

A Data Warehouse-Based Approach for Quality Management, Analysis and Evaluation of Intelligent Systems using Subgroup Mining

Martin Atzmueller and Frank Puppe

University of Würzburg
Department of Computer Science
Würzburg, Germany
atzmueller@informatik.uni-wuerzburg.de
puppe@informatik.uni-wuerzburg.de

Stephanie Beer

University-Hospital of Würzburg
Department of Gastroenterology
Würzburg, Germany
beer_s@klinik.uni-wuerzburg.de

Abstract

Quality management, analysis and evaluation of intelligent systems are important tasks. This paper proposes a data mining approach based on the technique of subgroup mining utilizing a data warehouse that contains data from the respective intelligent system to be evaluated and from other external sources. The context of our work is given by an intelligent documentation and consultation system in the medical domain of sonography. For demonstrating the applicability and benefit of the presented approach, we provide several real-world examples of a case-study applying the approach in the medical domain of sonography.

Introduction

Intelligent documentation systems are in wide-spread use, for example, in the service support domain for technical applications, or in the diagnostic domain for medical applications. For the latter, systems that cover both documentation and consultation features are often utilized. For evaluating the input – output behavior, and for quality management, manual evaluations by domain specialists are then usually applied. For these, the documented output of the system is compared with a (manually acquired) gold-standard. Then, the performance of the system can be concisely evaluated, even in complicated cases.

Another important quality parameter that is orthogonal to such evaluations is given by the performance of the users of such a system corresponding to the *input quality*. In order to tackle this issue we need to consider the persons that are entering the input data, e.g., the *examiners* in the medical domain that are documenting a case, before the system provides an output based on this data. If the entered input findings are not consistent, for example, then this may lead to incorrect conclusions that were not accounted for by the designers of the system: The specific situation may not be captured in the applied knowledge base which can significantly decrease the quality of the system. Then, either the knowledge base needs to be extended for this case, or the users need to be notified of such situations. In a way, both evaluations options can complement each other, since effects detected for one option can often lead to findings that were originally caused by the other option, respectively.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Since a purely manual evaluation for such purposes is usually time-consuming and costly, semi-automatic techniques are often a promising option. In this paper, we describe a data mining approach that is based on the implementation of a data warehouse (Kimball and Ross 2002) containing the data from the system to be evaluated and other external sources. The collected data can then be used for various automatic analysis options, and both the evaluation and the quality management objectives can easily implemented. This enables a multifunctional application of the data warehouse which is quite attractive for many domains and also rather cost-efficient. The applied data mining technique is subgroup mining (Wrobel 1997; Lavrac et al. 2004; Atzmueller, Puppe, and Buscher 2005) – a versatile method for data mining and knowledge discovery. The implementation is performed using the VIKAMINE system (Atzmueller and Puppe 2005), an open-source data mining system available at vikamine.sourceforge.net.

In this paper, we describe the application and implementation of the proposed approach in the medical domain of sonography (ultrasound). We present examples from the real-world system SONOCONSULT (Huettig et al. 2004; Puppe et al. 2008), which is a multifunctional knowledge-based system for sonography. SONOCONSULT has been in routine use since 2002 documenting more than 12000 patients in two clinics. An evaluation (Puppe et al. 2008) of the diagnostic accuracy, acceptance, and clinical impact indicated very good results. While the diagnostic accuracy was judged as highly accurate, the comparison with other clinical records showed discrepancies. These can be attributed to the two issues discussed above, i.e., considering the system and/or its users. These considerations motivated the development of the presented approach.

The rest of the paper is organized as follows: We first introduce some background about the medical application context and shortly discuss the SONOCONSULT system (Puppe et al. 2008). Next, we briefly describe subgroup mining as the core data mining technique. After that, we discuss the individual data mining scenarios of the proposed approach, describe the implementation of the technique, and finally provide first exemplary results of their application. Then, we discuss the presented approach. Finally, we conclude the paper with a summary and point out interesting directions for future work.

Preliminaries

In the following, we briefly describe the SONOCONSULT system (Puppe et al. 2008) and sketch its technical implementation. After that, we summarize subgroup mining as the applied core data mining technique.

SONOCONSULT

SONOCONSULT (SC) is a multifunctional knowledge system for sonography, which has been in routine use since 2002 documenting more than 12000 patients in two clinics. The system covers the entire field of abdominal ultrasound (liver, portal tract, gallbladder, spleen, kidneys, adrenal glands, pancreas, stomach, intestine, lymph nodes, abdominal aorta, cava inferior, prostate, and urinary bladder). It was developed with the knowledge system d3web (www.d3web.de), e.g., (Puppe 1998).

The system interacts with the user via dynamic questionnaires for all organs and generates two outputs: a structured report in a standard word processing system for the hospital information system and a data base of all cases for statistical analysis and data mining. An example of a screenshot of the SONOCONSULT dialog is shown in Figure 1. While the middle part shows the questionnaire, the right part shows inferences made by the diagnostic component.

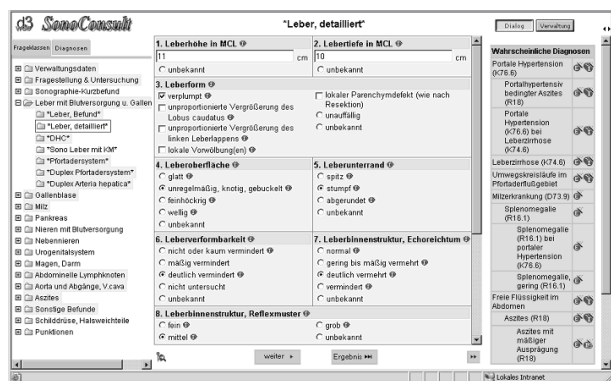


Figure 1: Screenshot (in German) of a section of an SC-questionnaire with part of the hierarchy of questionnaires (partially opened) (left panel) and the currently generated probable system diagnoses (right panel).

The terminology of SONOCONSULT is descriptive and follows that of standard textbooks and publications. Based on the completed questionnaires a textual report (see Figure 2) is generated using a rule based template. The knowledge base makes use of medical heuristics as a knowledge source (McDonald 1996) and was built according to the principles applied for the construction of HepatoConsult (Buscher et al. 2002). SC uses five main concepts: symptoms (input data), symptom classes (questionnaires grouping the input questions), symptom abstractions, diagnoses (output), and rules. Symptoms consist of a pair (attribute, value), e.g. "liver size" (symptom name) and e.g., "increased" (symptom value). In interactive settings, the attributes are questions and the values are the answers by the

user. There are two main types of attributes: choice and numerical. Choice attributes have a predefined range (e.g. for liver size: decreased, normal, increased) and are differentiated according to their cardinality as one-choice (exactly one value is allowed) or multiple choice. Symptoms are grouped into symptom classes if they are requested together most of the time. It is possible to define rules in a symptom class that specify which questions have to be asked in which order depending on the values of previously answered questions. Symptom abstractions are very similar to symptoms except that their values are inferred by rules. They allow a stepwise abstraction of the input data. Diagnoses are also inferred by rules from symptoms, symptom abstractions or other diagnoses. They usually aggregate uncertain evidence.

While d3web allows different reasoning mechanisms for inferring diagnoses, in SC a score-based (fuzzy) reasoning scheme is used, i.e. the rules are assumed to be independent and add or subtract points to the score of a diagnosis, which is rated by thresholds in one of the linguistic categories "probable", "possible" and "unclear or excluded". Rules consist of a condition, an action and exceptions. The condition may be a nested logical combination of criteria, e.g., "and", "or" and "not". Rule actions include, e.g., rating diagnoses, computing values for symptom abstractions, indicating symptom classes and (further) follow-up questions. Exceptions allow to differentiate between two types of negation, i.e., whether a fact is yet unknown or definitely wrong. For more details, see (Puppe 1998).

The diagnostic procedure of SC follows the hypothesis-and-test- and the establish-refine-strategy. The selection of a specific questionnaire (symptom class) depends on the overall clinical question and on the inferred diagnoses. Data gathering stops when (a) the user jumps to the conclusions or (b) all suspected diagnoses (category "possible") are either "probable" or "unclear or excluded" by means of the program's expertise or (c) there are no useful questionnaires left for clarification.

Age: 75; female
Clinical problem: uncertain abdominal complaint; gallstone disease
Findings: date 01/07/06; good condition of examination
Liver: height in medio-clavicular axis 10 cm; depth in MCA 11 cm; regular shape; smooth surface; caudal margin in shape of an acute angle; elasticity not or almost not reduced; structure of slightly to moderately elevated echogenicity; intermediate reflex pattern; regular vessel structure of the liver
Common bile duct: diameter 8.5 mm; regular
Gallbladder: normal size
Spleen: longitudinal size 8.5 cm; depth 3 cm; regular size, shape and structure
Portal system: diameter of portal vein 8 mm, regular
Pancreas: head and body visible; body diameter 1.3 cm; general structure of homogeneously increased echogenicity; duct system regular
Kidney: right kidney: orthotopic position; length 9 cm; thickness of parenchyma 1 cm; normal echogenicity of parenchyma; scarred retraction; pylon not dilated, not measured; regular calices left kidney: orthotopic position; length 9.5 cm; thickness of parenchyma 1.4 cm; normal echogenicity of parenchyma; regular renal structure; scarred retraction; pylon not dilated, not measured; regular calices
Abdominal aorta: partly visible; diameter 1.6 cm; regular
Vena cava: regular
Lymph nodes: in visible regions not detectable or not enlarged
Urinary bladder: not judgeable (insufficiently filled)
Gynecological tract: Uterus: not examined

Figure 2: Part of a generated exemplary SC-report.

For data mining, SONOCONSULT applies the subgroup mining tool VIKAMINE for knowledge discovery and quality control. We will discuss subgroup mining in more detail in the following section.

Subgroup Mining

In the following, we introduce subgroup mining, e.g., (Wrobel 1997; Klösgen 1996; 2002; Atzmueller, Puppe, and Buscher 2005), for discovering interesting patterns. Subgroup mining is a subfield of the general data mining approach. The mined subgroup patterns, often provided by conjunctive rules, describe 'interesting' subgroups of cases, e.g., "the subgroup of 16-25 year old men that own a sports car are more likely to pay high insurance rates than the people in the reference population." The main application areas of subgroup mining are exploration and descriptive induction, to obtain an overview of the relations between a target variable and a set of explaining variables, where variables are attribute/value assignments.

The exemplary subgroup above is described by the relation between the (explaining) variables (Sex = male, Age \leq 25, Car = sports car) and the (target) variable (Insurance Rate = high). The explaining variables are modeled by selection expressions on sets of attribute values. A subgroup pattern is thus described by a subgroup description in relation to a specific target variable. In the context of this work we focus on binary target variables.

Let Ω_A denote the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. An attribute-value assignment $a = v$, where $a \in \Omega_A, v \in dom(a)$, is called a *feature*. We define the feature space \mathcal{V}_A to be the (universal) set of all features. A single-relational propositional *subgroup description* is defined as a conjunction

$$sd = e_1 \wedge e_2 \wedge \dots \wedge e_n$$

of (extended) features $e_i \subseteq \mathcal{V}_A$, which are then called selection expressions, where each e_i selects a subset of the range $dom(a)$ of an attribute $a \in \Omega_A$. We define Ω_{sd} as the set of all possible subgroup descriptions. The subgroup size $n(s)$ for a subgroup s is determined by the number of instances/cases covered by the subgroup description sd . For a binary target variable, we define the true positives $tp(sd)$ as those instances containing the target variables and the false positives $fp(sd)$ as those instances not containing the target variable, respectively.

A quality function measures the interestingness of the subgroups and is used to rank these. Typical quality criteria include the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size.

Definition 1 (Quality Function) *Given a particular target variable $t \in \mathcal{V}_A$, a quality function $q : \Omega_{sd} \times \mathcal{V}_A \rightarrow R$ is used in order to evaluate a subgroup description $sd \in \Omega_{sd}$, and to rank the discovered subgroups during search.*

Several quality functions were proposed (cf. (Wrobel 1997; Klösgen 2002; Lavrac et al. 2004; Atzmueller and Puppe 2006)), for example, the functions q_{BT} and q_{RG} :

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}},$$

$$q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}, n \geq T.$$

p denotes is the relative frequency of the target variable in the subgroup, p_0 is the relative frequency of the target variable in the total population, N is the size of the total population, and n denotes the size of the subgroup. In contrast to the quality function q_{BT} (the classic binomial test), the quality function q_{RG} only compares the target shares of the subgroup and the total population measuring the *relative gain*. Therefore, a support threshold T is necessary to discover significant subgroups. The quality function is usually selected according to the application requirements. We have found that the *relative gain* quality function is easily interpretable and understandable by users.

We applied the VIKAMINE (Visual, Interactive and Knowledge-Intensive Analysis and Mining Environment) system (Atzmueller and Puppe 2005; Atzmueller 2007) for interactive and automatic subgroup mining. This tool is adapted to particularities of the medical domain like many missing values in the applied data records. VIKAMINE offers an efficient exhaustive and various heuristic search options with constraints for automatic subgroup mining and interactive visualizations for active user involvement. For more details, see (Atzmueller 2007).

Method: Implementation and First Results

In this section, we provide an overview on the methods for quality management and system evaluation, and describe their implementation exemplified by case studies in the context of the SONOCONSULT system. However, we first describe the setup of the applied data warehouse.

A data warehouse (Kimball and Ross 2002) typically contains data from different heterogeneous sources (Kerkri et al. 2001) that need to be accumulated, standardized, and finally imported into the data warehouse (Han and Kamber 2000). For the presented approach, we integrated several heterogeneous data sources into the clinical data warehouse ranging from structured data records containing the examination data from the sonographic records, various laboratory parameters, the final diagnoses for billing, but also unstructured data given by the textual discharge letters. From these, several additional data about further examinations is extracted. The design and implementation of the data warehouse required a lot of data cleaning efforts, since all the data neglecting the SONOCONSULT data needed to be extracted from legacy database systems. The data warehouse was completed after an initial design and several incremental refinement cycles for which the data sources and the selected data needed to be adapted and tuned.

As outlined above, we distinguish two objectives for the evaluation and quality analysis of intelligent systems:

1. System evaluation and analysis: Comparing and evaluating the input – output behavior of the system using external data for assessing the system solutions.
2. Quality management: Assessing the input *quality*: In the medical domain this corresponds to the *documentation quality* of the users (examiners).

In the following sections, we discuss these issues in detail. Furthermore, we exemplify the techniques by clinically interesting findings.

System Evaluation - Solution Profiling

The input and output relations for a given set of cases can be transparently evaluated by domain specialists that provide a gold standard for the solutions of the system. Then, the acquired cases are compared by a before-after strategy, that needs to be done manually by the domain specialists. Such a procedure usually works relatively well, e.g., (Puppe et al. 2008). However, providing the gold-standard takes a lot of time and is thus very cost-intensive. Therefore, other options that do not need to rely on a domain specialist are promising. Additionally, a continuous monitoring of all the possible test cases would potentially require an unlimited set of expert-rated solutions, which is rather unfeasible.

We provide a data mining method that relies on external data from other data sources (examinations). If all the available data has been integrated into a data warehouse, then the evaluation of the input – output relations is straight-forward using a gold-standard given by laboratory and/or other examinations, for example, magnetic resonance imaging or CT tomography. The accumulated gold-standard data is then automatically compared to the solutions of the system in order to identify potentially incorrect solutions that were documented using the intelligent system.

We can perform a rather simple evaluation just by comparing the gold-standard solution to the respective system solution. Then, we can acquire initial statistics that can already indicate problems with the intelligent system. For a more sophisticated approach we need to apply subgroup mining in order to implement more advanced analysis goals: Using the subgroup mining approach we can also consider the gold-standard solution as the target variable, and identify system solutions that are associated with this target, that is, other correct *and* incorrect solutions. Another analysis option is given by constructing a virtual target variable that is true, if the gold-standard and the system solution match, and false otherwise. In this way, we obtain a set of subgroup patterns indicating situations or combinations of factors that indicate potential causes for the observed discrepancies. Then, using explaining variables from the remaining data, we can try to identify explaining factors for the mismatch between the correct and the system solution.

However, the system solutions may be incorrect due to two different reasons. First, the solution of the system may be wrong, and secondly, the provided input findings may be wrong and/or inconsistent with respect to the true input description. In the first case, we need to refine the knowledge system, whereas in the second case we need to make sure, that the input findings are entered in the correct way, for example, by initiating tutoring sessions or by special training of the users. In order to clarify such situations, we therefore also need to consider the performance of the users that provide the input to the system, as discussed in the next section. Using the approach discussed above, that is, the analysis of the system solutions with respect to a gold-standard, we can only detect situation of the first kind, therefore both analysis options complement each other.

In the following we discuss exemplary results obtained from the SONOCONSULT data. By integrating different data sources into the warehouse it is possible to measure the con-

formity of sonographic results with other methods or inputs. In our evaluations, we applied computer-tomography diagnoses and billing diagnoses entered in the hospital information system as a gold-standard.

Total Case Number	SONOCONSULT Diagnoses	SAP Diagnoses	% Conformity with SONOCONSULT	CT/MR Diagnoses	% Conformity with SONOCONSULT	Discharge Letter Diagnoses	% Conformity with SONOCONSULT
Liver cirrhosis							
16	12	6	20	1	33	9	50
Liver metastasis							
28	16	11	65	15	87	17	94

Table 1: Conformity of system diagnoses with various sources of diagnosis input. The columns indicate the degree of correlation of the different sources with SONOCONSULT diagnoses measured by the number of covered cases.

Table 1 shows the correlation of SONOCONSULT based diagnosis with CT/MR, diagnoses listed in the discharge letter and diagnoses contained in the hospital information system for a first number of cases. It was quite interesting that the conformity between SONOCONSULT based diagnoses with the diagnoses contained in the hospital information system was quite low. Evaluating this issue it was obvious that various diagnosis were not listed in the hospital information system because they were not revenue enhancing. Therefore, we looked at the accordance with the discharge letters which were found to be highly concordant at least for the diagnosis of liver metastasis.

Liver cirrhosis is more awkward to be diagnosed with ultrasound and has to be in a more advanced stage. Therefore, some of the discharge diagnoses "liver cirrhosis" were only detected using histology or other methods. In one case liver cirrhosis was listed in the hospital information system but was neither found with ultrasound nor in the discharge letter. An analysis showed that in this case the input was performed by another department (neurology), for which documenting the disease was not really relevant.

These exemplary results of the correlations of diagnoses of various input sources indicate that there is a promising high conformity between SONOCONSULT and the discharge letters. However, for further quality improvement the correlation with other imaging techniques is very important. With a larger number of cases it should be possible to measure the sensitivity of different techniques for various diagnoses in more detail.

Quality Management

For the task of quality management or quality control we consider all the users individually that can enter findings for generating cases with a specific solution. In the context of the medical documentation system SONOCONSULT, the users are sonographic examiners that document the case by entering the specific findings they identify on the sonographic images.

Since sonographic examination is highly subjective and dependent on the experience of the examiner, the quality management is essential in order to identify examiners that deviate from the norm, if we assume a "similar" share of patients for each examiner. Ultrasound is a method which is strongly dependent on the examiner's degree of knowledge. Therefore, it is interesting to know how well the results of the examination agree between different examiners and with the results of other methods. Essentially, we can first discover circumstances under which examiners show a significantly different performance. Second, we can then try to identify explanations from these in order to train examiners, for example, beginners that are not very experienced.

For the quality management step we build user profiles statistically and test for all solutions, for which examiners there is a significantly different distribution: Essentially, we compare the frequency of documented diagnoses depending on the examiner. This method is easily implemented using subgroup mining by regarding the examiner as the target variable, and considering all diagnoses and/or findings as independent (explaining) variables. Table 2 shows the results of this analysis as a overview statistic for several clinically important diagnoses.

Diagnoses	All		Examiner 1		Examiner 2		Examiner 3		Examiner 4	
	F	%	F	%	F	%	F	%	F	%
All	2498	100	757	34	104	4,7	392	17,6	359	16,1
fatty liver	683	27,3	212	33,3	20	3,1	136	21,4	117	18,4
liver cirrhosis	42	1,7	22	52,4	0	0	15	35,7	0	0
aortic sclerosis, non calcified	29	1,2	3	10,7	0	0	0	0	18	64,3
aortic sclerosis, calcified	510	20,4	96	19,7	27	5,5	126	25,8	82	16,8
ascites	160	6,4	60	41,1	4	2,7	27	18,5	16	11
cholezystolithiasis	345	13,8	107	35,7	13	4,3	41	13,7	47	15,7
chron. deg. kidney disease	219	8,8	66	30,3	9	4,1	50	22,9	45	20,6
gut disease	35	1,4	10	28,6	2	5,7	18	51,4	1	2,9
mass/liver	119	4,8	39	33,9	3	2,6	21	18,3	25	21,7
obstructive cholestasis	15	0,6	3	23,1	0	0	6	46,2	3	23,1
lymphnode intraabdominal	33	1,3	16	59,3	2	7,4	3	11,1	1	3,7
pleural effusion	128	5,1	31	27	3	2,6	31	27	24	20,9
portal hypertension	35	1,4	16	45,7	0	0	11	31,4	1	2,9
prostate disease	334	13,4	55	18,1	4	1,3	50	16,4	79	26
liver size = very enlarged	178	7,1	25	15,8	3	1,9	26	16,5	35	22,2

Table 2: Diagnostic profiles for several examiners. The rows denote the frequencies of several different diagnoses, for which the column *F* specifies the absolute and % the relative frequency.

The results show that there are some major differences in the frequency of diagnosing a specific disease for the different examiners. For example, several examiners do not document certain diagnoses at all ("aorta sclerosis, not calcified" or "portal hypertension") – in contrast to the assumption that all examiners should encounter a "similar" share of patients (and diseases). The principal cause for this discrepancy may be due to the fact that the different examiner work in different departments and therefore the patients stock of each sonographic examiner has a different distribution of diseases. Since the data warehouse is continuously being extended, we plan to investigate these issues in more detail given a larger number of cases in more detail. Additionally, we can then also control the analysis for this situation and profile different departments individually.

Discussion

The use of test cases for system evaluation and validation is probably the method for the validation of intelligent systems (Preece 1998) that is most often applied. For this black-box testing method the intelligent system derives new solutions for previously solved test cases and compares the solutions stored in the cases with the derived solutions. Since the acquisition of test cases is usually time-consuming and costly, a number of methods have been proposed to decrease the acquisition costs of the test knowledge by automatically generating test cases, e.g., (Knauf, Gonzalez, and Abel 2002; Gupta and Biegel 1990): These approaches rely on an existing knowledge base, and they generate test cases based on the available set of derivation knowledge.

However, such a methodology can only cope with the first of our evaluation objectives, i.e., with the evaluation and analysis of potential errors within the intelligent system, i.e., contained in the applied knowledge base. Further analysis for detecting external explanations is then rather difficult, in contrast to our approach. Using subgroup mining, we can identify factors (e.g., diagnoses, findings, other parameters) that are associated with the occurrence of system errors. This can then indicate important spots that need to be carefully considered when trying to fix such situations. Additionally, the second application, that is, quality profiling and management, is usually not supported by the existing evaluation approaches. It is possible to use the test cases for system evaluation but not for "evaluation" of the users, that is, for building profiles of their behavior.

As we have seen, the evaluation and analysis objectives can be easily mapped to the subgroup mining approach. Since this data mining technique provides for a broad application spectrum, the detail analysis for the evaluation and for the quality management application can be easily implemented. The approach is flexible since it can be automated and a detail analysis can be applied if significant deviations are detected. Furthermore, it can also be performed by an analyst using appropriate tools, for example, the VIKAMINE system.

First results of the application of the presented approach demonstrate its effectiveness and applicability for the sketched scenario. The validation of solutions for system evaluation already shows a high share of correct diagnoses. However, due to the highly subjective documentation procedure of the input diagnoses, we also suspect a dependency on the experience of the examiner. If the results of the examiner profiling study are only attributable to this specific issue, or if they depend on other confounding factors, for example, different patient distributions due to different departments needs to be clarified using a larger number of cases.

The main advantage of the data warehouse-based approach using data mining techniques compared to the manual validation approach is its cost-effectiveness, ease of use, and potentially continuous application throughout the life-cycle of the intelligent system. Ultimately, it can be automated and can provide important feedback to certain types of user, for example, inexperienced examiners. Then, the quality of the documentation and of the input findings can be significantly increased.

Conclusion

In this paper, we have presented an approach for the evaluation, analysis and quality management of intelligent systems. The context of this work was given by SONOCONSULT, a documentation and consultation system in the medical domain, that provides a good example for a typical intelligent system to be evaluated. The approach is based on the availability of a data warehouse containing the system solutions and external data for validation. Then, subgroup mining is applied as a versatile and comprehensive analysis technique for the detailed analysis and discovery of evaluation patterns. In an incremental process these can then be applied for iterative adaptations.

We have discussed the design and implementation of the presented approach with respect to the setup of the data warehouse and the mining process enabling the evaluation, analysis and quality assessment aspects; the techniques were exemplified by the mentioned real-world system in the medical domain. Furthermore, we have shown, how the different objectives are implemented and have discussed their (clinical) impact in several exemplary case studies that are nevertheless clinically relevant.

The results indicate, that the presented approach is well suited for the evaluation and quality management. Furthermore, the knowledge collected within the data warehouse also provides for an ideal basis for further knowledge discovery by relating the different data sources.

For future work, we aim to integrate the system into the standard medical procedure. Furthermore, an extended integration of textual content contained in the different discharge letter should also provide for an increased benefit and utility of the presented approach. Additionally, we plan to extend the analysis to further knowledge sources. Of course, we also need to apply the techniques to a growing number of cases throughout the continuous increase in the size of the data warehouse.

Acknowledgements

This work has been supported by the German Research Council (DFG) under grant Pu 129/8-2.

References

- Atzmueller, M., and Puppe, F. 2005. Semi-Automatic Visual Subgroup Mining using VIKAMINE. *Journal of Universal Computer Science* 11(11):1752–1765.
- Atzmueller, M., and Puppe, F. 2006. SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In *Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, 6–17. Berlin: Springer.
- Atzmueller, M.; Puppe, F.; and Buscher, H.-P. 2005. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, 647–652.
- Atzmueller, M. 2007. *Knowledge-Intensive Subgroup Mining – Techniques for Automatic and Interactive Discovery*, volume 307 of *Dissertations in Artificial Intelligence-Infix (Diski)*. IOS Press.
- Buscher, H.-P.; Engler, C.; Führer, A.; Kirschke, S.; and Puppe, F. 2002. HepatoConsult: A Knowledge-Based Second Opinion and Documentation System. *Artificial Intelligence in Medicine* 24(3):205–216.
- Gupta, U. G., and Biegel, J. 1990. A Rule-Based Intelligent Test Case Generator. In *Proc. AAAI-90 Workshop on Knowledge-Based System Verification, Validation and Testing*. AAAI Press.
- Han, J., and Kamber, M. 2000. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher.
- Huettig, M.; Buscher, G.; Menzel, T.; Scheppach, W.; Puppe, F.; and Buscher, H.-P. 2004. A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. *Medizinische Klinik* 99(3):117–122.
- Kerkri, E. M.; Quantin, C.; Allaert, F. A.; Cottin, Y.; Charve, P.; Jouanot, F.; and Yetongnon, K. 2001. An Approach for Integrating Heterogeneous Information Sources in a Medical Data Warehouse. *Journal of Medical Systems* 25(3):167–176.
- Kimball, R., and Ross, M. 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. New York, NY, USA: John Wiley & Sons, Inc.
- Klösgen, W. 1996. Explora: A Multipattern and Multi-strategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press. 249–271.
- Klösgen, W. 2002. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, New York. chapter 16.3: Subgroup Discovery.
- Knauf, R.; Gonzalez, A. J.; and Abel, T. 2002. A Framework for Validation of Rule-Based Systems. *IEEE Transactions of Systems, Man and Cybernetics - Part B: Cybernetics* 32(3):281–295.
- Lavrac, N.; Kavsek, B.; Flach, P.; and Todorovski, L. 2004. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research* 5:153–188.
- McDonald, C. J. 1996. Medical Heuristics: The Silent Adjudicators of Clinical Practice. *Ann. Intern. Med.* 124:56–62.
- Preece, A. 1998. Building the Right System Right. In *Proc. KAW'98, 11th Workshop on Knowledge Acquisition, Modeling and Management*.
- Puppe, F.; Atzmueller, M.; Buscher, G.; Huettig, M.; Luehrs, H.; and Buscher, H.-P. 2008. Application and evaluation of a medical knowledge-system in sonography (sonoconsult). In *Proc. 18th European Conference on Artificial Intelligence (ECAI-08), Prestigious Applications of Intelligent Systems (PAIS-08)*, 683–687.
- Puppe, F. 1998. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3. *Intl. Journal of Human-Computer Studies* 49:627–649.
- Wrobel, S. 1997. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st Europ. Symp. Principles of Data Mining and Knowledge Discovery*, 78–87. Berlin: Springer.