A WRAPPER SELF-MANAGEMENT VIABLE SYSTEM FRAMEWORK FOR INFORMATION EXTRACTION IN HEALTH CARE

SAMEER QASIM AMEEN

UNIVERSITI KEBANGSAAN MALAYSIA

A WRAPPER SELF-MANAGEMENT VIABLE SYSTEM FRAMEWORK FOR INFORMATION EXTRACTION IN HEALTH CARE

SAMEER QASIM AMEEN

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION TECHNOLOGY AND COMPUTER SCIENCES
(FTSM)
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

KERANGKA KERJA PENYALUT SISTEM KEBOLEHTAHANAN PENGURUSAN-KEKENDIRIAN DALAM PENGEKSTRAKAN MAKLUMAT KESIHATAN

SAMEER QASIM AMEEN

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

03 October 2013

SAMEER QASIM AMIN

P34977

ACKNOWLEDGMENT

First and foremost, I'd like to thank my supervisor and mentor, Prof. Dr. Khairuddin Bin Omar for his technical advice, moral support and the friendship that I have developed over five years in UKM. He has left a lasting impression on me as a fine advisor who is full of patience and tolerance yet having his high expectations met.

I would also like to thank Prof. Dr. Shahrul Azman Mohd. Noah , Prof. Dr. Azurallza Abu Bakar and Dr. Mohd Zakree bin Ahmad Nazri for their support and valuable advice throughout the course of my study.

Last but not least, I would like to dedicate this thesis to my family, my wife Maram, my daughter Dalia and my sons Yazan and Zaid for their love, patience, and understanding that allowed me to spend most of my time on this thesis.

ABSTRACT

This study presents a novel infrastructure of an information extraction system underpinned by two existing models, namely; Distributed Object Oriented Programming and Viable System Model (VSM). Autonomic systems infrastructure has been used successfully in many global computing systems like grid and cloud. Such models for computing environments are required to exhibit self-managing capabilities to reduce cost and complexities and to improve dependability. This research aims to investigate interaction scenarios which enable humans and machines to collaboratively solve problems, applying the interaction results into an autonomic system infrastructure in medical domains by using wrappers and Meta knowledge to control the process of information extraction from patients' records. The main approach in this research for solving the problem of information extraction is, using wrapper approach which relies on semi-structured documents. This research also presents the generic requirements for tools, services and frameworks to facilitate the design and development of self-managed systems dealing with the problem of information extraction with the help of wrappers. This study used consistent query as an approach for interaction between the users to provide a collaborative architecture to extract medical care information data for users to share their experience and knowledge. The methodology of this research consists of two main parts, namely; Wrapper Retrieval and Wrapper Generator. Wrapper Retrieval consists of a user interface and a Wrapper Key Retrieval to recognize the users' needs, search and present a suitable wrapper to the user from the pool of wrappers. In case that none of the wrappers in the pool meets the user's needs, the Wrapper Generator starts functioning by creating a new wrapper which suites the user's inquiries. An autonomic infrastructure system is used to build an information extraction system with the help of wrapper automation for health information where the output from the system is a text document (report) extracted from patients' medical records using automatic extracting wrappers. To implements this methodology, a working system was build, tested and installed in two hospitals for test running. A monitoring model has been designed in this research to provide the situated autonomic computing with feedback and context information from their environment. The implication of this approach represent a significant step forward and may serve as a road sign indicating that this path is an efficient yet easy to implement and control. In support of the monitoring and awareness services, the researcher also conducted a survey questionnaire to collect responses from the users about the performance of the system. The findings which were evaluated using quantitative and qualitative method indicate maximum users' satisfaction.

ABSTRAK

Kajian ini mencadangkan infrastruktur baru untuk sistem pengekstrak maklumat yang berasaskan dua model sedia ada iaitu; Pengaturcaraan Berorientasikan Objek Teragih dan Model Sistem Berdaya Maju (VSM). Infrastruktur sistem autonomi telah berjaya digunakan dalam banyak model komputeran global seperti grid dan awan. Penggunaan model seumpamanya sebagai persekitaran komputeran adalah diperlukan untuk mempamerkan keupayaan pengurusan kendiri dalam menurunkan kos dan kompleksiti serta meningkatkan kebolehharapan. Penyelidikan ini bertujuan untuk menyiasat senario interaksi yang membolehkan manusia dan mesin menyelesaikan masalah secara kolaboratif, mengaplikasi keputusan interaksi ke dalam suatu infrastruktur sistem autonomi untuk domain perubatan menggunakan pembalut dan pengetahuan meta bagi mengawal proses pengektrakan maklumat rekod pesakit. Kaedah kajian terdiri daripada dua komponen utama, iaitu: bahagian capaian pembalut dan bahagian penjana pembalut. Kajian ini turut menunjukkan keperluan generik alatan, khidmat dan kerangka kerja untuk memudahkan reka bentuk dan pembangunan sistem pengurusan kendiri yang berurusan dengan masalah pengekstrak maklumat menggunakan pembalut. Kajian ini juga menggunakan pertanyaan konsisten sebagai pendekatan interaksi antara pengguna sebagai suatu seni bina kolaboratif bagi pengektrak maklumat penjagaan kesihatan dengan pengguna dapat berkongsi pengetahuan dan pengalaman. Pendekatan dalam penyelidikan untuk menyelesaikan masalah pengekstrak maklumat adalah menggunakan pendekatan pembalut yang berkait separa-terstruktur dokumen. Kaedah kajian ini terdiri daripada dua bahagian iaitu capaian pembalut dan penjana pembalut. Capaian pembalut terdiri daripada antara muka pengguna dan capaian kunci pembalut untuk mengecam keperluan pengguna, mencari dan membentangkan pembalut yang sesuai bagi pengguna daripada lubok pembalut . Sekiranya tiada pembalut dalam lubok yang sesuai dengan jangkaan pengguna, bahagian penjana pembalut bertindak menghasil pembalut baru yang sesuai dengan keperluan pengguna. Infrastruktur sistem autonomi telah digunakan untuk membina sistem pengekstrak maklumat menggunakan automasi pembalut bagi maklumat kesihatan. Output bagi sistem adalah dokumen teks (laporan) yang diekstrak daripada rekod perubatan pesakit menggunakan pembalut pengekstrak binaan automatik. Bagi melaksanakan kaedah ini, sistem sebenar telah dibangunkan, diuji dan dipasang pada dua hospital untuk ujian larian. Model pemantauan telah direka bentuk dalam penyelidikan ini bagi menyediakan komputeran autonomi terkondisi dengan maklum balas dan konteks maklumat daripada persekitaran. Implikasi pendekatan ini menggambarkan satu langkah kehadapan yang ketara dan dapat melayan sebagai tanda jalan yang menunjukkan bahawa langkah ini adalah efisien, mudah diimplementasi dan dikawal. Untuk menyokong perkhidmatan pemantauan dan kesedaran, penyelidik turut menjalankan kajian soal selidik untuk mengumpul respon dari pengguna mengenai prestasi sistem. Penemuan yang dilakukan secara kuantitatif dan kualitatif menunjukkan kepuasan pengguna yang tinggi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
ASIDL	Assembly Services and Infrastructures Description Language
ASIF	Assembly Services and Infrastructures Framework
BMU	Best Matching Unit
CIM	Common Information Model
DISCO	Discovery
EHMS	E-Health Monitoring System
GoF	Gang of Four
JSU	Job Schedule Unit
KNN	K-Nearest Neighbourhood
MOWS	WSDM Management of Web Services
MSDL	Monitor Session Description Language
MUWS	WSDM Management Using Web Services
ODS	On-Demand Service
OGSA	Open Grid service Architecture
QoS	Quality of Service
SADL	Sensor and Actuator Description Language
SAF	Sensors and Actuators Framework
SLA	Service Level of Agreement
SM-VSM	Self-Management Viable System Model
SOAP	Simple Access Description Language
SOM	Self Organising Map
SRU	Service Reservation Unit
SSM	Soft Systems Methodology
SVM	Support Vector Machine
UDDI	Universal Discovery, Description and Integration
UML	Unified Modelling Language
VServer	Virtual Servers
VSM	Viable System Model
WS	Web Service
WSDL	Web Service Description Language
WSDM	Web Services Distributed Management
XML	Extensible Markup Language

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

This chapter contains an explanation of the various components of this research. This introductory chapter seeks to set the context by which the study is framed and encompasses an explanation of the background to the problem, the problem statement, research objectives, research scope, research methodology, research significance and the contribution of the research to the existing body of knowledge and to existing environments and its importance thereof. In general, this chapter not only gives a preliminary depiction of the research, but also provides an executive summary of the entire work.

1.2 RESEARCH BACKGROUND

Information Extraction (IE)a represents a rich research area within the domain of Natural Language Processing (NLP). NLP possesses the potential for addressing many real-world applications such as database access, question answering, and information extraction. However, in-depth natural language processing is an expensive endeavour that can strain computational resources. As an alternative to full-blown natural language processing, some researchers in the NLP community have turned their attention to adopt (IE) techniques.

The goal of an IE system is to extract specific types of information from text. For example, an IE system might extract the names of all persons, their age distributions, physical addresses etc. This simplifies the job of the NLP system

considerably. IE is less computationally expensive than full-blown natural language processing because many phrases and even entire sentences can be ignored if they are not relevant to the domain. Since the system is only concerned with the domain-specific portions of the text, some of the most difficult problems in NLP are simplified (e.g., part-of-speech tagging, ambiguity resolution, etc.)a. IE is a more practical and robust technology than in-depth NLP and has met with notable success in recent years.

IE systems are capable (at least partially) of understanding the content in the input. This capability is realised by cutting out the uninteresting portions of the text and extracting information from the remainder. This information is then structured and provided as output (Vincent Breton 2005). Different techniques are proposed for information extraction methodologies. These techniques are summarised by the following approaches (Hercules 2007) : Named Entity Recognition, Co-reference Resolution, Template Element Construction, Template Relation Construction and Scenario Template Production. These approaches are chosen according to the domain characteristics. Due to the variety of information to be extracted, one single method may not be suitable for all cases.

Integrated Information Extraction for Health Care (IIEHC) as depicted in Figure 1.1 uses NLP techniques to combine information from multiple sources represented as text data and automatically extracts patient-information from them (databases, text), then the system allow the user to use these information to produce different types of outputs. This is done by using programs called wrappers; these are programs that extract contents of a particular information source and put it into structured data form.

There are two main approaches to wrapper generation: wrapper induction and automated data extraction. Wrapper induction uses supervised learning to learn data extraction rules from manually labelled training examples. Automated wrapper generation using unsupervised pattern mining. Automated extraction is possible if the system follow fixed templates. Discovering such templates or patterns enables the system to perform extraction automatically (Liu 2011).

Wrapper generation is an important problem with a wide range of applications. Extraction of such data enables one to integrate data/information from multiple sources to provide value-added services.

For any health care system, the most important factor is the efficiency of extracting accurate patient information this information are unique to each patient and constitutes as the backbone of the health care system upon which all subsequent medical actions depend.

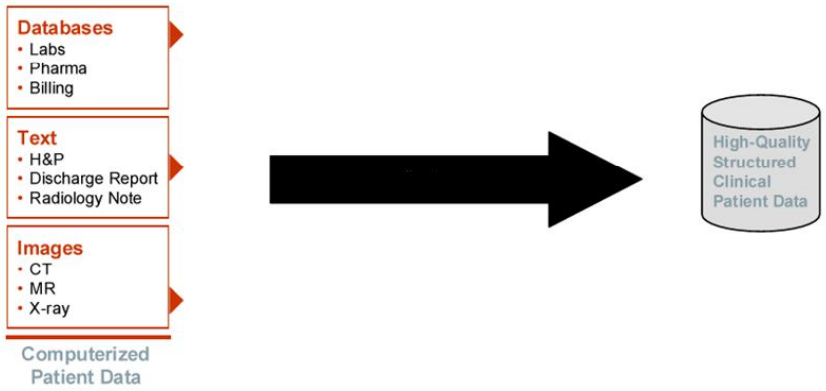

Figure 1.1 The basic approach behind IIEHC

Complex tasks in health care systems are often solved in teams as no one individual has the collective expertise, information, or resources required for effective performance. Among the more notable benefits of working in teams is the combination of a collective genius and its application in problem solving, which by nature involves a multitude of activities such as gathering, interpreting and exchanging information, creating and identifying alternative courses of action, by arriving at a common resolution amongst the differing and often conflicting preferences of team members, and ultimately implementing a choice, determining how incremental progress will be measured, and monitoring its evolution. Such efforts and initiatives rest on the solid assumption that the collective genius is preferable to that of any given individual.