Creating Content and Use Cases for Testing Semantic Correlation Rules in Content Delivery

Sophie Bischoff[1], *Jule* Fuchs[1*]

[1]Karlsruhe University of Applied Sciences, Faculty of Information Management and Media, 76133 Karlsruhe, Germany

Abstract. Information retrieval (IR) systems like content delivery portals (CDP) help users to complete processes like searching tasks. However, these systems are facing difficulties like a lack of context or an abundance of content within the required information. An approach to solve this problem is the creation of microDocs with semantic correlation rules (SCR). The objective of this paper is to examine the realization of SCR in CDP according to use cases by focusing on the impact SCR have on the effectiveness of CDP for IR. SCR can be implemented in various creation systems. Furthermore, the paper focuses on the exemplary development of SCR in a component content management system (CCMS). Therefore, a system-based evaluation was conducted. In addition a test collection, following an ontology and a corresponding metadata architecture was created to evaluate the impact of SCR on CDP. The evaluation includes three systems, representing the different types of CDP. It uses methodology precision and recall, as well-fitting methods for the intended purpose. To conclude and finalize, testing results will be summed up, interpreted and the findings will be provided with an outlook.

1 Introduction

In recent years, content delivery has become an important aspect for companies and businesses. The purpose is to distribute information in a more dynamic and specific way than it has been in the past. In order to create technical requirements, it is important to establish logic rule-based networks. This ensures the provision of appropriate knowledge context for each individually selected module. One approach to realize this is the use of microDocs [1].

The technical realization of microDocs is achieved by the definition and implementation of SCR. These rules enhance content delivery concepts to create even more intelligent content [1]. The systems are relevant when referring to a semantic information and metadata architecture and can be divided into three categories: CCMS, Semantic Modelling Systems (SMS) and CDP. For the interaction of the systems, different scenarios are conceivable [2].

Several vendors have started implementing SCR in their systems already. The different types of systems and various levels of integration offer the opportunity to analyze the impact of SCR on the functionality of a CDP [3]. This paper focuses on the scenario of a CCMS with an attached CDP. Figure 1 shows the interaction between two systems after the implementation of SCR. The semantic relations are managed within the system. After implementation, this system is evaluated and compared with two other CDP.

Fig. 1. Interaction of CCMS and CDP after implementation of SCR.

2 Approach and implementation

2.1 Context of research

This study was conducted within the scope of a collaborative project of Karlsruhe University of Applied Sciences, Germany, and the University of Aizu, Japan.

The course semantic information management (SIM) deals with the question of how and to what extent SCR and microDocs are able to increase the intelligence of IR using a CDP. The students' task was to examine the different aspects and potential of SCR.

Simultaneously, a research team from Aizu has been working on a project dealing with smart toilets. For the Karlsruhe project team, those findings worked as exemplary content for smart homes. Using the newly gained research knowledge, an ontology was developed to provide a base for further examination.

[*] Email: biso1012@hs-karlsruhe.de

[**]Email: fuju1012@hs-karlsruhe.de

2.2 Analogies of information retrieval

CDP are a common type of IR systems in technical communication (TC) and enable delivery-based content based on intelligent information. In order to allow context- and target group-oriented indexing, a classification of content is necessary. CDP are able to use content from different creation systems or other document sources [2].

Regarding CDP, microDocs are an approach to provide reduced, but sufficient information realized by SCR. Necessary preconditions for microDocs would be an existing metadata architecture or an additional semantic model. SCR describe correlations between information objects, which can be interpreted dynamically in IR systems [1].

The correlation rules define the relationship between a primary object and several secondary objects. These objects are defined and labelled, using InRule and OutRule classification. The correlation from InRules to OutRules can be characterized as untyped and describes the binding of secondary objects to primary objects within a specific delivery scenario. If the primary information object is displayed, secondarily correlated information objects are displayed in the CDP as microDocs additionally [1].

2.3 Creating a test collection

The initial point of any IR experiment is a retrieval task. To meet this need and provide an experimental environment, a test collection has to be created. Test collections are used to model exemplary users with specific information needs. They consist of three components: a corpus of documents to search, a set of user information needs and relevance assessments [4].

2.3.1 Creating the document corpus

The corpus represents the largest part of the test collection. The collection of documents is the result of an internet research in the field of smart homes. To enable IR, the gathered data consists of different information types such as scientific articles, conference papers and manufacturer information.

To fulfill user information needs, different types of content must be provided. For this reason, different data types, for example, PDFs, websites, images and videos were collected.

To allow relevance assessment, all documents within the corpus were classified manually and based on the ontology developed previously. For unique identification, each document was classified by an ID.

2.3.2 Identifying user information needs

In order to develop microDocs and test the functionality of CDP according to the implemented SCR, the development of use cases is required. To capture even slight differences in performance, user information needs must be constructed carefully. As there was no authentic data available, the use cases were derived from

the document corpus, created beforehand. However, this proceeding is not the common way [4, 5].

Generally, use cases are delivered from the web, sales, service, administration or even support environment. Based on semantic and linked up information, they can then be modeled and optimized analytically [5].

To define user-centered use cases, various aspects have to be considered. Those are relevant and defined for the use cases within this examination and include the definition of the user's role, his or her situation as well as prior knowledge or experience. Additionally, the user's general information needs were derived from the definition of the aspects mentioned above. The information needs represent the actual retrieval task, which can be evaluated in the systems.

Other factors to be defined and considered when developing use cases are, for example, the occurring product situation as trigger or the product lifecycle [5]. Since this study used exemplary content only to test the concept of SCR, those aspects are irrelevant according to the use case definition.

Each of the developed use cases represent one retrieval task. This allows to conduct several individual assessments to gather more significant testing results for each system. Within the use of SCR, relevant topics or documents could be defined. For each retrieval task, one InRule document and a set of correlating OutRule documents were determined. This enables technical implementation of the developed use cases in the tested systems.

2.3.3 Conducting relevance assessment

The relevance assessments specify the importance of each document for the defined user information needs. For each use case all relevant documents of the document corpus are identified. Evaluating every single document in the corpus for the respective use case is crucial to guarantee that all relevant documents are noted, and the evaluation of IR can be carried out completely. Time and budget pressure often make this impossible to implement. Referring to this study, however, the process was feasible due to the limited number of documents. The approach formed a pool of content capturing enough relevant documents to ensure sufficient depth within the IR. Thereby, the relevance assessments provide a base for the evaluation of information retrieval, using recall and precision.

2.4 Implementation in CCMS

Since the main aspect of this study is the creation of SCR in a CCMS, the implementation does not represent a standard proceeding, but points out only one of numerous possibilities to create SCR. The implementation was realized through the web based CCMS Smart Media Creator (SMC) by Expert Communication Systems (ECS).

CCMS are well suited for the development of SCR. In most use cases, classified content modules, an essential aspect for the implementation of SCR, are

already available. Furthermore, implementation in a CCMS is appropriate since SCR are mainly applied in TC and the considered content is typically created in a CCMS.

2.4.1 Implementing content

The content for this study was implemented in the CCMS for the time being. For this purpose, the entire content gathered in the document corpus had to be considered. Since the user should be offered different document types in a search query, the corpus appears with this option. One issue lies within the implementation. At the time of implementation, the CDP is only capable of giving out single modules. Different media types can be assigned to the same semantic classification, but are not considered within the output of the CDP.

The content of the document corpus was, therefore, created in the CCMS as individual modules. Basically, the modularity describes the purpose of creating content in a CCMS. However, for content resulting from the issue of non-functional presentation of other media types, the entire content of a document was created as a module. The content was only used to end the subsequent testing of the SCR.

2.4.2 Creating a metadata architecture

Semantic classification is a key aspect when using SCR to create microDocs. One of the main principles of CCMS is to classify content using semantic metadata to address the individual modules and realize the single sourcing approach. Since these are mainly used for rule-based document creation, the logical consequence is to use this for rule-based linking with SCR as well.

For systems that do not provide a standard classification using metadata similar to a CCMS, a different type of semantic classification is required. Therefore, at the beginning of the SIM course, an ontology regarding smart homes was constructed. The content of the document corpus was classified using this ontology. Since this already provided a full content classification, the metadata architecture for the CCMS was adopted as an exact copy of the ontology.

2.4.3 Creating the ontology

Ontologies and metadata architectures are both semantic classification tools. However, only one of the two approaches is necessary for the use of SCR. The modules in CCMS will be classified using the metadata architecture. Nevertheless, for technical implementation of SCR, ECS has decided to additionally build up an ontology.

To realize SCR, the ontology was also created within the SMC. The system offers several ontological components. Since content and SCR are both classified by different semantic models, the two structures have to be combined. During the creation of the classes, subclasses and instances of the ontology, the elements were assigned an Internationalized Resource Identifier

(IRI). Through this IRI, the elements of the ontology were linked to the corresponding element of the metadata architecture. Figure 2 visualizes this approach.

However, duplicate maintenance of elements should be prevented. Therefore, a different approach would have been to build the SCR using the existing metadata architecture.
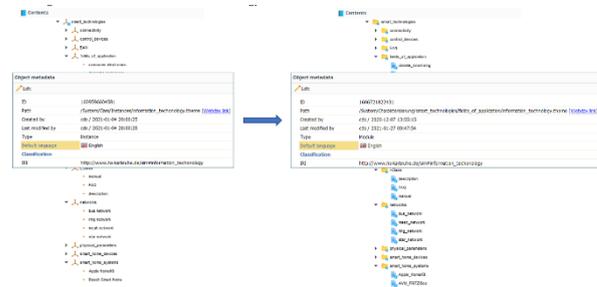

Fig. 2. Linking of metadata architecture and ontology via IRI. (Expert Communication Systems)

2.4.4 Implementing SCR

Semantic classification and content are the basis for SCR to be created. After classifying content and establishing the correlation between metadata and ontology, SCR were implemented into the system. For this purpose, two classes were created, one for the InRules and one for the OutRules. The classes were assigned specific properties, which they inherit to their instances. In this case, those are the individual InRules and OutRules. The class InRule assigns the properties *has correlation* and *select*. Whereas OutRules can also be used as *equals* in addition to *select*.

Using *has correlation*, the primary objects addressed by the InRule are correlated to the secondary objects using the respective OutRule. With the *select* property, classes or instances to be searched for are selected for both, the InRules and the OutRules.

The OutRules class also offers the possibility to set an *equals* relation instead of a *select*. *Equals* allows OutRule to be aligned with a specific class of InRule and automatically adopts the selection from InRule for a particular class.

3 Testing methods

3.1 Comparing of CDP

The existing SCR formalism has already been implemented in several systems in initial research and development initiatives at different levels of integration [3]. This offers the possibility to evaluate and compare different stages of development. In all systems considered, the technical implementation of microDocs is realized by the formal definition of rules in the standardized SCR format. The visualization of microDocs depends on the system. Considering the systems tested and the current stage of development, these are displayed as dynamically generated link lists.

The testing was carried out based on three CDP. All three systems are based on different initial systems. To

be able to obtain significant results about the functionality and effectiveness of the SCR in the individual systems, the same initial situation has been established in all three systems. For this purpose, all documents of the test collection were implemented in the CDP and provided with the same semantic classification.

Since the focus of this paper was not only on testing, but also on the implementation in a CCMS, the remaining systems were provided with content by other members of the research team. The results of the testing process only refer to the impact of the implemented SCR in the CDP. However, in all systems tested, the CDP only forms an extension to the actual creation systems.

The first system to be tested was the CDP from ECS, presented in figure 3. The system is primarily a CCMS to which a CDP can be attached. In addition, the CCMS interface is extended by ontological components. The creation and administration of all content is done in the CCMS itself, while the CDP can only adopt and display the semantic structures and content.

Fig. 3. CDP attached to CCMS. (Expert Communication Systems)

The second system is the Knowledge Graph Platform from I-Views. This ontology system offers the possibility to map complex semantic structures. These capabilities can be transferred to the attached CDP, thus being visualized in Figure 4. The content was imported into the CDP and classified using the ontology created in the Knowledge Graph Builder. The SCR were also developed in the ontology system itself. In the CDP, the semantic dependencies for the representation of the content were retrieved and interpreted [7].
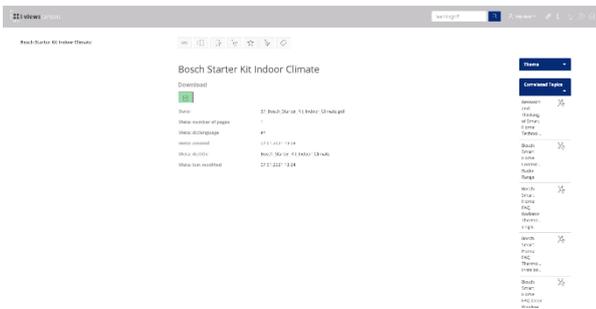

Fig. 4. CDP attached to ontology system. (I-Views)

The third tested system is the ONTOLIS software. Similar to the previous system, this is primarily an ontology system. ONTOLIS is a system for mapping

information models as an ontology with classes, properties, and relations. The system can then be extended by various extensions. Using the CMS and CDP extension, content can first be classified and then be given out to the created semantic structure [8]. The interface of ONTOLIS is shown in Figure 5.
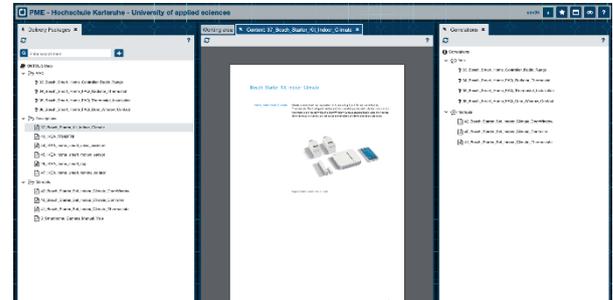

Fig. 5. CDP attached to ontology system. (ONTOLIS)

3.2 Recall and precision

The assessment of effectiveness in the evaluation of IR systems is closely related to the relevance of elements found. Recall and precision describe the result set of one user query by considering the relevance and retrieval of content [6].

The four aspects that need to be considered when evaluating the functionality of SCR in CDP are:

- the number of relevant documents found,
- the number of non-relevant documents found,
- the number of relevant documents not found and
- the number of non-relevant documents not found.

Recall r is considered as a measure to what extent the query exhausts. It describes the ratio of relevant documents found to the total quantity of relevant documents. The possible results of recall are $0 <= r <= 1$. Precision p describes the measure for the target accuracy of the query: the ratio of relevant documents found to the total number of documents found. The result set of precision is $0 < p <= 1$. The closer the result is to 1, the better [4]. In the following, Figure 6 displays the interpretation of recall and precision.
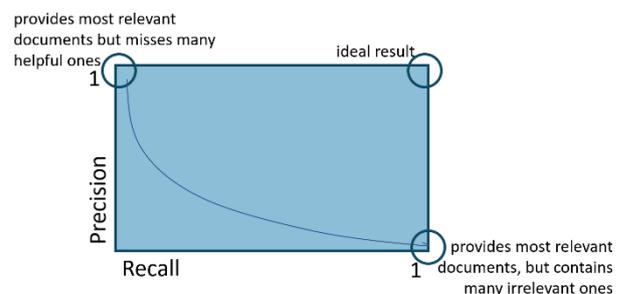

Fig. 6. Visualization of the result set.

3.3 Proceeding

The created test collection describes the experimental environment and the initial point of the conduction of

the evaluation simultaneously. For orientation during testing and to capture testing results, a table was created. This table includes the use cases with the IDs of their corresponding relevant documents, the defined OutRule documents.

The three systems were evaluated simultaneously as well. This proceeding offers the opportunity to test in both directions: system-side implementation issues and modeling inconsistencies can be recognized. Accordingly, it was possible to detect potential error sources, which are self-contained from the implementation of the SCR. This includes, for example, inadequate use case definition, insufficient SCR modelling or incorrect classification, which has to be taken into account during the interpretation process.

For testing the accuracy of IR, all four developed use cases and the resulting retrieval tasks were tested individually and in each system. This allows each use case to be considered and compared individually. For this purpose, the InRule document was selected manually within the systems. After selecting this InRule document, the displayed OutRule documents were examined. The entirety of found documents was listed in the overview table. Afterwards, it was compared with the defined OutRule documents, to register the specific kind and quantity of relevant documents found by the CDP.

Based on the captured data, the recall and precision pair for each use case in each system was calculated. This provides the advantage that the systems can be compared using only two numbers.

3.4 Testing results

At first, it can be determined that the results of the testing show a good elaboration of SCR implementation into the systems. The testing approach offered the important advantage of identifying both, implementation and modelling issues. If the CDP performed contrary to the expectations, the chosen approach made it possible to decide whether this happened due to the implementation of the SCR or if the problem occurred during the development of the test scenario. Both of these reasons occurred.

For the first use case, an ideal was achieved in all three systems. The values for recall and precision are 1. This indicates, that all documents defined as relevant for the use case were found by the CDP and all displayed documents are relevant. Since the values are 1 in all three systems, there is no implementation issue in any of them. The CDP reacted the way they were expected to. Furthermore, it can be assumed that the use case is elaborated in detail and the SCR are defined adequately. All relevant documents were completely classified.

The unexpected test results of use cases 2 and 4 can be attributed to issues that are not linked to the implementation of SCR in the system. When testing these use cases, all relevant documents for the use case were retrieved in all three systems. In addition, one further document was displayed that was not defined as relevant in the use case definition. For this reason, the value of precision is slightly below 1, while recall is 1.

In all three CDP, this was an identical document. When examined in detail, it was determined that the metadata of this document corresponded to the metadata defined for one of the OutRules of the use case. This led to the conclusion, that the reason was not an implementation issue. Accordingly, the reason for finding the non-relevant document can be found in the modeling of the use case and the corresponding SCR or in the classification of the document. Possible solutions would either be to extend the use case to include the displayed document or to adjust the classification of the document for it being no longer addressed by any of the OutRules. In this case, the subject of the document is not necessarily relevant for the use case. Due to the manual classification, the document was tagged with too much metadata. To avoid this, it is important to ensure that only necessary and relevant metadata is assigned when classifying documents.

During the testing of the third use case, the CDP by I-Views achieved the result predicted beforehand. The system retrieved all documents defined as relevant for the use case and there were no irrelevant documents in the result set. The values for recall and precision are 1. This indicated that the modeling of the use case and the classification of the documents were both complete and correct. The other two systems reacted different to the expectation.

The CDP by ONTOLIS also retrieved all documents relevant for the use case, thus the value of recall is 1. In addition, another document irrelevant for this use case was displayed. Therefore, the value for precision is slightly below 1. The reason for this could not be determined. This provides an opportunity for further examination of the implementation.

Additionally, the CDP of ECS behaved differently than expected. The system did not find two documents of relevance for the use case. Therefore, the value for recall is below 1. However, since all displayed documents were relevant for the use case, a precise value of 1 could be achieved. Again, the cause could not be identified, and further investigations will be conducted.

It can be concluded that in total, all three CDP achieved satisfactory testing results. This confirms, that the technical implementation of SCR has already been successful. No fundamental errors were identified in the technical implementation. The differences in the testing results can be attributed to the different stages of development and depth of integration into the systems. This is due to the fact that the system vendors have been working on the implementation of SCR for different periods of time. In general, an unexpected behavior of CDP and the variations in the testing results provide a useful ground for further testing of reference cases in order to identify the actual causes.

4 Findings and Outlook

4.1 Findings

According to this study, the use of SCR can clearly be presented as an advantage in the area of content

delivery. Concerning the regular creation of context reduced to relevant content, SCR can decrease a large administrative effort in TC.

The work of SCR in a CCMS does not represent the standard for implementation. Working with a CCMS had various advantages but also disadvantages. Basically, a major benefit of choosing a CCMS as a creation system is that in this type of system the content is already semantically classified, which reduces the effort required to work with the system. However, in the actual study this factor was not beneficial.

Since the CCMS requested working parallel to a metadata architecture and the created ontology, the creation effort was even twice as high. Maintenance was more complex, as every adjustment had to be executed in two different positions.

In the future, the SCRs in CCMS ought to be created directly using the metadata architecture. Since semantic classification is a precondition for standard creation of documents in CCMS, the logical consequence would be to use this metadata architecture as a semantic classification for the SCR as well. The metadata architecture provides the exact same hierarchical structure and consequently also the relations of one object with another.

In general, it can be captured, that the ontology has been useful for understanding the complex structures of the content in the topic area. However, the creation of an ontology is not mandatory for the use of SCR. This would require a different type of semantic classification to make content available when using SCR.

Another advantage results within the fact that a CCMS is a system for content creation. Some of the other systems tested required more effort by the research team to implement the content in the system.

For the creation of SCR, self-contained from the system environment, explicit definitions are essential. Inaccurately defined preconditions can lead to potential errors in the SCR definition. Unexpected results can also be attributed to the different implementation methods of the system vendors. However, most of the testing issues analyzed are found within use case and SCR modelling.

Inaccurate use case definition was one of the reasons for poor results. However, since use cases were created by the research team and do not represent real use cases, this problem can be reduced significantly for use cases in companies, because mostly authentic use cases are available.

In addition, an exact classification of the media to match content is important. Basically, it can be summarize that both too small and too large classifications significantly reduce the effectiveness of SCR. Therefore, a completely manual classification is recommended. Classification using AI is likely to be more effective for a large dataset.

4.2 Outlook

One main issue is the export feature of SCR to make them available for import in other systems. This would offer the opportunity to develop standardized SCR sets. Furthermore, it reduces workload in the implementation process, because InRules and OutRules can be imported and used immediately. Even within the current state of development, several system vendors are implementing SCR in their systems already.

The future of SCR lies in its further development into a standardized format that can be used in various systems through export and import functions.

5 References

1. Ziegler, W. *Semantic Correlation Rules as a Logic Layer between Content Management and Content Delivery,* Proceedings of the ETLTC ACM Chapter International Conference. Aizuwakamatsu, Japan. (2021)

2. Ziegler, W. *Rules-Based Content Correlations* tcworld online conference, Germany (2020)

3. Ziegler, W. *Regelmäßig Verbindungen schaffen*, Technische Kommunikation, **6**, 18-25 (2020)

4. Carterette, B., Voorhees, M. *Overview Information Retrieval Evaluation* (2010)

5. Ziegler W. *Extending intelligent content delivery in technical communication by semantics: microdocuments and content services*, Proceedings of the ETLTC ACM Chapter International Conference. Aizuwakamatsu, Japan. (2020) https://doi.org/10.1051/shsconf/20207703009

6. Quasthoff, U. *Information Retrieval,* University Leipzig, Institute for Computer Science, 2-15 (2009)

7. Baro, D., Neundörfer, L. *Contrasting Ontology Modeling with Correlation Rules for Delivery Applications,* Proceedings of the ETLTC ACM Chapter International Conference. Aizuwakamatsu, Japan. (2021)

8. Nägele, D., Vobl, P. *Ontology modelling and standardized information exchange with Content Delivery Applications in Technical Communication*, Proceedings of the ETLTC ACM Chapter International Conference. Aizuwakamatsu, Japan. (2021)

9. Ziegler, W. *Man muss auch austeilen können,* Technische Kommunikation*, **04***, 15-21* (2018)