International Journal of Information Technology and Knowledge Management
January June 2009, Volume 2, No. 1, pp. 159-162

EXTRACTING & REPRESENTING CONCEPTS AS M-INFORMATION

Rajan Vohra* **& Gunjan Pahuja***

A Large number of methods are available to represent the acquired knowledge in machine learning. Decision Tree is one of the methods used to capture the knowledge. In this research paper an alternative method of concept extraction and validation is proposed.

***Keywords**:* M-Information, OPD, Machine Learning, Queuing Model

1. Introduction

The information extraction from documents is an increasingly urgent problem of enterprise knowledge management. Knowledge sources may be internal like text files and forms of business administration processes or external like HTML pages, e.g. When the number of knowledge sources is paramount, substantial computer support is inevitable, and machine learning techniques play a crucial role. The automatic generation of – hypothetical – wrappers for information extraction through the invocation of machine learning techniques is raising several questions. What can we expect of a wrapper generated in case it is not yet completely correct? Can we generate wrappers in a properly incremental fashion? For answering these practically relevant questions, a new algorithm for extracting and representing the concepts as M-information is being introduced in this research paper. Although existing ID3 algorithm is used for building a decision tree from a fixed set of examples. ID3 uses information gain to help it decide which attribute goes partitioning into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from a domain expert, inducing decision rules from the set of examples. In fig 1, the functioning of the concept extraction algorithm is shown, where concepts are represented as production rules and are dynamically added to an evolving database.

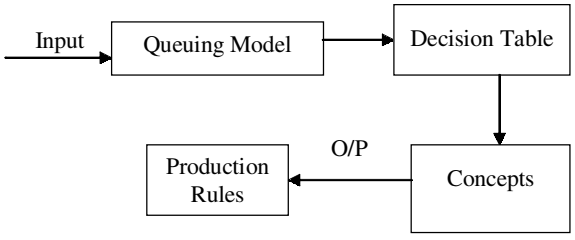

Figure 1: Concept Extraction and Addition

* CSE & IT DEPTT., ITM, Gurgaon, *E-mail: rajanv12@yahoo.com, gunjanpahuja_04@yahoo.com*

2. Review of Literature

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.[1]

A paper [2] studies all three components: creation of synonym, less vocabulary, text transformation, representation of concepts by their models, Optimization of these models, and concept identification with Dynamic Logic that is a general method of reducing computational complexity of combinatorial problems.

Knowledge-based systems must represent information abstractly *so* that it can be stored and manipulated effectively. Schema for learning representations, or concepts from examples promise domain experts direct interaction with machines to transfer their knowledge. This paper describes an integrative framework to develop that describes concept learning techniques that enables their relevance to knowledge engineering to be evaluated. The framework provides a general basis for relating concept learning to knowledge acquisition First concept learning is framed in the context of knowledge acquisition. Then the general forms of input and concept representation are discussed such as logic, functions and procedures [3].

A research effort represents an inquiry into an important problem of automated acquisition, indexing, retrieval, and effective use of knowledge in diagnostic tasks. The principal tool is INC2, an incremental concept formation system, which automates both the design and use of diagnostic decision-support systems by a novice [4].

The process of automatically extracting metadata from an experiment's dataset is an important stage in efficiently integrating this dataset with data available in public

bioinformatics data sources. Metadata extracted from the experiment's dataset can be stored in databases and used to verify data extracted from other experiments' datasets. The extracted metadata can be mined to discover useful knowledge as well as integrated with other information using domain ontology to reveal hidden relationships. The experiment's dataset may contain several kinds of metadata that can be used to add semantic value to linked data [5]

This paper [6] presents a new learning classification algorithm for machine learning. Although the decision-tree classification algorithms have been widely used as the machine learning theory in artificial intelligence, there has been little research toward evaluating the performance or quality of the current classification algorithms.

Nelson [7] used Queuing theory to determine the optimal number of emergency rooms needed in a hospital. By doing sensitivity analysis of 5 key variables namely-average length of queue, average length of the system (patients in the system),average waiting time in queue, average waiting time in system and probability of all servers being idle, the optimal no. of rooms is determined.

3. THE PROPOSED CONCEPT EXTRACTION ALGORITHM

3.1 Queuing Model

The organization of an OPD in a hospital is shown in the figure 2. While a hospital information system generates reports and automates the different departments i.e. OPD, Pharmacy, diagnostics, there is a need to determine optimal allocation of resources. This can be determined by M-information.

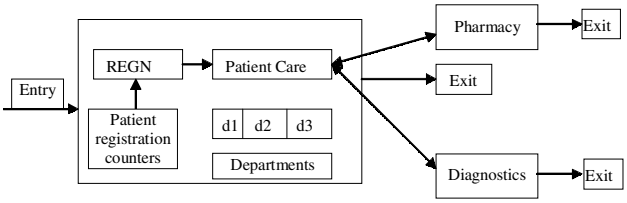

Figure 2: Organization of OPD

The extraction of Meta-information can be obtained through: Models, Heuristics, Data Mining Techniques, and Algorithms etc. as shown in fig 3. In particular, a template for storing such M-information shown in fig 4.

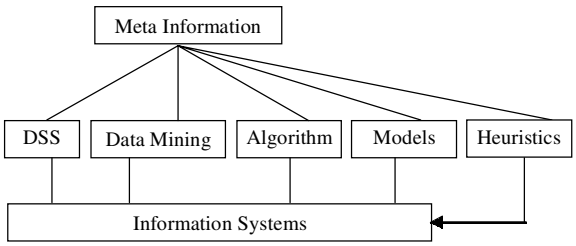

Figure 3: Meta Information Extraction Techniques

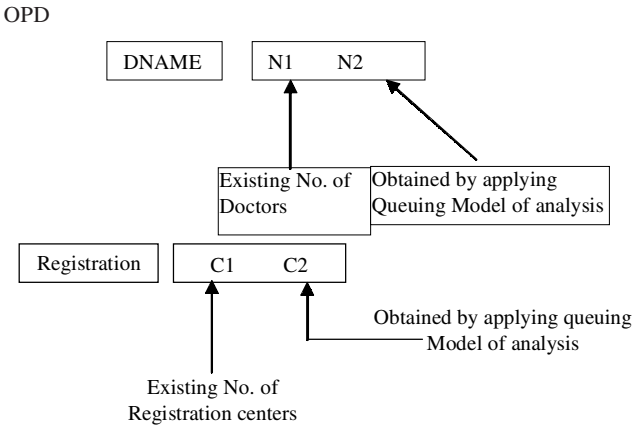

Figure 4: A Template for Storing M-Information

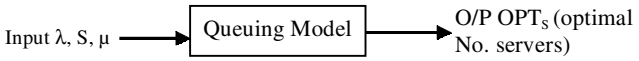

Figure 5: Queuing Model

The parameter values are obtained by sensitivity analysis using Queuing Model. The Meta information (conceptual information) is generated from the inferences made by Queuing Model and can be illustrated below.

The process of concept extraction and validation can be illustrated through a small example. The M/M/S queuing model [8] is applied for decision analysis in a hospital, with the objective of optimizing the number of rooms for emergency care. The patient arrival and servicing along with the existing number of rooms is input to the queuing model, which produces the following decision table, through sensitivity analysis:

Table 1
Sensitivity analysis using M/M/S Queuing Model

	C = 3 Rooms	*C = 4 Rooms*
L_q	17.30	1.80
L_s	20.15	4.64
W_q	1.92	0.20
Ws	2.24	0.52
P_o	0.0118	0.0839

It can be seen that the addition of another room positively impacts the system performance. This Decision table is input for concept extraction and then their subsequent validation. The concept can be represented through IF….THEN rules and are obtained by traversing through the above table. Accordingly the following five concepts can be obtained:

(1) Traffic intensity $\rho = \lambda/\mu > 1$ for M/M/S Queuing model.

With no. of servers S, $S >= 2$ for M/M/S model.

(2) As resources (No. of Servers) increases, P_o (probability of all servers being idle) increases.

(3) As no. of servers increases the value of W_q and Ws reduces.

(4) As no. of servers increases the value of L_q and L_s reduces.

(5) If $C = 3$, the corresponding value of Wq, Ws, L_q, Ls and P_o can be retrieved suitably.

These concepts can be represented by IF….THEN type of rules in a knowledge base, as follows:

R1 C1 IF $S >= 2 \wedge \rho > 1$
THEN VALID M/M/S MODEL.

R2 C2 IF $S < S1 < S2$ …
THEN $P_{00} < P_{01} < P_{02}$ …

R3 C3 IF $S < S1 < S2$…
THEN $((W_q > W_{q1} > W_{q2}\ldots))$ and
$Ws > Ws_1 > Ws_2\ldots)$

R4 C4 IF $S < S1 < S2$…
THEN $((L_q > L_{q1} > L_{q2}\ldots..) \wedge (L_s > L_{s1} > L_{s2}\ldots)$

R5 C5 IF C = 3
THEN $(L_q = 17.30 \wedge L_S = 20.15 \wedge W_q = 1.92 \wedge WS = 2.24 \wedge P_0 = 0.0118)$

where $R1$ to $R5$ are rules and $C1$ to $C5$ are the concepts.

These production rules $R1$,$R2$, $R3$, $R4$, $R5$ represent the conceptual information $C1$, $C2$, $C3$, $C4$, $C5$. These 5 concepts represent the M-information obtained, which can be used for reasoning at M-level i.e.

C1	C2	C3	C4	C5

Figure 6: Structure of M-Information

The validation of these concepts can be done by referring to the corresponding rules $R1$, $R2$, $R3$, $R4$, $R5$ respectively. e.g.

Query: Is model is a valid M/ M/ S model.

Reasoning: Use production rules $R1$.

Calculate value of traffic intensity ρ.

For valid $R1$, $\rho > 1$, this means only positive instances of concept $C1$.

3.2 Data Representation

The table generated by *M/M/S* Queuing model is shown in Table 1. In order to extract the concepts, by table traversal, it is essential to translate this data to a knowledge base. Translating into the relations in the knowledge base, we obtain relations of the following type:

$p(Vc, Wq, Ws, Lq, Ls, Po)$.

where Vc, Wq, Ws, Lq, Ls and Po corresponds to their respective parameter values in the table. So, for the above table we have 2 facts as follows:

$p(3, 17.30, 20.15, 1.92, 2.24, 0.0118)$.

$p(4, 1.80, 4.64, 0.20, 0.52, 0.0839)$.

The concept extraction process involves traversal through these facts and then assertion of the concepts, as dynamic clauses, through dynamic updating of the knowledge base.

3.3 Program

The corresponding production rules, in the form of Prolog clauses can be written as follows:

Concept 1:

S > = 2,

Rho >1, assert (C1).

Concept2:

bagof (Vc, $Lq \wedge Ls \wedge Wq \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L1$), asc_order(L_1),

bagof (Po, $Lq \wedge Ls \wedge Wq \wedge Ws \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, L),

asc_order (L),

assert(C2).

Concept3:

bagof (Vc, $Lq \wedge Ls \wedge Wq \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L1$),

asc_order(L1),

bagof (Wq, $Vc \wedge Lq \wedge Ls \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L2$),

bagof (Ws, $Vc \wedge Lq \wedge Ls \wedge Wq \wedge P_o \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L3$),

des_order($L2$),

des_order($L3$),

assert ($C3$).

Concept4:

bagof (Vc, $Lq \wedge Ls \wedge Wq \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L1$), asc_order(L1),

bagof (Lq, $Vc \wedge Ls \wedge Wq \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L2$),

bagof (Ls, $Vc \wedge Lq \wedge Wq \wedge Ws \wedge Po \wedge p(Vc, Lq, Ls, Wq, Ws, Po)$, $L3$),

des_order(L2),

des_order(L3),

assert(C4).

Concept5 :

read (X),

p(*X*, *Vc*, *Lq*, *Ls*, *Wq*, *Ws*, *Po*),

assert (C5).

where asc_order(L) represents the predicate and returns true if list *L* is in ascending order.

des_order(L) represents predicate and returns true if list L is in the descending order.

The concept mapping for C1 to C5 can be thus obtained.

C1	*C2*	*C3*	*C4*	*C5*
No. of servers S >= 2, ρ > 1, valid M/M/S model.	As no. of servers increases (Vc), probability of all servers being idle increases (Po).	As no. of servers increases (Vc), Wq and Ws reduces.	As no. of servers increases (Vc), Lq and Ls reduces.	If Vc =3, values of Lq, Ls, Wq, WS, Po are retrieved from the knowledge base.

4. DISCUSSION

The process of knowledge acquisition is an important step and researchers have attempted various ways of automating this process to enable the automatic capture of knowledge. The strategic benefits of such capture of Domain specific knowledge are huge.

The representation of such knowledge can be done through the use of production rules. This is an important research topic in machine learning and algorithms have been written to capture such knowledge. In particular Decision Trees have been used to capture such knowledge. The traversal of such a tree results in Decision Rules which represents the Knowledge acquired. Such Decision Rules can be used to validate a Concept if it is represented by such a Decision rule.

There are well known algorithms to construct a Decision tree from an input set of examples. ID3 is a well known example of such a tree generating algorithm. In this research paper, we take an alternative approach – we traverse the decision table generated by the Queuing model. This is analogous to common sense reasoning that would also involve traversal through the decision table. Thus row traversals are simulated in our concept extraction algorithm. A positive instance of the Concept is validated by a valid row traversal, which is specified by the corresponding production rule (e.g Rules *R*1, through *R*5). The concept is then dynamically asserted as a clause in the evolving knowledge base.

The gained knowledge in the form of the Concepts C1 through C5 is represented in the dynamically updated knowledge base. These are validated through corresponding Prolog Clauses, Concept1 through Concept5.

This research paper has proposed an alternative method of Concept extraction and Validation. Concepts are an important part of Meta Information. Such Meta Information is extremely important in generating inferences and as a reference point for future computations.

In future work, segregation between primary and secondary concepts can be done. Primary Concepts can be used to Validate Secondary Concepts or Derived Information, which can again be stored for use in further computations.

Further, a combination of Theories, Heuristics, Models, Concepts, Meta Knowledge and facts and rules can be made to derive inferences. This is also a valid usage of Meta Information. Again, a negative instance of a Concept can be added to show the negation of the Concept. It can be seen that Meta information does Abstraction and therefore reduces the Complexity of reasoning and inferencing at a higher level of analysis and thought, which can be called the Meta level.

References

[1] Nilsson Online Book, "*http://ai.stanford.edu/~ nilsson/mlbook.html*"

[2] Sergey Petrov, Ilya Muchnik, Simon Streltsov, Leonid Perlovsky, "Identification of Concepts in Text Using Dynamic Logic", *IEEE* (2007).

[3] Mircea Ionescu and Anca Ralescu, "Extension of the Concept Representation in Conceptual Spaces", *IEEE* (2006).

[4] Bruce A. Macdonald, "A Framework for Knowledge Acquisition through Techniques of Concept Learning" *IEEE*, (June 1989).

[5] Mirsad Hadzikadic, "Medical Diagnostic Expert Systems: Performance Vs. Representation", *IEEE* (1992).

[6] Tzung-Pei Hong Shian-Shyong Tseng, "Comparison of ID3 and Its Generalized Version" *IEEE* (1999).

[7] Nelson Carl W, "Operational Management in the Health Services-Planning, Restructuring and Control", Elsevier Publication, North Holland,1982.

[8] Gupta and Sharma, "OR for Management" Second Edition, (2001), Mayur Paperback, Queuing Model 454-495, M/M/S Queuing Model, 477.

[9] Pea-Lei Tu Jen- Yao Chung, "A New Decision-Tree Classification Algorithm for Machine Learning" *IEEE* (2000).

[10] Badr Al-Daihani, Alex Gray, Peter Kille, "Extracting Metadata from Biological Experimental Data", *IEEE* (2006).