ISSN: 2321-2152

# IJMECE

**International Journal of modern electronics and communication engineering**

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152www.ijmece .com
Vol 6, Issue 1 March 2018

# The Problem with Automatically Annotating Solutions Documents

Shaikh Noor, MdMutte Khan, M  MakaraJyothi

## INTRODUCTION

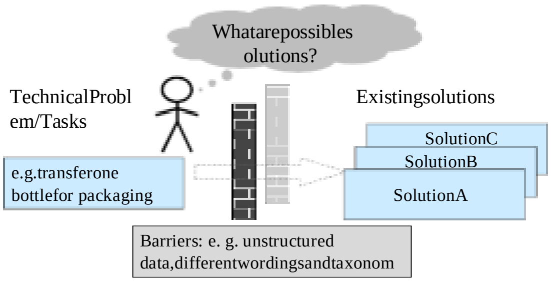
In product development, the access to existing knowledge about previous solutions may reduce the amount of development cycles and conception rework and therewith reduce the efforts of time and costs. Principally, various sources exist for supporting this knowledge. In a study in the German automation industry, sources for the search of existing solution knowledge were identified (Ponn et al., 2006). Besides direct personal communication, organisation-internal knowledge sources (e.g. project folders or databases), construction catalogues, internet portals, and publically available marketing documents were identified as mostly used.

However, an engineer who wants to retrieve existing solution knowledge may face several barriers (see Figure 1). First of all, solution knowledge is mostly unstructured and the access to unstructured data is often insufficient (Blumberg et al., 2003). Secondly, different wordings are used by

the involved developers (Dylla, 1990). This different wording hinders the access via a normal full-text search (Pocsai, 2000). Furthermore, varying taxonomies and classifications due to different viewpoints in sales, marketing, and engineering (Hepp, 2003) contribute to the barrier that hinders the access to needed solutions.

Figure 1: Barriers in the process of retrieving technicalsolutions.

Improving this process of retrieving existing solutions is one of the main goals of the use case




PROCESSUS as one part of the German project THESEUS. Within PROCESSUS, an ontology has been developed, that is used for capturing the knowledge of technical solutions (Gaag et al., 2009). The instances of the ontology and the modelled relations can also be used as a vocabulary for automated annotation of solution documents. This annotation should help in the later retrieval of the documents.

This paper focuses on improving the process of annotating unstructured text data stored in publicly available

Department of Mech
Shaiknoor.jntu@gmail.com, mdmak.khan98@gmail.com, jyothi123@gmail.com
ISL Engineering College.
International Airport Road, Bandlaguda, Chandrayangutta Hyderabad - 500005 Telangana, India.

solution documents of the automation industry. In these documents, companies provide information about previously installed solutions (e.g. a bottling and filling line for beverages). They are mostly used for marketing purposes to give references of previous work. Furthermore, they are useful in generating first ideas how to approach an engineering task.

In engineering design theory, technical solutions can be described by their functions - typically composed of an object and an operation performed

USAGE OF THE ONTOLOGYINTHEPROTOTYPE

Aprototypewasimplementedthatusesthedeveloped ontology(asanOWLontology)tosupport the automated annotation and the subsequentsearchfor solution documents. For the on the object (Ponn et al., 2008). Given a solutiondocumentwithacertainnumberofdifferentfu nctions,anannotationtoolthatidentifiesmostof thesefunctionsbutnotthereallyimportantonesissurel ynotthebestone.Duetotheseuncertainties, generallyappliedmethodsofrankingliketerm

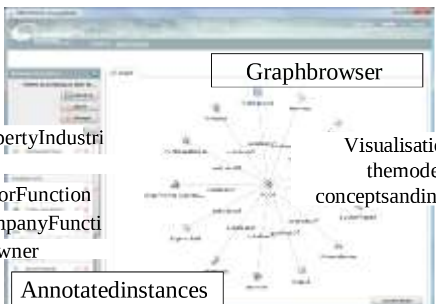
operation

frequencyortheevaluationofannotationswithprecision andrecallcan hardlybeappliedhere.

Toevaluateandimproveannotations,itisnecessary to get a deeper insight into the content ofthe existing solution documents. For this purpose,solutiondocumentsareanalysedbycomparing manualannotationsmadebydifferentpersons.Thesema nualannotationsaremergedandbyapplying ranking numbers the most relevant contentof the document concerning the technical functionsofthesolutionisidentified.Subsequently,thes erankingcanbeusedtoevaluatetheautomatedannotatio n.Thisprocedureisexemplarilytestedwithsixsolutiond ocumentsandappliedonthedevelopedannotationtoolof our prototype.

The paper is organised as follows: First we willprovide a short overview of the ontology (the mainconceptsandtheirrelations)anditsuseinthedevelo ped prototype. The technical functionalities ofthe prototype will not be described in detail and onlyas far as it is relevant for this work. Second, wedescribe our methodology. Then,wewill describethe case study and results in detail. This is followedby a review of related work. We will conclude thepaper with a discussion and summary of our findingsandprovideanoutlookonthenextstepstotake.

sector,etc.).Ontherightside,agraphbrowseroffersthep ossibilitytonavigatethroughtheontologyandaddingfur Figure4givesashortoverviewofthemergingprocedure.




automatedannotation, the ontology serves as a vocabulary andprovides the needed information about the existingrelationsofelementsbelongingtotechnicals olutions. The base structure with the core conceptsof this ontology is shown in Figure 2. The functionhas the central position. It is realised by a technicalsolution, used in a special industrial sector, executedbyafunctionowner,andperformsacertaino peration on a decent object. Existing solutions canbe described by instantiating these concepts with theappropriateinstances.

techn
ical

industrialsector
uses

performs workson

object

Figure 2:Base structure ofthedomain-specificontologyto supportsolutionretrieval.

For the automated annotation, the prototype usesthe label property of the instances in the ontology torecognizetheappropriatewordsandattachthecoresp onding concept to the document. Linguisticfeatures as word stemming and flexion of words areconsidered. Also, linguistic algorithms are supposedto analyze the syntax of a sentence and to determinerelationsbetweenthewordsinasentence.The annotation process will be illustrated by a simpleexemplary sentence“Theconveyorbelttransportstheboxes”takenf romoneofthesolutiondocuments. “Conveyor belt” is the function ownerwhichperformstheoperation“transports”ontheo bject “box”. If these instances are available in theontology, the corresponding concepts are annotated.Withthehelpofthelinguisticalgorithms,thec ombinationof“transport”and“box”inonesentencelead stotheannotationofthefunction“transportbox”.

Figure 3 shows screenshot of this prototype withan exemplary result of an automated annotation. Onthe left side, the annotated instances of a documentarelistedaccordingtotheconceptschosenasa nnotationfilter(propertyofasolution,industrial

therannotationsmanually.

Manualannotation1

| Object | Location inthedocument |
|---|---|
| bottle | line1 |
| bottle | line2 |

Manualannotation2

| Object | Found inannotation |
|---|---|
| bottle | 1and2 |
| bottle | 1 |
| box | 1and2 |

Mergedannotation–2nd step

| Object | Location inthedocument |
|---|---|
| bottle | line1 |

| Object | Found inannotation | Found inannotation |
|---|---|---|
| bottle | 1and2 | 1 |

Figure3:Screenshotoftheprototype.

# 1 ANALYSINGMANUALANNOTATIONS

Thissectionshowstheprocedureofmanuallyannotatingthedocumentsandmergingtheseannotations.Afterwards,theappliedrankingnumbers for the annotated instances and their use fortheevaluationofautomaticannotationareexplained.

## ManualAnnotationofDocuments

Togetadeeperinsightintothecontentofthesolutiondocuments,theyareanalysedbyacomparison of manual annotations. Test participantswere asked to identify all function owners and thecorresponding functions. There was no limitation ofthenumberofmaximumfunctionownersorfunctions annotated in each document. It was alsoallowed to annotate only function owners without acorrespondingfunctionorviceversa.

As a result of the single manual annotation, a setoffunctionownersandcorrespondingfunctions(operationandobject)emerges.Additionally,theposition ofthesourcefortheannotationinthedocument was marked in order to identify where theannotationstemsfrom.

## Merging of the ManualAnnotations

Subsequently,themanualannotationsofonedocument are merged to give an overview over thesimilaritiesanddifferencesofthesinglemanualannotations.Themanualannotationsaremergedaccordingtotheirappearanceinthedocument.

Figure4:Mergingofthemanualannotations.

When,forexamplethesamepartofthedocument is annotated by more than one person (inthis example the objects “bottle” and “box” in line 1and 6 by person 1 and person 2), it is only addedonce to the merged annotation. While merging thedocuments,boththenumberofoveralldifferentannotations of one concept and the number of equalannotation between thesinglemanual annotationsbecomes evident. In the example, the instance “box”wasannotatedoncebytwopersons,whiletheinstance “bottle” was annotated twice (one time bytwopersons, the other timebyone person).

Therewith,themergedannotationscanbeinterpreted as a very precise annotation as possiblymissed annotations of one single annotation can befound inthe annotationofanother person.

## Applianceof RankingNumbers

The number of equal annotations within the singlemanual annotations gives a first impression aboutmajororminorimportantinstances.Whenaninstance of a concept or two instances of two relatedconcepts are annotated by a high number of people,theycanbeinterpretedasimportantforthedocument.Intheexamplepresentedabove,theinstance “bottle” is annotated by two people in line 1but only once in line 2. This indicates that in thesecondline,theonepersondidnotinterpretthe“bottle” in this sentence as an important object forthissolution.

Additionallytothissimplemeasurement,twomore ranking numbers are proposed. These rankingnumbersshowsimilaritiestothetermfrequencyused in information retrieval (Salton et al., 1986). Incontrast to the term frequency, not the importance ofawordinadocument,buttheimportanceofanannotationaccordingtoallannotationsoftherespectivedocumentisfocusedhere.Furthermore, the ranking numbers combine the amount of overallannotationsofaninstancewithinallmanualannotationswith the number of annotations of aninstanceafterthemergingofthemanualannotations. By this combination, the error rate of asinglemanualannotationisdecreasedwhilethe“over all intelligence” of several manual annotationsisincreased.

Thefirstrankingnumber*R(i)*considerstheannotation ofinstancesofsingleconcepts.Itiscalculatedbythemultiplicationoftheoverallnumber of similar annotatedinstances of a singleconcept *N(io)* with the number of similar annotatedinstancesofasingleconceptaftermergingthemanual annotations *N(im)*. To normalise the number,theproductisdividedbytheproductofthemax imumsof*N(io)*and*N(im)*overallinstances(equation1).

## Interpretation

The ranking numbers take values between 0 and 1.These ranking numbers, applied to each instance orrelated instances,give theweighting according totheoverallandmergedmanualannotationsandtherewith the reference for the expected result of theautomatic annotation. The automatic

R(i)=

annotation hastoidentifyatleastthehighestrankedinstances.EspeciallyR(r)canbeusedforevaluatingthequalityoftheannotationofrelated instances.

With the help of the ranking numbers, precision andrecall measures for the evaluation of the automaticannotationcanbecalculatedwithahighergranularity.Itismoreimportanttofindhigherranked instancesthanlower ranked ones.

$$\frac{N(io)*N(im)}{max(N(io))*max(N(im))} \quad (1)$$

## 2 CASESTUDY

Thissectionshowstheapplicationoftheabove

In the example in Figure 4, the ranking numbersarecalculated asshowninTable 1.

Table1:Calculationoftherankingnumbers.

| Object | N(io) | N(im) | R(i) |
|---|---|---|---|
| bottle | 3 | 2 | 1 |
| box | 2 | 1 | 0,33 |

The instance "box" was annotated twice in line 1andonceinline2,sotheoverallnumberofannotations *N(io)* is 3. It is annotated in line 1 and 2which makesthe*N(im)*equalto 2.

The second ranking number *R(r)* - and from theontological point of view the more interesting one - considerstheannotationofinstancesofrelatedconcepts.Similarto*R(i)*,itiscalculatedbythemultiplication of the number of overall annotationsandthenumberofannotationsaftermergingthemanual annotations. This time, the numbers are onlycountedwhentheannotationcontainsapairofinstancesbelongingtoconceptsthatarerelatedinthe

$$R(r)=\frac{N(ro)*N(rm)}{max(N(ro))*max(N(rm))} \quad (2)$$

Therewith,therankingnumberR(r)providesinformation about the mutual annotation of instancesthat are relatedaccordingto the ontology.

overallannotationsaftermerging;thefollowingcolumnsshowthenumberofannotationsoftheindividualma annotations.

Table3:Numberofannotations.

| Concept | all | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Operation | 30 | 28 | 14 | 12 | 14 |
| Object | 27 | 24 | 14 | 9 | 14 |
| Functionowner | 16 | 13 | 11 | 9 | 8 |
| Function | 27 | 24 | 14 | 9 | 14 |

Table5:Comparingwiththeautomatedannotation.

ontology. Once again, it is normalised by themaximumofthesenumbers*N(ro)*and*N(rm)*asshown inequation2.

describedstepsofannotatingandmergingthedocuments. The annotated documents and the resultsoftheannotationsarepresentedandfinallycomparedwiththeautomaticannotationofthedevelopedprototype.Fourpersonsofdifferentbackground(marketing,computerscienceandmechanicalengineering)wereaskedtomanuallyannotatesixsolutiondocumentsconcerningthecontained function owners and their correspondingfunctions.

Thedocumentsdescribetechnicalsolutionsinthe field of automation technology (see Table 2 for ashortoverview ofthecontentofthedocuments).Their length varies between 2 and 8 DIN A4 pagesandtheirnumberofwords lies between 343and912.

Table2: Overviewoftheuseddocuments.

MergingofManualAnnotations

Bytheexampleofonedocument(packagingofmedical tablets), the results of the manual annotationand the merging shall be explained. Table 3 showsthe numbers of annotations of instances of the fourconcepts.Thefirstcolumnshowsthenumberof

nual

| Functionowner | *R(i)* | *Autom.annotation* |
|---|---|---|
| Robot | 1,00 | found |
| Machine | 0,20 | notfound |
| Conveyorbelt | 0,03 | found |
| Barcodereader | 0,03 | found |
| Operator | 0,03 | notfound |

The four participants differ in the number of annotations made. In subsequent interviews it was identified that this can be explained due to the different professional backgrounds. A mechanical engineer did not consider every function as "important". He focused on the core functions. In a subsequent search, he expects these functions to be ranked higher than other functions.

By merging these annotations the number of different annotations of instances of the four concepts can be identified. In this document, 26 different operations, 14 objects, 7 function owners, and 26 different functions were identified.

Table 4 gives an exemplary overview of the instances of the concept "function owner". The corresponding values of *N(io)* and *N(im)* are presented and the resulting *R(i)*-values shown.

Table 4: Annotations of the concept "function owner".

| **Function owner** | ***N(io)*** | ***N(im)*** | ***R(i)*** |
|---|---|---|---|
| Robot | 17 | 7 | 1,00 |
| Machine | 8 | 3 | 0,20 |
| Conveyor belt | 4 | 1 | 0,03 |
| Barcode reader | 4 | 1 | 0,03 |
| Operator | 2 | 2 | 0,03 |

The instance "robot" was annotated 7 times in the document and was mentioned 17 times altogether by the four annotators. This identifies this instance as most relevant for the annotation of function owners.

### Evaluation of the Automatic Annotation

With the help of these ranking, numbers, the automatic annotation can be evaluated. Table 5 shows exemplary which terms have been annotated as functions owners in the document by the automated annotation process. As illustrated, the most important function owner (*R(i)* = 1) has been identified. Nevertheless, some instances have not been automatically annotated.

The results of the evaluation of function owners, operation und object over the six documents were quite similar. Only the annotation of technical functions did not achieve the expected results. This result can be explained by the fact, that the automated annotation. Secondly, the merging of the manual annotation and its later validation is useful for obtaining a set of well-annotated documents for further evaluation of automatic annotations.

The findings of this work can be used in other domains of knowledge where unstructured data has to be annotated using a domain-specific linguistic algorithms do not properly recognise when an object and an operation constitute a technical function.

## RELATED WORK AND DISCUSSION

An overview of general methods and tools for semantic annotation is given by Uren et al. (2006). Uren et al. proposed seven requirements for ontology-supported annotation and evaluated twenty-seven annotation tools. Especially automatic annotation was mentioned as an important field for further improvement. Corcho (2006) compared different annotation approaches (ontology, thesauri and controlled vocabulary) for supporting the process of creating metadata. He identified ontology-based annotation as the most powerful annotation approach concerning the annotation of relations between the instances of a document and also emphasized the meaning of improving automated annotation. A domain ontology as knowledge base for information retrieval is used to improve search over large document repositories by Vallet et al. (2005). In their approach, Vallet et al. also used a label property to identify potential occurrences of instances in the annotated documents. The high amount of work for manually annotating and the following merging make this approach only limited applicable for a larger number of documents and questionable concerning its statistical validation. Furthermore, the influence of the personal background has to be considered when interpreting the results of the manual annotations. Nevertheless, in addition to the identification of ranked instances for the annotation, this approach is twofold useful: First of all, by analysing and verifying the manual annotations, linguistic and syntactic properties of the solution documents can be identified. In a next step, these can be used to deduce typical linguistic schemes (e.g. the syntax of sentences) of solution documents for improving the ontology.

In this context, it has to be considered, where the needed knowledge stored using only instances for the annotation, the ontology could become huge. For example, if every function owner should be part of the ontology, huge classifications or standards have to be integrated. For instance, transferring the products and services categorization standard seCl@

ssinOWLyielded75,000ontologyclassesplusmoretha n5,000properties(Hepp,2006).Alternatively,youmay useacombinationofontologicalknowledgeandlinguisti cpatterns(orrules) for annotation. For example,modelling onlyonthe(technical)operationsintheontologyanddefi ning patterns to annotate a technical function incombination with an identified noun in the sentencewoulddecreasethesizeoftheontology,asthenu mber of technical operations is limited. However,thenumberofrulestobedefinedwillincrease. What works best has to be judged considering therelevant domain and the complexity of the modelledknowledge.

# 3 CONCLUSIONSAN DOUTLOOK

Theanalysisofsolutiondocumentsdoneinthisresearchp ermitsaninsightintothecontentofsolutiondocumentsin thefieldofautomationtechnology. With the help of the proposed rankingnumbers,importantinstancescanbeidentifieda ccordingtothemanualannotationsmadebydifferentper sons.Thisrankingnumberscanbesubsequentlyusedfort heevaluationofanautomatedannotation.Theevaluatio n of theusedprototype showed need for improvement concerningtheannotationofrelatedinstancesintheontol ogy.

Toimprovethisannotation,furtherworkwillfocus on the interpretation of the made analyses foridentifyingpatternsinthesyntaxorlayoutofsolutiond ocuments.Furthermore,thepersonalbackgroundofthe manualannotationswillbeconsideredforthepurposeofi dentifyindividualrequirements on the annotation. This will improvetheautomaticannotationandmayalsobeinstru mental to identifying the “core functions” of atechnicalsolution.

ACKNOWLEDGEMENTS

This work has been funded by the German FederalMinistryofEconomyandTechnology(BM Wi)throughTHESEUS.Theauthorswishtoacknowl edgegratitudeandappreciation toallthePROCESSUS project partners for their contributionduringthedevelopmentofvariousideas andconceptspresentedinthispaper.

# REFERENCES

Blumberg,R.,andAtre,S.(2003).TheProblemwithUnst ructuredData.In *DMReview*, 13(2).

Corcho, O. (2006). Ontology based document annotation:trends and open research problems. *Int. J. of Metadata,SemanticsandOntologies*,1(1),47–57.

Dylla,N.(1990).*Denk-undHandlungsabläufebeimKonstruieren*,PhDthesis, TechnischeUniversitätMünchen.

Gaag, A., Kohn, A., and Lindemann, U., (2009). Function-basedSolutionRetrievalandSemanticSearchinMechan icalEngineering.In*ICED’09,17thInternationalConfer enceonEngineeringDesign*,Stanford, California,USA.

Hepp,M.(2003).*GüterklassifikationalssemantischesS tandardisierungsproblem*.Wiesbaden:DeutscherUniv ersitäts-Verlag.

Hepp, M. (2006). Products and Services Ontologies: AMethodologyforDerivingOWLOntologiesfromInd ustrialCategorizationStandards.*Int.J.onSemanticWeb &InformationSystems*,2 (1),72-99.

Ponn, J., and Lindemann, U. (2008). *KonzeptentwicklungundGestaltungtechnischerProdu kte*.Berlin:Springer.

Pocsai, Z. (2000). *OntologiebasiertesWissensmanagementfür die Produktentwicklung*. PhD thesis,TechnischeUniversität Karlsruhe.

Ponn,J.,Deubzer,F.,andLindemann,U.,(2006).Intellig entSearchforProductDevelopmentInformation-anOntology-basedApproach.In:*DESIGN’06,9thInternationalDesi gnConference*,Dubrovnik, Croatia.

Salton,G.andMcGill,M.J.(1986).*IntroductiontoMode rn Information Retrieval*. New York: McGraw-Hill.

Uren, V., Cimiano, P., Iria, J. e., Handschuh, S., Vargas-Vera,M.,Motta,E.,andCiravegna,F.(2006).Semantica nnotationforknowledgemanagement:Requirementsan dasurveyofthestateoftheart.*Journal ofWeb Semantics*,4(1),14-28.

Vallet, D., Fernández, M., and Castells, P., (2005). AnOntology-BasedInformationRetrievalModel.In*ESWC’05,TheSe manticWeb:ResearchandApplications,SecondEurope anSemanticWebConference.* Heraklion, Crete,Greece.