobat [Ado19] for optical character recognition (OCR) and Apache Pdfbox [Apa19b] for text extraction. For further processing we used Apache Lucene [Apa19]. The tool is an environment for indexing and searching textual documents that comes with some pre-built classification, NLP methods as well as support of German language. The text-files of the corpus are converted to a Lucene-index. The index stores the content of the documents as well as additional information. The Lucene-index is created by stop-word removal and stemming of the remaining words. The term- vector of frequent terms based on a tf-idf-measure [Berg02, Giu10] with term frequencies and position information is stored for each document. In the previous work we recognized that the textual structure and content of security documents can be quite different even if the described situation is very similar. We also

recognized, that the “spelling style” of the author is very influential.

	Pages	17	30	31	41	58	72	26	10	54	4	1	9	2	47	16
	Coverage	11,90%	23,70%	23,70%	18,00%	22,30%	14,40%	15,50%	16,20%	30,22%	13,31%	12,23%	13,57%	16,91%	25,18%	23,38%
	Event	christm	wine	wine	folk	city	carne	folk	music	carne	fair	fair	running	camp	arena	campus
Event	Case	ecla0	ecla1	ecla2	ecla3	ecla4	ecla5	ecla6	ecla7	ecla8	ecla9	ecla10	ecla11	ecla12	ecla13	ecla14
christm	ecla0	x	1-2-3-4	1-2-3-4				0			0	0-1-2-3				
wine	ecla1	1-2-3	x	0-1-2-3-4		2		0			0	1-3				
wine	ecla2	1-2-3	0-1-2-3-4	x				0			0	1-2-3				
folk	ecla3				x	0-1-2	0	1-2-3-4	0-7	0-2	1-3					
city	ecla4				1-2	x	0-2-3	1	0	0-2-3	1		3			4
carne	ecla5		1			0-2-3	x		0-1-2	0-1-2-3			3			4
folk	ecla6	0	0-2	0	1-2-3-4	1-2		x	3		1-3				0	
music	ecla7				3	0-2-4	0-1-2	3	x	0-1-2			3		1	
carne	ecla8		1			0-2	0-1-2-3-4		0-1-2-3	x			3			
fair	ecla9	0-1	0-2	0-2		1-3-4		1			x	0-2-3	3			
fair	ecla10	0-1	0-1-2-3	1-2-3							0-2-3-4	x				
running	ecla11				0	0-2-3	0-2-3		0-2	0-3-4	1		x	1	0	1
camp	ecla12			0		4	4				0-1-2-3	0-2-3	1-2	x	4	0-1-3
arena	ecla13	0			0-2	4	1	0-2	1	1		3	2	3	x	3
campus	ecla14		4	0-4		1-2-4	4				1-3	0-2-3		0-1-3	0-2	x

Fig. 11: Evaluation results comparing case-based ranking, document-based ranking and expert ranking. 0=cbr, 1,2,3=domain experts, 4=term vector.

Figure 10 shows the pairwise post-mortem analysis for the 15 security documents. In an analogous way to the case-based post-mortem analysis we computed the cosine distance for each pair of documents of the corpus where 1 means very similar and 0 means not similar. The outcome of the experiment was discussed in a group of three domain experts. The events *ecla1/ecla2*, *ecla3/ecla6* and *ecla9/ecla10* were written by the same author this expresses as assumed and observed in a very high textual similarity for those security documents. In fact, for security assessment one cannot rely on the high textual similarity. The events *ecla1* and *ecla2* are indeed very similar but due to a significantly higher attendance and some differences in the architectural construction the *event1* comes with different and much higher needs for security measures as the event *ecla2*.

This emphasizes that there is a need for a finer assessment of the event scenario what should lead to an improvement of retrieval as theoretically described before. In the follow up this similarity measure is compared to the case-based similarity as shown in the following Figure 11. The cipher 0 signifies which events were selected as most similar by the case-based ranking. The ciphers 1,2,3 depict the ranking the three different domain experts did for the same events.

For the facilitation of the evaluation we aggregated the ranking of the domain experts and merged them into one by neglecting multiple classifications and just considering whether an event was rated by one of the three experts as can be seen in Figure 12. The comparison with the previous similarity assessment was also discussed with the domain experts. They rated the method suitable for decision support as a supplementary key figure to the case-based approach for retrieval of similar documents out of a corpus. The results where estimated not sufficient for creation and adaptation of security documents as needed in a real world scenario. A clustering of terms into relevant topics leading to topic related text-based similarity measures was proposed as methodological improvement

and refinement.

	Pages	17	30	31	41	58	72	26	10	64	4	4	9	2	47	16
	Coverage	11,90%	23,70%	23,70%	18,00%	22,30%	14,40%	15,50%	16,20%	30,22%	13,31%	12,23%	13,67%	16,91%	25,18%	23,38%
	Event	christm	wine	wine	folk	city	came	folk	music	came	fair	fair	running	camp	arena	campus
Event	Case	ecla0	ecla1	ecla2	ecla3	ecla4	ecla5	ecla6	ecla7	ecla8	ecla9	ecla10	ecla11	ecla12	ecla13	ecla14
christm	ecla0	x	1-2	1-2				0			0	0-1				
wine	ecla1	1	x	0-1-2		1		0			0	1				
wine	ecla2	1	0-1-2	x				0			0	1				
folk	ecla3				x	0-1	0	1-2	0-1	0-1	1					
city	ecla4				1	x	0-1	1	0	0-1	1		1			2
came	ecla5		1			0-1	x		0-1	0-1			1			2
folk	ecla6	0	0-1	0	1-2	1		x	1		1				0	
music	ecla7				1	0-1-2	0-1	1	x	0-1			1		1	
came	ecla8		1			0 1	0 1 2		0 1	x			1			
fair	ecla9	0 1	0 1	0 1		1 2		1			x	0 1	1			
fair	ecla10	0 1	0 1	1							0 1 2	x				
running	ecla11				0	0 1	0 1		0 1	0 1 2	1		x	1	0	1
camp	ecla12			0		2	2				0 1	0 1	1	x	2	0 1
arena	ecla13	0			0-1	2	1	0-1	1	1		1	1	1	x	1
campus	ecla14		2	0-2		1-2	2				1	0-1		0-1	0-1	x

Fig. 12: Aggregated evaluation 0=cbr, 1=aggregated domain experts, 2=term vector.

To improve the text-based classification. The existent corpus was enriched by a list of words or phrases indicating certain classification tendencies. The list was created manually by the domain experts. The integration into the ontological structure was done by defining them as *ecla:Component* a subclass of *skos:Concept* and linking with the property *ecla:classificationRelation* a sub-property of *skos:semanticRelation*. The list was exported to Apache Lucene and queried to the index resulting in the output which documents mentioned each element of the queried list. Due to the pre-processing for the index-creation the concept of stemming was respected for the query. In the following we describe the experimental results by the classification component *ecla:AlcoholAndDrugs* which is a part of the visitors social and demographic behavioral sub-classification as summarized in *ecla:SocialDemographic* and its broader concept *ecla:Visitors*. Words indicating that a security document covers the concept *AlcoholAndDrugs* were e.g. alcohol, drugs, beer, wine, liquor, and drink. The corresponding part of the event-classification ontology can be seen in Figure 13.

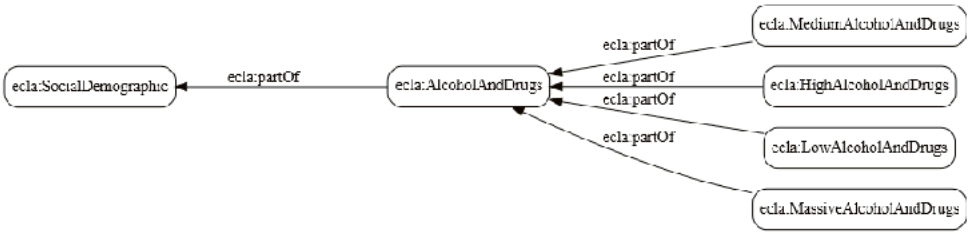

Fig. 13: Excerpt of the event-classification showing the social and demographic parameter *AlcoholAndDrugs*.

The limits of this approach are as follows. The classification attribute *Alcohol-AndDrugs* splits into sub-concepts *Low-*, *Medium-*, *High-* and *MassiveAlcoholAndDrugs*. Just by considering appearance of words it is impossible to decide whether an event belongs to one of those narrower classes. Therefore, a distinct analysis of word-co-occurring or appearance of patterns will be necessary. For instance, to identify phrases like "visitors with massive consumption of alcohol are expected" and distinguish it from e.g. "visitors with massive consumption of alcohol are NOT expected". Additionally, the ontological classification basing on refined typicality could be used. A typical rock concert naturally comes with a high consumption of alcohol, a typical heavy metal concert comes with a massive consumption of alcohol. Same holds for other classification parameters like *Attendance* where it is necessary to extract a number of estimated visitors. The extraction of the event-type showed good evaluation results. A sport event is easy to distinguish from a Christmas fair often only by the name of the event or title of the security document.

5. Conclusions

In this work we presented an approach for the adaptation of a hierarchical case-based retrieval structure. We showed the efforts to integrate natural language processing into the existing case-based classification structure. For the later use of textual case-based reasoning mechanisms ontological structures where established. Aiming for natural language generation of security documents the formalization of the corpus was pushed further.

There is some closely related work we adapted and combined for the needs in this scenario. For the integration of the case-based similarity concept into SKOS we were inspired by the work of Giuliano et al. [Giu10]. They describe the exploitation of lexical substitution of terms in a scenario of similarity assessment. Ground-laying work has been done by them for the acquisition of thesauri from textual data. An issue they consider is the question, how new terms are integrated into a knowledge organization system like SKOS. The development of adaptational changes to the retrieval structure using typicality in a case-based scenario was supported and inspired by Gaillard et al. [Gai15]. A very similar problem statement to ours is reported by Metcalf and Leake [Met18]. They describe a new way of combining structural and textual similarity assessment in the medical domain. Their work also gives a good overview of the corresponding issues of state-of-the-art textual case-based reasoning. Delir Haghighi [Hag13] introduces the ontology DO4MG (Domain Ontology for Mass Gatherings), that describes mass gatherings and case-based reasoning to give decision support for medical emergencies. The author mentions the problem of not having official standards in the domain of mass gatherings. This problem is mitigated by a unified vocabulary covering synonyms to improve case retrieval. She mentions that a classification system can be improved greatly if synonymy is covered. This should hold even more for respecting in general semantic relatedness as also emphasized for ontology population by Furth and Baumeister [Furth14]. Wiratunga et. al.

present basic principles for unsupervised feature selection we adapted for the needs in this domain [Wir06].

What we left for future work is the improvement of the textual classification. We so far focused on entity-based techniques, which shall be extended by relational techniques e.g. relation extraction. To cluster the ontological concepts the usage of an appropriate topic model seems promising. We want to test whether the methodologies of sentiment analysis can be used for distinguishing narrow classification concepts.

References

[Ado19] Adobe: Acrobat: https://acrobat.adobe.com

[Apa19] Apache: Lucene: http://lucene.apache.org/

[Apa19b] Apache: PDFBox: https://pdfbox.apache.org/

[Bach12] Bach, Kerstin, Althoff, K.-D.: Developing case-based reasoning applications using mycbr. In: Agudo, B.D., Watson, I. (eds.) Case-Based Reasoning Research and Development. pp. 17–31. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[Bach14] Bach, K., Sauer, C., Althoff, K.D., Roth-Berghofer, T.: Knowledge modeling with the open source tool mycbr. In: Proceedings of the 10th International Conference on Knowledge Engineering and Software Engineering - Volume 1289. pp. 84–94. KESE'14, CEUR-WS.org, Aachen, Germany (2014)

[Bau14] Baumeister, J., Reutelshoefer, J.: The connectivity of multi-modal knowledge bases. CEUR Workshop Proceedings 1226, 287–298 (01 2014)

[Bau11] Baumeister, J., Reutelshoefer, J., Puppe, F.: Engineering intelligent systems on the knowledge formalization continuum. International Journal of Applied Mathematics and Computer Science (AMCS) 21(1), 27–39 (2011)

[Bau11b] Baumeister, J., Reutelshoefer, J., Puppe, F.: KnowWE: A semantic wiki for knowledge engineering. Applied Intelligence 35(3), 323–344 (2011)

[Bau13] Baumeister, J., Striffler, A., Brandt, M., Neumann, M.: Towards continuous knowledge representations in episodic and collaborative decision making. In: CEUR Workshop Proceedings. vol. 1070 (01 2013)

[Berg02] Bergmann, R.: Experience Management. Springer, Berlin, Heidelberg (2002)

[Furth14] Furth, S., Baumeister, J.: Telesup: Textual self-learning support systems. In: Proceedings of German Workshop of Knowledge and Experience Management at LWA'2014. vol. 1226 (01 2014)

[Gai15] Gaillard, E., Lieber, J., Nauer, E.: Improving case retrieval using typicality. In: Hüllermeier, E., Minor, M. (eds.) Case-Based Reasoning Research and Development. pp. 165–180. Springer International Publishing, Cham (2015)

[Giu10] Giuliano, C., Gliozzo, A.M., Gangemi, A., Tymoshenko, K.: Acquiring thesauri from wikis by exploiting domain models and lexical substitution. In: Aroyo, L., Antoniou, G., Hyvo-

nen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tu- dorache, T. (eds.) The Semantic Web: Research and Applications. pp. 121–135. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)

[Hag13] Haghighi, P.D., Burstein, F., Zaslavsky, A., Arbon, P.: Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings. Decision Support Systems 54(2), 1192 – 1204 (2013)

[Jones72] Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)

[Kang90] Kang, K.C., Cohen, S.G., Hess, J.A., Novak, W.E., Peterson, A.S.: Feature-oriented domain analysis (foda) feasibility study. Tech. rep., Carnegie Mellon University (1990)

[Kor18] Korger, A., Baumeister, J.: The secco ontology for the retrieval and generation of security concepts. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR. Lecture Notes in Computer Science, vol. 11156, pp. 186–201. Springer (2018)

[Luhn57] Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1(4), 309–317 (Oct 1957)

[Met18] Metcalf, K., Leake, D.: Embedded word representations for rich indexing: A case study for medical records. In: Cox, M.T., Funk, P., Begum, S. (eds.) Case-Based Reasoning Research and Development. pp. 264–280. Springer International Publishing, Cham (2018)

[Mik13] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)

[Mik13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013)

[Sal75] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (Nov 1975)

[W3C09] W3C: SKOS Simple Knowledge Organization System Reference: http://www.w3.org/TR/skos-reference (August 2009)

[W3C13] W3C: PROV-O: The PROV Ontology: http://www.w3.org/TR/prov-o (April 2013)

[Wir06] Wiratunga, N., Lothian, R., Massie, S.: Unsupervised feature selection for text data. In: Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A. (eds.) Advances in Case-Based Reasoning. pp. 340–354. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)