AN ONTOLOGY-DRIVEN METHODOLOGY TO DERIVE CASES FROM STRUCTURED AND UNSTRUCTURED SOURCES

by

SELVAKUMAR MANICKAM

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

DECEMBER 2013

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor, Associate Professor Dr. Cheah Yu-N from the School of Computer Sciences, Universiti Sains Malaysia whose guidance and help, stimulating suggestions and encouragement helped me in the research for and writing of this thesis. I also thank Professor Dr. Sureswaran Ramadass for pushing me to complete this thesis.

I am extremely grateful to my wife, Tilagavaathy Santharan for supporting me during my struggle to complete this thesis. I would also like thank to my parents: Mr. Manickam Nadeson and Mdm. Vasentha Kalimuthu for pushing me to pursue a post-graduate degree and for being everything anyone could wish their parents to be - the perfect gift nobody else can replace. I also thank my brother: Manivannan Manickam and sisters: Sujatha Manickam and Kavitha Manickam for being patient with me and being there whenever I needed them.

I would like to thank Dr. P. Susila Devi for supporting me in verifying the work and the results from a medical perspective. I would also like to thank Hemananthan Palakarnim and Christopher Ooi Chiang Lun for assisting me in formatting, compiling and printing the thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

		Page
ACCS	Automated Case Creation System	23
AI	Artificial Intelligence	24
AIAI	Artificial Intelligence Applications Institute	131
ASTM	American Society for Testing and Materials	36
BP	Blood Pressure	73
BPNN	Back Propagation Neural Network	120
CADD	Clustering Algorithm based on object Density and Direction	99
CARE-PARTNER	Computerized knowledge-support system for stem-cell post-transplant long-term follow-up on the World-Wide-Web	52
CB	Case Base	4
CBR	Case-Based Reasoning	4
CC	Clinical Case	56
CCV	Clinical Case Values	100
CEN	Committee for European Standardization	38
CPT	Current Procedural Terminology	30
CRF	Conditional Machine Fields	26
DTD	Document Type Definition	83
EDI	Electronic Data Interchange	30
EMR	Electronic Medical Record	16
ETL	Extracting Transforming Loading	22
ETPL	Extracting Transforming Predicting Loading	22
EV	EMR Values	100
FHCRC	Fred Hutchinson Cancer Research Center	52
GATE	General Architecture for Text Engineering	69
HL-7	Health Level Seven International	67
HTML	Hyper Text Markup Language	14
ICARUS	Intelligent Case-based Analysis for Railroad Uptime Support	13
ICD	International Classification of Diseases	28
IHTSDO	International Health Terminology Standards Development Organisation	33
IR	Information Retrieval	28
kNN	k-Nearest Neighbors	55
LGP	Link Grammar Parser	69
LOINC	Logical Observation Identifiers, Names, and Codes	36
MESH	Medical Subject Headings	101
MUSCATI	Multi Source Case Acquisition and Transcription Info-Structure	17
MUSTANG	Medical UMLS-based Terminology Server for Authoring, Navigating and Guiding the Retrieval from Heterogeneous Knowledge Sources	35
NLM	National Library of Medicine	29
NLP	Natural Language Processing	23

NN	Neural Network	54
OIL	Ontology Inference Layer or Ontology Interchange Language	48
ONIONS	ONtologic Integration Of Naïve Sources	41
OPS	Operationen- und Prozedurenschlüssel	28
SAP	Systems, Applications, and Products in Data Processing	27
SGML	Standardized General Markup Language	30
SHOE	Simple HTML Ontology Extensions	48
SM	*Standardisation Measure*	117
SNOMED	Systematized Nomenclature Of Medicine Clinical Terms	32
SNOMED-RT	SNOMED Reference Terminology	32
SPSS	Statistical Product and Service Solutions	29
SQL	Structured Query Language	29
SVM	Space Vector Machine	27
TAMBIS	*Transparent Access to Multiple Bioinformatics Information Sources*	35
UMLS	Unified Medical Language System	26
UNESCO	United Nations Educational, Scientific and Cultural Organization	33
UPML	Unified Problem-solving Method description Language	48
WF-C-means	Weighted C-Means Clustering	55
WWW	World Wide Web	67
XML	eXtended Markup Language	14
XOL	XML-Based Ontology Exchange Language	48

LIST OF APPENDICES

KAEDAH BERPACUKAN ONTOLOGI UNTUK MEMPEROLEH KES DARI SUMBER BERSTRUKTUR DAN TAK BERSTRUKTUR

ABSTRAK

Kebolehan penyelesaian masalah sistem Penaakulan Berasaskan Kes (PBK) bergantung kepada kekayaan pengetahuan yang terkandung dalam bentuk kes, iaitu pangkalan kes. PK patut mengandungi volum besar kes-kes terbaru yang kaya dengan penyelesaian yang selalunya dibina oleh pakar-pakar domain teriktiraf dalam bidang masing-masing. Usaha mengisi dan seterusnya memastikan kandungan PK sentiasa mengandungi bilangan kes yang mencukupi adalah suatu aktiviti yang manual dan menjemukan yang memerlukan banyak sumber manusia and operasi. Penyelidikan ini bertujuan untuk membentuk pengetahuan dari pelbagai sumber dan struktur. Tesis ini mengemukakan Infostruktur Perolehan dan Transformasi Kes dari Pelbagai Sumber (IPTKPS). IPTKPS telah dilaksanakan sebagai senibina pelbagai lapisan dengan menggunakan peralatan terkini yang boleh dianggap sebagai suatu lanjutan fungsi kepada sistem PBK tradisional. Secara prinsipnya, IPTKPS adalah bebas domain dan bidang kesihatan dipilih. Rekod Perubatan Elektronik (RPE) digunakan sebagai sumber untuk menjana pengetahuan. Keputusan eksperimen menunjukkan volum dan kepelbagaian kes meningkatkan kebolehan penaakulan enjin PBK. Eksperimen yang dijalankan juga menunjukkan bahawa pengetahuan yang terkandung dalam rekod perubatan (tanpa menghiraukan struktur) sememangnya boleh digunapakai dan dipiawaikan untuk menambahbaik pengetahuan (perubatan) dalam sistem PBK tradisional. Seterusnya, enjin pencarian Google adalah kritikal dalam pembetulan and pengkayaan ontologidomain dengan segera.

AN ONTOLOGY-DRIVEN METHODOLOGY TO DERIVE CASES FROM STRUCTURED AND UNSTRUCTURED SOURCES

ABSTRACT

The problem-solving capability of a Case-Based Reasoning (CBR) system largely depends on the richness of its knowledge stored in the form of cases, i.e. the CaseBase (CB). Populating and subsequently maintaining a critical mass of cases in a CB is a tedious manual activity demanding vast human and operational resources. The need for human involvement in populating a CB can be drastically reduced as case-like knowledge already exists in the form of databases and documents and harnessed and transformed into cases that can be operationalized. Nevertheless, the transformation process poses many hurdles due to the disparate structure and the heterogeneous coding standards used. The featured work aims to address knowledge creation from heterogeneous sources and structures. To meet this end, this thesis presents a Multi-Source Case Acquisition and Transformation Info-Structure (MUSCATI). MUSCATI has been implemented as a multi-layer architecture using state-of-the-practice tools and can be perceived as a functional extension to traditional CBR-systems. In principle, MUSCATI can be applied in any domain but in this thesis healthcare was chosen. Thus, Electronic Medical Records (EMRs) were used as the source to generate the knowledge. The results from the experiments showed that the volume and diversity of cases improves the reasoning outcome of the CBR engine. The experiments showed that knowledge found in medical records (regardless of structure) can be leveraged and standardized to enhance the (medical) knowledge of traditional medical CBR systems. Subsequently, the Google search engine proved to be very critical in "fixing" and enriching the domain ontology on-the-fly.

CHAPTER 1
INTRODUCTION

1.1 Overview

Conceptualized codifications of knowledge, far beyond what already exists in manuscripts and in human brains, constitute a practical possibility. Taking into consideration the current sophisticated nature of computer technology, one is led to assume that computer scientists, equipped now with unlimited data and improved access to human intellectual resources, might soon achieve complete codification of knowledge for any particular domain. But, research observations suggest the contrary; the reality is that there exist too few consolidated codified 'knowledge assets', yet there are so many knowledge resources still to exploit! Although there has been many knowledge sources that are stored in structured form, yet multitude other (approximately 80%) (Das & Kumar, 2013) are still in unstructured form.

The problem of knowledge acquisition to some extent can be attributed to the complex epistemology, nature and make-up of knowledge. Put simply, human knowledge is regarded as '*a body of facts and principles accumulated by mankind in the course of time* (Clarke, 1999), yet philosophically the issue is still under debate. However, for practical purposes, one can argue that knowledge includes but is not limited to information, advice, experiences, best practices and lessons learned. More so, knowledge is differentiated along the lines of *Explicit Knowledge and Tacit Knowledge*. Explicit knowledge can best be described as canonical knowledge, i.e. knowledge formalised within databases, business rules, manuals, protocols and procedures and so on. Explicit knowledge is about *how things should work*. Tacit knowledge is non-articulated knowledge, more appropriately it can be referred to as

non-canonical knowledge—knowledge about *what really works*. Tacit knowledge does not manifest as rules, rather it exists as the domain expert's skills, common-sense and intuitive judgment whilst solving problems (Holsapple & Joshi, 2011). Such a dichotomy of views and beliefs about the very nature of knowledge renders the problem of knowledge acquisition in a computational paradigm not only challenging but at the same time quite interesting.

Knowledge acquisition (Bernardi et al., 2011) is a research topic that is vehemently pursued by computer scientists from different perspectives, each group of researchers practicing a different methodology to acquire different modalities of knowledge that is subsequently applied to knowledge-based systems for decision-support tasks. Prominent fields related to knowledge acquisition include *Knowledge Engineering Knowledge Discovery* and *Knowledge Management* (Holsapple & Joshi, 2011).

Traditionally, knowledge acquisition issues have been addressed by the field of knowledge engineering (Motta, 2013). Knowledge engineers have been involved with the acquisition and formalisation of knowledge owned by human experts, leading to the development of knowledge bases. Lately, the emergence of the field of knowledge discovery has presented an alternate, yet interesting, dimension to knowledge acquisition practices, whereby knowledge is inductively derived from vast volumes of collected data. There is interest in the field of knowledge management as it provides a framework that not only supports the capture of both explicit and tacit knowledge but also the operationalization of derived knowledge within an enterprise.

Organizations are increasingly interested in accessing knowledge stored in unstructured sources, in addition to structured sources. Unstructured data consists of freeform text such as word processing documents, e-mail, Web pages, and text files, as well as sources that contain natural language text. Although unstructured data also includes audio and video streams as well as images, this will not be considered in this thesis, as the focus is knowledge discovery from textual sources.

Knowledge stored in a structured format is inherently record-oriented; it is typically stored with a predefined schema, which makes it easy to query, analyze, and integrate with other structured data sources. Unlike structured data, however, the nature of unstructured data makes it more difficult to query, search, and extract, complicating integration with other data sources.

Regardless of the complexity in manipulating and integrating unstructured content, there is a strong need to build tools and techniques for managing such data. As mentioned earlier, some 80 percent of the data residing in an organization is in unstructured format (Das & Kumar, 2013). Knowledge discovered solely based on the structured data (which constitutes a small percentage of the organization's data) may not be accurate as it does not take into account of the majority of knowledge found in unstructured data.

The knowledge hidden or stored in unstructured data can play a critical role in making decisions, understanding and complying with regulations, and conducting other functions. Integrating knowledge discovery to cover data stored in both structured and unstructured formats can add significant value to an organization.

1.2 Case-Based Reasoning: A Brief Overview

Case-based reasoning (CBR), broadly construed, is the process of solving new problems based on the solutions of similar past problems (Riesbeck & Schank, 2003). CBR is a computer technique, which combines the strength of rule-based system with a simulation of human reasoning when past experience is used, i.e. mentally searching for similar situations which occurred in the past and reusing the experience gained in those situations. In the same way, in CBR, the knowledge cases are structured and stored in a Case Base (CB), which the user queries when trying to solve a problem. The system *retrieves* a set of similar cases and then *evaluates* the similarity between each case in the database and the query. The most similar case(s) are presented to the user as possible scenarios for the problem at hand. The user has to decide if the solution retrieved is applicable to the problem, i.e. the system does not make the decision, it only supports the decision making process. If it cannot be *reused*, the solution is *adapted* (manually or automatically). When the user finds a solution, and its validity has been determined, it is retained with the problem as a new case in the database (the case is "*learned*"), for future reuse.

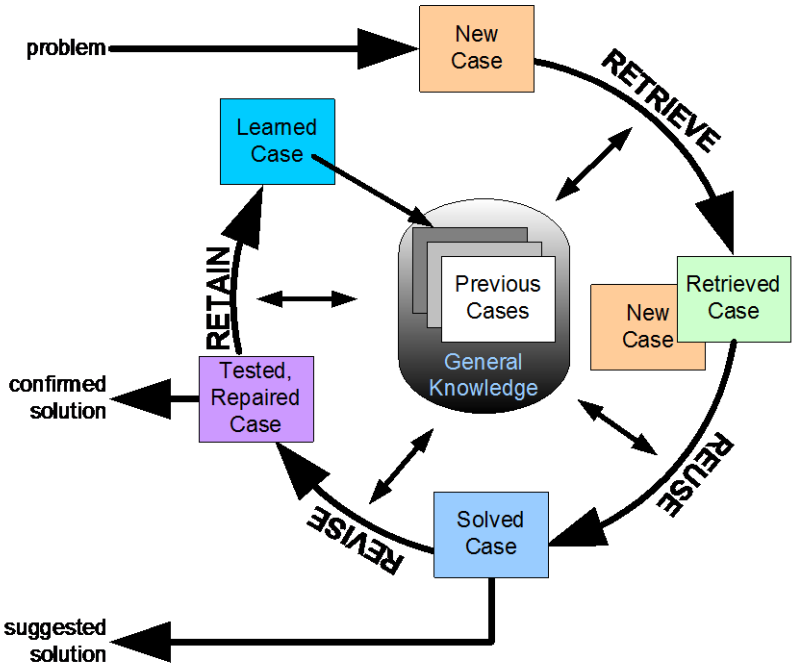

Figure 1.1: The CBR Cycle (Leake, 2003)

Leake (2003) describe the CBR process as being cyclic and comprising the four "RE"s as shown in Figure 1.1:

(a) *Retrieve*: Given a target problem, relevant cases are retrieved from the CB. A case consists of a problem, its solution, and, typically, annotations about how the solution was derived.

(b) *Reuse*: This step maps the solution from the previous case to the target problem. This may involve adapting the solution as needed to fit the new situation.

(c) *Revise*: Having mapped the previous solution to the target situation, the new solution in the real world (or a simulation) is tested and, if necessary, revised.

(d) *Retain*: After the solution has been successfully adapted to the target problem, the resulting experience is stored as a new case in memory.

The knowledge in a CBR system is stored in the form of cases. Cases are a collection of attribute-pair values divided into two sections, i.e. “problem” and “solution” as opposed to knowledge stored in a Rule-based system. In a typical Rule-based reasoning system or expert system, the knowledge used by the reasoning engine is stored in the form of “if..then..” rules.

In order for CBR to be successful, the following issues need to be handled:

(a) A representation form for cases has to be determined,
(b) An appropriate retrieval algorithm has to be selected and
(c) An infinite growth of the CB has to be avoided e.g. by clustering cases into prototypes and removing redundant cases or by restricting the CB to a fixed number of cases and updating the CB during an expert consultation session.

The adaptation (revision) of retrieved cases is a component where little research has been undertaken. Even if there is an adaptation method available, it is more likely that it is specific to a certain domain and that a generic adaptation model is still not available. In current approaches, adaptation basically involves the use of constraints and rules acquired from experts. Due to the process of knowledge engineering and the subjective nature of adaptation, alternative approaches need to be considered:

(a) Focus on retrieval: An approach to avoid the adaptation problem is to build retrieval-only systems. These are programs that only retrieve similar cases and

present them as information to the user. Some of them additionally point out important differences between current and similar cases.

(b) Use of generalised cases: One reason for the adaptation problem is the extreme specificity of individual cases. Therefore, an approach to address this is to generalise individual cases into abstracted prototypes, abstract or classes (Bichindaritz & Marling, 2006). Although the main ideas for generalisation are to structure the CB, to decrease the storage amount by erasing redundant cases, to speed-up the retrieval, and sometimes to learn more general knowledge, additionally it can at least partly help to solve the adaptation problem.

The CBR problem-solving strategy bears a close similarity with how healthcare practitioners solve clinical problems. Cases can be deemed as the most specialized form of knowledge representation. The knowledge of medical practitioners comprises *objective knowledge* acquired from medical books and journals, plus *subjective knowledge* in terms of clinical experiences in the form of past cases that they would have treated themselves or those experienced by colleagues. In diagnosis, the problem-solving thoughts of healthcare practitioners tend to revolve around typical cases—they would consider the differences between a current patient and past treated patients (or cases). The importance of medical case was highlighted by Khan (2011) and Pantazi et al. (2004) who proposed an extension of the definition of biomedical evidence to include knowledge in individual cases, suggesting that the mere collection of individual case facts should be regarded as evidence gathering (see Figure 1.2).

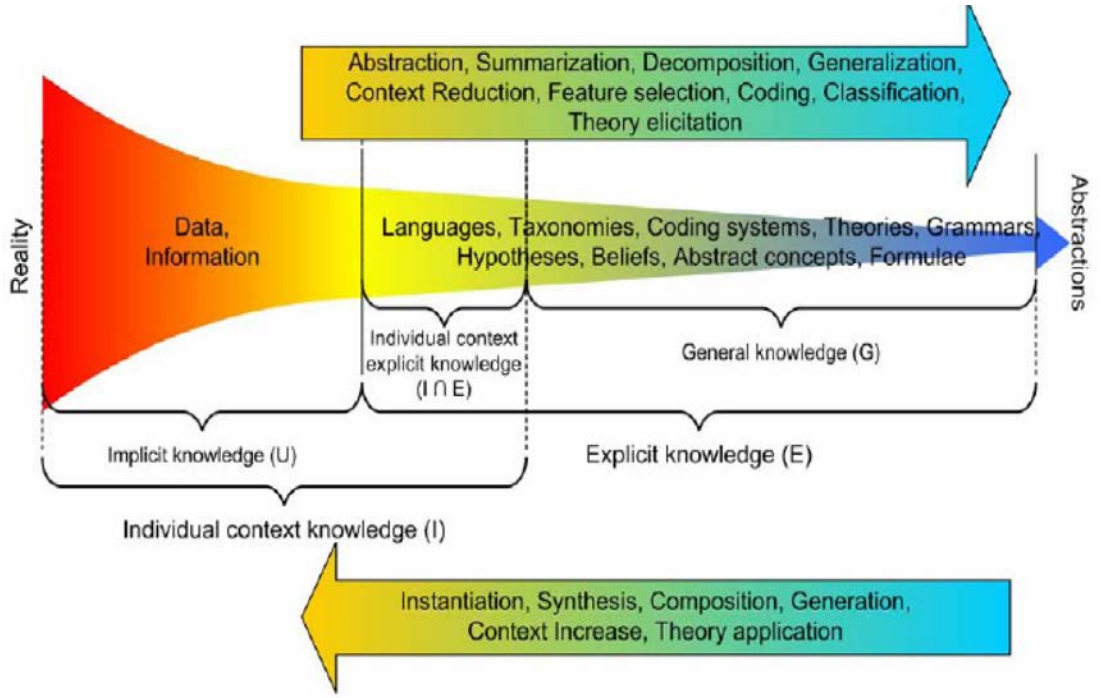

Figure 1.2: The Knowledge Spectrum (Pantazi et al., 2004)

For diagnostic tasks, cases are usually described by a list of symptoms that describe the problem-situation and the outcome or prescribed treatment as the problem-solution. CBR provides a mechanism to manipulate the healthcare practitioner's tacit subjective knowledge to derive experience-mediated solutions. Hence, there are parallels between CBR and healthcare diagnostic reasoning and recommend the application of CBR in healthcare along the following lines:

- Reasoning with cases corresponds with the decision making process of healthcare practitioners.
- The incremental nature of subjective knowledge can be achieved with the addition of new cases to a CBR system.
- Objective and subjective knowledge can be clearly separated.
- As clinical encounters are routinely recorded and stored, it brings to relief the possibility of integrating them into routine healthcare diagnostic systems.

1.3 Issues Affecting the Incorporation of CBR in Healthcare

Attempt to introduce any knowledge system, i.e. CBR, into healthcare poses various challenges. This includes gathering background knowledge, adherence to specific standards and other issues.

1.3.1 Case Procurement

The issue of case procurement has always been at the forefront of CBR implementation. Aligned with the case procurement issue is the problem of case representation as they both directly impact each other. Case procurement, as it is achieved now, involves domain experts who are trained on how to transcribe cases in a conversational setting (see Figure 1.3). Note the obvious difficulties in this scenario: (a) the domain experts need to be engaged, which is not only expensive but is resource-intensive; and (b) the domain experts are required to map their experiential knowledge, which is organized with respect to their cognitive models, to an alien and even artificial (especially from the domain expert's point of view) representation formalism. For example, to populate a CB pertaining to a particular disease, a medical expert needs to meticulously create cases manually based on his/her experience which is expensive, time-consuming and sometimes inconsistent.

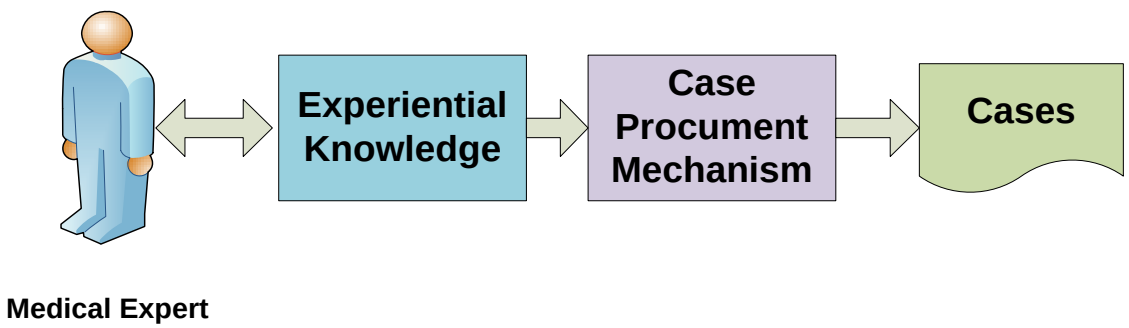

Figure 1.3: Case Procurement Framework

The issues pertaining to case procurement may compromise the efficacy of CBR systems for real-world applications, and there is a need to strategize or devise alternate mechanisms for case procurement.

1.3.2 Case Terminology Standardization

To enforce consistency of data across a CB, the terminology used in describing the cases has to be specific and standard. A lack of consistent terminology can lead to problems with case matching for the case similarity function—the most relevant cases can be missed due to text-based similarity calculations. Ideally, knowledge facilitators—i.e. the domain experts—must use the same terminology when describing the same concepts, yet there is usually no mechanism to ensure such standardization. This is because case procurement is a distributed activity and the domain experts have their own preferences when it comes to describing the problem situation. It should be appreciated that imposing standards on domain experts does not solve the problem; rather it merely discourages domain experts to get them involved in case procurement activities.

A case with the term *"heart attack"* and another case with the term *"myocardialinfarction"* although conceptually the same, it would be rendered a non-match since they are syntactically different. This leads to inconsistent reasoning and inaccurate outcome.

A case standardization—both at the terminological and conceptual levels—should be independent of the case procurement exercise and not involve domain experts. The work put forward suggests an (almost fully) automated case

standardization mechanism. This can be achieved by leveraging domain ontologies or taxonomies, which may not only define the correct terminology but also the conceptual relationships with the problem domain.

1.3.3 Perpetual Growth of Domain Ontology via "Self Learning"

It is suggested in Section 1.3.2 to leverage on domain ontology to standardize cases. Nevertheless, this domain ontology is only dependent on a static corpus of knowledge that may not cover the full depth of the relevant domain. In most new and unique situations, a domain expert is required to recommend and add new terms and concepts to the domain ontology. In some cases, the efficacy of the recommendation by the domain expert may be flawed due to the simple fact that humans tend to make mistakes.

The Internet is now regarded as a new and unique medium as a source of information about health and medicine (Berg, 2011). The Internet is an inherently interactive environment that transcends established national boundaries, regulations and distinctions between professions and expertise. By leveraging on the Internet, especially the Google search engine and online dictionaries, new and unknown situations can be handled (to a high degree) and at the same time enrich the relevant domain ontology. This reduces the dependency on human experts and eventually allows the system to self-sustain.

1.3.4 Feature Weighting

A source of uncertainty in the design of cases is the required evaluative calculation (in order to assign a relative importance to the items of information)

included in a case representation. Since all the case-defining attributes are not equally significant, with some attributes asserting more importance than others, current case representation do not reflect the relative importance of each attribute in the diagnostic and treatment process. In the computing community, the importance of feature significance/weights (with respect to a problem) is widely acknowledged (Sun, 2007) and a number of techniques, such as neural networks, fuzzy sets, statistical techniques, etc. (Begum et al., 2011), are presently applied to determine feature weighting.

In order to improve case representation, in particular in a healthcare context, it is important to establish the relative importance of case-defining features, more attractively in an inductive manner as opposed to asking domain experts to 'rank' the case features.

1.3.5 Knowledge Validation

Validation of knowledge-based systems is an important aspect as it directly impacts the efficacy of the system (Gupta, 2009). However, the majority of the reported validation work to date has centered around rule-based systems, notwithstanding the fact that the cases (representing the reasoning knowledge) in a CBR system also need to be validated. In its purest form, CBR validation requires a domain expert to validate the entire set of cases in a CB, which of course is not possible. O'Leary (2000) addresses the problem of CBR validation, and provides a valuable insight into the problem by discussing the issues involved. Researchers have worked to address this important issue. For instance, Ou et al. (2007) describes methods that enable the domain expert, who may not be familiar with machine

learning, to interactively validate the knowledge base of a Web-based tele-dermatology system. The validation techniques involve decision tree classification and formal concept analysis. Meanwhile, ICARUS (Varma & Roddy, 2004) is a CBR used for diagnosing locomotive faults using such fault messages as input. In this system, historical repair data and expert input for case generation and validation is used. Additionally, other published validation efforts for CBR systems, Protos, HYPO, and Clavier (as discussed by O'Leary (2000)) made extensive use of domain experts which turned out to be extremely expensive.

Knowledge validation for CBR systems should leverage the experiential knowledge which they encode, i.e. the cases are procured from validated sources, comprising standardized experiential knowledge.

1.4 Problem Statement

The discussion highlights some of the issues pertaining to the incorporation of CBR systems in real-life applications, in this case, in healthcare. From an operational point of view, it may be apparent that the ‘weakest link’ in the development and deployment of CBR systems is the domain expert factor! The reliance on domain experts to both provide and validate a critical mass of CBR-specific knowledge raises serious issues that impact the efficacy of CBR systems towards critical, real-life problem-solving applications. A lot of this domain expert knowledge can be found in structured and unstructured sources. In this context, some key constraints involved in the manual CB enrichment are noted as follows:

1. Domain experts are required to transcribe real-life situations into a CBR-system compliant case structure. In most operational settings, the case structure is likely to be different from the domain expert's data recording format. Hence, one can believe that domain experts, who are already quite busy, may find it difficult to perform the transcription of real-life situation-action information into case structures. Medical knowledge sources available over the Internet need to be pre-processed. With the advent of the Internet, the operating database environment may be distributed across multiple sites and the data may be represented using a multitude of formats including HTML, XML and other formats. Even if data is represented in the same format, i.e. XML, data procured from heterogeneous sources tend to have different data definitions. Hence, there are serious usage constraints when one chooses to incorporate Internet-mediated data.
2. A large volume of up-to-date domain-specific cases from multiple domain experts (who may be dispersed at various sites) needs to be routinely sourced for and collected. This calls for a dedicated service, whereby the knowledge engineer or 'knowledge scout' is required to routinely check for new knowledge, which indeed is a resource consuming activity. Since most knowledge is still stored in the form of unstructured documents, without the appropriate techniques, the task of explicating knowledge from these sources would render to be a difficult task.
3. Due to the heterogeneous origins of the cases, the knowledge engineer is required to perform a structural, terminological and conceptual standardization of the collected cases as per the CBR-system's information representation standards. Static ontologies may cause new terms not to be recognized. Ontologies used in standardization need to grow with the demand as to increase the accuracy of

matching. The automation of ontology enrichment with new concepts and terms is necessary in ensuring the ontology is always complete and up-to-date.

4. The knowledge engineer in conjunction with the domain expert is required to judge the importance of each case-defining attribute towards the associated outcome, and then assign a *weight* to it. The numerical value of each attribute's weight is commensurable withits influence towards the associated solution and in operational terms, the weight value is used to determine inter-case similarity.

Despite the natural propensity of CBR technology to effectively provide decision and diagnostic-support to a variety of domains, the need to satisfy the kind of aforementioned constraints tends to compromise the overall acceptance and deployment of CBR-based systems in adaptive real-world environments.

Henceforth, this thesis attempts to address the issues by providing a technical solution to CB enrichment, in particular the automation of the CB enrichment lifecycle in an effort to minimize (but not to eliminate) the involvement of human domain experts and knowledge engineers.

1.5 Research Objectives

This research puts forward a systematic methodology to realize an automated knowledge acquisition environment that allows the acquisition of previously conceptualized domain knowledge to be used for CBR applications. In essence,the methodology is grounded in the principle of acquiring knowledge from generic information resources (such as databases) and transforming 'raw' information (in the

form of EMRs) to CBR-specific knowledge. In addressing this, the objectives of this research are as follows:

1) To devise a mechanism to automatically generate cases. This involves the automation of EMR transformation (both structured and unstructured) into standardized *case* representations by procurement of domain-specific situation-solution type information. This is done by leveraging various Internet-mediated databases or structured XML documents. Cases are extracted from unstructured knowledge sources employing linguistic relation parser and part-of-speech tagger by automatically generating corpus-based co-occurrence thesaurus of semantically related concepts. These relationships and concepts will be used to re-build the records into a structured form.
2) To build self-perpetuating medical ontology using Google's underlying web semantic and online medical dictionaries. Existing medical ontology do not constitute the complete body of knowledge required to handle all standardization requirements and need to be updated on-demand basis. This will improve and increase the knowledge corpus of the ontology by correcting erroneous values and adding previously unknown terms and concepts to the ontology.
3) To automatically estimate an attribute's sensitivity towards an inferred conclusion. Each attribute in a case representation can therefore be ranked with respect to its relative impact factor on the overall inferred decision. This is achieved by inductively determining the influence—i.e. the *weight*—of each case-defining attribute towards the associated outcome via the application of NN based feature sensitivity analysis techniques applied to a cohort of cases.

4) To facilitate the automatic generation of CBR-system compliant cases derived from situation-action information collected from heterogeneous information sources. Our autonomous case generation methodology features multi-level equivalence—at the structural, numerical, terminological and conceptual levels—between the source EMR and the target case representation standards.

1.6 Research Contributions

Automatic and tool-supported knowledge maintenance procedures—note that knowledge creation procedures are not at the same level as knowledge acquisition procedures, rather they operate on already acquired knowledge—are available from dispersed CBR research for very specific knowledge types for certain task and domain types (Leake and Wilson, 2011).

None of the available systems, such as INRECA (Bergmann et al., 2004) or DISER (Tautz, 2000), ascribe to an automated knowledge acquisition and extraction methodology as ours, and their functionalities are rather limited.

This thesis will impact the field of CBR and Health Informatics. Significant impacts of this work are noted as follows:

1) The operationalization of static data objects (structures and unstructured sources) to yield decision-support knowledge. Typically, documents such as medical records are used for clinical administrative and recording purposes. Nevertheless, placid information objects—i.e. Electronic Medical Records (EMR)—can be used as a knowledge resource.

2) The automated enrichment and refinement of medical ontology. An ontology with agrowing corpus of knowledge will provide standardization with improved accuracy by leveraging the large body of medical and healthcare knowledge embedded in literatures found on the Internet.

3) The move towards the 'recorded' experiential knowledge of domain experts as the source of knowledge as opposed to the recruitment of domain experts as a knowledge resource implies a change in the knowledge engineering outlook.

4) The procurement of CBR-specific knowledge (i.e. cases) from routinely collected information will enhance the practicability of CBR-systems in real-life applications, in particular for healthcare applications where a large corpus of medical data (in terms of EMR) is routinely collected for clinical tasks.

5) From a healthcare perspective, the transformation of generic knowledge objects into specialized cases will lead to (a) abstracting general knowledge for medical topics that are well-understood and can thus improve the domain corpus of knowledge; and (b) abstracting experiential information that may not necessarily be available in medical publications—i.e. the abstracted information can be used to strengthen the knowledge content of the existing medical domain.

The research contributions outlined are formulated via Multi Source Case Acquisition and Transformation Info-Structure (MUSCATI).

1.7 Research Scope

The CBR system development lifecycle involves an active interplay between domain experts—the source of problem-definitive *cases*—and knowledge engineers

who are responsible for representing domain expert supplied real-life *cases* into CBR-system compliant computational formats. Indeed, the problem-solving capability of any CBR-system largely depends on the richness of its CB—notwithstanding the importance of CBR algorithms employed to derive an 'analogy-based' solution—which for maximum effectiveness should contain a large volume of up-to-date, decision-quality cases, collected from an ensemble of acknowledged domain experts. Cognizant of the problems associated with knowledge acquisition from domain experts, manual collection of problem-specific knowledge demands vast human and operational resources, which at times compromises the implementation and maintenance of CBR systems.

Premises form the basis upon which this research rests. Delimitations define the scope of the research. The premises of this research are:

a) Automating the process of CB enrichment—starting all the way from case procurement to case generation/transcription to case storage in the CB.
b) Leveraging alternate resources of real-life situation-solution information (akin to cases), other than domain experts, that can subsequently be automatically transformed to resemble real-life CBR-system compliant cases. For instance, there is a rationale for transforming causal information contained in databases, knowledge bases or structured documents represented in eXtensible Markup Language (XML).
c) Making use of intelligent agents to pro-actively seek Internet-accessible data/information repositories as possible resources for automatic case generation and subsequent CB enrichment.

d) The ontology can be extended by providing "learning" capabilities that learns new concepts and terms from the Internet using Google's underlying semantic and other online medical dictionaries.

Nevertheless, there are certain delimitations of this research. They are:

- This research assumes the body of knowledge provided by Google search engine and online medical dictionaries are sufficient to demonstrate their facilitation in improving the transformation of EMRs into standardised cases.
- This research does not cover the safety aspect of the correctness of the transformed cases. Healthcare/medicine was chosen merely as a demonstrative domain.
- This research will not consider the efficiency of the EMR to Clinical Case (CC) transformation since it cannot be tested in a production environment due to the privacy issues involving EMRs. Rather, the research focuses on the efficacy of the transformation using crafted dataset (with the help of a medical doctor) in a controlled environment.
- This research assumes that the engineering design process at the level researched herein is generalizable to other domains such as law and education.

1.8 Theoretical Framework

Knowledge can be seen as integrated information, including facts and their relations, which have been perceived, discovered, or learned as "mental pictures" (Bao, 2005). In other words, knowledge can be considered data at a high level of abstraction and generalization. The process of knowledge discovery inherently

consists of several steps as shown in Figure 1.4 and MUSCATI follows these principles.

Figure 1.4: The Knowledge Discovery Process

Although there are many mechanisms for populating a CB, the ground reality is that populating the CB demands an active involvement of domain experts. In reality, domain experts are required to transcribe real-life situations to a CBR-system compliant case structure. Indeed, this is a tedious and resource-intensive activity which results in a lack of 'decision-quality' cases, which in turn adversely impacts the efficacy of real-life CBR systems.

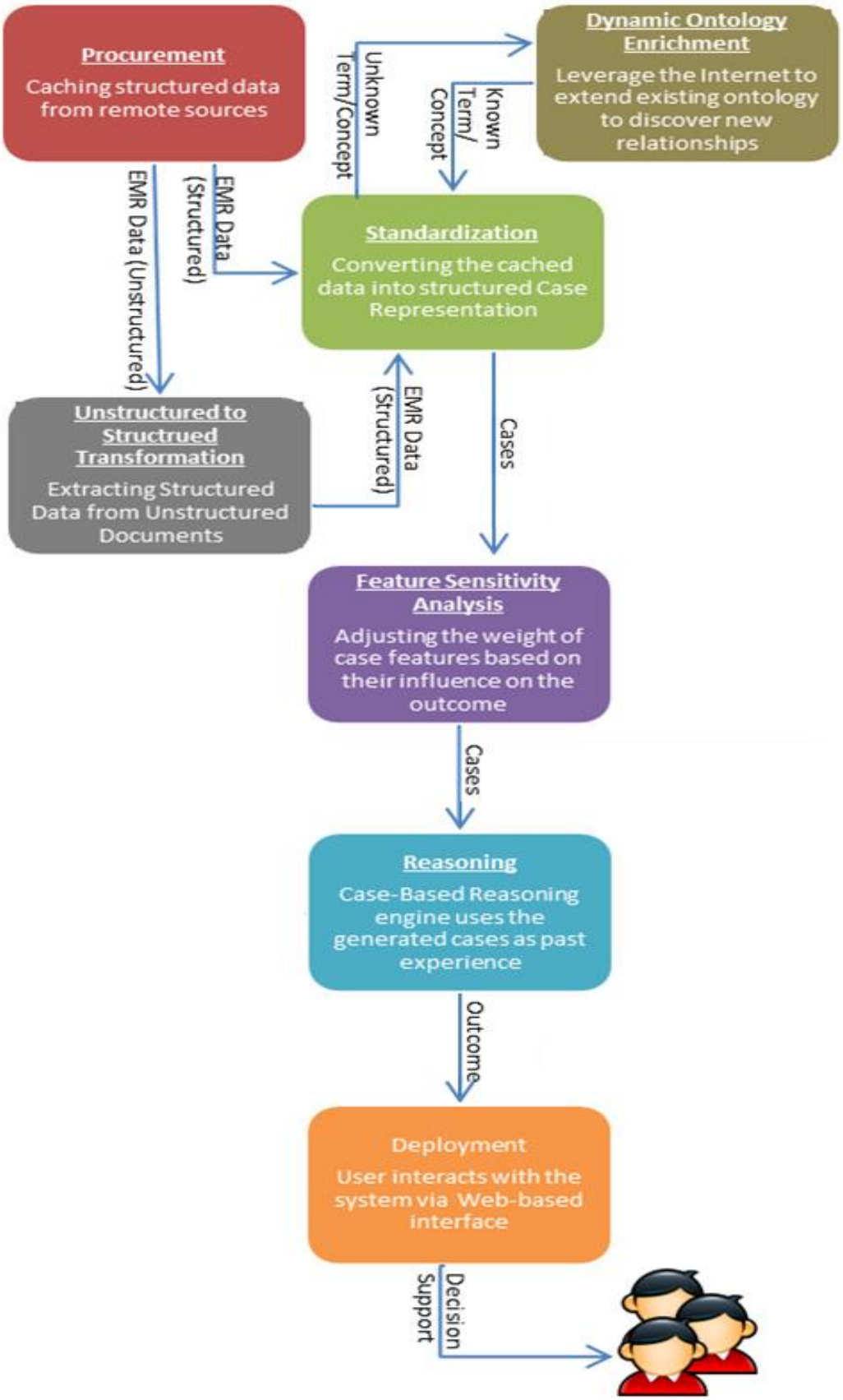

Figure 1.5: Theoretical Framework of MUSCATI

The work presented in this thesis aims to address the above-mentioned suggestions via the formulation of a methodology for the automation of CBR system development, in particular the automation of the CB enrichment lifecycle at the expense of minimizing (but not eliminating) the involvement of human domain experts and knowledge engineers. The theoretical framework of the work is shown in Figure 1.5.

1.9 Thesis Organization

Chapter 1 is the introduction to this work, and provides a summary of the background of this thesis. The task description on CBR and the issues surrounding the creation of knowledge are also provided.

Chapter 2 examines the current literature in the fields of CBR, Knowledge Extraction (and case generation), Ontology (and its enrichment) and Feature Weighting. The motivation for this work is also presented.

Chapter 3 introduces the MUSCATI infostructure. The conceptual framework that addresses the problem statements which highlights salient features of the methodology is presented as a pipeline.

Chapter 4 explains the methodology presented in Chapter 3 in a granular manner. Details of MUSCATI's infostructure are presented by explaining the functionalities of each module and the mechanisms used in achieving the goals of this thesis.

Chapter 5 illustrates two EMR to Case transformation scenarios and highlights the experiments that have been carried out to measure the efficacy of the proposed methodology. This chapter also presents the results of these experiments and the explanation for the outcome.

Chapter 6 states the conclusions drawn from this work and suggests possible directions for future research.