Quality Assessment of Service Providers in a Conformance-Centric Service Oriented Architecture

Gareth Shercliff

A thesis supervised by J. Shao, W.A. Gray and N.J. Fiddian, sponsored by EPSRC and BT, and submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Cardiff University School of Computer Science and Informatics*

December 2009

*Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA, U.K.
Tel: +44 (0)29 2087 4812 Fax: +44 (0)29 2087 4598 Email: office@cs.cardiff.ac.uk

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate)

Date ....

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

Abstract

In a Service Oriented Architecture (SOA), the goal of consumers is to discover and use services which lead to them experiencing the highest quality, such that their expectations and needs are satisfied. In supporting this discovery, quality assessment tools are required to establish the degree to which these expectations will be met by specific services. Traditional approaches to quality assessment in SOA assume that providers and consumers of services will adopt a *performance-centric* view of quality, which assumes that consumers will be most satisfied when they receive the highest absolute performance. However, adopting this approach does not consider the subjective nature of quality and will not necessarily lead to consumers receiving services that meet their individual needs.

By using existing approaches to quality assessment that assume a consumer's primary goal as being optimisation of performance, consumers in SOA are currently unable to effectively identify and engage with providers who deliver services that will best meet their needs. Developing approaches to assessment that adopt a more *conformance-centric* view of quality (where it is assumed that consumers are most satisfied when a service meets, but not necessarily exceeds, their individual expectations) is a challenge that must be addressed if consumers are to effectively adopt SOA as a means of accessing services.

In addressing the above challenge, this thesis develops a conformance-centric model of an SOA in which conformance is taken to be the primary goal of consumers. This model is holistic, in that it considers consumers, providers and assessment services and their relationship; and novel in that it proposes a set of rational provider behaviours that would be adopted in using a conformance-centric view of quality. Adopting such conformance-centric behaviour leads to observable and predictable patterns in the performance of the services offered by providers, due to the relationship that exists between the level of service delivered by the service and the expectation of the consumer.

In order to support consumers in the discovery of high quality services, quality assessment tools must be able to effectively assess past performance information about services, and use this as a prediction of future performance. In supporting consumers within a conformance-centric SOA, this thesis proposes and evaluates a new set of approaches to quality assessment which make use of the patterns in

provider behaviour described above. The approaches developed are non-trivial – using a selection of adapted pattern classification and other statistical techniques to infer the behaviour of individual services at run-time and calculating a numerical measure of confidence for each result that can be used by consumers to combine assessment information with other evidence.The quality assessment approaches are evaluated within a software implementation of a conformance-centric SOA, whereby they are shown to lead to consumers experiencing higher quality than with existing performance-centric approaches.

By introducing conformance-centric principles into existing real-world SOA, consumers will be able to evaluate and engage with providers that offer services that have been differentiated based on consumer expectation. The benefits of such capability over the current state-of-the-art in SOA are twofold. Firstly, individual consumers will receive higher quality services, and therefore will increase the likelihood of their needs being effectively satisfied. Secondly, the availability of assessment tools which acknowledge the conformance-centric nature of consumers will encourage providers to offer a range of services for consumers with varying expectation, rather than simply offering a single service that aims to delivery maximum performance. This recognition will allow providers to use their resources more efficiently, leading to reduced costs and increased profitability. Such benefits can only be realised by adopting a conformance-centric view of quality across the SOA and by providing assessment services that operate effectively in such environments. This thesis proposes, develops and evaluates models and approaches that enable the achievement of this goal.

Acknowledgements

This work has benefitted from guidance and advice provided by many individuals. First and foremost, I would like to thank my PhD supervisors, Dr. Jianhua Shao, Professor Alex Gray and Professor Nick Fiddian, for their support throughout my PhD and for their hard work in reviewing and suggesting improvements to this document. Thanks also go to Vikas Deora and Patrick Stockreisser, with whom I had significant discussions about the concepts and principles which formed the basis of this work, and which were invaluable in its early stages. This work was part of the CONOISE-G project, funded by EPSRC and BT, and I would finally like to thank the members of this team for their considered reviews and opinions at various stages of this work.

Contents

Table of Abbreviations

C (Service) Consumer

CA Clearing Agent

CONOISE-G Constraint Oriented Negotiation in an Open Information Sharing Environment for the Grid

CS Candidate Service

D (Space) Delivery (Space)

DES Discrete Event Simulation

DS Directory Service

E (Space) Expectation (Space)

$\mathbf{E}_c$ The expectation of consumer C

EU Expected Utility

N The number of “actual” service classes for a specific service

QA Quality Assessment

QAS Quality Assessment Service

QDB Quality Database

QoS Quality of Service

QoSA Quality of Service Assessment (Agent)

QoSC Quality of Service Consultant (Agent)

R (Space) Rating (Space)

SC Service Class

SLA Service Level Agreement

SOA Service Oriented Architecture

SP Service Provider

T_i Target Level of Service (for Service Class i)

VO Virtual Organisation

VOM Virtual Organisation Manager (Agent)

YP Yellow Pages (Agent)

List of Figures

List of Tables

Chapter 1

Introduction

It is well established that people are goal-seeking and that most goals can be achieved by combining a variety of resources, whether they be the raw materials required to complete a task, or the time and energy consumed during its enaction. In many instances the level of resource required to carry out a task may not be immediately available. In such cases, it is often necessary to interact with a provider who is willing to supply the resources or the service required at a cost to the consumer. When there is sufficient consumer demand for a product or service, multiple providers may exist each offering identical functionality or similar products. In this instance, the consumer must make a decision of which provider to approach to obtain the service.

In a real world situation, it is not usually necessary to consider externalising exactly what properties of particular services make them more attractive than others, nor to determine and document the exact process involved in making that decision. However, the current trend in computing is towards distributed open environments, where autonomous entities acting on behalf of consumers and providers interact with each other in order to attain their goals. Most distributed computing environments can be seen as marketplaces, composed of consumers and providers of services and resources. If the software components are to effectively support a human consumer, they must be able to represent and reason with all the factors that are important to the individual consumer that they represent.

A consumer's decision of which provider to choose will be based upon a number of requirements. The most obvious of these are those relating to functionality and the cost of the service, both of which are easily expressed objectively. A consumer will also have requirements of a provider in terms of *how* it provides the service. Although often overlooked, these non-functional properties often play an equal or more significant role in a consumer's decision making process than their functional counterparts. Some of this set of non-functional features of a service are commonly referred to as the *quality* of the service.

Incorporating the concept of quality into an agent's decision making process requires more than providing an additional "quality variable" within the software agent. In order to make effective decisions concerning quality, tools must be deployed within the environment's infrastructure to gather and disseminate information regarding the behaviour of service providers in a form which is useful to consumers. By providing such information, agents are able to determine the likelihood and extent to which any provider will meet the needs of the consumers whom they represent. The major contribution of the work presented in this thesis is embodied in the development of approaches which provide support for consumers in determining the quality of services within a distributed computing environment. This is achieved through the use of QoS assessment tools which use historic performance information to predict future service performance.

Measuring the quality of a service is subjective - the same object viewed by different consumers may be seen as having very different degrees of quality. This difference in attribution may be explained by the fact that consumers may be using the object for different purposes or may simply be a reflection of the differences in "harshness" or "idealism" of the consumers concerned. It is evident that the *degree of conformance* of a particular service to a consumer's wants and needs forms the basis for such evaluative judgements. Adopting a "conformance-centric" view of quality directly and significantly affects the behaviour of consumers and providers in a distributed computing environment and as such is the defining driver of the characteristics of the approaches to quality assessment developed in this thesis.

1.1 The Role of Quality in Service Oriented Computing

The last decade has seen considerable growth of interest in the area of Service Oriented Architectures (SOA) [58] - a paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains [82]. The SOA concept is based on information or computational services that are offered by autonomous providers and which can be discovered and utilised by end users (referred to in this context as *service consumers*). The principle idea behind the concept is one of automation, such that consumers and providers are both represented by software agents within the environment. These agents are aware of their owners' goals and capabilities and are able to act in order to achieve these goals by interacting with one another. Typically, within such an environment the goal of the service consumer is to satisfy a need for services in terms of functional and non-functional requirements at the lowest cost; and the goal of service providers is to maximise the profit from their resources through the provision of services to consumers. Enabling the seamless and autonomic interaction of agents in order to enable the achievement of consumer and provider objectives is one of the overriding goals of service oriented research.

The principles of SOA identified above appear under various guises in current research. Web services [122] currently focus on the development of standards for the representation and exchange of information in such contexts. Grid computing is concerned with co-ordinated resource sharing and problem solving in virtual organisations [42, 149]. Agent-based computing [166] is concerned with the development of techniques for engineering complex, distributed software systems [65]. Finally, autonomic computing [74] seeks to develop a set of high level capabilities in software components which enables them to be self-organising, self-aware and self-healing.

Although the frameworks implied by the research areas described above differ in terms of their history, vision and technologies they share a set of common characteristics due to their inherent service oriented nature. Each characteristic presents requirements and challenges for the assessment of quality of services within an SOA. These challenges are discussed below.

1.1.1 Openness

Within an SOA, control is usually distributed throughout the environment and hence no one agent has the ability to force other agents of the system to behave in a particular way. The inherent unpredictably of agents means that agents which rely on each other will need to establish to what degree their potential counterpart is likely to meet their expectations in them. If agents are unable to directly control the actions of those who they interact with, then using information relating to their potential interactee's past performance may provide the best indicator of their likely future performance. In order to achieve this, service oriented environments must be endowed with tools and resources which allow for the monitoring and propagation of information regarding the quality aspects of the services. That is, agents must be endowed with the capability to make decisions about which other agents to enter into relationships with, based on evidence regarding quality, that may be elicited from peers or from third-party information brokering services. This thesis focusses on the development of approaches for quality assessment (QA), to be provided as third party services in SOA environments, to allow agents to obtain information in order to discover services which best meet their needs.

Current work within the area of assessing quality of services is fragmented across many disciplines. A consideration of quality is required in establishing the *trust* to be placed in providers - a key area of research in Agent-based computing. Similarly, in order to establish a set of recommendations of services for a particular consumer – defined as the goal of *recommender systems* – it is useful to assess the quality of all potential services available. Independently of architecture or domain, all approaches to quality assessment share a common aim in seeking to minimise the inherent risk to consumers when identifying the services that will offer them highest quality. That is, there exists an implicit need for quality assessment within any open SOA, if consumers are to be able to use the environment productively. The remaining characteristics, presented below, serve to define or otherwise constrain the way in which this necessary assessment takes place.

1.1.2 Individual Differences

Individual agents will differ in the type and level of resources available to achieve their owners' goals. Similarly, the goals themselves may vary in that different individuals may have different preference and priorities. Individual differences will exist between consumers in terms of their preference of both the functional and non-functional aspects of the services that they decide to use [30]. Existing QA approaches tend to assume a view of quality on the part of consumers, which equates quality with performance, or assumes that all consumers share a common set of preferences. Whilst some degree of *performance* on the part of the provider is necessary to satisfy the needs of consumers, such approaches fail to effectively capture or consider the importance of consumer expectations prior to their use of a service, in order to determine the *conformance* of the provider to the consumer's needs. That is, by ignoring a consumer's preference (which may not equate to the highest level of performance) QA approaches may mislead consumers and give a poor indication of what services are most likely to meet their expectations. For example, many approaches may consider exceeding consumer preference in a particular direction (such as increasing video resolution) as desirable. However, this may actually impede the quality of service experienced by a consumer who is using a portable low-powered device. The lack of effective support for discovering high quality services (as distinct from high performance services) is a significant issue which must be addressed before SOA will be widely adopted. Quality assessment approaches should consider a conformance-centric view of quality on the part of consumers in resolving this issue.

More fundamentally, whilst exhibiting heterogeneity in the needs and capability of its component entities, SOAs must be supported by shared protocols to allow the exchange of information. A major component of such exchange must be concerned with the representation of quality aspects of services. In order to allow the effective and meaningful exchange of information, a common model and understanding of the concept of quality must be adopted across the entities of the environment and embodied in its protocols. In addition to affecting the details of protocols, such a shared understanding will define the rational behaviour of the consumers and providers within the environment.

1.1.3 Heterogeneity of Information Sources

Heterogeneity is also evident in an SOA in terms of the sources of information available to consumers. In existing SOA environments, a myriad of approaches to gathering information about services exists, and are embodied in the research fields of *service representation and discovery* [159], *collaborative filtering* [12], *trust & reputation* [175] and *performance assessment* [63]. The information obtained from each of these sources may vary in its accuracy, and agents within an SOA will seek to obtain a more reliable indicator of the quality that might be received from a particular provider by consulting multiple information sources.

In terms of the information that they provide and require, the large majority of existing approaches can be considered inadequate in supporting a consumer in combining information from various sources and determining the quality that might be received in choosing a particular service. Many approaches, especially those relating to collaborative filtering provide a set of *recommendations* to consumers which indicate the services that the assessment service believes that the consumer would consider as high quality. Such recommendations are difficult to combine with other sources of information, as they do not usually provide any quantified value that could be used to infer the degree of conformance with a consumer's requirements. Approaches which do provide information representing a quantitative assessment of a service's likely future behaviour are an improvement, but are also flawed in that they do not usually provide an indication of the likelihood that the assessment is correct.

Allowing for the combination of assessment results with other sources of information and minimising assumptions will lead to better support for consumer decision making. In order to achieve this, quality assessment approaches would most usefully produce both a prediction and confidence measure, indicating the likelihood that the prediction is correct.

1.1.4 Service Provider Dynamicity

The behaviour of services in terms of the level of service delivered to consumers may change due to uncontrollable factors or as result of the provider's direct intention. Any such change in level of service provided will affect the degree of conformance of the service to the consumer's expectations, and hence the quality of the service delivered. Taking a conformance-centric view of quality has two direct implications for the behaviour of service providers. Firstly, if providers are aiming to satisfy consumers, they should aim to meet those consumer's expectations and should not assume that an increase in performance will necessarily lead to increases in quality for all consumers. Secondly, since providing higher levels of performance is generally more costly than lower levels, it is cannot be deemed rational to expend valuable resources on maintaining high levels of performance to consumers who would not perceive a higher quality from doing so. As such, in adopting a conformance-centric view, providers may segment services into service classes, that are provided to consumers with particular expectations. It is likely that such policy will be kept private, since it effectively captures the individual business model of the provider.

If providers provide different levels of service to consumers based on their expectation, it is not justifiable for assessment approaches to consider all historic performance of a particular provider as equally relevant when making an assessment. In doing so, such assessment would produce a result which represents an average performance of the provider over consumers in all service classes, rather than the performance of the provider in terms of service consumers who are in the same service class as the one requesting the assessment. Existing approaches are flawed in this respect, and as such would be ineffective in an environment where providers adopt a conformance-centric view of quality. A number of approaches do collect expectation information prior to assessment (such as collaborative filtering approaches and expectation-based assessment techniques) but do so in order to identify users with similar tastes, whilst still assuming that consumers will be treated equally by each individual provider regardless of their expectation.

In an environment where providers adopt a conformance-centric view of quality, whereby consumers are segmented by a provider based on their expectation, approaches which identify a subset of relevant data from the set of all historic performance data for use in assessment will allow better predictions to be produced. Such approaches will therefore lead to consumers in these environments discovering and using services which are of higher quality than would be used if existing approaches to QA were used.

1.2 Aims and Contribution

Integrating conformance-centric quality assessment into an SOA to support the decision making process of consumer agents poses a set of distinct issues described above. The aims of the work described in this thesis are directly aligned with the achievement of this integration. In realising these aims the work described in this thesis makes the following specific contributions to the discipline of Computer Science.

1.2.1 A Comprehensive Model for a Conformance-Centric SOA

This thesis presents the first comprehensive model of an SOA in which the principles of a conformance-centric view of quality are adopted by consumers, providers and third parties. In such an environment, consumers are most satisfied when their requirements are met, but not necessarily exceeded, and the rational behaviour of providers is guided by the principle of providing services to consumers that meet these needs. The model contrasts with existing SOA in that performance is typically taken to be the underlying principle towards which both providers and consumers aspire. The role of a quality assessment approach is defined in such an environment to allow consumers to experience high quality services, through supporting their decision making by providing information relating to a provider's historic performance.

Developing a definition of quality from first principles then taking a holistic approach in understanding how the entities within an SOA might behave if such a model of quality were adopted provides a framework for the development and evaluation of QA approaches. By developing a framework of agent behaviour based on a shared understanding of quality approaches to assessment may be robustly and defensibly evaluated within a well defined set of assumptions. Such a framework also provides a means to evaluate different policies of consumers and providers with respect to quality though demonstration of such a capability is outside the immediate scope of this thesis.

A complementary conceptual model is defined to be used to describe and contrast approaches to quality assessment from a range of traditionally disparate fields – including collaborative filtering, trust & reputation and performance assessment – in terms of how data relating to service provision are combined in order to generate predictions. The model is used to provide a taxonomy of existing approaches. This thesis seeks to demonstrate its contribution by using the model to define and describe a new type of quality assessment approach that is different to existing approaches at a conceptual level.

1.2.2 Conformance-Centric Approaches to Quality Assessment

A conformance-centric view of quality implies that the quality of a particular service may only be judged when the expectation of the consumer making the quality judgement is known. In terms of QA, there are potentially two architectures for supporting this process. Firstly, consumers might state their expectation to an assessment service that would then carry out an assessment of available services and determine the one which best meets the consumer's expectation. This method is considered as being most appropriate in many current approaches to assessment including collaborative filtering and many reputation based approaches. However, consumers are likely to consult a number of sources of information relating to the performance of services (see Section 1.1.3) and will have the capability and desire to combine these results in making an informed selection. Approaches which adopt service ranking and subsequent recommendation of a single (highest ranked) service will therefore be ineffective in supporting consumer decision making as no indication of degree is provided, which would be necessary in order to combine the results with other evidence.

In the second approach, making a clear separation between the role of QA as a source of information and the role of consumer as decision maker will allow a wider range of consumers to be served by the QA service. In this model a prediction of the *level* of service that is most likely to be received by the consumer from a service is provided along with an associated *confidence* in the prediction allowing consumer agents to easily combine QA information with other sources using existing techniques.

This thesis develops and evaluates a range of approaches that seek to fulfill the role of QA in the conformance-centric SOA. The range of approaches share a common principle in recognising patterns in service behaviour that are due to a provider adopting a conformance-centric policy in managing its services. The approaches differ in terms of their complexity and the degree to which they assume the availability of information relating to provider behaviour. Such approaches are demonstrated to be necessary and effective in SOA where a conformance-centric view of quality is adopted by consumers and providers. In contrast, existing QA approaches which assume a non conformance-centric perspective are shown to be ineffective in such environments.

In contrast to existing models of quality a conformance-centric model considers highest quality as being experienced when a consumer's expectation is met and that exceeding expectation cannot be assumed to be desirable in general. By developing conformance-centric approaches to the assessment of services it can be shown that consumers will be better supported in identifying and using the services which best meet their needs. Improvements in the quality received by consumers will lead to greater confidence in SOA as a means of effectively supporting consumers.

1.2.3 Environments for the Empirical Evaluation of Quality Assessment Approaches

The conceptual model of a conformance-centric SOA described above constitutes a significant change in terms of the way in which consumers and providers behave especially in the way that information relating to quality of services is represented, gathered, processed and exchanged. In order to demonstrate the contribution of the new architecture and to facilitate the rapid development and evaluation of new conformance-centric components two implementations of a conformance-centric SOA have been developed.

Constraint Oriented Negotiation in an Open Information Sharing Environment on the Grid (CONOISE-G)
The CONOISE-G architecture (developed as part of the CONOISE-G project[108]) provides an agent-based framework for the formation and management of virtual organisations (VO) in a grid environment. The conformance-centric quality assessment architecture developed in this thesis is embodied in the Quality of Service Assessment (QoSA) and Quality of Service Consultant (QoSC) components of the architecture and may be used to support consumer agents in discovery and selection of services based on quality requirements during the formation and perturbation VO phases [111].

Discrete Event Simulator for Conformance Centric Quality Assessment
A software simulation platform has been developed which constitutes a dynamic model of the conformance-centric SOA described above. The platform can be used to develop, test and compare approaches to QA, supporting fine-grained control over scenario parameters and a range of tools for the monitoring and evaluation of QA performance. The use of the platform removes the need to establish and tune a live environment for the evaluation of approaches, which would be costly in terms of effort and is inflexible.

1.3 Thesis Structure

A review of the background to the concept of quality and existing work in the field of quality assessment in SOA is given initially to place the work of this thesis in context and to refine the aims of the approaches to be developed. This is followed by the presentation and discussion of a model of a conformance-centric SOA and presentation of approaches to QA including a qualitative and empirical evaluation of these approaches in a range of scenarios. Specifically, the thesis is structured as follows:

- **Chapter 2 - Quality Assessment in Service Oriented Architectures:** provides a survey of the current work in QA and related fields. The survey identifies the extent of current research and the limitations of existing approaches with respect to the aims identified.

- **Chapter 3 - A Conformance-Centric Service Oriented Architecture**: provides a formal conceptual model of concepts relevant to supporting a conformance-centric view of quality assessment in a service oriented environment. A formal definition of Quality of Service (QoS) is proposed within the context of SOA. This definition is used as the basis for explaining the role and behaviour of service consumers, service providers and infrastructure services. The chapter concludes with a statement of the QoS assessment problem in terms of how assessment may take place in the given context.

- **Chapter 4 - Quality Assessment in a Conformance-Centric Service Oriented Architecture:** develops a range of approaches to QA based on statistical and clustering techniques. The approaches developed are designed to operate in SOA where a conformance-centric view of quality is taken, with details of the policy of providers being partially or fully hidden. The approaches developed generate predictions and validity scores which may be used by consumers to infer the expected utility (and hence quality) that might be experienced when using a particular service.

- **Chapter 5 - Evaluation of Quality Assessment Approaches:** describes two implementations of the Conformance-Centric SOA defined in Chapter 3. A qualitative evaluation is provided through demonstration of the role of quality assessment in the CONOISE-G architecture, using a practical scenario of VO formation and perturbation. The simulation application used to produce the empirical results described in Chapter 6 is described in terms of its architecture, behaviour and interfaces.

- **Chapter 6 - Results:** presents a set of criteria for the evaluation of QA approaches, and defines a set of experiments to demonstrate the performance of the approaches described in Chapter 4, under a range of realistic scenarios. The simulation described in Chapter 5 is used to carry out these experiments and a discussion of the empirical results and their significance is presented.

- **Chapter 7 - Conclusions:** presents the conclusions which have been drawn as a result of this research, including consideration of potential future work.

1.4 Publications

The work developed as part of this thesis has resulted in a number of publications in academic peer-reviewed conferences. A selection of significant publications is provided below, along with a description of their relationship to the material contained in this thesis.

- **Incorporating QoS Specifications in Service Discovery [29]**: described early work in developing a basic Quality Assessment service to allow discovery of providers based on quality expectations of consumers.
- **Managing Quality of Service in Virtual Organisations [147]**: first proposed the conformance-centric architecture (see Chapter 3), comprising Quality Assessment and Quality Monitoring tools.
- **Supporting QoS Assessment and Monitoring in Virtual Organisations [134]**: included discussion of the basic requirements for a Quality Assessment approach, that have been formalised in Section 3.6.
- **Agent-based Virtual Organisations for the Grid [107]**: described the CONOISE-G architecture (see Chapter 5), including the role of conformance-centric Quality Assessment in supporting the formation and perturbation of Virtual Organisations.
- **A Multiple Quality-Space Mapping Approach to QoS Assessment [135]**: first described the concept of quality-spaces (discussed in Section 3.7) as a means of modelling quality assessment methods, proposing and evaluating an approach which formed the basis for the development of the "pattern passive" approaches described in Chapter 4.
- **QoS Assessment of Providers with Complex Behaviours: An Expectation-Based Approach with Confidence [136]**: proposed a more advanced approach to quality assessment based on quality space mapping and demonstrated the role of confidence in supporting discovery of high quality services (using early versions of the techniques defined in Section 4.3.1).

Chapter 2

Quality Assessment in Service Oriented Architectures

The act of quality assessment (QA) is carried out in order to support the consumer in determining the degree to which a particular service is likely to meet their needs. In this broad definition of the area three fields can be considered as constituting the current research into the assessment of the quality of services in SOA. The first of these, *Trust and Reputation* [114] uses consumer feedback to distinguish good services from bad ones [162] - reputation being used to describe the collective opinion of a particular provider. In a related area *Collaborative Filtering* [59] is used to predict the "worth" of a particular product or service to an individual based on observation and analysis of trends in feedback ratings. Finally, *Performance Assessment* [80, 96] utilises monitored performance information in order to predict the likely future behaviour of a provider in terms of the level of service which might be received. Although differing in terms of terminology and the information and techniques used, at a conceptual level these three fields can be considered as having the same objectives.

This chapter begins with a discussion of the concept of quality from a traditional viewpoint. The section draws out a definition for quality that will be used in this thesis. The survey which follows this discussion encapsulates work from the three identified research fields as well as work from other relevant fields. Within this survey existing models, methods and techniques are presented and classified in order to demonstrate the extent of the research and to identify current research issues and challenges. Finally, a summary of the state of the art, in which the research challenges identified throughout the section are collated into a coherent form is presented. These challenges form the basis for developing a conceptual model and set of approaches described in the following chapters.

2.1 The Meaning of Quality

In "Zen and the Art of Motorcycle Maintenance" [110], Robert Pirsig relates the story of one man's vain attempt to separate the concept of quality from the subject which it describes. That is, to find an objectively defining characteristic which make some things "good" and others "bad". The ensuing struggle, which leads to the character's eventual insanity never quite provides an answer to the original question, but this remains an important point which warrants further elaboration in order to understand the principles on which this thesis is based. What is Quality? And furthermore, why is it that a word which is used frequently and confidently, in many different contexts proves so elusive in its definition?

When faced with such challenges about the fundamental concepts of a subject it is usual to begin by examining the definitions used by authorities of the area in question. However, Quality is not a concept unique to a particular discipline - it is ubiquitous across all walks of life - and those organisations and communities of practice which profess authority on the concept [2, 113, 62], on closer inspection reveal themselves to be concerned with quality within a particular domain. For example, the American Society for Quality (ASQ) [2] focus primarily on the quality inherent in manufacturing processes and products (though it does apply these principles to other domains) whilst the Quality Assurance Agency (QAA) [113] is concerned with maintaining and improving the academic standards and quality of higher education in the UK. Such organisations adopt a statement of quality which reflects aspects of the discipline concerned - the ASQ defining quality as being "the characteristics of a product or service which bear on its ability to satisfy stated or implied needs". Even within particular industries, competing standards and approaches disagree on what constitutes quality. Whereas the ISO9000 standards [62] (which cover "Quality Management" within organisations) define quality in terms of a customer's quality requirements, customer satisfaction and increased performance, the Six Sigma approach [112] aims to maximise organisational quality by eliminating defects in products, services and processes using statistical analysis to quantitatively define the degree of quality achieved.

One might also consider the philosophical viewpoint when examining an abstract concept such as quality. Indeed, several branches of philosophy including *ethics* [60] and *moral philosophy* [70] are concerned with the consideration of what makes objects "good". Plato developed theories in which he believed that objects which we perceive could be judged against "The Forms" - objective ideals of everything that we perceive which could be seen as representing perfection [48]. The "degree of *goodness*" of a perceived object he maintained, could be judged objectively by comparing it with its Form. Such a definition captures the idea of quality as being a comparison against an ideal. However, Plato's interpretation of the concept sees the ideal as objectively knowable (and real), being independent of any individual perspective.

Later philosophers reject this purely objective view in favour of views which capture the purposeful aspect of quality such that the quality of an object becomes an *attributable* characteristic that stems from the separation of the object from its purpose. In achieving such a distinction Savery [126] states that there is one *generic* meaning and multiple *specific* meanings of the concept. Generic goodness is usually judged, based upon the achievement of a task or goal, and specific meanings of goodness are the properties that an individual might reflect upon at the "end of action". Once a particular reflective property is chosen or stated, the meaning of goodness is changed from generic to specific. In slightly less formal terms, Savery states that although goodness (and therefore quality) is subjective in its measurement, there is commonality (and therefore generality) in the goodness of all objects in the sense that the judgement is carried out in relation to an end. In later work by Rohr [120] the following formal definition of goodness is proposed:

X is a good K iff X is a K and X answers our interests in Ks.

Such a definition captures the concept of goodness as being an attributive rather than a predicative property of an object. That is that the goodness (and therefore quality) of an object cannot be judged until the purpose to which the object will be put has been defined. Through the "answering of our interests" such a definition also hints at the subjective nature of quality which is a fundamental principle in many contemporary views of quality from the domains of business and manufacturing.

The lack of agreement on a definition of quality across and within disciplines does not negate the need for providing approaches which support its use in a particular context. Each day one makes decisions which incorporate a substantial "quality judgement" and so it would be unrealistic to consider any computational model of such situations as complete without a suitable representation and appreciation of quality. There is agreement at a high level within the definitions discussed that quality pertains to the degree to which aspects of an object *conform* to a particular set of criteria or to another object. Figure 2.1 provides a summary of these. Throughout this thesis, therefore, the general definition of quality provided in Definition 1 is adopted.

Source	Domain	Conformance Criteria
ASQ	Manufacturing	stated or implied needs
ISO9000	Organisational and Manufacturing	needs and expectations of consumers
Plato	Philosophy	The Form
Rohr	Philosophy	our interests
Six Sigma	Manufacturing	zero defects

Table 2.1: Conformance Criteria

Definition 1
Quality is conformance of an object to an ideal

In this definition the ideal may be defined by an individual as an explicit statement or implicitly; or by an organisation, industry or discipline. Such a definition is consistent with those specific definitions provided within the domains described in Table 2.1.

This definition of quality is general, in the sense that it may be applied to any domain or context without contradiction. However, applying the definition usefully to any specific domain would require a definition of the "ideal" criteria against which conformance might be judged. When seeking to apply quality in a particular domain, it may be possible to borrow or adapt existing definitions from similar domains. This is specifically applicable to the application of quality within the domain of SOA, which shares similarities with real world marketplaces. This similarity is explored in more detail when a definition of Quality of Service is provided in Section 3.1. While applying the above general definition to the work developed throughout this thesis it is important to be continuously and objectively aware of the problems of subjectivity and domain dependence, described above.

In taking a *conformance-centric* approach to quality there are effectively of aspects for consideration. Firstly, what is it that the object is conforming to. Secondly, to what degree is it conforming (this can be taken as being a measure of the current "state" of the object). The survey presented in this chapter can be considered as seeking to establish the state of the art in terms of these two aspects. That is, research will be considered as relevant to the field if it seeks to *establish*, *capture* or otherwise *represent* the features which will be in the service representation; or if it seeks to *determine the state* (through direct measurement or by other means) of the service whose quality is in question.

2.2 Abstract Quality Assessment Model

In order to provide a survey of approaches it is useful to present a model of an abstract quality assessment process (Figure 2.1). This model can be used to describe a generic assessment process in terms of four high-level concepts and the interaction between them. Each approach to assessment discussed in the next section describes specialisations of one or more of these high-level concepts. The way in which these specialisations address the problem are used as the basis for the classification of assessment approaches. Whilst there are a small number of existing surveys within the area of quality assessment these either tend to focus on one particular architecture such as agent-based computing [139, 162, 66], or web service discovery [46, 183], or focus on the application of QoS within a specific domain, such as mobile computing [20]. Existing surveys do not provide a model of the assessment process that is general enough to allow comparison of techniques between fields. Instead they rely on architecture models specific to their own field. Using the model below to frame the assessment problem at an abstract level allows direct comparisons to be made between techniques in different fields, in terms of the degree to which they support consumers in discovering high quality services.

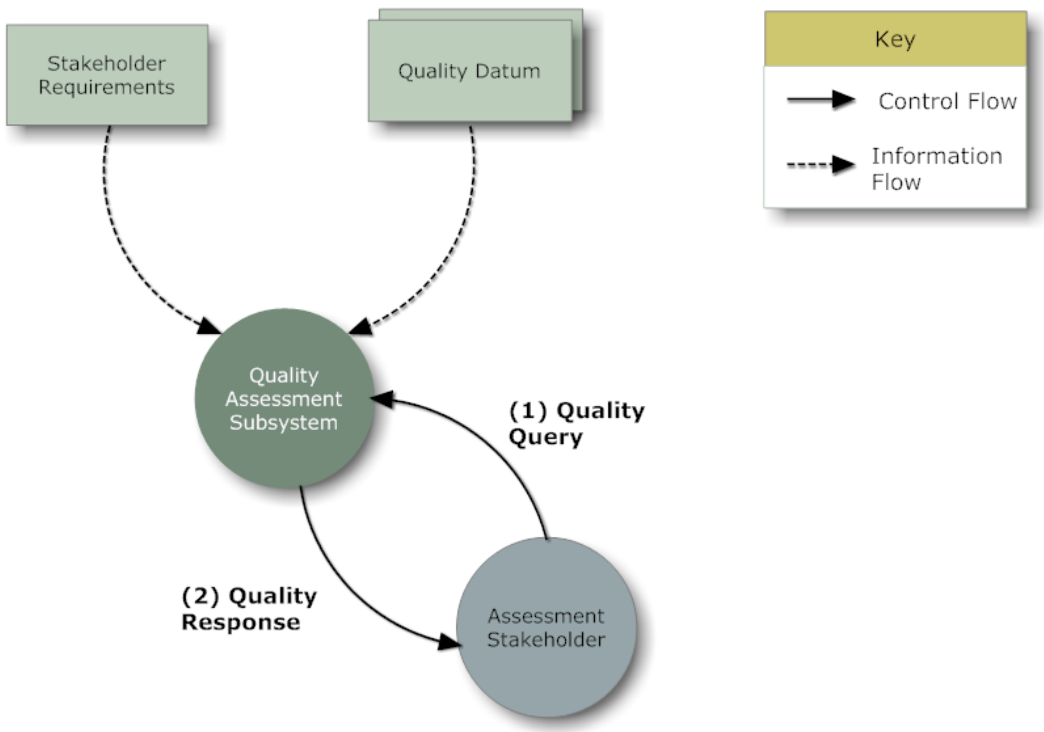


Figure 2.1: Abstract Quality Assessment Model

This abstract quality assessment model illustrates the four fundamental principles of quality assessment that are common to all frameworks and models, namely *Assessment Stakeholder*, *Stakeholder Requirements*, *Quality Datum* and *Quality Assessment Subsystem*.

2.2.1 Assessment Stakeholder

For an assessment to take place the Quality Assessment Subsystem must be invoked by a party interested in gaining information about the quality of a particular service. The content and format of the Quality Query and Quality Response messages sent between the Assessment stakeholder and Quality Assessment Subsystem vary between approaches being dependent on the method by which requirements are represented and information is gathered. The assessment stakeholder is the entity that initially invokes the quality assessment subsystem and can usually be considered as being the entity that receives the most direct benefit from the information gained as a result of that assessment. Since stakeholder goals may differ it may assess different aspects of individual services in order to verify adherence to service level agreements (SLA) [15]. If the assessment stakeholder is a *provider* of services it may assess the performance of services in order to ensure that it is able to meet the requirements of its customers or as a means of supporting their advertisment [131]. Alternatively, a service *consumer* may wish to determine the expected quality of a service in order to minimise the risk when deciding which service to choose and to maximimse their own individual utility. Finally, in closed or centrally controlled marketplaces the assessment stakeholder may be the *environment* owner. In such cases, the goal of assessment is to ensure that some global utility – such as the total profit or income across all services, or the average quality experience across all consumers – is maximised.

2.2.2 Stakeholder Requirements

Stakeholder requirements represent the preference of the Assessment Stakeholder. This is used by the Quality Assessment Subsystem as a template for expressing or inferring the stakeholder's ideal (see Section 2.1). Services are compared to this template in order to determine their overall value to the stakeholder. In certain assessment models, the stakeholder is not responsible for explicitly stating requirements - requirements in this case being assumed by the quality assessment subsystem. The preference represented by consumer requirements can be considered as an indicator of the ideals in the definition of quality (Definition 1) against which a service may be judged as conforming.

Differences in how the semantics and syntax of these requirements are expressed will influence the design choices of the approaches to quality assessment. At a high level, approaches may be separated into those, that require an *explicit* statement of the assessment stakeholder's requirements and those that use *implicit* methods to deduce the stakeholder's requirements. The approaches to quality assessment developed in this thesis adopt a point-based model to represent explicit consumer requirements (as a single numerically-expressed concept). The discussion of the full range of requirements expression provided in this chapter is included to illustrate the practical significance of differences between explicit and implicit methods.

2.2.3 Quality Datum

A single Quality Datum represents a piece of information that may be useful in determining the quality of a service. The Quality Assessment Subsystem will always attempt to use at least one Quality Datum in order to make an assessment of a provider - in fact, it will usually collect and store a set of Quality Data in order to make an assessment. Quality Data provides *evidence* to determine the most likely behaviour of a service in the future. This may be used in turn to determine the likely degree of conformance of a service to a consumer's ideals. At a high level individual items of data used by approaches to quality assessment may be considered as representing an *advertisment or certificate*, *consumer rating* or *objective measurement* of a service.

2.2.4 Quality Assessment Subsystem

The purpose of the Quality Assessment Subsystem is to respond to the Assessment Stakeholder's initial query and in doing so provide information regarding the service that is the subject of that query. This information will be a function of both the set of Quality Datum available to the subsystem and the Stakeholder Requirements though the techniques and assumptions used when combining this information vary widely. In practice, in addition to carrying out the assessment the Quality Assessment Subsystem is usually responsible for maintaining the quality knowledge that is gathered from the environment.

Consumers are interested in determining conformance of a service to their ideals, to determine the level of quality they will experience. However, the quality assessment subsystem may not conduct the comparison between ideals and the service itself instead providing information which is useful to the consumer in making the decision. As discussed in Section 1.1.3 in certain situations it may be more appropriate for an assessment service to return a prediction of service performance rather than a prediction of quality in order to allow the consumer to determine conformance themselves.

2.2.5 Distribution of Assessment

All assessment approaches must take ownership of and utilise information from a number of sources in carrying out the process of assessment. The act of assessment necessitates some processing of the information - a process which must be controlled so that it is carried out effectively. However, the information, processing and control aspects of assessment are not necessarily dependent upon each other and thus, each may be carried out by different entities at different locations and under different conditions in the environment. It is important to note that in the above abstract model, four key components are distinguished in order to demonstrate their interaction. In many models, individual components may be integrated - for example, the Assessment Stakeholder and Quality Assessment Subsystem may be part of the same entity.

2.3 A Taxonomy of Quality Assessment Techniques

Having established the purpose and scope of quality assessment research, the taxonomy presented in this section is used to categorise current approaches based upon a number of criteria, typically related to one or more of the high-level concepts described in the abstract model, namely: the role of the assessment stakeholder, the nature of quality information used in an assessment, the distribution topology and the method of capturing stakeholder requirements. For each criterion, a classification including an overview of each class is presented highlighting existing research in the area and identifying relevant research challenges. In general, each individual framework, approach and technique described in this review will not cover all aspects of the abstract model. That is, it is more likely that each will define a method which covers only one aspect such as the specification of stakeholder requirements. In such cases only the aspects that are directly relevant or applicable to the area of quality assessment are identified. A classification based upon the concepts defined above provides an effective and useful partitioning of existing approaches to assessment that may be used as the basis for drawing out and clarifying the research challenges that are addressed by this thesis in later chapters.

2.3.1 Assessment Stakeholder Role

The volume, type and quality of information available to and required from the quality assessment process is largely dependent on the assessment stakeholder. This has implications for the design of approaches to assessment, discussed below.

2.3.1.1 Provider as Stakeholder

A key issue for providers in providing services to consumers across an SOA is being able to build confidence that each service "can deliver a function that is able to deliver the best [quality] to consumers" [15]. For a provider the goal of using information gained from assessment is in informing the provider to what degree they are meeting (or are likely to meet) their quality guarantees to consumers. This information may then be used by the provider to change their policy, or as a supporting piece of evidence which may be used in advertising the service. Ran [115] describes a framework, where providers may use a trusted third-party service to certify the level of QoS provided. In a similar approach proposed by Serhani et al. [131] a QoS broker generates and executes a set of test cases which aim to test the functional and QoS aspects of a service at runtime. In both approaches the assessment results in the production of a certificate, a copy of which is stored with the third-party broker. Such certificates may be referred to by the provider when advertising their services in existing web service registries such as UDDI [102]. By using an independent third-party to verify their service, a provider may be able to gain a competitive advantage over others who use a self-certified advertisment [131]. However, the level of performance delivered by any provider is likely to be dynamic [66] and vary over time and as such confidence in the certificate could diminish with time. Furthermore, it is unlikely that the certifier will be using the same network configuration as either the provider of the service, or the potential consumer of the service [15]. This means that the level of service provided on the certificate may not be an accurate indication of the level of service that would be received by an individual consumer on a particular network. For the same reason the independence or trustworthiness of any third-party might reasonably be called into question by the consumer [54].

In contrast to gaining leverage with consumers through certification a provider may choose to use service simulation to test the robustness of their services. By simulating the use of its services as if it were a consumer [86] the provider may increase its confidence in being able to meet consumer requirements. This activity may be done prior to the release of services [109], or after their release [14]. In early work Hafid et al. [52] propose and evaluate an approach whereby the provider invokes its services in multiple instances using a range of scenarios to evaluate its own performance. Within the web-services domain the WebSob approach [86, 87] uses existing unit testing tools to automatically generate and execute a set of black-box tests which aim to reveal robustness problems within a provider's web services. Similarly, DiPenta et al propose the use of genetic algorithms in automated web services test case generation and execution [109]. The reliability of the results of such testing of services by providers is limited. Since the network configuration of the testing platform is likely to be different to that of individual consumers there is no guarantee that the results of the performance based tests will be relevant to consumers using the service in a live environment. However, such testing is appropriate for identifying those SLA violations that may occur as a result of particular combinations of inputs [15].

2.3.1.2 Consumer as Stakeholder

In an SOA individual consumers will have very limited knowledge of the services available prior to using them and thus will seek to use tools within the environment to discover and select the best services to use [104]. A significant criteria for selection will be the level of service that the provider is realistically able to commit to [97] since this will in turn influence the level of quality that the consumer perceives. The current work in QoS assessment that considers the consumer as the main stakeholder focusses on two key areas. Firstly, on the use of ontologies and languages to represent consumer QoS requirements and service properties. Secondly, on the development of techniques to use historic performance information to determine the likely behaviour of a service in the future. Many approaches combine these two features since they are so closely dependent. As part of the G-QoSM framework [4], UDDIe [132] proposes annotating existing service advertisments in a UDDI registry with QoS information to support consumers in locating service providers who advertise particular QoS characteristics. In similar work the WS-QDL (Web Services Quality Description Language) [171] allows consumers to discover services based on QoS criteria using a language which enables the description of QoS concepts including those related to the "dynamic, statistical and static aspects of the service". Padovitz et al. [103] propose an approach using mobile agents which retrieve peformance information directly from providers in order to carry out dynamic run-time selection of services. In each of the above approaches the assessment process results in the collation of a list of providers who meet a minimum set of consumer supplied QoS criteria. These approaches effectively classify all service providers into a set that are able to meet the consumer's requirements and a set of those that are not.

A number of approaches to the modelling of QoS semantics have been proposed which provide a richer representation of a consumer's QoS requirements using ontologies. Such approaches provide a more fine-grained matchmaking capability so that the degree of match may be improved. The WS-QoS framework allows service clients and providers to specify requests and offers with QoS properties [156], providing the means for providers to update their offerings over time. A *Web service broker* acts as intermediary, comparing consumer requests with provider advertisments to determine the offering which best matches the consumer's requirements. dvQL [144], a DAML-based [53] query language provides an ontology, and query and manipulation language for QoS in web services. As with UDDIe [132], existing provider advertisments have QoS-specific information attached. The associated query language may be used by consumers to discover and utilise information regarding particular aspects of the advertisment. The DAML-QoS [180, 181] ontology augments the principles of QoS semantics by including a matchmaking algorithm which is designed to indicate a *degree of match* between a consumer's requirements and a provider's QoS advertisment. By doing so the manual effort required on the part of the consumer in evaluating and interpreting results in order to evaluate the quality of a provider is reduced. In such cases the degree of the value serves as an indicator of the likely quality of service that a consumer would receive. A comprehensive ontology and matchmaking approach is proposed by Maximillien and Singh [92, 95, 96]. In their approach statistical and logical relationships between particular QoS attributes are represented as part of the QoS ontology. Such relationships might represent the fact that some QoS properties are related (for example bandwidth and framerate), or that others are semantically similar (for example availability and mean time to failure). By increasing the richness of the semantic descriptions of web services the approaches outlined above allow consumers to discover the services which based on provider advertisments appear to best meet their needs. However, the use of provider-supplied advertisments as the sole means of determining which is the best service to use is inherently limited - regardless of the complexity of the matchmaking or richness of semantics - since it is likely that some providers may inaccurately state their ability to provide services (intentionally or unintentionally). Furthermore, even if providers are completely honest about their abilities since the QoS of individual services varies with time it is unlikely that provider supplied advertisments will represent the most up-to-date view of the services they provide.

Thus, a significant amount of work in the literature is concerned with using information about the past behaviour of providers, in order to improve the reliability and timeliness of information regarding their future behaviour. Histos and Sporas [175], propose two complementary approaches aimed at discovering good quality services utilising ratings provided by consumers to determine the reputation of individual providers. Such ratings are provided after the consumer has used a service and are stored in a centrally accessible database. The calculated reputation score is used as an indicator of the trust that a consumer may have in a provider satisfying its requirements. Both approaches include methods to disincentivise changing identities, and feedback from reputable consumers is given a higher weighting when calculating the score for a service. In order to establish the reputation of consumers, it is assumed that both providers and consumers are rated after provision. In such cases, the usefulness of the ratings of reputable consumers is not fully utilised, until that consumer has supplied a significant number of ratings. In contrast to using a central repository of ratings the approach proposed by Wang and Singh [161] assumes that consumers (or agents) operate within a completely open environment where no trusted third party may exist. They propose trust operators which allow the aggregation of reputation information from sources in an effective way. Other approaches to reputation calculation use ratings from across a network of agents or peers [179, 176, 47]. A comparative analysis of these approaches is provided in Section 2.3.3.

Billhardt et al. [9] separate the idea of trust from that of confidence. The former being taken as an external recommendation of the degree to which a provider is "good",and the latter as an internal measure of a consumer's direct experience with the provider. The approach proposed considers that both pieces of information will be combined when the consumer is making a final decision of which provider to choose.

Other notable approaches to reputation calculation include the Beta Reputation system [61] which utilises a beta-distribution model based partially upon ratings supplied by consumers, a neural network based reputation model proposed by Song and Phoha [142], and a fuzzy-logic based reputation model proposed by Rubiera et al [123].

Many decentralised collaborative filtering approaches in which items are recommended to consumers based upon ratings provided by people with similar tastes [133] where the ratings and recommendations are stored and calculated across the network can be considered as having consumers as the primary stakeholders. Such approaches differ from traditional recommender systems, which tend to operate in closed environments for the major benefit of the environment owner. The PocketLens recommender system, addresses issues of portability and trust through the use of a collaborative filtering mechanism distributed across a peer-to-peer network, evaluating a range of methods for judging consumer and item similarity. Similarly, Weng et al [161] propose an approach in which multimedia content across a P2P network is recommended to individual consumers based on preferences implied by previous download activity. The approaches to quality assessment described above result in the production of a reputation score for the provider in question. Such a score may be considered as representing either the likely rating that the consumer might give to the service if they were to use it, or the likelihood of the consumer having their requirements met.

In contrast a large number of approaches base assessment on the use of historic *objective* performance information. Such assessments tend to result in the production of a prediction in terms of the likely delivered performance that might be received by a consumer if they were to use the service. Such an approach proposed by Gao et al. [45] utilises objective performance information to predict the future performance of a service in terms of a set of QoS attributes. A consumer may in turn use such a prediction to determine the expected degree of conformance of the service to his ideals and hence the level of quality he is likely to receive. Utilising an Artifical Neural Network (ANN) for each provider the approach requires objective performance information to be provided in order to train the model. When a consumer requests an assessment of a provider the parameters of the ANN are set based on the consumer's specified QoS requirements and results are obtained. Similar approaches are proposed by Zhou et al. [180] in which a federated database of performance information is used to allow consumers to determine the likely behaviour of providers in their local domain. Liu and He [79] and Juhua et al. [67] have developed approaches to the selection of grid services using objective performance information and Cardoso et al. [18], Anselmi et al. [5], Silver et al. [137] and Cao et al. [16] have developed approaches for QoS aware composition of web services based on objective performance information.

The results of the approaches utilising reputation and objective performance information present a value (or values) which can be treated by the consumer as an indication of the likely level of quality that might be received from the provider. Such approaches assume that the consumer will use the information to decide which provider is preferred. In doing so the consumer may use internal models of their own preference and may consult additional sources of information before making a final decision.

In contrast a final set of approaches to assessment which consider the perspective of the consumer as a stakeholder take a further step by acting as not only a source of information regarding providers but a tool to *select* the best service for a consumer. The method of assessment itself is similar in purpose to those discussed above however rather than returning a prediction for a given provider, such approaches will carry out an assessment of all providers and return the one which they deem as best in meeting the consumer's needs. The approach developed by Manikrao and Prabhakar [84], using a DAML document to specify functional and non-functional requirements, which are semantically matched against service descriptions registered by providers results in a selection of the service that is deemed to be the closest semantic match. This approach also incorporates a recommendation system using an item based collaborative filtering approach [125] which aims to ensure that as well as being semantically well-matched the chosen service would likely be rated highly by the consumer. The QoS_IC framework proposed by Taher et al. [148] also considers semantic matchmaking as a means of selecting an optimal service based on a consumer request. In the QWCS architecture [182, 174] a QoS broker is deployed between clients and web service providers. Objective performance information regarding the response time, service cost, network bandwith and service availability of each service is monitored with the results being stored in a centralised database. In addition to calculating the expected performance of each service based on a consumer's assessment request the QoS broker handles the negotiation and binding of the service consumer and service provider. The DQoS model [56], a general decision model for choosing between services based on historic performance information is similar in approach to QWCS although it utilises multiple attribute decision making (MADM) techniques across an extensible set of QoS criteria rather than considering a simple matchmaking of a consumer's individual requirements to the highest performing service. Guan et al. [51] propose a service selection method based on *trust force* which utilises social networking concepts and coulombs law in determining the best service based on objective performance information. The approaches proposed by Lamparter et al [76], Hossain et al [55], Talwar et al. [150] and Maximillien and Singh [91, 96, 94] also fall into this category, and are considered in detail in Section 2.3.2.

The final set of approaches considered above, which carry out a selection process to recommend a particular service to the consumer exhibit a number of weaknesses. Firstly, since such approaches remove the consumer from the decision making process in order to be effective they must be confident of the exact needs and preference of the consumer. This is evident in the observation that some assessment frameworks taking this approach propose rich ontologies [95, 180] in order to represent consumer QoS preference to support the assessment and selection process. However, the decision making process of a consumer is very complex and it is unlikely that even the most comprehensive ontologies proposed will be able to capture all relevant aspects. Furthermore, such approaches are based on the underlying assumption that a consumer is capable of expressing their QoS preferences effectively. Current research would suggest that this is not the case and that most consumers perform poorly at expressing requirements especially when talking about service specific and low-level QoS criteria [63]. Secondly, in an open environment where there may be many different sources of information regarding a provider and its services it is unlikely (and less effective) for a consumer to base a decision on a single information source which may not be completely trusted. Rather, as in a real-world marketplace the consumer would prefer to consult multiple sources and make a final decision based on a balanced opinion. By returning an evaluation or prediction of a specific provider's performance approaches to assessment are effectively specifying a *degree* of recommendation. This is in contrast to approaches which directly recommend services and where the consumer is unable to see how much better the recommended provider is than the others.

2.3.1.3 Environment as Stakeholder

In completely open environments which are characterised as containing entities that are autonomous, unpredictable and heterogeneous in their structure and preferences it is useful for consumers to be able to assess individual providers in order to determine their likely performance. In such environments there is no one entity with environmental omniscience and therefore consumers may need to consult multiple information sources in order to make the best possible decision. The purpose of such environments is merely to provide the infrastructure and protocols to serve the mutual needs of its component entities. However, in many instances where the environment is closed i.e. owned by a particular organisation, or where there is a central point of control the role of the environment may be more than structural. In such cases the owner of the environment may have a preference as to how its component entities behave and may desire to control their actions to a greater or lesser extent. In effect, where a marketplace existing between consumers and providers is controlled by a central authority, the environment itself may be seen as being a stakeholder in any assessment process. By controlling the assessment process the environment owner may be able to encourage consumers to use particular services over others thereby maximising their profit or maximising the usage of these resources. The process of resource brokering within the fields of distributed [52, 141] and Grid computing [13, 129] may be considered as constituting this type of marketplace. A resource owner responsible for a set of resources (services) must manage a set of incoming requests (consumers) who have some requirement which may be satisified in terms of the qualities exhibited by the resources. In such a case, the goal of the environment owner will be to coordinate the system components to maximize system-wide utility [75]. A comprehensive example of processes supporting such an environment is proposed by Dan et al. [26] in which consumer requests expressed in terms of SLAs are mapped to a set of resources under the control of a service provider. A dynamic monitoring and periodic assessment process takes place across all resources of the system to ensure that overall resource usage is efficient whilst still meeting consumer demand. A similar approach is also proposed by Cardellini et al. [17] to manage SLAs which may refer to composite web services.

The approach proposed by Bennani and Menasce [7] to managing the resources of large autonomic data centres demonstrates similar goals to those employed by Dan et al. though in this case the resources to be controlled may be owned by a number of different service providers. The management process used contains a *prediction model solver* which periodically assesses each resource in the data centre and based on observed workload measurements predicts future performance.

Research into using utility functions as a means of representing the optimality of particular environmental states (as carried out by Kephart and Das [75]) allows a data centre or resource administrator to specify preferable system states which an autonomic process seeks to achieve by a process of continuous reassessment of resources and readjustment of their allocation. In a web services environment two approaches consider the role of a QoS broker which services assessment requests from consumers to support environmental optimisation [98, 24]. Both approaches maintain that endowing the broker with the capability to determine the load on all providers in the environment might allow services to be distributed more evenly so that the best services do not become overloaded quickly. The stakeholder in such approaches might be considered as being the consumer (who is trying to discover the service which will offer the best quality), the provider (who is trying to gain custom from as many consumers as possible without overloading the resources), and the environment (which seeks to maintain a workable balance of requests across services). However, it is unlikely that such an approach could be practically deployed since it relies on the openness and honesty of providers when stating their resources and will result in some consumers receiving lower qualities of service than others. Much work in the area of collaborative filtering focusses on the provision of recommender systems to a set of consumers within a closed environment with recommendations generally being bounded to the products or services of a single service provider. Such environments are typically operated for the direct benefit of the environment owner though individual consumers may also indirectly benefit through the provision of recommendations which better meet their needs. Collaborative filtering systems can be broadly broken down into two types: user-based filtering [12] and item-based filtering [125]. In the user-based filtering approach items are recommended which have been rated highly by other consumers who are similar to the current consumer. In the item-based filtering approach, items are recommended which are similar to other items which have received high ratings from this consumer, the item-similarity to be judged is based on some set of known characteristics.

The Amazon.com recommender system [78] utilises an item-based collaborative filtering algorithm in order to target particular products and services to consumers. The approach utilises a two stage process in which item-similarity tables, containing the average similarity in ratings between each pair of items available on the site are first produced. The generation of such tables is a computationally intensive task and as such is carried out offline. In the second stage, each time an item is viewed by a site user the similarity table is queried for the items which are most similar to the current item these items being subsequently displayed in various ways to the consumer. In such systems users are generally expected to provide ratings of items that they would like or have already used or own. The power of such approaches is embodied in the omniscience of the environment owner who is able to observe all actions taken by consumers in the environment.

2.3.1.4 Summary

The assessment of services by a provider is done to support their internal assurance processes in order to build confidence in their ability to provide services. In such cases it can be reasonably assumed that all information regarding the performance of the services is available to the assessment process. Furthermore, providers will usually have detailed knowledge of the structure of their services and will be able to generate a comprehensive set of test cases. Although the methods employed by providers may result in accurate and comprehensive appraisals of their services these benefits are not immediately transferable to consumers or independent third party assessment services. This is because by doing so providers would reveal details of the internal operation of their services which they do not wish to share with their rivals for business reasons. Also, when the stakeholder is the environment it can be assumed that the information required by the assessment process is reliable and complete. In achieving its objectives of maximising some overall utility the environment will usually endeavour to increase the satisfaction of the consumers. This may be because the environment is benevolent, or because it is seeking to maximise profits through offering services to consumers which meet their needs. In such approaches consumers can generally be seen as benefitting from the assessment process though they may not be confident that the services or products being recommended by the environment are those which maximise their individual utility.

In open environments consumers (or their agents) may have a large degree of autonomy in terms of how they may behave. In such environments consumers are able to consult multiple sources of information including third-party brokering services and their peers in order to determine how well particular services might meet their needs. Approaches in these environments can be assumed to have the consumer as the only stakeholder in assessment. The information provided by individual assessment approaches varies widely in terms of the assumptions which are made regarding what the consumer requires of an assessment process and also the capabilities of consumers in evaluating the information provided by the process. Many approaches assume that consumers are limited in terms of their ability to reason and that the assessment process assumes a further role of recommending a particular service for the consumer to use. Such a view of the consumer is not consistent with the visions of grid computing, agent-based computing and autonomic computing in which consumers are seen as decision makers capable of complex reasoning.

2.3.2 Nature of Collected Information (Quality Datum)

The abstract model described in Section 2.2 stated that quality assessment is carried out using two types of information: *stakeholder requirements* and *quality datum* - the latter being collected from various sources then aggregated and stored prior to use in assessment. In this section the type of information that might be used in assessment is considered from the perspective of identifying relevant issues regarding its collection, storage and dissemination (as opposed to the comparative *value* of particular types of information to the consumer).

2.3.2.1 Advertisments and Certificates

Service advertisments give an assessment approach an un-accredited indication of the level of performance a service is capable of providing. However, it is reasonable to assume that a provider may knowingly overstate their capability in an advertisment especially if the advertisments are not legally binding as is the case in most Western countries. Due to this inherent unreliability third-party certification services may be employed by concerned providers prior to advertising the service in order to verify the advertisment's claims [115, 131]. Having been certified by such a service the provider may attach a reference to the certificate in the advertisment [131]. Alternatively, obtaining a certificate may be a prior requirement for registering services with particular registries [115]. In a completely open environment however, the trustworthiness of certifier and service registry may also not be assured so that the use of certification may do little to increase a consumer's confidence in a service compared to one which uses an advertisment alone. Various approaches to establishing the level of trust that a consumer may place in an agent in a distributed system [19, 162, 3] have been proposed which may mitigate this problem. Other approaches describe methods by which the trust in a third party might be assured [116]. Service consumers in an SOA are likely to have wide ranging requirements of services that will be evident in terms of the criteria used in assessing the suitability of a service. Thus, it is unlikely that by using advertisments alone a provider will be able to cover the needs of such a wide range of consumers. A service whose performance may be specified in terms of n attributes where each attribute could take one of m values would require $n * \times m$ advertisments in order to be specified fully. In addressing this issue Lamparter et al. [76] propose the use of multi-attribute utility functions to specify the complete range of qualities offered by a provider. Such functions may be used by consumers to accurately determine the cost of invoking a service provider with a particular combination of attributes and values, without needing to negotiate or otherwise contact the provider.

2.3.2.2 Consumer Ratings

Agents of an SOA have the capability to share information with others in the environment. Leveraging this assessment approaches may obtain feedback from consumers after they have used a service as an indication of their degree of satisfaction in using that service. The feedback mechanism employed typically involves providing a scale on which the consumer indicates their chosen value for a service [117]. After collecting feedback from multiple consumers of the same service assessment approaches may aggregate this information in various ways [66] to calculate an overall value intended to indicate the likely quality of the service for future consumers. The way in which rating information is aggregated differs between specific approaches but most approaches can be considered as falling into one of two classes dependent upon the semantics of consumer ratings. Firstly, aggregation or weighted aggregation [118, 128, 39] may be used to produce a prediction of the rating that a new consumer might give the service if it were selected. Secondly, approaches which employ Bayesian analysis or beta distribution models [61, 100, 152, 165] to produce a confidence score indicating a level of trustworthiness in the service rather than a predicted rating. Research within the field of multi-agent systems [166] refer to approaches to assessment which are based on the use of consumer feedback as *reputation systems* [66, 143]. Two issues related to obtaining reliable feedback from consumers are well recognised. Consumers as a general rule do not like to provide feedback [119, 23] and if they do so there is no guarantee that it represents a reliable indicator of the actual satisfaction they perceived [175, 21, 173, 179] in using the service. Various approaches to eliciting feedback from consumers have been proposed [36, 178], including those which provide financial incentives for feedback [68, 99]. Rubiera et al. propose the use of fuzzy logic in addressing this issue [123] effectively distributing the weight of a user's rating across the whole range of possible ratings. Chen and Singh propose and evaluate a similar approach [21] whereby a consumer's rating for a particular point on the scale *endorses* other points on the scale though to a lesser extent. Both approaches have the effect of lessening the skew of data towards the extremes of the scale. In an alternative solution Hu et al. propose and verify the *Brag-and-Moan Model* in which consumers are assumed to leave feedback only if they are very satisfied (brag) or very disgruntled (moan) with the service [57].

The issues regarding the provision of ratings by consumers described above assume that the consumer is attempting to provide an honest appraisal of the service they have received. However, in many instances rational consumers will deliberately attempt to mislead the rating system by providing information which is intentionally inaccurate [130, ?, 175]. To address the problem of dishonest feedback, Zhang et al. propose an approach in which a centrally controlled assessment service with a database of reputations distributed across multiple agents, periodically selects consumers at random and asks them to confirm the scores they have provided in order to verify the honesty of particular components of the system [179]. In a different approach Yu and Singh [172] use a *weighted majority* technique in which rating outliers are identified and ignored to reduce the effect of deceptive activity. Other approaches propose solutions whereby consumer ratings are weighted based on the level of trust that the assessment system has in the consumer [175]. In the approach proposed by Chen and Singh [21], the level of trust in an individual consumer is determined by grouping consumers based on their rating pattern and then carrying out statistical analysis to determine the similarity of a consumer's ratings to the norm. In doing so, the approach does not require the consumer to have used a large number of services before their ratings become relevant.

The cold start problem [127] – assessing a service for which there are no, or few ratings available – is also an issue for any ratings based approach to assessment. Zhang et al. [179] propose a solution in which agents in the environment join *social groups*. When a provider offers a new service the members of its social group are permitted to provide it with a set of initially high ratings, effectively providing an endorsement of the service. Similarly, Billhardt et al. propose a solution [9] whereby a rating for the new service is established by considering the ratings by consumers of other services offered by the provider and of ratings provided to identical services offered by other providers. In contrast to the cold-start problem issues also arise with this approach when a service has been available for a long period of time and has collected many ratings [90]. In such cases it may be possible for a provider to abuse their position by suddenly lowering the level of service provided [179]. In such cases it will take a considerable time for the new negative ratings received to affect the overall reputation of the service. Similarly, it may be difficult for the reputation of a poor service to be improved - a potentially significant problem since in such cases it may be more rational for the provider to withdraw their service and return under a new identity [175, 33]. The Histos/Sporas reputation mechanism [175], as others [165, 38, 94, 179], uses a form of *damping* by assigning a lesser weight to older ratings in order to alleviate these effects.

2.3.2.3 Objective Measurements

The collection and aggregation of ratings as a means of deriving reputation is intended to be a direct measure of the quality offered by a service. Such approaches maintain that the point of most significance in measuring quality is what level of quality the consumer *perceives* they have experienced. Conversely, due to the limitations described above regarding the unreliability and subjective nature of consumer provided ratings a second set of approaches exist which take a contrasting point of view. Namely, that since quality is the degree of conformance of a particular service to a consumer's ideals the level of quality which a consumer might experience can be ascertained by comparing the objectively measured level of service delivered by a provider in the past to a consumer's requirements. Some approaches go even further by implying that in many instances the requirements of consumers can be assumed such that quality is equated to the actual level of service delivered. The objective performance information used in these approaches is assumed to be collectable through some means. This usually takes place through the direct monitoring of services using agents or tools which either reside locally to the server providing the service [6, 24] locally to the client consuming the service [182, 180] or at some point in between [81, 73, 79]. The information tends to refer to a particular attribute (QoS attribute) of the service being delivered such that for each service there will be a number of measurements obtained. As with ratings within reputation systems, once objective performance information has been collected it is stored and then aggregated by some means in order to provide an assessment. In most approaches the result of an assessment based on objective performance information will be a prediction of the most likely level of service that will be delivered by the service to the consumer in the future. Aggregation is carried out by various means, ranging from simple averages [80, 180] and weighted averages [45, 57, 24, 182] of performance information to more complex techniques including the application of neural networks [45], and Bayesian learning techniques [167].

Early approaches to quality assessment using objective information were domain dependent (e.g. for specific multimedia applications) in terms of both the criteria used for assessment and the means of measurement and evaluation. In recognition of the need for flexibility and extensibility most current work takes one of two approaches. Firstly, by developing a comprehensive ontology to capture what are deemed to be all relevant concepts that could possibly be needed to represent service quality with the ontology being populated with quality attributes specific to a particular application when needed [96]. Secondly, by developing approaches based on a generic set of quality attributes such that any future requirement may be integrated into the approach [80, 169, 24, 56, 16]. The framework proposed by Maximilien and Singh [90, 95], combines both approaches by developing a multi-level ontology that captures a rich set of QoS attributes whilst providing for additional concepts to be added if necessary. Two significant issues regarding the use of objective performance information as a means of evaluating the performance of providers relate to the differences in service received by different consumers across the environment. Firstly, the underlying network of the environment in which services are delivered to consumers may be very large and since the physical resources of the end-to-end connection between the service provider and consumer are not under the direct control of the provider the state of the network in terms of its quality is difficult to guarantee. For this reason consumers at different points in the network may receive different levels of service. This means making an assessment of a service based on objective performance collected from services delivered to different points in the network may be deemed inaccurate or unfair. In addressing this problem a number of approaches proposed a federated architecture of assessment brokers [182, 79, 180] who maintain objective performance information for services delivered to consumers in their "domain". Domains may be fixed by administrative or other boundaries [182, 180], or may be dynamic, defined by grouping disparate groups of consumers who have similar contexts [79]. A second issue regarding the level of service received by different consumers relates to the behaviour of service providers. In order to make a profit providers must supply services to their consumers which meet their requirements (such that penalty charges defined in SLAs are not imposed) whilst minimising their resource usage. Since the quality of service experienced by consumers (and hence satisfaction) is maximised when the level of service is closest to their ideal it is logical for the provider to aim to meet this requirement without expending more resources than are necessary. Recent approaches to provide SLA management maintain that providers will manage a set of service levels or service classes for each type of service they offer [182, 179]. Offering services through multiple levels has implications for the quality assessment of services since it means that providers are effectively treating consumers differently based on the requirements recorded in their SLAs. This means that aggregating all objective performance information about a service will produce a misleading result since it will capture the delivered level of service across all service classes offered by that service.

2.3.2.4 Summary

The most basic form of quality assessment involves the use of advertisements or certificates to demonstrate the capabilities of a particular provider. Such approaches could be considered as more appropriately belonging to the domain of service discovery rather than quality assessment since they assume complete trust in the statements of a provider or service registry and do not depend on any historic information about the service. The collection of ratings from users is a justifiable solution to quality assessment in environments in which consumers are able to freely exchange information with one another. The use of ratings information in assessment is fraught with problems caused by the free-will of consumers. These are embodied in the issues of subjectivity, inaccuracy, collusion, shilling and dishonesty. However, in cases where a consumer is unable to sufficiently trust independent third parties or in situations where such third parties do not exist an approach based on reputation systems appears to be the most effective means of assessing the potential quality of services. In environments where it is assumed that large centralised or federated information brokers are practical and whose trust may be assured or established the use of objective performance information is an effective quality assessment approach. Based on the definition of quality proposed in Section 2.1, approaches which utilise objective performance information in order to make predictions about the future performance of a provider are not strictly making an observation about the future quality of a service since this is also dependent upon the ideals of the consumer. However, it is assumed that in such approaches the prediction returned may be used by the consumer in inferring the quality that they might receive based on the stated performance. Recent research has suggested that it is reasonable to assume that providers will provide a level of service to consumers based on their prior expectation (or SLA defined requirement) of the service but existing approaches to assessment have not fully considered the implications of such behaviour.

2.3.3 Distribution Topology (Quality Assessment Subsystem)

In this section the way in which *information*, *process* and *control* of quality assessment might be distributed across the entities of an SOA is described along with the implications of such a distribution and how it might affect the requirements or constraints on the assessment process. The topology of all assessment approaches can be considered as belonging to one of the following three types:

- **Fully Centralised Approaches:** that inherit their principles largely from the fundamental ideas of distributed computing and are designed to be built in much the same way as centralised registries of services [102].
- **Fully Distributed Approaches:** whose background is from the fields of agent-based [65] and peer-to-peer [138] computing, assume that there is no central point of control or reference in the architecture and that peers must act autonomously and co-operatively in order to succeed [166].

◦ **Federated Approaches**: A spectrum of approaches which exist between the first two approaches (which are the extremes) adopt a *federated approach* whereby information and processing are decentralised whilst control remains centralised.

2.3.3.1 Centralised Broker

As discussed in Section 2.3.2, early work in quality assessment of services focussed on the inclusion of QoS information in advertisments within existing service registries such as UDDI [102]. Similarly, the set of approaches to quality assessment which utilise monitored performance information about services grew out of the work into resource brokering [141] (which usually occur in controlled environments). As such, it is unsurprising that a majority of approaches to quality assessment make use of a centralised broker [4, 30, 115]. In these approaches the centralised broker is responsible for storing the information to be used in assessments carrying out the assessment process and maintaining control, over both who provides and has access to the information stored. Wang [162] maintains that it is impractical for all information in an environment to be stored and accessed on one node especially when operating on scales such as those envisaged by autonomic [74], agent-based [65] and grid computing [42, 41]. Wang uses this argument to justify the adoption of a peer-based assessment approach (as discussed later in this section). However, the problem of ensuring the scalability of massive multiuser databases is not new having received extensive attention within the database community [37]. As such, a *federated* approach is likely to somewhat mitigate the issues highlighted above.

2.3.3.2 Fully Distributed Approaches

In each of the centralised approaches to assessment entities in the environment are distinguished as *providers* (entities providing services), *consumers* (entities consuming services) and *agencies* (entities providing information about providers to consumers). In such environments it is necessary for some control to be maintained centrally since the role fulfilled by the agencies is essentially architectural. In contrast in peer-based approaches no distinction is made between consumers, providers or agencies - each agent being treated as an autonomous entity, which may on some occasions be a source of information or services and on others be a consumer [170, 179]. Such a concept of equality is fundamental to the fields of agent-based computing [65, 166] and peer to peer computing [140], from which the majority of the approaches discussed below grew. Peer approaches assume that each peer in the environment maintains information locally regarding the behaviour of services and is able to process this information in order to infer the quality it might receive from a particular service. The beta reputation system [61] models the reputation of each peer in the environment as a probability density function where the resultant expected value can be considered an indication of the likelihood that a particular peer will behave as promised. Such a value may be used by the assessing peer as an indicator of the quality of the peer, the model being populated through recording a series of binary events which correspond to instances where a peer has succeeded or failed in delivering what they promised. In the TRAVOS-C reputation system Teacy [152] proposes and evaluates an improved version of the beta reputation system which takes into account the *degree* to which a peer delivered what it promised, rather than considering its behaviour as a set of binary events. Yolum and Singh [170] propose an approach in which the expertise of an agent's immediate neighbours in providing reputation information are evaluated. The expertise of each peer is updated after a service has been used by comparing the difference between the referral of each peer and the actual level of service received from the provider. Over time each peer builds up knowledge about the most reliable peers in their neighbourhood and can take this into account in future assessments. Golbeck and Hendler [47] propose a similar approach which extends the reputation mechanism to include peers outside the immediate neighbourhood using *inferred reputation*. In this approach, as in others [38, 161, 158, 124], the reputation of peers outside the requesting peer's immediate neighbourhood can be inferred if they are also in the neighbourhood of one of its neighbours - thus allowing the requesting peer to gain knowledge of peers with whom it has had no previous direct interaction. Other notable approaches to establishing reputation between peers include Carter et al. who incorporate a multi-aspect view of reputation in which each peer is evaluated in terms of its interactivity, content, longevity and referrals [19]; and Chen and Singh, who evaluate an approach to reputation management, which takes into account a peer's expertise in multiple contexts. A number of approaches are specifically designed to deal with the problems that may occur in a peer-based reputation network including the problem of shilling [25], changing identities and collusion [175], and the formation of cliques [27], all of which may potentially affect the fairness of the assessment process.

2.3.3.3 Federated Broker

The federated broker approach [129, 168] represents a compromise between the centralised broker approach and the peer-based approaches discussed above. Taking this perspective, information, control and processing are distributed across the nodes of the environment. In doing so to a greater or lesser extent two issues are addressed. Firstly, as discussed above the reliance of information and processing load on a single centralised node of the system is reduced significantly. In an approach proposed by Song and Phoha [142] a centralised *global master agent* is the first point of contact for consumers requesting a quality assessment of a provider. The master agent controls a set of distributed *slave* agents who maintain reputation information regarding a subset of the users of the environment. The approach is flexible in allowing the balance of load between the global and slave agents to be adjusted and can be considered as distributing the information and processing aspects of assessment, whilst still maintaining the control aspect centrally. A similar approach is proposed by Zhang et al. [179] in which reputation information about users is stored in a distributed hash table (DHT) across multiple nodes in the system whilst a centralised server is responsible for maintaining a map of the network and facilitating the addition of new users to the system. In addition to overcoming the issues stated by Wang with respect to centralisation of processing and storage a number of approaches adopt a federated structure to leverage the locality of peers as a means of improving the quality of the assessment process. In the UX architecture [180, 182] each local domain (which is intended to refer to the set of users within a single organisation) maintains an UX assessment server. The responsibility of the local UX server is to maintain knowledge collected locally about services across the entire network and handling assessment requests from users within the domain. Since users within the same local domain will likely have similar end-to-end connections with services across the network locally acquired quality knowledge can be seen as more relevant than that collected from foreign domains. However, if the local UX server does not have sufficient information to assess the service in question it will contact UX servers in other domains in order to respond to the local request. The approach proposed by Liu et al. [80] utilises a similar architecture within the domain of ad hoc networks.

2.3.3.4 Summary

The information storage, processing and control aspects of quality assessment are not necessarily dependent upon each other and may be distributed across entities in an SOA to a lesser or greater extent. Peer based approaches assume a completely decentralised model of assessment in which each peer has the capability to maintain its own information regarding the behaviour of other peers. A peer is able to process the information in various ways in order to determine the likely quality that might be received from a service of a peer in the environment. The majority of research in this area concentrates on the modelling of trust and reputation and on how peers can combine information from other peers which they may not have direct contact with. The peer based approach to assessment utilises solely consumer feedback ratings as a means of establishing the quality of a service. Hence the drawbacks of consumer provided ratings, as discussed in Section 2.3.2, apply equally to these approaches. At the other end of the scale the architecture of a centralised approach usually consists of a broker service which maintains all information, processing it when requested to do so by a consumer. Amongst the reasons stated for using a peer-based approach utilising ratings over a centralised approach using monitored information is that the latter architecture puts too much of a load on particular nodes in the environment. However, as discussed above adopting a federated approach to information distribution or utilising various approaches from the database community reduces the effect of these issues. In describing approaches to assessment it is useful for authors to adopt a centralised architecture since it enables them to concentrate on describing the novelty of techniques for processing information rather than focussing on the information storage and co-ordination aspects that would be required of a federated architecture. It is likely that the majority of the proposed approaches assuming a centralised broker architecture might in practice be structured as a federation of brokers with no significant change in the effectiveness of the approach.

2.3.4 Stakeholder Requirements

At a high level approaches may be separated into those that require an *explicit* statement of the assessment stakeholder's requirements from those that use *implicit* methods to deduce the stakeholder's requirements.

2.3.4.1 Implicit Requirements

In reputation systems it is generally assumed that ratings are equivalent to the quality perceived by the consumer. The assessment process is therefore based on an implicit assumption that consumers will prefer the services which have the highest aggregate rating, meaning that the specific requirements of the consumer are not captured. In the definition provided in Section 2.1 quality was described as being conformance of a service to the ideals of someone in a position of appraising it. Reputation systems are based on the principal assumption that the reputation of a service may be equated to the quality that a consumer might be expected to receive from that service in the future. Therefore, in keeping with the definition of quality as conformance to ideals the indicators of quality (ratings) which are provided by consumers about services must be at least partially dependent upon their ideals, which may also be interpreted as a reflection of the consumer's expectation of the service prior to using it [150, 30]. For example, two consumers who use the same service and receive exactly the same level of service may potentially (and in all likelihood probably will) provide a different rating for that service [150]. The problem of prior expectation in determining the significance of measures of perceived quality such as ratings is well established in the field of marketing [106, 105, 153]. A number of approaches to quality assessment maintain that if the prior expectations of consumers are not taken into account during the assessment process, that the results will be misleading at best [30, 150, 167, 69]. In addressing these issues Deora et al. propose an approach [30] whereby consumers state their prior expectations of the service at the same time as providing a quality rating. During the assessment process, the expectation of the consumer requesting the assessment is also obtained. The expectation is used to determine a subset of ratings which were made by consumers with *similar* expectations following which a standard aggregation of ratings is carried out. In order to reduce the burden of collecting further information from the user Talwar et al. [150] propose an approach whereby the expectations of the consumer are calculated automatically by considering the trend of information about the provider that the consumer has already been exposed to. Similarly, Kalepu et al. [69] propose an approach in which a consumer's expectation is deemed to be equivalent to the level of service specified in the service level agreement between consumer and provider.

User-based collaborative filtering [117] shares many similarities with expectation-based approaches to service assessment. In this process ratings by consumers about particular products are captured. In addition, a set of information about individual consumers is also known or derived. Recommendations are made based on the assumption that consumers who are similar in terms of their background, context or previous rating policy are likely to provide similar ratings for products. When objective performance information is used in making an assessment it is assumed that the quality of a service may be judged on the conformance of the expected level of performance of the service to a consumer's ideals. For those approaches which return a prediction of the likely level of service a statement of consumer requirements may appear unnecessary. In these cases it may be assumed that quality is determined by the consumer after assessment has taken place by comparing the prediction with their expectation or ideals. However, since providers may base the level of service that they provide to a particular consumer on the requirement stated in an SLA (see Section 2.3.2) it is necessary to capture requirements in order to determine to which provider-defined class the consumer belongs.

2.3.4.2 Explicit Values

In approaches where few assumptions regarding the stakeholder's QoS requirements are made a method of explicit representation is used. The most common approach taken in expressing explicit requirements is to represent them as a list (or table) of *distinct attributes* (see Table 2.3.4.2) along with acceptable values corresponding to each attribute; the acceptable value giving an indication of the optimal value(s) that a stakeholder might receive. Approaches differ in terms

Attribute	**Point**	**Range**
Availability	95%	95%-100%
Framerate	24fps	18fps-28fps
Latency	3ms	1ms-3ms

Table 2.2: Requirements as a set of Distinct Attributes

of the richness of the method used to express these values - in general, as the semantics of expressing the stakeholder's acceptable values becomes richer the assumptions regarding stakeholder requirements that might be made at assessment decrease.

The most basic approach used in representing acceptable values of distinct attributes - used in a number of approaches [141, 155, 115, 90, 160] - is to state a *point* in the range of possible values for the attribute. Although as the most basic form of representation point-based approaches are commonly used within quality assessment models, the semantics of using point descriptions for expressing acceptable values are often unclear since they can be interpreted in two ways. Firstly, the point may be an explicit representation of the optimal value for that attribute. In this case, assessment algorithms treat providers that exhibit QoS levels both lower *and higher* than the point as inferior to a provider that exhibits QoS at exactly that point [30]. Secondly, the point may actually represent a *threshold value* - a minimum or maximum that the stakeholder is willing to accept for the attribute. In this case, the value acceptable to the user is actually an unbounded range (or a range bounded by the upper or lower limit of the attribute range). In addition to point-based approaches some approaches allow the expression of *bounded ranges* of acceptable values [180, 44]. The semantics of such an expression define that stakeholders are only willing to accept values that fall into the specified range. Furthermore, some approaches allow the *weighting* of values within the acceptable range through the use of utility curves [88, 93]. In these approaches utility, which can be thought of as the *value* that would be received by the consumer from a provider offering service at that level and normally stated on a scale from 0 to 1, is expressed for specific points in the range (other values being interpolated to form a complete curve). A hybrid approach proposed in the WSAF framework [92] uses a combination of range and utility representations allowing the user to specify a range of acceptable values along with an optimal value within that range.

In contrast to the approaches that consider attributes distinctly some approaches maintain that relationships between a consumer's individual requirements in terms of quality attributes exist and that these are not preserved when limiting the expression of an assessment stakeholder's requirements to a tabular format regardless of the richness of method used to describe the acceptable values. The stakeholder may have a preference in satisfying their requirements for certain attributes or sets of attributes over others or may be willing to *tradeoff* particular sets with others. The methods of expressing this *stakeholder preference* vary between approaches with the most basic using the concept of attribute weighting to take into account the preference between single attributes [80, 90, 177]. Models such as the G-QoSM framework [4] allow the construction of arbitrary boolean expressions such as $((Q1 > 30 \bigwedge Q2 < 6) \bigvee (Q1 > 20 \bigwedge Q3 > 40))$ to represent the stakeholder's preference over the attribute set. The QoSPath [89] architecture addresses the issue of logical dependencies between a stakeholder's QoS requirements by using multi-dimensional QoS utility curves each dimension representing a single attribute (the nature of utility in this model, however, is to determine an effective QoS degradation path of a service should a service failure occur though the concept is relevant to quality assessment). In an extension to the WSAF framework [93] the issue of statistical and ontological *attribute dependence* is addressed. This considers the relationships that exist between attributes independent of the stakeholder's belief of what QoS it is possible to receive. For example, framerate and bandwidth are positively correlated, hence, it may not be possible to satisfy a stakeholder's requirement for very high framerate whilst also satisfying their requirement for low bandwidth.

2.3.4.3 Summary

Although assumed to be reasonable in many approaches the use of implicit stakeholder requirements is in all cases unjustifiable for several significant reasons.

(a) **Optimal Service Levels:** It cannot generally be assumed that there is consumer preference for the maximum or minimum possible level of service for all attributes. Approaches which assume that this is the case may recommend services which are unsuitable for the purpose for which they are intended. However, there are some attributes (such as reliability, availability and latency) where such an assumption may be reasonable.

(b) **Ratings are based on Expectation:** Since the ratings provided by consumers are partially based on the prior expectations of consumers, an aggregation of all ratings of a service will be misleading since it will incorporate ratings from consumers with wide ranging expectations. In order to address this issue, a reputation system must also capture consumer expectations in order to select the ratings of consumers who are similar for use in the assessment process.

(c) **Provider Behaviour:** As discussed in Section 2.3.3 providers are likely to provide different classes of service to consumers based on their stated requirements. As such, approaches which take an implicit view of consumer requirements will be unable to determine the class to which the consumer belongs and will therefore produce a misleading assessment of the service.

Issues similar to those discussed above apply to the use of explicit point based representations of consumer requirements. In many instances a point is used to refer to a threshold value which implies that the consumer will be equally satisfied with all levels of service beyond the threshold. This assumption has the same implications for the consumer as the optimisation assumption discussed above. Alternatively, a point may be used to refer to a target value of service which is optimal for the consumer. This representation is consistent with the definition of quality provided in Definition 1 since conformance (or compliance) may be judged based on distance from the target. The use of ranges as a means of representing consumer requirements is also consistent with this definition of quality. Utility curves provide a comprehensive means of representing consumer preference, in that they map every possible value of delivered level of service that might be received by the consumer to a *utility* value which can be taken as a direct indicator of the quality of the service in each case. However, for the same reasons as discussed in Section 2.3.2 regarding the difficulty of gathering ratings from consumers, establishing utility curves which accurately represent the preferences of consumers may be difficult and will prove an arduous task for a consumer.

2.4 Summary of Research Challenges

A summary of the approaches to quality assessment described above is provided in Table 2.4. Based on the analysis of existing approaches provided in this chapter, a number of themes emerge as being of particular importance in achieving this goal. These are described in this section.

Stakeholder Role	Provider	[15] [115] [131] [102] [66] [86] [14] [87] [109] [52]
	Consumer	[4] [132] [103] [171] [53] [144] [180] [181] [92] [95] [175] [161] [179] [176] [47] [9] [61] [142] [123] [133] [161] [45] [79] [67] [5] [18] [137] [16] [84] [125] [148] [182] [174] [56] [51] [76] [55] [91] [96] [94]
	Environment	[75] [17] [26] [7] [98] [24] [12] [125] [78]
Quality Information	Advertisments	[115] [131] [19] [162] [3] [116] [76]
	Ratings	[117] [66] [118] [128] [39] [61] [100] [152] [165] [166] [143] [119] [23] [175] [21] [173] [179] [36] [178] [123] [21] [57] [130] [175]
	Objective	[6] [24] [182] [180] [81] [73] [79] [80] [45] [57] [167] [169] [56] [16] [90] [95] [179]
Distribution	Centralised	[102] [141] [4] [30] [115] [162]
	Distributed	[179] [166] [140] [61] [152] [170] [47] [38] [161] [158] [124] [19] [25] [175] [27]
	Federated	[129] [168] [142] [179] [180] [182] [80]
Requirements	Implicit	[150] [30] [167] [69] [117]
	Explicit	[141] [155] [115] [90] [160] [180] [44] [88] [92] [80] [90] [177] [89]

2.4.1 Quality as Conformance

The use of conformance as a means of judging the quality of services is applied inconsistently across the field of QA. A number of approaches use compliance of delivery to consumer requirements as a means of judging quality. Similarly, some approaches which use ratings to directly infer the quality of a service incorporate the use of expectation into the assessment process. However, there is currently no comprehensive and holistic view of how consumers, providers and assessment services might be expected to behave if a conformance-centric view of quality is taken.

2.4.2 Ratings vs. Objective Measurements

There is significant correlation between approaches in that those which are peer-based are also likely to use ratings as a means of inferring quality. Similarly, approaches which utilise objective performance information are likely to take a centralised or federated broker approach. Such correlation is partially due to the respective backgrounds of the approaches in that rating-based approaches have arisen from domains that assume open environments. It is also largely due to a difference in assumptions - peer-based approaches assume that centralised stores of information and direct monitoring of services are impractical. Both types of approach are relevant and justifiable in particular domains. However, the use of ratings is affected by the issues of subjectivity, collusion, identity and maliciousness, as discussed above. Thus, in environments where it is possible to do so it is desirable to take an approach which is able to utilise objective performance information about services to infer their quality.

2.4.3 Supporting Consumer Decision Making

The goals of grid computing, agent-based computing and autonomic computing require that consumers are capable of making decisions based on complex and heterogeneous information. Approaches to assessment which return a prediction to the consumer in terms of the level of service or quality of service they might receive from a particular service help towards the achievement of this goal. However, in these approaches there is no explicit consideration of the exact requirements of consumers in terms of the information that might be useful to them in supporting decision making. Furthermore, none of the approaches are evaluated in terms of how well they support the goals of the consumer in obtaining high quality services.

2.5 Summary

In this chapter a survey of the research related to the assessment of the quality of services was presented based on a formalised definition of the concept of quality and an abstract quality assessment model. At a high level the approaches were classified in terms of the assessment stakeholder, the nature of the information used in assessment, the distribution of information, process and control; and the means of representing stakeholder requirements. The chapter concluded with a summary of the relevant research challenges in supporting consumers in finding services which offer the highest quality. The research challenges identified are used for the development of the conceptual model of a conformance-centric SOA in the next chapter and of the specific approaches to quality assessment proposed in Chapter 4.

Chapter 3

A Conformance-Centric SOA

In Chapter 2 the role of QA was considered at a high level in order to provide the scope for inclusion of approaches to assessment in the literature review presented. It was established that a number of existing approaches to assessment and models of consumer preference could be loosely described as taking a view of quality that was conformance-centric. However, such approaches are limited by the lack of a general model of an SOA in which a conformance-centric view of quality is adopted by all participants.

The goal of this chapter is therefore two-fold: firstly, to develop a model of a conformance-centric SOA; and secondly, to use this model to define a set of principles for the establishment of the role of QA that will be used to guide the development of specific approaches to assessment described in Chapter 4. In this chapter a formal definition for QoS is derived based on the previous discussions of quality and the overview of service oriented computing provided in Chapter 2. This is followed by the definition of a conformance-centric SOA that supports the activities of service consumers, in terms of allowing the monitoring, discovery and assessment of providers. A discussion of the means by which consumers and providers interact within this architecture is provided. Next, the ways in which service consumers and providers might utilise such a conformance-centric view of QoS to achieve their goals are identified, by considering the likely behaviour of both entities in an SOA environment. The QoS Assessment Problem is formally defined with a set of principles for development of specific implementations. The chapter concludes with the description of a conceptual model of quality assessment which will be used to describe a class of quality assessment approaches defined in Chapter 4.

3.1 Quality of Service

A general definition of quality as "conformance" was introduced in Chapter 2 where it was stated that this concept would require further refinement so that it could be represented and used computationally. Two important facets of quality were identified that must be considered whenever it is used, namely that:

(a) Quality is context-dependent with *specific* definitions of quality existing within each context; and

(b) Quality is subjective in that quality judgements made of a single subject by individuals may differ even when those individuals are within the same context.

As stated in Chapter 2 the context of the work presented in this thesis is a SOA in which consumers (or their agents) discover and consume services to achieve their goals. The goal is to help consumers discover services that will provide them with the best quality of service (QoS) for a task. Since the approaches within this thesis are concerned with discovering services to meet the needs of consumers within a service oriented architecture, it is clear that the subject of the statement "quality is conformance" must be the specific services offered by providers and that the measure of conformance will be judged as the degree to which the level of service provided to a consumer met the consumer's ideals. Such an understanding of QoS can be effectively formalised in the following definition:

Definition 2
Quality of Service (QoS) is the degree of conformance of the level of service delivered by a service to the ideals of a consumer.

For example, for a streaming movie service the consumer may have an ideal of 36 frames per second while the service currently delivers 20 frames per second. The degree of conformance is taken to be the relationship between these two values.

Based on the above definition conformance might be interpreted in two distinct ways. Firstly, conformance might be considered as representing the amount of time that the delivered level of service matches the consumer's ideal as a proportion of the total time that the consumer uses that service. However, such a measure of conformance considers all instances where a service fell short of meeting a consumer's ideals as equal regardless of the degree of shortfall. In an extreme case a service that consistently provides a level of service that is very close to the ideals of a consumer but never quite reaches the ideal will be defined as delivering a worse quality than a service which consistently delivers a level of service well below the consumers ideal but manages to meet the ideal for a very small portion of time. This interpretation is flawed since it is reasonable to assume that nearly meeting a consumer's ideal offers nearly as much utility to the consumer as exactly meeting that ideal.

In contrast the relationship between the delivered level of service and the ideal at a particular instant in time could be considered as the definition of conformance. Such a definition is acceptable since the measure implies that at that instant in time a delivered value that is closer to a consumer's ideal constitutes a higher quality. For this reason in this thesis the latter view of conformance is adopted, namely:

Definition 3
Conformance represents the relationship between the delivered level of service to an individual consumer's ideal at a particular instance in time.

Considering the movie service example again, conformance might be judged by comparing the framerate delivered over the previous 3 seconds rather than over the entire duration of service provision. It is important to note that this definition implies that a particular service does not have an inherent property of "Quality of Service". Statements such as "Service X is of good quality" which are reasonable and commonly used, therefore require some explanation. Quality of service is an attributive property of a service which emerges when it is used by a particular consumer - the level of quality being dependent upon the degree to which the level of service delivered by the provider conforms to the ideals of a particular individual consumer. Statements which appear to attach a predicated property of *quality* to a service - such as the one above - are considered as referring to a subjective opinion regarding the consistency of a provider to offer and provide high quality services to individual consumers.

Definitions 1 and 2 together provide a model of quality within a SOA environment which considers conformance as of central importance and which specifies the meaning of conformance in this context. However, such a model is incomplete without information about how the individual differences between consumers can be fully considered. In addressing this issue fully it is useful to consider a hypothetical (yet familiar) example and a real world service.

Consider two holidaymakers, *"Lenient Lena"* and *"Fussy Fred"* who have independently taken advantage of the same all-inclusive package holiday from a travel agent. They stay in the same hotels over the same period of time, take the same trips, eat the same meals - during their trip they receive exactly the same level of service from their hosts. When they return home Lenient Lena tells her friends that even though there were a few glitches she had a wonderful time. Fussy Fred, however, complains that nothing was to his liking and swears that he will never use the same travel agent again. This scenario poses the question "Why, if Fred and Lena received exactly the same service are their opinions of the holiday so different?"

The question may potentially be answered by considering the definition of quality of service proposed above. Although both Lena and Fred receive the same level of service from the provider they have different ideas of what constitutes their *ideal* holiday. In terms of conformance Fred might have had higher expectations than Lena and was therefore less happy than she was with the same level of service. However, assuming for the sake of argument that this is not the case and that Fred and Lena both had the same ideal holiday in mind when they took the trip then the differences of opinion cannot be explained by a difference in ideals. Nor, as has already been established can the differences in opinion be explained in terms of the difference of delivered service. The only way to explain such an anomaly is if Lena and Fred have a different appreciation of *conformance* to their ideal. In this case, it can be inferred that Fred treats the same drop in service from his ideal more harshly than Lena and is therefore less satisfied with the provider. This scenario illustrates that the quality perceived by an individual is based on conformance to their ideal. Furthermore, the degree to which a particular delivered level of service is perceived to conform to that ideal differs on an individual basis.

Each individual's method of determining quality based on degree of conformance can be expressed satisfactorily as a conformance function (Equation 3.1).

$$QoS_c^s = f_c(d_c^s, \beta_c^s) \tag{3.1}$$

Where s is a service , c is a specific consumer, d_c^s is the level of service delivered by s to c at a particular instant in time, β_c^s is c's ideal level of service from s, f_c is a conformance function, performed by C; and QoS_c^s is the quality attributed by c to s when d_c^s is received. This function is a restatement of the definition of conformance provided in Definition 1, but which explains individual differences between consumers by capturing the ideals of the consumer, C.

Defining a representation of QoS and consumer conformance constitutes a complete and usable model of the concept of quality within an SOA that meets the criteria developed in the introduction. Such a model may be used to determine the quality experienced by a particular consumer when using a particular service, at a specific point in time. The above representation is general in that it may be used to capture a consumer's requirements expressed in multiple dimensions (e.g. as a vector of values representing multiple attributes of a service). However, as will be discussed later in this section the approaches to quality assessment developed in this thesis use a single point to represent a consumer's requirement in terms of a single attribute. This is consistent with the above representation and allows the focus of this thesis to remain on the core contributions of applying techniques to identify patterns in provider behaviour. Whilst adopting this point-based model of requirements, the specific techniques implemented in the approaches of this thesis are chosen to be technically extensible to operate with multiple dimensions.

The remainder of this chapter considers how consumers and providers within an SOA behave with respect to the above conformance-based definition of QoS. Before these aspects of an SOA are examined, however, it is important to briefly discuss the attributes of a service which might constitute a consumer's *ideal* of a service.

3.1.1 QoS Attributes

Numerous models have been proposed (see Section 2.3.2) to provide a classification of the set of QoS attributes that might be used by a consumer to judge the quality of a service. However, there is still no general consensus of a definition of a QoS attribute as distinct from other service attributes such as cost and functional specification.

In the context of this thesis an attribute of a service is considered to be a QoS attribute if it fulfils the following three criteria:

(a) **Totally Ordered w.r.t. Delivered Values**: The performance of the attribute should be quantifiable such that the unit of measurement is defined as using a scale that is monotonically increasing. Attributes of a streaming movie service such as *latency* (milliseconds) and *framerate* (frames per second) have scales that are inherently totally ordered; whereas *programme type* taking a value from the set {news, drama, entertainment} cannot generally be considered as totally ordered. However, in the latter case it may be possible to transform the set into one that is totally ordered by applying another function.

(b) **Dynamic over the Service Lifetime**: The attribute must be expected to (or be capable of) change during the duration of service provision to a consumer. Attributes of a cluster-based grid service such as *mean time to recovery* and *response time* are dynamic; whereas *maximum processors available* may be fixed for the duration of service provision.

(c) **Measure of Delivery from Provider to Consumer:** An attribute of service can only be considered a Quality of Service attribute if it relates to a measurement taken of something delivered by provider to consumer.

Thus, aspects of a service such as "query rate" cannot be considered as QoS attributes. .

Property 1 has a useful implication in that since the scale of measurement is monotonically increasing with respect to delivered performance it is generally possible to normalise any measurement taken onto a domain $[0, 1]$, where 0 represents the theoretical minimum delivered value; and 1 the theoretical maximum delivered value. This is a useful property since it means that any approaches to assessment developed may be treated as being general solutions to the assessment of any attribute which meets the above criteria. There are a class of attributes for which no useful theoretical lower or upper limit exists (e.g. duration) or for which the limits change with time (e.g. bandwidth). In such cases an arbitrary limit must be imposed on the measurement to allow for the normalisation to take place - such a limit is based on the currently available minimum or maximum value for the attribute.

Services will clearly have more than one associated QoS attribute and consumers will want to describe their expectations in terms of multiple attributes. However, the goal of this thesis is to consider how a conformance-centric view of QoS affects the behaviour of consumers and providers and the development of approaches to support service discovery in such a context. For this reason only the case where a single QoS attribute is associated with a service is considered here since this will allow the research to be focussed more fully on exploiting the conformance-centric view of QoS within SOA. Such abstraction does not reduce the general applicability of approaches developed. That is, the specific techniques employed – such as those relating to pattern classification (see Section 4.6) – are chosen to be easily adaptable for addressing the equivalent multi-dimensional problems which would arise from considering assessment of a service using multiple QoS attributes.

3.2 The QoS Lifecycle

To be used effectively in supporting a consumer in finding services which best meet their needs, an appreciation of the QoS aspects of services must be considered across the whole SOA. Whenever services or requirements are described functionally they must also be described non-functionally in terms of quality. If this is not the case providers will be unable to effectively provide high quality services to individual consumers. Such integration of QoS with existing SOA is necessary in order to support the next generation of web, cloud and grid services and their consumers. This section describes the entities necessary to support QoS across an SOA and describes the capabilities which they must have in order to do so effectively.

The standard publish-find-bind process on which most SOA are based provides a useful starting point for deciding on the most effective way of integrating QoS into an SOA. It is apparent that at each of the stages involved in this process, QoS information might be included along with the functional aspects of the services that are being advertised, discovered and bound. However, the inclusion of support at only these stages ignores some fundamental issues surrounding QoS. Firstly, the performance of a service in terms of the levels of service that they deliver may change frequently and rapidly. Thus, providers are unlikely to reflect recent, up-to-date information about the performance of services. Secondly, it cannot be assumed that the advertisment is accurate since providers have the incentive and means to over-state their capability. Finally, the advertisments issued are unlikely to sufficiently capture the behaviour of providers towards consumers with particular requirements.

In order to overcome these issues QoS information regarding the provision of services must be monitored, captured and stored. The information may then be utilised by consumers in the future to gain an indication of how well a particular provider might meet their needs. Such a capability is achievable through the inclusion of three architectural components that are responsible for monitoring service provision, storing service provision data and providing information in a useful form for consumer decision-making. It is this feedback that leads to the use of the term "QoS lifecycle", as illustrated in Figure 3.1.

The conformance-centric SOA in which the QoS lifecycle operates can be formally defined as consisting of the following entities:

- **Service Consumer** - An SOA contains a set of Service Consumers (SC) where each service consumer ($sc_i \in SC$) has a set of functional ideals and a set of QoS ideals that they are aiming to satisfy by interacting with providers in the environment. The ideals of consumers may not be provided directly - instead, consumer expectation representing an acceptable level of service in terms of functionality (FE_i) and QoS (QE_i) are used. The nature of a consumer's QoS ideals and how these are used to derive expectation, and evaluate and choose services is discussed in Section 3.3.

- **Service Provider** - An SOA contains a set of Service Providers (SP) where each service provider ($sp_r \in SP$) offers a set of services (S_r) offering varying types of functionality to consumers. Section 3.4.1 considers the behaviour of service providers in terms of the way in which they allocate resources in order to manage their interaction with consumers.

- **Directory Service (DS)** - The directory service contains a set of advertis-

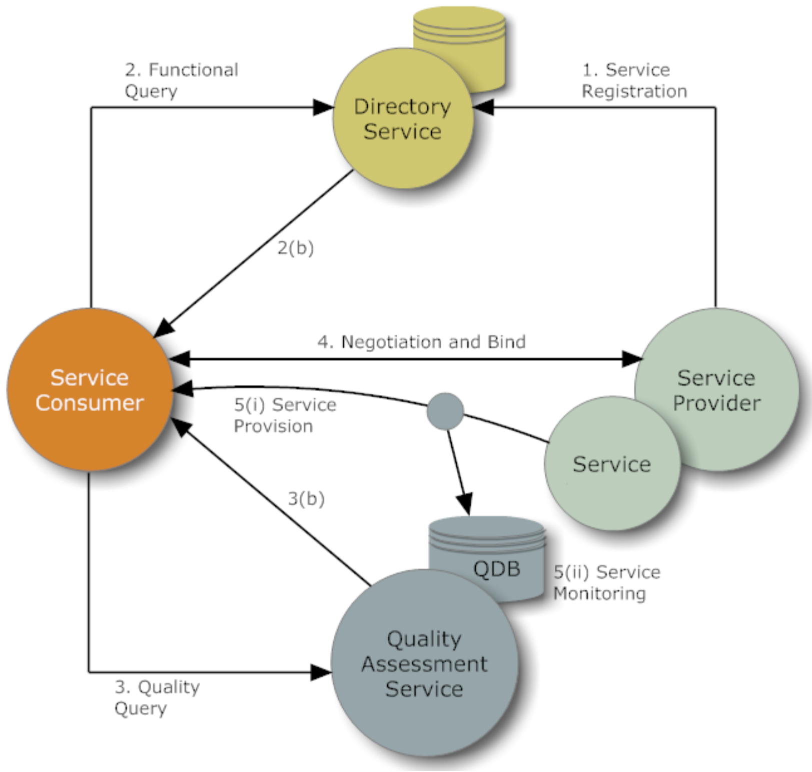


Figure 3.1: The QoS Lifecycle

ments from individual providers which specify the services offered by that provider. Each service advertisment consists of a provider (sp_i), service (s_r) and a pair of capability measures, $\mathrm{AF}_{i,r}$ and $\mathrm{QF}_{i,r}$ which represent the advertised capability of s_r in terms of its functional and QoS aspects.

In the QoS lifecycle (3.1) the DS is included as a centralised component. In reality the DS may be distributed though its underlying function in this case remains unchanged.

- **Quality Assessment Service (QAS)** - The Quality Assessment Service acts as a source of information about the likely future behaviour of services and may be queried by service consumers. The QAS makes use of information stored in a quality database (QDB) that records historic performance information about service providers. The information requirements and exact role of the QAS in supporting decision making are discussed in depth in Section 3.6. The structure and role of the QDB are discussed in Section 3.5.

In the above architecture a deliberate decision is taken in separating the Quality Assessment and Directory services into distinct components. Although similar in interaction (in that they are both used by consumers to find information relating to services) the components fulfil subtly separate roles. Importantly, the information contained about services within the DS will be gathered entirely from provider advertisments and therefore a consumer cannot assume that these advertisments are an honest reflection of actual performance. The level of trust that a consumer places in a Directory service will therefore be based upon the level of *recall*of the results rather than accuracy. That is, whether the DS finds all of the services that claim to match the consumer requirements. In contrast, the QA Service maintains control (and responsibility) for the accuracy of the objective and unbiased information gathered about service performance, which is the basis for subsequent assessment. Whether a consumer will place trust in a QAS is contingent on its ability to provide high precision results. That is, assessment results which accurately reflect the performance of the assessed service. By separating the two stages of service discovery (via the QAS and DS), the consumer is able to establish their reputation independently, based on the attributes of information described above.

At a high level the flow of QoS information through the elements of the SOA constitutes a lifecycle. However, from the perspective of an individual consumer the discovery, choice and consumption of a service, in order to fulfill a need is a linear process consisting of the following steps:

3.2.1 Service Registration

Before services may be discovered and used in a SOA environment providers must develop, implement and deploy the services. When each provider has decided which services to offer, they will advertise their availability by issuing an advertisment for each offered service to the DS.

3.2.2 Service Discovery

In order to discover services that best meet their needs a consumer contacts the directory service specifying their functional and quality expectations. From the set of all registered advertisments the DS returns a subset of the services matching the consumer expectations.

3.2.3 QoS Query

The list of services obtained from the DS consists of a set of self-certificates from the provider of each service claiming to be able to satisfy the consumer's needs. The consumer then consults the QoS Assessment Service to determine the likely outcome in terms of QoS of engaging the services of each of these providers based on the consumer's QoS expectations. The criteria of an assessment approach which is effective in meeting these consumer needs by using the available information in the environment is considered in depth in Section 3.6.3. Once the consumer has assessment results for each of the providers they must select which service to use. The decision making process which the consumer goes through in this respect is discussed in Section 3.3.

3.2.4 Bind

When the consumer has decided upon which provider to approach to provide the service needed they will contact the provider and state their expectations. If the provider accepts their proposal a contract is created - usually in the form of a Service Level Agreement (SLA) - detailing the obligations of both provider and consumer in the contract. Such a contract will be used to determine when the agreement should terminate and to resolve potential conflicts with regard to aspects of service provision.

3.2.5 Service Provision

Once an SLA is in place the consumer may begin to use the service instance. During service provision the provider monitors and adjusts the resources allocated to the service instance in order that the SLA or other objectives are satisfied, based on defined policies, which are discussed in Section 3.4.

Service provision is monitored and data regarding the performance of the service collected. Such information is stored in the QDB, so that it can be used by the QAS to predict the performance of providers. The nature of the monitoring process and the type and format of data collected is discussed in Section 3.5.

The following sections discuss aspects of the QoS lifecycle in relation to the QoS information exchanged between entities and how each entity utilises and processes such information. Firstly, the way in which the ideals, conformance functions and expectations of consumers may be modeled is proposed. Next, the policy of service providers in terms of how they manage their resources in order to deliver high quality services to consumers is considered. This is followed by a description of the type of information which might be made publicly available for the use of third parties. Finally, a set of criteria are defined for an effective quality assessment process based on the information available from monitoring, the expected behaviour of providers and the needs of consumers established in the previous sections. This section concludes with a statement of the QoS assessment problem in terms of the factors discussed.

3.3 Service Consumer Model

The model of QoS derived in the previous section is consumer-centric in that it considers quality as an appraisal of the services delivered for the overall benefit of consumers. In this section, an understanding of how consumer ideals are integrated into the service selection decision process is developed. As will be described the mode of rational decision making employed by the consumer has requirements in terms of the type of information obtained from the Quality Assessment Service.

3.3.1 Consumer Preference

Conformance has been introduced as the means by which consumers judge quality of service, and conformance functions as a means of representing the preference of those consumers for particular delivered levels of service. A conformance function f_c can be considered as a specialisation of a traditional utility function $c : D \to Q$ in which the consumption set D is the set of all possible delivered levels of service and Q is the utility or quality received from a particular $d \in D$. A conformance function is intended to represent the preferences of an *individual* consumer or *group* of consumers rather than representing a general view of all consumers.

3.3.1.1 Types of Attribute

Though conformance functions will be unique for individual consumers there are particular QoS attributes where there may be a general consensus as to what constitutes an ideal value for the attribute in question. Such consensus emerges when it is desirable to minimise or maximise the attribute (see Figure 3.2). Examples of

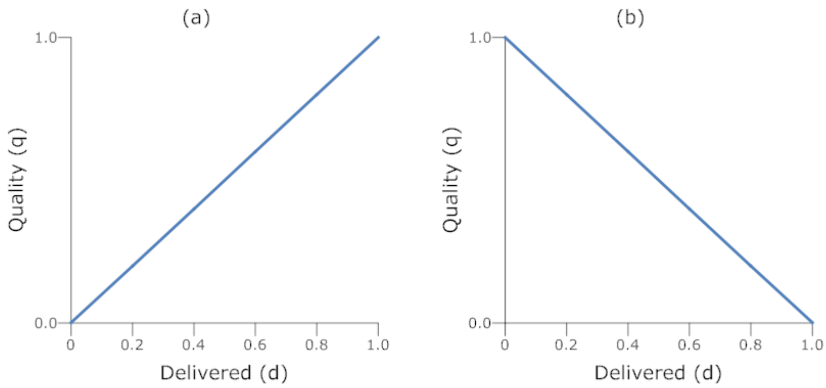


Figure 3.2: Conformance Graphs - Low and High Optimal Attributes

QoS attributes for which maximisation is generally desirable include availability, reliability and Mean Time To Failure (MTTF), whereas for QoS attributes such as latency, jitter, delay, noise and Mean Time to Recovery (MTTR) it is usual for the minimum to be desirable.

3.3.1.2 Marginal Quality and Tolerance

In the conformance functions illustrated in Figure 3.2 the level of quality is defined as being directly correlated to the delivered level of service in each case (exhibited by the straightness of the curves). Such a function implies that for every marginal increase in delivered value the consumer receives the same marginal increase in quality. The difference in opinion illustrated in the holidaymaker example can be captured effectively by changing the shape of the curve to reflect the differing tolerance of the two consumers (see Figure 3.3)

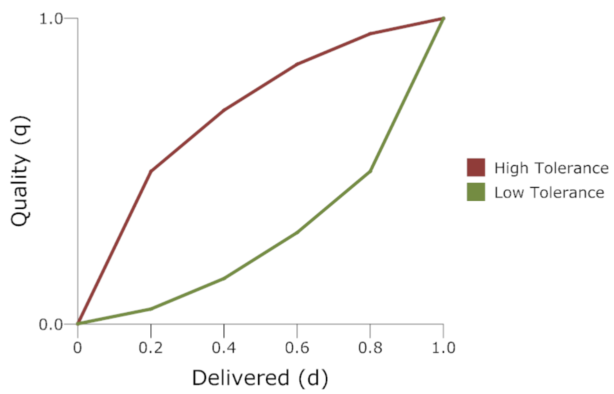


Figure 3.3: Conformance curves for consumers with (a) low, and (b) high tolerance

3.3.1.3 Consumer Context

There are a number of factors related to the context of the consumer which may mean that the consumer ideal differs from the minimum or maximum delivered value:

- **Software Limitations:** Software installed on the consumer system may not be capable of passing on any increase in performance. For example, a consumer may not be able to benefit from the increased sample rate of their internet radio service if they do not have a particular audio compression codec installed.

- **Hardware Limitations:** The physical resources of a consumer's system may present a performance bottleneck. For example, the rate of message receipt of a stock quote feed service may be limited based on the speed of the processor on the consumer's system.

- **Network Limitations:** The network used by the consumer to access the service may mean that a consumer is not able to receive improvements in performance beyond a particular level. For example, most mobile networks are technically limited by the maximum bandwidth available to consumers.

In such situations the consumer is satisfied with a particular delivered level of service leading to a "flattening" effect on the conformance curve at a threshold point (Figure 3.4(a)). Furthermore, in some cases an increase in delivered level from the provider (for an attribute in which maximisation is logically desirable) may actually cause a decrease in quality as perceived by the consumer. For example, consider a consumer using a personal digital assistant (PDA) with a slow processor to watch a streaming news service. An increase in the framerate of a streaming movie service may increase the perceived quality initially as the video becomes more fluid but an increase in framerate that requires a level of buffering and memory use above the capability of the client device may lead to increased jitter in the presentation of the video to the consumer and hence a reduction in quality. In such cases, the consumer becomes satisfied at a certain point. This is followed by a decline in quality for any further enhancement in the attribute (see Figure 3.4(b)).

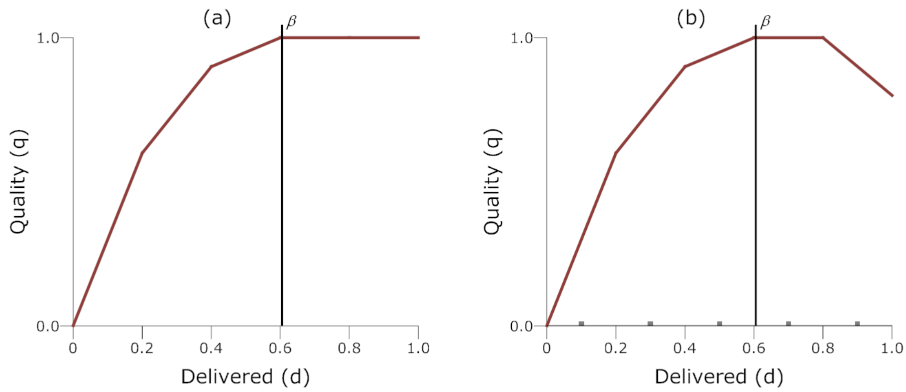


Figure 3.4: Satisfaction with (a), and without (b) decline

3.3.1.4 Effect of Consumer Behaviour on a Conformance-Centric SOA

The above discussion of the conformance functions of individual consumers illustrates the following points:

(a) Although many consumers will share the same view that maximisation of performance of a particular attributes delivers maximum quality this cannot be assumed for all attributes and all consumers.

(b) The differences in marginal utility illustrated through the varying tolerance of consumers means that conformance functions may be different

(c) The context of consumers may also lead to satisfaction in terms of particular attributes. When this will occur and for what attributes is beyond the knowledge of the provider or third parties.

These points indicate a requirement for providers in a conformance-centric SOA to make no implicit assumptions regarding the preference of consumers with different levels of performance unless explicitly stated by the consumer. The same observation may be applied to any QAS in a conformance-centric SOA which should refrain from making recommendations based on expected levels of quality if aspects of consumer preference are not known.

3.3.2 Consumer Expectations

As described in the QoS lifecycle (Section 3.2) at some point before using a service a consumer will inform the provider of the level of service which they expect to receive. In an ideal world consumers would state their ideals and providers would then attempt to provide a service as close as possible to this ideal in order to satisfy their customers. However, in most situations there are factors which mean that this ideal will not be expressed by the consumer.

- **Uncertainty:** A consumer's ideal and their perception of it will almost certainly be different. Consumers may be inexperienced in using a particular service or unfamiliar with a particular attribute and will not have sufficient understanding to be able to express their ideals.
- **Availability:** The consumer's ideal may not be offered by any service provider within the marketplace. The consumer may choose to engage a provider that comes closest to matching the ideal and request the provision of the best possible service from this provider.
- **Affordability:** The consumer may not be able to afford the services of a provider who will meet their ideals in terms of all aspects of the service due to prohibitive cost. In these cases, the consumer will tend to engage the services of a provider who offers a level of service at a lower cost but with a correspondingly lower quality.

The effect of the above factors is that the value stated by consumers to inform a provider of the acceptable level of service may not be the same as their ideal. The level of service actually stated to the provider is referred to as a *consumer expectation.*

3.3.3 Consumer Decision Making

In order to consume services a consumer must first make a decision about which one to use. This decision will be guided by a consumer's appreciation of their ideals and expectations as discussed above. As previously stated the role of a QoS assessment service is to support the decision making process of consumers. In order to determine a set of requirements for such a process it is first necessary to consider the nature of the decision making process. Consumers are faced with substantial uncertainty in making decisions, since the sources of information which they might consult are not omniscient and may be malicious in nature. Thus, in trying to find the provider that will offer the best quality, the consumer will seek to reduce the risk in the process. Expected utility theory, from the field of economics provides a reasonable method for decision making in such cases. Briefly, the theory of expected utility seeks to provide an underlying model for decision making, which assumes that the decision maker has a set of possible actions (A) to choose from, and that there is uncertainty surrounding the real state of the world, which is taken to be in one of a number of mutually exclusive states (S). It is also assumed that the state of the world will affect the utility gained from taking a particular action. In such cases, the expected utility (EU) of a consumer (c) taking a particular action (A_i) can be expressed as:

$$EU_c(A_i) = P(S|A_i) \times Utility(S|A_i) + P(\neg S|A_i) \times Utility(\neg S|A_i) \qquad (3.2)$$

That is, the expected utility of the consumer taking action A_i is a function of both the gain expected if S is true, given that A_i is chosen, the utility of S being false, given that A_i is chosen and the probability that S is true.

The decision of choosing between n candidate services based on information from the QAS can be considered as choosing the service that exhibits the best expected utility. In such a decision the two potential states are that the prediction of the QAS regarding the performance of candidate service i (CS_i) is correct or incorrect. Formally the EU of a consumer c choosing CS_i can be expressed as:

$$\begin{aligned} EU_c(CS_i) = {} & P(QASisCorrect|CS_iischosen) \times U(QASisCorrect|CS_iischosen) \\ & + P(QASisincorrect|CS_iischosen) \times U(QASisincorrect|CS_iischosen) \end{aligned} \qquad (3.3)$$

Solving the above equation for each candidate service CS_i will result in an EU value for each choice. Rationally, the consumer will choose the candidate service that corresponds to the action offering highest EU.

Since the role of the QAS is to provide information to support consumer decision-making the above model of consumer decision making raises the following criteria for a QAS in a conformance-centric SOA:

(a) The QAS should provide a value that represents a *prediction* of service performance which allows the consumer to determine the level of utility that they may receive from the candidate service if they choose it. Since the QAS will be unaware of the consumer's conformance function the prediction must be in terms of a *delivered* level of service rather than by predicting an actual level of QoS. The consumer will combine a prediction in terms of delivered level of service with their own conformance function in order to determine the utility from such a prediction. It should be noted that in the context of this thesis, the term prediction is taken to refer to an "indication" of the future performance of a provider, though formal forecasting methods will not necessarily be used to achieve this.

(b) The QAS should provide an indicator of the level of *confidence* with which it is offering the prediction. This will support the consumer in attributing a probability that the QAS is correct. Furthermore, the inverse probability - the probability that the QAS is incorrect - is easily derived from the same information

Assuming that the QAS is able to provide the above information, the EU equation may be simplified to:

$$EU_c(CS_i) = (Confidence \times Prediction)$$
$$+(1 - Confidence) \times U(QASisincorrect|CS_iischosen) \quad (3.4)$$

Since utility, if the QAS is correct, is the predicted level of service and the probability of QAS being correct is the measure of confidence returned by the QAS. The remaining factor, $U(QASisincorrect|CS_iischosen)$ represents the utility that will be received if CS_i is chosen and the prediction provided by QAS is incorrect. The value attributed to this parameter by a consumer may be considered as reflecting a consumer's optimism about how well they believe the service will overperform or underperform. It can be reasonably assumed that without specific prior knowledge of the performance of any candidate service the consumer will not be biased towards attributing higher levels of utility to particular services. Hence, for each decision the constant U_R is used across all expected utility functions to represent the utility that the consumer believes they will receive if the QAS is incorrect, weighted to reflect their *optimism*. The expected utility equation may be further simplified as:

$$EU_c(CS_i) = (Confidence \times Prediction) + (1 - Confidence) \times U_R \qquad (3.5)$$

It should be noted that the decision making process above considers only the QoS aspects of a service which might be received by the consumer if they choose CS_i. There are of course many other considerations that might be taken into account by the consumer in addition to QoS aspects. For example the cost or reputation of the service. The above expected utility may be combined with other sources of information in making an overall decision though this thesis does not consider the details of such a combination.

3.4 Service Provider Model

Service providers have almost complete control over the delivered level of service that they provide to consumers but have little control over the expectations of those consumers and no control over their ideal. However, by adopting a conformance-centric view of quality providers will be better able to manage their interactions with consumers during both the negotiation and service provision stages (see Section 3.2). This section considers how taking a conformance-centric view of quality which presumes that consumers behave as described in Section 3.3 influences the policy and behaviour of a service provider.

Understanding the behaviour of service providers with respect to QoS is necessary since the QAS will be assessing these providers in terms of their predicted performance. Acknowledging the existence of patterns in provision data which might occur as a result of particular types of policy and discovering these patterns at run-time is a key aspect of the role of a QAS and one which necessitates particular design decisions, that will be considered in Section 3.6.

3.4.1 The Nature of Service Providers

In achieving their goals service providers have a set of resources at their disposal, a key function of service providers being the management of these resources. One can think of a service provider's resource consumption in terms of the QoS attributes described previously. Providers may choose to provide a particular delivered level of service to consumers in terms of these QoS attributes but their choice will affect the amount of resources "spent" on the consumer. Although the parameters which a provider may adapt in order to ensure SLAs are met may be "system based", such as *processor time*, *network bandwidth* and *memory al-*

location, the change to be effected by such adaptions will be visible in terms of a changing level of service for individual consumers in terms of QoS attributes. Since the actual management of resources is likely to be carried out by software agents, service providers are faced with a problem in choosing and implementing policies that can be used by agents to allocate resources such that both global (strategy based) and local (SLA-based) objectives are met.

3.4.2 Service Provider Policy

In deciding upon a particular policy a provider is influenced by a number of factors. The aspects of minimising cost reduction through resource efficiency and meeting the SLA with consumers have already been identified though there are several other factors that must also be considered. Taking a conformance-centric view of quality is a key consideration – the major implication from such a view being that it is desirable to meet consumer expectations as closely as possible in order to minimise the gap between the delivered level of service and the consumer's ideal. This has two consequences for service providers. Firstly, that providers must recognise that exceeding the delivered level of service stated as a requirement in the SLA does not guarantee an increase in QoS as perceived by the consumer. Also, an increase in the delivered level of service requires an increase in the resources allocated to the service in question. If there is no increase in consumer QoS from such an increase in delivered level of service then the resources are wasted and may be better used by being allocated elsewhere.

Business strategy and policy will also influence the way in which providers treat their consumers. Many service providers may seek to offer services to the whole market but may treat consumers differently based on which market segment the provider defines them as falling into. In such a case, it is desirable that the policy of allocating consumers to segments be kept private since public knowledge may mean less custom.

Finally, the volume of custom which the service provider can expect in the marketplace will influence the chosen policy. In an SOA a provider can realistically expect to be offering tens or hundreds of services to many thousands of consumers concurrently across the same shared resources. Managing SLAs on an individual basis in such a situation is hugely complex. Service providers will need to develop general strategies for managing groups of SLAs sharing similar characteristics so that the overhead and resource requirements are reduced.

3.4.2.1 Service Classes

A policy which meets all the above criteria involves the segmentation of consumers such that they are assigned by the provider into one of n service classes based on a consumer's stated expectations in terms of QoS. An individual service class SC_i is defined by a lower ($SC_{i,low}$) and upper ($SC_{i,hi}$) boundary so that any consumer having an expectation e_c in terms of a particular attribute where $SC_{i,low} \leq e_c < SC_{i,hi}$ will be classed by the provider as belonging to service class SC_i.

The choice of segmentation policy in terms of the number of service classes and their boundaries will depend on the policy of the provider though the set of all service classes SC* will cover the complete range of possible expectations (i.e. [0,1]). For example, one such segmentation policy may be defined based on a service being hosted across multiple physical platforms (servers) each with varying capability (in terms of bandwidth, processor speed and memory, etc). In this case, a service class segmentation emerges based on a provider's policy defined to ensure that a consumer is allocated to be provided service from the single platform which is most likely to meet their expectations. Figure 3.6 illustrates one such segmentation of the expectation range in which the range is split into four service classes (corresponding to four physical or virtual servers). By controlling the level

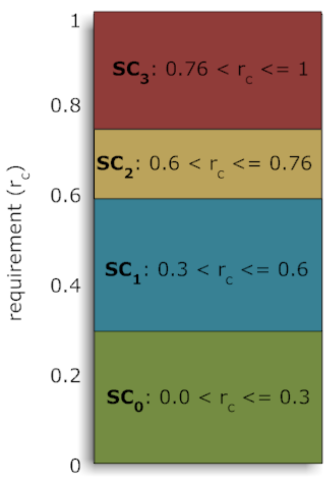


Figure 3.5: Segmentation of the Expectation Range into Service Classes

of resource assigned to services at the system level the overhead in managing individual SLAs is reduced. All consumers within the same service class will receive the same delivered level of service from the provider with respect to a specific QoS attribute.

3.4.2.2 Target Level of Service

At a particular instant in time, a service provider will attempt to meet a target level of service T_i for each QoS attribute relevant to each service class to ensure that SLAs are met. In the example above, this may be achieved by adjusting the level of resource (e.g. processor speed, memory and bandwidth) available to each physical server. It is realistic to expect that the provider will not be able to consistently deliver the target to every consumer since a number of factors may prevent this level of service being received, including network issues outside the control of the service provider, unexpected increases in load on the server etc. The distance between the level of service delivered to a service class SC_i at time t, d_i^t and the target level of service at that instant T_i^t can be considered as an error ϵ_i^t that will vary with time. This means that the delivered level of service d_i in each of the service classes will fluctuate about T_i (Figure 3.6(a)). Such fluctuation will likely cause points sampled from the delivered level of service to approximate a normal distribution since the fluctuation is caused by multiple random changes on the target level of service.

d_i over time will also be affected by the fact that the service provider may change T_i in order to achieve better global or local management of resources (Figure 3.6(b)).

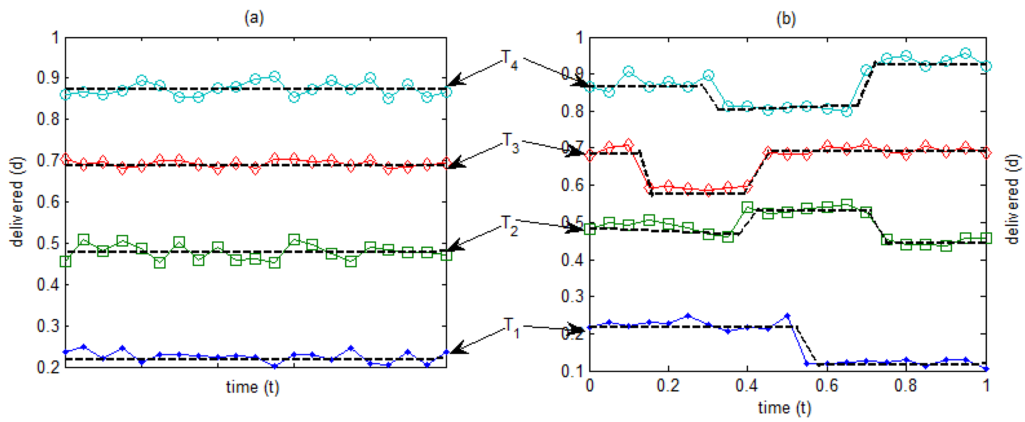


Figure 3.6: Target Levels (T_i) and Error (ϵ_i)

3.4.2.3 Effect of Consumer Behaviour on a Conformance-Centric SOA

The policy and behaviour of service providers in terms of the level of service that they offer to consumers over time has implications for the operation of the QAS when considering the likely performance of a provider during assessment. This issue is complicated due to a number of issues. Firstly, service providers will provide a different level of service to consumers based on their expectation.

Secondly, the target level of service within each service range is not provided consistently. Error occurs at each instant in time and providers may adjust the target over time. Finally, the policy used by providers in defining service classes may change with time resulting in new service class boundaries. These issues are further complicated since the information used in defining service class boundaries, setting and changing target levels and changing service class boundaries is private.

For a QAS, it is necessary to consider that a provider may be adopting a service-class based approach to provision. As such, assessment strategies used in any assessment approach must consider that the data used in each assessment may be different as it depends upon the service class into which the current consumer is placed by the provider. The issue is further complicated by the fact that providers will likely keep details of service classes private and as such their definition may need to be inferred from the data available to the QAS.

3.5 Data Collection

Monitoring will take place during service provision so that the SLA may be checked and enforced. Such monitoring must necessarily be carried out by a mutually trusted third party so both parties have confidence in its findings. Monitoring information may be used both by the consumer to ensure that their expectations are being met; and the provider to ensure that they are delivering their obligations so as to not face penalty charges. Monitoring of all the QoS attributes stated as consumer expectations in the SLA is necessary for this to be achieved.

During monitoring performance information is stored in a Quality Database (QDB) which may be consulted by the QAS when it is assessing a particular provider. Each time a monitoring result is received by the QDB, a new *quality tuple* qt, is added to the database.

$$\mathrm{qt} = \{\mathrm{t},\ \mathrm{s}_i,\ \mathrm{c}_m,\ \mathrm{e}_c,\ \mathrm{d}\}$$

Where t is a time stamp stating when the recording was taken, $s_i \in S$ is the service, $c_m \in C$ is the consumer consuming the service, $e_c \in [0..1]$ is c_m's expectation in terms of the QoS attribute and $d \in [0, 1]$ is the level of service delivered in terms of the QoS attribute.

3.6 The QoS Assessment Service

Having considered a specific definition of quality within a Service Oriented Architecture along with the way in which service consumers and service providers behave with respect to a conformance-based view of QoS and the nature of information available during assessment, it is now possible to formally define the role of the Quality Assessment Service with respect to its available inputs, desirable outputs and a set of criteria which should guide the specific approaches to assessment developed in the next chapters.

3.6.1 QAS Inputs

The QAS accepts as input a consumer expectation (e_c) expressed in terms of a single QoS attribute and with domain [0,1] and a candidate service ($s_i \in S$) for whom the consumer requires an assessment. These pieces of information are provided by the consumer when a QoS query is initiated. In making an assessment, the QAS has access to the performance history of the provider which is obtained from the QDB and contains all information pertaining to the past behaviour of s_i gathered as a result of monitoring.

3.6.2 QAS Outputs

The outputs from the QAS are defined by the requirements of the consumer in supporting decision making, as discussed in Section 3.3. An assessment result consists of two components:

- **Performance Prediction (Pr):** indicating the level of service that the QAS believes would most likely be delivered to the the consumer from s_i if they were to ask them to provide e_c. $Pr \in [0, 1]$. This prediction necessarily refers to a likely *delivered* level of service rather than a likely *quality* of service since quality is defined in terms of the ideals of the consumer which are not known to the QAS.

- **Confidence (Conf):** which represents the level of certainty with which the QAS is making the prediction. $Conf \in (0, 1]$ where 0 indicates no confidence, and 1 indicates complete confidence.

3.6.3 Assessment Approach Criteria

When making an assessment of a provider the design and operation of the QAS must consider the following points relating to the behaviour of providers:

- **Service Classes:** a provider's policy will mean that individual consumers will be treated differently by that provider based on the consumer's QoS expectations (see Section 3.4.2). Whilst the provider will seek to keep details of its service class policy private it is important that the QAS is able to determine whether particular information gathered from monitoring is relevant to the consumer requesting the assessment. That is, the QAS should seek to only use information that is from consumers in the same service class in making an assessment.

- **Changing Intra-Class Service Level:** the level of service delivered by the provider may change with time due to changes in resource allocation priorities by the provider. That is, a service provider may decide to reallocate system level resources between service classes therefore changing the level of service provided to consumers within these classes. Whilst these changes are unpredictable the QAS should consider the temporal relevance of provider historic performance data within the QDB.

- **Changing Service Class Boundaries:** a provider's policy in relation to service classes may change with time. This means that the way in which segmentation of consumers into service classes in terms of their QoS expectations occurs will also change. As above, these changes are unpredictable and the QAS should therefore consider the temporal relevance of service class boundaries that it has established for particular providers.

3.6.4 The QoS Assessment Problem

Considering the requirements of supporting consumers in decision making and the constraints imposed by service provider behaviour, the QoS assessment problem may be formally defined as: using historic performance information and based on a consumer expectation e_c in terms of a single QoS attribute and a specific provider sp_c, determine the most likely level of service which will be delivered by sp_c to the consumer if they were to ask them to provide e_c, indicating the level of confidence in the prediction and taking into account patterns in the historic performance information due to the hidden and changing policy of providers.

3.7 Quality Spaces

The requirements for the QAS defined above have a number of possible implementations. In the next chapter, a set of approaches to fulfil the role of QAS are defined. Before doing so, it is useful to first introduce a notation which describes the general approaches to assessment. This will be used in the next chapter to demonstrate the conceptual and technical differences between new approaches and existing approaches.

In the abstract assessment model introduced in Section 3.2 a general process to which all existing assessment techniques and approaches adhere was described. At the core of this model a Quality Assessment Subsystem operates on behalf of a stakeholder using a set of quality data to determine the suitability of a service. All approaches to service assessment can be considered as utilising the three types of data described in the model: *consumer expectations*, *delivered values* and *ratings*. As discussed previously, the process of assessment is concerned with analysing and aggregating such data in order to make predictions about the future performance of a service.

It is possible to describe and classify the various high-level approaches to assessment of service quality - QoS assessment, trust and reputation systems; and collaborative filtering - in terms of which of the three data types described above are used in the assessment and what processing is carried out upon the data. The goal of classifying assessment approaches in this way is to describe the conceptual differences between each approach so as to provide a context in which to introduce the approaches to assessment developed in this thesis. Whilst sharing most similarity with existing QoS assessment techniques by using delivered values and consumer requirements data, the approaches developed in this thesis are novel in that at a conceptual level the existence of different types of pattern within the data is utilised.

In order to describe approaches to quality assessment a conceptual notation has been developed (as part of this thesis) in which data is distributed across multiple linked "quality spaces" representing the *expected*, *delivered* and *rated* values of each quality datum recorded in the quality database. There is a one-to-one mapping between spaces, each set of three points representing a single instance of provision (3.6).

$$QD = (e, d, r) \tag{3.6}$$

For example, using "Q-Spaces" notation the information regarding a service (S_3) in the quality database whose contents are described in Table 3.1 may be represented as illustrated in Figure 3.7. Here, there are three records of provision indicated by the three points in each of the spaces. The location of the point within each space is defined by its value and the lines between points indicate that they refer to the same record.

In the remainder of this section, existing classes of approach to quality assessment are defined in terms of processes within and between particular quality spaces. In Chapter 4 a novel high level approach to quality assessment is described using the quality spaces notation in order to illustrate its conceptual differences when compared with existing approaches. The high level approach forms the ba-

ID	t	SP	C	E	D	R
0	1	S_1	C_1	0.3	0.1	0.4
1	3	S_1	C_2	0.8	0.8	0.9
2	7	S_2	C_2	0.5	0.6	0.8
3	10	S_3	C_2	0.7	0.3	0.6
	.	.	.	.	.	.
	.	.	.	.	.	.
45	81	S_2	C_1	0.9	0.5	0.6
46	87	S_3	C_3	0.1	0.5	0.3
47	88	S_3	C_1	0.9	0.7	0.3

Table 3.1: Partitions of e-space and Associated Service Behaviour

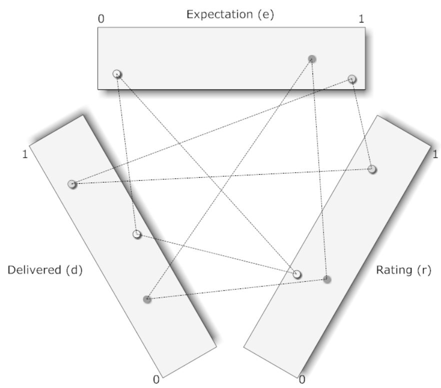


Figure 3.7: S_3 Information represented as Quality Spaces

sis for the development of the specific approaches described and evaluated in the remainder of this thesis.

3.7.1 Traditional D-Space QoS Assessment

Traditional, non-conformance based approaches to QoS assessment which utilise objective measurements consider all points in the QDB to be equally relevant when carrying out an assessment of a service. This is based on an implied assumption that providers will deliver the same level of service to a consumer regardless of the consumer's expectation. The process of assessment of this approach in terms of quality spaces consists of the stages:

(a) **D-Space Selection**: Select all points from D-space that correspond to the past performance of this service.

(b) **D-Space Aggregation**: Carry out some form of aggregation, using the values of the points from D-space.

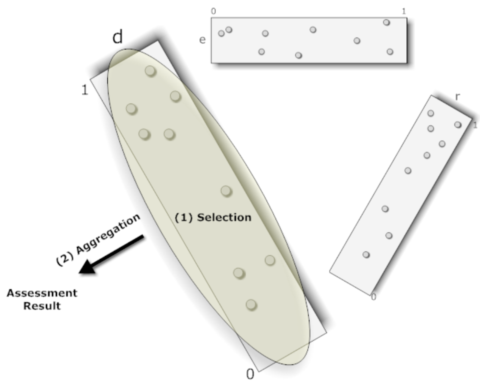


Figure 3.8: Traditional D-Space Assessment

Figure 3.8 illustrates Traditional D-Space assessment in terms of selection and aggregation of d-space points. Such assessment results in a prediction of the expected performance of the provider in terms of the delivered level of service offered to a consumer. This prediction may be used by the consumer to infer the level of quality that they might receive by using the service.

3.7.2 Traditional R-Space Assessment

Approaches to assessment of trust and reputation of providers use subjective consumer provided ratings rather than objective delivered values. In this case, assessment consists of a set of processes similar to that described for Traditional D-Space Assessment but using data from the R-space rather than D-space. Ratings are assumed to be synonymous with the quality perceived by consumers. As with Traditional D-Space assessment, these approaches assume that the ratings provided by consumers are not affected by their prior expectations of the service. That is, that ratings are provided by consumers based solely on the level of performance which they received from the service. In such approaches the process of assessment mirrors that of Traditional D-Space Assessment:

(a) **R-Space Selection**: Select all points from R-space that correspond to past ratings of the current service.

(b) **R-Space Aggregation**: Carry out some form of aggregation using the selected points from R-space.

Figure 3.9 illustrates traditional R-Space assessment in terms of the selection and aggregation of r-space points. The result of a Traditional R-Space Assessment is a prediction in terms of how highly the consumer making the assessment would rate the service. Under the assumptions of the approach such an assessment can be considered as a direct prediction of the level of quality that a consumer could

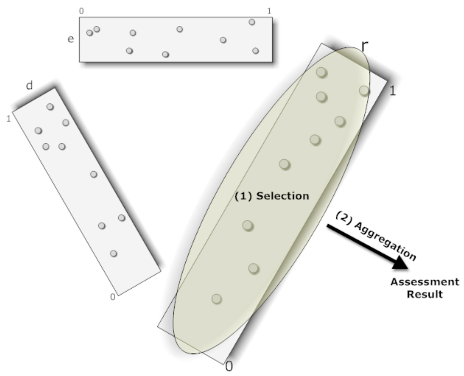


Figure 3.9: Traditional R-Space Assessment

expect from the service. That is, the higher the predicted rating the higher the quality of service that would be received by the consumer.

3.7.3 Collaborative Filtering

The goal of collaborative filtering can be defined as making predictions of how a consumer would rate a particular item if they were to use it. Such predictions are based on a collection of user votes over a population of users (U) and items (I) where information about each item and user is known. Such a situation can be represented in Q-spaces where the votes provided by users form the rating space and the properties of the items used form the delivered space. Each rating-delivery pair can be interpreted as a rating (r) provided by a consumer after using a product with certain measurable characteristics (d). The two types of collaborative filtering – *user based* and *item based* – function by recognising different patterns within and between the spaces.

User-based Collaborative Filtering User-based collaborative filtering is based on the principal assumption that consumers who are *similar* to each other will have preference for the same items. In terms of Q-spaces (Figure 3.10) the process of user-based filtering to establish the expected quality of an item I for an active consumer C_a consists of the following stages:

(a) **D-Space Selection**: Select all points in d-space which correspond to item I. These represent all instances where I has been used by any consumer.

(b) **D-R Space Mapping**: Select the corresponding points in r-space. These are the ratings provided for the selected items by each consumer.

(c) **R-Space Aggregation**: For each point in R-space R_i:

(i) Identify the consumer C_i who made the rating

(ii) Determine the similarity $S_{i,a}$ of C_i to C_a

(iii) Aggregate R_i into the overall rating for I weighted by $S_{i,a}$

(d) **Return Result**: Return the aggregate rating to C_a as a prediction of their likely rating of I.

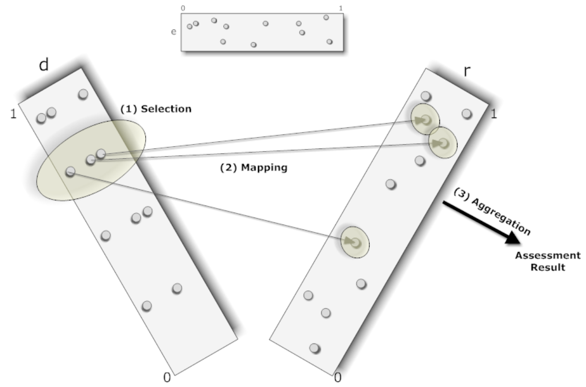


Figure 3.10: User-based Collaborative Filtering

Item-based Collaborative Filtering Item-based collaborative filtering uses the assumption that a consumer who rated particular items highly will also rate *similar items* highly. In terms or Q-spaces (Figure 3.11) the process of item-based filtering to establish the suitability of an item I for an active user C_a (assuming the characteristics of I are known and represented as a vector D_I) consists of the steps:

(a) **D-Space Selection**: Select all points in d-space which correspond to all other items used by C_a

(b) **D-R Space Mapping**: Map the selected items to the corresponding points in r-space. These are the ratings provided by the consumer for the used items.

(c) **R-Space Aggregation:** For each point in d-space d_j

(i) Determine the similarity $S_{I,j}$ of D_I and d_j. That is, determine how similar the current item, j, is to I, by comparing the characteristics of the two items (represented by the vectors d_j and D_I).

(ii) Determine the rating R_j provided by the consumer for item *j*.

(iii) Incorporate R_j into the overall rating for I weighted by $S_{I,j}$. This "rating component" will be higher if the consumer rated item *j* highly and if item *j* is similar to the item being assessed (I).

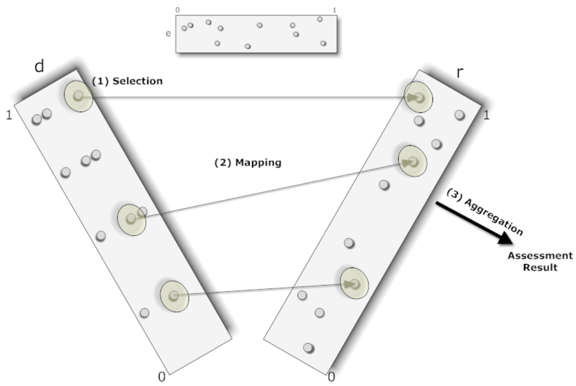


Figure 3.11: Item-based Collaborative Filtering

The result is an aggregate rating which is taken as the rating that consumer C_a would provide for item I if they were to use it. This aggregate rating is considered to be the likely quality of the service.

3.7.4 E-R Space Assessment

User-based collaborative filtering works on the basis that consumers who are similar will have provided similar ratings of items in the past and that the opinions of similar consumers should be weighted more in the decision than consumers who are dissimilar. The principle of expectation-based assessment provides a way of establishing similarity by assuming that consumers provide an "expectation" of how happy they expect to be with the service before they use it. Since these consumer ratings will be based on prior expectation of the service similarity between consumers may be judged by comparing their expectations before assessment. In terms of quality spaces such an approach assumes that there is a correlation between the points in E-space and R-space. The assessment across these spaces is undertaken by the following steps:

(a) **E-Space Selection**: Select all points in e-space which correspond to consumers who have similar expectations to C_a.

(b) **E-R Space Mapping**: Map the selected points to the corresponding points in r-space. These are the ratings provided of the service by the consumers who are similar to C_a.

(c) **R-Space Aggregation**: Aggregate the points in r-space.

As with collaborative filtering the result represents an aggregate rating that is taken as the rating that consumer C_a, would give the service if they used it.

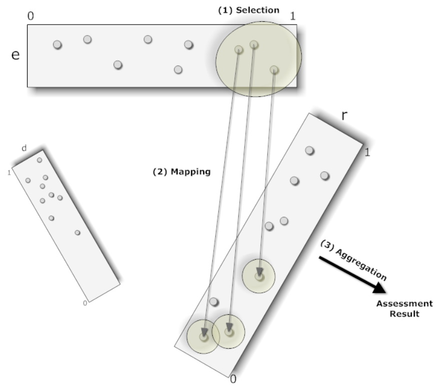


Figure 3.12: E-R Space Assessment

3.8 Summary

In taking a conformance-centric view of quality consumers decide which service to choose based on the degree of match between a provider's predicted service and their expectations. Providers' behaviour is affected by their resource allocation policies which mean that based on their stated expectation consumers will be treated differently in terms of delivered level of service. Due to the monitoring and control overhead involved such policies may result in segmentation of consumers into service levels where all consumers in each service level receive the same level of service. In supporting consumer decision-making these approaches return a predicted level of service for a particular provider and an associated confidence value. The approaches must also consider aspects of service provision which are dynamic with respect to consumer expectations and time.

The quality spaces notation was introduced as a means of describing at a conceptual level the methods of assessment used by existing approaches. In the next chapter, a new class of assessment approach is defined for a conformance-centric SOA. The quality spaces model is used to define the approach and emphasise the differences with existing approaches. The defined requirements for QAS are satisfied by a range of approaches to assessment that may be implemented in the QAS role described in this chapter.

Chapter 4

QoS Assessment in a Conformance-Centric Service Oriented Architecture

In the last chapter, a set of criteria for an effective approach to QoS assessment of service providers in an SOA was specified. In this chapter, these criteria are used to develop a set of approaches to QoS assessment that support consumer decision making within a conformance-centric SOA. This chapter begins by defining a new class of approach to quality assessment that acknowledges the existence of patterns between the expectations of consumers and the classes of service delivered by providers which may be present in performance data if providers adopt a service-class based approach. Following the conceptual definition of the new type of approach four specific implementations of the approach are defined utilising statistical and clustering techniques to identify patterns due to service classes in performance data. Approaches of this type are appropriate for fulfilling the role of the Quality Assessment Subsystem (QAS) in the conformance-centric SOA defined in 3.6.

4.1 Quality Assessment Utilising E-D Space Patterns

At a conceptual level, all approaches described by the Quality Space notation in Section 3.7 consist of a common sequence of stages (see Figure 4.1). In the first stage, points are selected from a particular space based on some criteria. In the second (optional) stage, points are mapped to another space to select the data that will be aggregated. In the final stage, the selected points are aggregated resulting in a prediction of the likely future behaviour of the service being assessed. This is

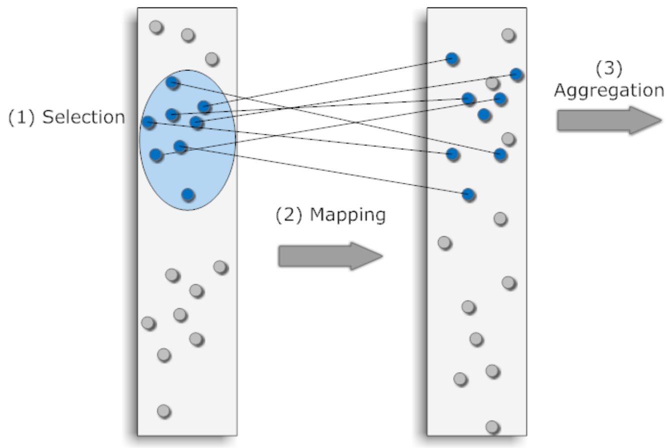


Figure 4.1: The Stages of Quality Assessment

a prediction in terms of the delivered level of service or a rating dependent on the approach. In each case, the additional step of mapping points is justified by the existence of patterns within or between quality spaces. Item-based collaborative filtering recognises that consumers are likely to provide similar ratings to similar items (a correlation between R and D space). User-based collaborative filtering recognises that there is a relationship between types of consumers and the ratings that they provide (a pattern in R space). E-R Space Assessment defines a similarity between consumers in terms of how similar their expectations are, justifying this by explaining that the consumer's rating of a service will be based upon their expectation prior to using it (a pattern between E and R space). The approaches to assessment developed in this thesis are conceptually different from existing approaches in that the relationship between a consumer's expectation and the level of service delivered by a service (a pattern between E and D space) is considered as being of central importance in the assessment process.

4.1.1 The Emergence of ED Space Patterns

A provider can be expected to place an individual consumer into a pre-defined service class based on that consumer's stated expectation (see Section 3.4.2.1). In terms of quality spaces each service class may therefore be considered as a non-overlapping region in e-space. Figure 4.2 illustrates recent data recorded in a QDB and represented using quality spaces for a service offered via two service classes (SC_0 and SC_1). The provider-defined e-space boundaries of these two actual service classes are indicated in Figure 4.2 as shaded regions of e-space. In this example, consumers within service class SC_0 have been provided with a delivered level of service of approximately 0.2 and those consumers within SC_1 with a delivered level of service of approximately 0.7. These levels of service provision have caused the two groups of points at each end of the d-space range to emerge.

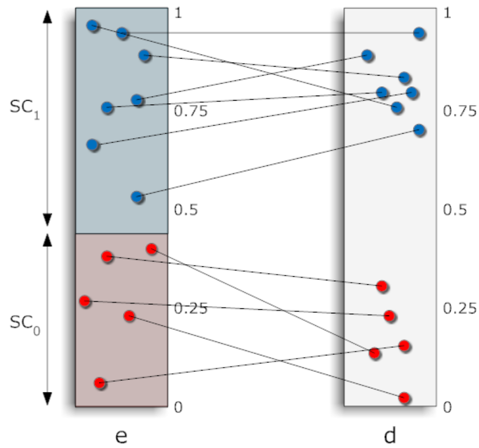


Figure 4.2: E-D Space Correlation

Traditional D-space assessment approaches use an aggregation stage in which all points within d-space are considered as equal. That is, they consider that the data recorded through all instances of service provision are relevant in making a current assessment. In assessing a provider within a conformance-centric SOA (leading to behaviour such as that illustrated in Figure 4.2), this approach is ineffective since the delivered level of service received by individual consumers is dependent upon their expectation. Hence, including all d-space points in the aggregation stage of assessment would produce misleading results. As an example, based on the behaviour illustrated in Figure 4.2 aggregation of all d-space points using an arithmetic mean function would result in a prediction of approximately 0.5. This would be misleading since based on past performance it is very unlikely that a consumer would receive this level of service, regardless of their expectation. Rather, the consumer is likely to receive a level of service relating to one of the two groups of d-space points. Furthermore, which of these levels of service the consumer will receive is dependent upon whether their expectation falls into SC_0 or SC_1. Determining this is a non-trivial problem since the e-space service class boundaries will be hidden (see Section 3.4.2.1).

4.1.2 A High Level E-D Space Quality Assessment Approach

In carrying out assessment which recognises the existence of service classes defined in e-space an initial selection of points from e-space is required in order to determine those which are relevant to the expectation of the current consumer (E_C) – relevance may be judged as any e-space point which is deemed as being in the same service class as E_C by the provider. The goal of the initial stage is to find boundaries in e-space that match the boundaries defined by providers. Throughout this thesis the phrase *service class* is used in three contexts:

- **Actual Service Class:** is used to refer to a service class defined by a provider
- **Potential Service Class:** is used to refer to an e-space range defined by an assessment approach which is intended to correspond to an actual service class.
- **Active Service Class:** the service class into which E_c falls is referred to as the *active service class.*

The first stage of such an E-D Space assessment approach is to identify (or otherwise define) the boundaries of the potential service classes in e-space. More specifically the approaches developed can be considered as consisting of the following stages:

(a) **Provider Policy Definition (E-Space Selection)**: All points in e-space within the same service class as E_c are identified. This step is non-trivial since the boundaries of actual service classes are unknown.

(b) **E-D Space Mapping**: The identified e-space points are mapped to d-space to determine the level of service delivered in each of these cases. This is a trivial step which can be considered as selecting all d-space points which are paired with the e-space points selected in the previous stage.

(c) **D-Space Aggregation**: The identified d-space points are aggregated to predict likely future behaviour of the service for consumers with expectations within this service class. This is a non-trivial step which must consider how the behaviour of the service in the past might be an indicator of the service's behaviour in the future.

Figure 4.3 is an example of the E-D space assessment process. The provider illustrated offers three service classes exhibited in E-Space as SC_0, SC_1 and SC_2, which offer services to consumers whose expectation falls inside the expectation bounds of the class. The consumer's expectation of 0.82 is contained within the boundaries of SC_2 and so all e-space points from this class are identified as relevant. Following mapping of these points to their d-space equivalents, the d-space points are aggregated in order to determine a prediction. In this example, it is assumed that the e-space boundaries of the actual service classes are known and correct and that a d-space aggregation method has been defined though as described above these are non-trivial problems. The remainder of this chapter describes a set of

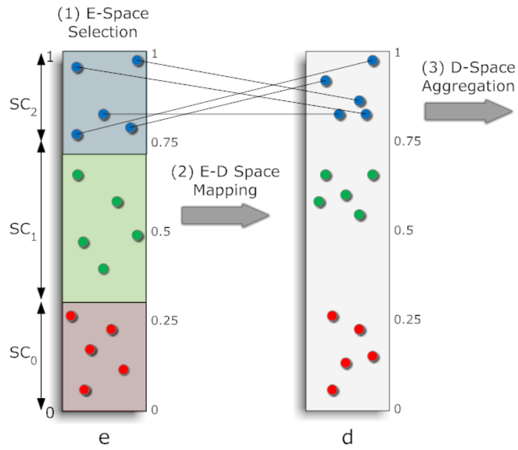


Figure 4.3: E-D Space Assessment Stages

methods and techniques for the solution of the two non-trivial problems described above. Since the method of aggregation is common to all approaches the final aggregation of data points in d-space (Stage 3) is considered first. This is followed by a discussion of a set of approaches for defining potential service classes.

4.2 Aggregation of D-Space Points

The set of d-space points selected following the provider policy identification stage and mapping to d-space represents the delivered level of service for all past provision instances for the relevant potential service class. In this section, it is assumed that this data has been identified and that the goal is solely to produce a prediction of the likely current behaviour of a service based on past performance.

Fundamental to any method of prediction is an assumption of correlation between time (t) and delivered level of service(d). Providers may change the target level of service for any class and therefore, at a particular point in time the relationship between d and t may be in one of three possible states: d is constant, d is increasing, d is decreasing. In the latter two cases in which d is changing in order to determine the likely current behaviour of the service it is necessary to determine the ratio of proportionality between d and t (which can be considered as the gradient of the line) in order to make a prediction. The first case in which d is constant can be considered as a special case.

The problem of prediction may be considered as determining the degree of slope of the line through the collected d-space data and then extrapolation of the line to the current point in time. This is confounded by a number of issues:

(a) **Line of Best Fit:** Observed past performance data in d-space is very unlikely to produce a perfectly straight line. Thus, the parameters of a line of best fit must provide a suitable model of the current state of the relationship between t and d.

(b) **Recency of Data:** Monitored data will become less relevant to a projection as it ages. Including old past performance data in this projection using a line of best fit may lead to a line which is not able to detect recent and ongoing changes since they may be obscured by longer term trends. Conversely, if the time period used is too small the line of best fit may be adversely influenced by noise.

(c) **Uncertainty:** Since it is possible to establish a line of best fit through any set of points, predictions based on extrapolation of such a line are subject to a degree of uncertainty. This may increase due to sparseness of the data set used, noisy data due to inaccurate monitoring, or high levels of variation in delivered level of service.

In the remainder of this section a method for the prediction of future performance through the establishment of a line of best fit is developed addressing the issues identified above. This algorithm is used by each of the quality assessment approaches in this thesis.

4.2.1 Predicting Future Performance

Performance prediction can be broken down into three parts. Firstly, linear regression is used to determine a line of best fit through the set of d-space points. Having identified the line parameters the equation of the line may be used to extrapolate the current delivered value of the service (d_l) at the current time (t_l). Finally, a level of uncertainty in the data used to generate the line of best fit is determined. In order to reduce the impact of problems relating to recency of data a *decay function* is first applied to the data which selects the most recent data points when making an assessment. The number of points selected is definable as a parameter within each of the approaches described later in this chapter.

4.2.1.1 Linear Regression and Extrapolation

Linear regression is used to establish the correlation between a set of variables [50]. When establishing a straight line of best fit through the set of observed data points $O = \{(t_0,d_0),...,(t_n,d_n)\}$, the two line parameters, a and b, in (Equation 4.1), must be determined.

$$d = at + b \tag{4.1}$$

Using the least squares method, the parameters of the line which minimise the sum of the squared errors when applied to the set of d-space points D={d_0,d_1,...,d_n)} are determined such that

$$\arg\min \sum_{i=0}^{n} (d - d_i)^2 \tag{4.2}$$

is satisfied. Hence, a and b may be determined to satisfy Equation 4.2 as:

$$a = \frac{\sum (t_i - \bar{t})(d_i - \bar{d})}{\sum (t_i - \bar{t})^2} b = \bar{d} - a\bar{t} \tag{4.3}$$

Having identified the line parameters the delivered value (d_l) at the time of assessment (t_l) may be extrapolated by substitution of the values into

$$d_l = at_l + b \tag{4.4}$$

d_a is a prediction of the most likely level of service that a consumer would receive if they were to use the service immediately following the assessment.

4.2.1.2 Measuring Uncertainty

Linear regression may be used to determine a line of best fit for any set of data regardless of the degree of real correlation between the underlying variables. In cases where the set of d-space points used in the aggregation has been incorrectly determined there may be no real correlation between the variables (see Figure 4.4). It is very difficult to assure that the data used in aggregation is from a single service class (see Section 4.3.1) so it is desirable to determine the uncertainty associated with the parameters calculated.

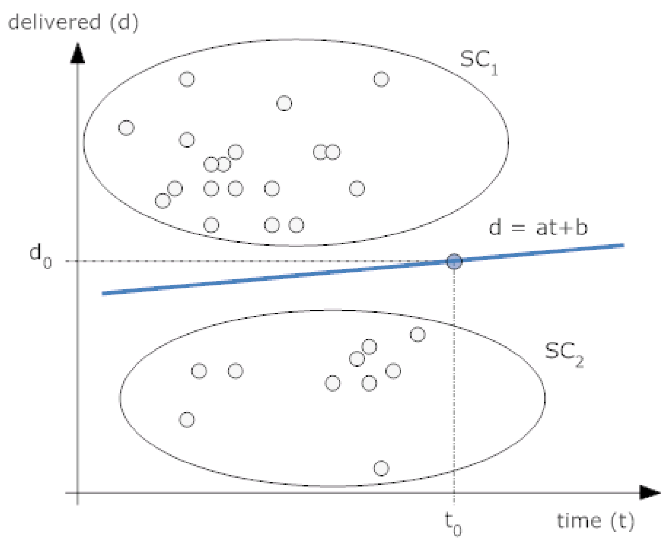


Figure 4.4: Line of Best Fit through Non-Correlated Data

A *correlation coefficient* (r) may be calculated and used to measure the strength and direction of a linear relationship between two variables (examples including Pearson's Product-Moment Coefficient [49] and Spearman's Rank Correlation Coefficient [145]). Such a measure may be considered as a validity measure of the line of best fit established by the least squares method defined above. A standard correlation coefficient (derived from an algorithm defined by Wolfram [164]) is provided in Equation 4.5.

$$r = \frac{n\sum(td) - \sum t \sum d}{\sqrt{[n\sum(d^2) - (\sum d)^2][n\sum(t^2) - (\sum t)^2]}} \tag{4.5}$$

The value of r falls into the range (-1,1), where -1 indicates a perfectly negative correlation between the variables and 1 indicates a perfectly positive correlation. A value of 0 indicates that there is no correlation between the variables.

A more useful measure in determining confidence in the established line of best fit is the *coefficient of determination* (R^2) [50], which falls into the range [0,1] and is a measure of the proportion of variability in the data that may be explained due to a relationship between the two variables. As such, R^2 may be considered as a direct indicator of the uncertainty in the line of best fit.

4.2.2 A Projection Algorithm for Quality Assessment

Based on a consideration of the three stages described in this section, a projection algorithm may be specified. This is described as Algorithm 1 and is used by all approaches developed in this thesis to carry out a final d-space aggregation. The projected value is returned as the assessment result to the consumer (line 17), the uncertainty associated with the parameters of the line of best fit (line 18) being combined with another measure of validity (described in Section 4.3.1) before being returned to the consumer as an overall measure of the validity in the prediction. The projection algorithm is derived wholly from information found in the existing sources defined above [50, 49, 164].

The projection algorithm begins by calculating the components that are required to determine the coefficients and projections: firstly by iterating through the set of points to calculate the sums and sums of squares (lines 1-7) and a second iteration to determine the sum of square errors for each of the points (lines 10-13), based on calculated averages for the d and t components (lines 8-9). Finally, the coefficients (line 17-18) and prediction (line 19) are calculated using the values.

Algorithm 1: Projection Algorithm - derived from [50]

Input: D={(t_0,d_0),...,(t_n,d_n)} the set of D-Space points and associated time stamps relevant to the current service and service class. t_a, the time of the assessment

`/* Calculate sum and sum of squares for d and t */`
foreach $(t_i, d_i) \in D$ **do**
  $sumD = sumD + d_i$;
  $sumD2 = sumD2 + d_i * d_i$;
  $sumT = sumT + t_i$;
  $sumT2 = sumT2 + t_i * t_i$;
  $sumDT = sumDT + d_i * t_i$;
end
`/* Calculate average d and t */`
avgD = sumD / n;
avgT = sumT / n;
`/* Calculate sum of squared error */`
foreach $(t_i, d_i) \in D$ **do**
  $sumNum = (t_i - avgT) * (d_i - avgD)$;
  $sumDen = (t_i - avgT) * (t_i - avgT)$;
end
`/* calculate gradient and intercept */`
a = sumNum/sumDen;
b = avgD - (a * avgT);
`/* Use gradient and intercept to make prediction for time t */`
$Pr = a * t_a + b$;
`/* Calculate correlation coefficient and coefficient of determination from components */`
r = ((n * sumTD) - (sumD * sumT)) / (sqrt [((n*sumD2)-(sumD*sumD))*((n*sumT2)*(sumT*sumT))]);
$C_{agg} = r * r$;

Output: $0 \leq Pr \leq 1$, a prediction of the likely current level of service provided by the service level. $0 \leq V_{agg} \leq 1$, validity in the prediction

4.3 Detecting Provider Policy

The aggregation process described in Section 4.2.2 requires a set of d-space points to be identified. In order to identify this set of points it is necessary to first determine or otherwise define boundaries approximating the actual service classes in e-space, before identifying the active service class and mapping to d-space, from the relevant e-space points.

Defining potential service class boundaries in e-space is a non-trivial problem since a provider is unlikely to publish detailed information about their policies and how they offer differentiated services to consumers (as described in Section 3.4.2).

Ideally, potential service classes defined in e-space by a QAS will match the actual service classes defined by a provider (Figure 4.5(i)). However, it is very difficult to ensure that this is always the case. Characteristics of the data used in making an assessment may mean that in many situations the potential service classes identified in e-space may be a poor match to the actual service classes (Figure 4.5(ii)). Since potential service class characteristics must be inferred from the underlying data, there will always be uncertainty in how well they match the actual service classes. The problem of identifying service classes is therefore two-fold. Firstly, in defining a set of potential service classes in e-space which correspond to the set of actual service classes. Secondly, in evaluating the *validity* of the potential service classes in terms of the quality of match to actual service classes.

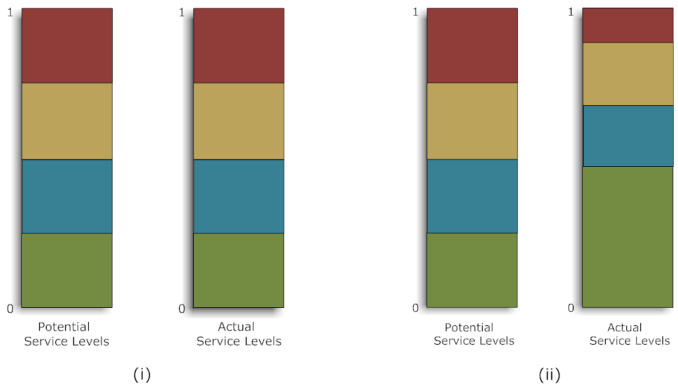


Figure 4.5: Potential vs Actual Service Classes

There are multiple techniques that might be applied in determining the boundaries of e-space service classes and in evaluating the validity of this choice. One major contribution of this thesis is in defining and evaluating a set of approaches for defining potential service class boundaries in order to allow a QAS to select the historic performance data that is relevant in making an assessment of a provider for a particular consumer. These approaches utilise a range of existing pattern classification techniques with tradeoffs being made between predictive effectiveness and computational complexity in each.

This section begins with a description of a set of validity measures that may be used to determine the validity of a particular set of potential service classes (validity being a measure of the likelihood that the potential service classes match the actual service classes). This is followed by the development of a range of approaches to identifying potential service classes. The appropriateness of each validity measure is dependent upon the specific approach employed in identifying potential service classes. That is, each approach selects from the set of available validity measures based on its particular characteristics – though the role of a validity measure in establishing confidence in a particular mapping is consistent across all approaches. Specifically, the validity measure of *Cohesion* is used in approaches to assessment in which multiple, valid, yet overlapping potential service classes are identified. The *Silhouette Coefficient* [121] is used for situations where approaches will generate potential service classes that will be non-overlapping.

4.3.1 Measuring Validity

The set of d-space points which are used in the final aggregation (see Section 4.2) should ideally correspond to all points within the service class that are relevant to the expectation of the active consumer. However, this may not be the case and it is useful therefore to establish a validity measure indicating the degree to which the points used belong to a single actual service class. Dependent upon the approach used once this validity measure has been established it could potentially be used for three purposes: to reject a set of potential service classes in favour of another, as an indicator of quality of the set of potential service classes, or to supervise the algorithm in finding a good set of potential service classes.

Any validity measure should possess three desirable characteristics:

(a) The measure should be an *internal* validity measure meaning that it is calculable solely by the analysis of the data which it is evaluating.

(b) The measure should be proportional to the accuracy of the predictions generated by the approach since the role of the validity measure is to allow the consumer to establish confidence in the result.

(c) The measure should not be affected by the size of the potential or actual service classes in terms of the extent of the boundaries of the corresponding points in d-space.

Three types of validity measure are now described and their characteristics evaluated in terms of meeting the above criteria. These validity measures will be referred to in the definition of specific assessment approaches defined in the remainder of this chapter, to fulfill the three purposes described above.

4.3.1.1 Statistical Hypothesis Testing

Statistical hypothesis testing may be used to prove or refute a particular hypothesis which usually relates to a statement made regarding the underlying distribution of a sample of data [35]. The end result is interpreted as a “degree of significance”, that loosely translates as the likelihood that the data is not generated from a specified data model.

As described in Section 3.4.2, in the short-term the distribution of data within a service class approximates a normal distribution due to the effect of multiple random interference variables (the central limit theorem). As such, statistical hypothesis testing might be used as a validity measure to refute the null hypothesis that *"the data in the potential service class is generated from a normal distribution"*. In this case, the level of significance at which the null hypothesis might be rejected can be considered as an indicator of the likelihood that the data is from a normal distribution and hence from a single service class.

A chi-square test [1] might be used to compare the number of observed data points (O) in various ranges within the distribution against the number of data points that would be expected under the assumed normal distribution (E) from the probability distribution function. In order to establish E, the parameters of the normal distribution (μ and σ) must first be estimated from the data using Maximum Likelihood Estimation (MLE). Furthermore, determining the statistical significance of the chi-square test requires the PDF of a test statistic for the given hypothesis[154] (i.e. that the points are from a single actual service class). This may potentially be calculated using Monte Carlo simulation.

The resulting measure of significance would adhere to all the characteristics defined above since it may be applied to data from a single class, is calculated solely from the sample data, is proportional to the accuracy of predictions and since it is calculated as a likelihood would be unaffected by range size. The measure is also attractive since it is based upon sound statistical foundations. However, whilst theoretically appropriate the measure is likely to be practically infeasible for use in the current context. This is due to the fact that the value of the significance measure returned would likely deteriorate rapidly following a small decrease in the predictive accuracy of an approach due to a potential service class partially overlapping two actual service classes. That is, if data used in the validity testing has just a few percentage of d-space points from a second actual service class the measure will be disproportionately affected. This feature of the approach is desirable when applying statistical hypothesis testing to scientific data which requires a robust and sensitive measure. However, although technically a desirable characteristic for use in establishing cluster validity, this degree of sensitivity to data is undesirable here.

Statistical hypothesis testing adheres to the desirable characteristics of a measure and is a conceptually sound method. However, following some preliminary experimental analysis of the technique the approach was established as being too sensitive to very small degrees of inconsistency between the potential and actual service class boundaries. This is unsurprising since the technique is designed for establishing the likelihood that a dataset is generated from a specific parameterised probability distribution function. Given that approaches to identifying potential service classes will generally lead to matches which are sub-optimal (i.e. that do not correlate strongly with even the function of maximum likelihood), statistical hypothesis testing is discounted as an appropriate validity measure.

4.3.1.2 Cohesion and Separation

Cohesion and separation are internal measures used in pattern classification for establishing cluster validity[151], being based on the principle that a high quality model exists where the difference between points in the same cluster is minimised and difference between points in different clusters maximised. Cohesion measures the former, usually being defined as the sum of pair-wise distances between points in the same cluster; and separation measures the latter, defined as the sum of pair-wise distances between points in a single cluster and the set of points in all other clusters.

$$Cohesion = \sum_{x \in CL, y \in CL} similarity(x, y) \tag{4.6}$$

$$Separation = \sum_{x_1 \in CL, x_2 \in D} similarity(x_1, x_2) \tag{4.7}$$

As a local measure low separation of a cluster from its neighbours will not necessarily be an indicator of low accuracy predictions. In general, if separation of an identified cluster is low this will tend to be an indicator that the points of the cluster are very similar to those in a neighbouring cluster suggesting that the two clusters may be better merged. Whilst in many clustering applications this situation is undesirable, in the context of assessment the situation should not result in significantly lower accuracy of predictions. Accuracy is adversely affected when points from two actual service classes are included in a single potential service class. In the case of low separation potential service classes will still only contain points from a single cluster. Using the same example, if the d-space points identified in clusters C_2 and C_3 are actually from a single d-space cluster when mapped to e-space they will correspond to a single service class. Hence, all points in C_2 and all points in C_3 can be assured to be from a single service class and any predictions based upon them will therefore likely be accurate.

Conversely, cohesion of clusters will likely serve as a good indicator of accuracy. Lower accuracy of predictions results when points from two actual service classes are contained within a single potential service class. If this is the case the d-space points from each actual class will very likely be less similar than those points from a single actual service class. Since cohesion is high if points within a cluster are close together the inclusion of points from multiple actual service classes will lead to lower cohesion.

Whilst being a conceptually useful measure the basic cohesion measure does not meet the third desirable characteristic of being comparable across multiple assessments since it is largely affected by the size of the cluster. Thus, in order to be applied effectively the measure must be normalised. One such weighting is based on cluster size.

Cohesion might be effectively applied as an indicator of predictive accuracy since it is proportional to the distribution of points in the cluster and is diminished by inclusion of points from multiple actual service class in a potential service class. However, the results of the measure are only temporarily relevant and are affected by the size of clusters. Thus, the measure would need to be normalised by weighting in terms of cluster extent to be applied effectively.

4.3.1.3 Silhouette Coefficient

The *Silhouette Coefficient* combines cohesion and separation of clusters [151]. However, rather than operating at a cluster level the silhouette coefficient takes as its basic unit an individual point. For each point (i) in the model, the silhouette coefficient (s_i) is calculated as a relationship between the average distance between the point and all other points in the cluster (a_i) and the average distance between the point and all points not in the cluster to which the point is assigned (b_i).

$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{4.8}$$

The Silhouette coefficient for a cluster may be defined in terms of an aggregation of the silhouette coefficients for its member points. Similarly, the silhouette coefficient may be used to calculate the validity of a cluster model by aggregating the values of individual clusters.

The silhouette coefficient [121] measures the degree to which the point i belongs to the cluster to which it is currently assigned where the value for the coefficient ranges from -1 to 1. Whilst the silhouette coefficient works well in practice its calculation is not based on statistical foundations and the semantics of the various values of the measure are therefore difficult to interpret. Based on analysis of the measure Kaufman and Rousseeuw [72] propose a range of partitions of the value. Significantly, they interpret that a coefficient value for a cluster of less than 25% suggests that no cluster exists and a value of greater than 75% suggests a strong likelihood that the cluster is correct. There is little evidence to support the assumption that this might be generally applied across domains, though silhouette coefficient has been applied successfully as a validity measure in the domains of recommender systems [28] and image classification [77] amongst others. Thus, it would be necessary to validate that the values from silhouette coefficient might be interpreted similarly when applied in other domains.

Since silhouette coefficient is partially based on the concept of cluster separation it is likely that it would not correlate well to predictive accuracy (as described above). However, combining separation and cohesion is a useful measure for global validity of a cluster model. As such, it might be useful in providing an objective function for guiding the iterative clustering algorithms described in later sections.

Although not suitable for combined use in assessing local validity the measures of cohesion and separation are conceptually appropriate in judging the overall quality of a clustering model. The silhouette coefficient offers a combination of the basic measures that has been applied effectively in a number of different domains. Such an approach might be used in supervising a clustering algorithm. However, since it is based upon a point-wise separation measure it may not be effective in measuring local validity in particular situations. Furthermore, it would be necessary to evaluate the correlation of such a measure with the predictive accuracy of results in order to establish significance.

4.4 Approaches to Selecting Service Policy Ranges

As described in Section 4.1.2, the structure of quality assessment approaches developed in this thesis consist of three stages:

(a) **Provider Policy Identification (E-Space Selection)**: select all points in e-space corresponding to consumers with an expectation in the same service class as the expectation of C_a. This involves the identification of a set of potential service classes.

(b) **E-D Space Mapping**: map points to d-space to determine the level of service delivered by the provider in each of these cases.

(c) **D-Space Aggregation**: aggregate points in d-space to predict likely future behaviour of the provider for consumers in this service class. As part of this final stage validity measures are calculated to indicate the quality of prediction.

Sections 4.2.1 and 4.3.1 have discussed the means by which the final d-space aggregation takes place and how to establish validity.

The problem addressed in this section is the identification of boundaries of the provider-defined service classes in e-space. The approaches each define a set of potential service classes in e-space which are deemed to correspond to the provider-defined actual service classes. Appropriate measures of validity selected from those described in the Section 4.3.1 are used in each approach as an indicator that the potential service classes are a good match to the actual service classes by examination of d-space points. The selection of a particular validity measure is described for each algorithm with justification of its choice.

The arrangement of potential service classes in e-space is usefully described in terms of a set of bounded ranges which may be characterised by three parameters:

- The number of classes (N)
- The placement of the class centroids (C)
- The size of the classes (S)

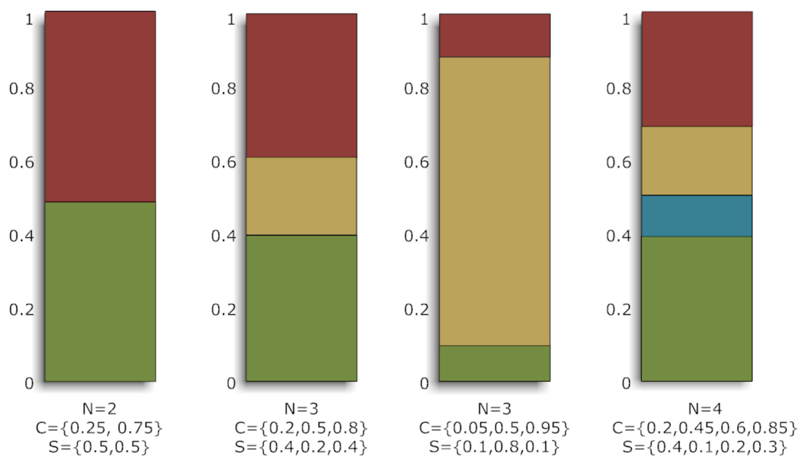


Figure 4.6: E-Space Service Level Range Policies

Figure 4.6 illustrates four arrangements of service classes defined in terms of these three parameters.

Approaches to establishing potential service classes may be classified in terms of whether they actively attempt to identify service classes by determining all three parameters at run-time based on analysis of the data (dynamically) or whether they assume the values of parameters at design time (fixed). From such a classification emerges a spectrum of possible approaches, from those where the parameters are all fixed through approaches where a combination of N, S and C are determined dynamically with the remainder fixed to the other extreme where all parameters are determined dynamically. Table 4.1 illustrates these combinations.
Combinations which are not present in Table 4.1 correspond to situations which are not possible, due to dependencies between the three parameters. Specifically, the placement of class centroids and the size of classes are dependent upon the number of classes defined. That is, if the number of service classes changes, the size of the service classes and their centroids will necessarily change, with centroids being repositioned and class size increasing or decreasing to take into account

Approach	No. of Classes	Centroid	Size
(a)	Fixed	Fixed	Fixed
(b)	Fixed	Dynamic	Fixed
(c)	Fixed	Dynamic	Dynamic
(d)	Dynamic	Dynamic	Dynamic

Table 4.1: Potential Combinations of Service Class Parameters

the removal or addition of a service class. As such, any approach which determines the number of classes at run-time must also determine the other two properties at run-time meaning that certain combinations are ommitted from Table reftab:spcombination. In contrast, fixing the number of service classes at design-time still allows for the centroids and size of classes to be determined dynamically at runtime. For example, Figure 4.6 illustrates two potential choices of centroid and class size for the case where the number of service classes is fixed as two. The techniques employed by the approaches detailed in the table above increase in terms of complexity as the criteria become more dynamic though in doing so the effectiveness of the approaches in terms of predictive accuracy is likely to increase. In Chapter 6, approaches are evaluated based on both the predictive accuracy of approaches and in terms of the degree to which the effectiveness of consumer decision making is improved.

In the remainder of this chapter, four approaches to determining potential service classes that are based on the possible combinations defined above are presented. Although the approaches primarily differ in terms of the means of definition of ranges in e-space they may have implications for the basic aggregation of d-space points discussed previously in Section 4.2. Thus, each approach is explained as a complete approach to quality assessment incorporating the three stages defined in Section 4.1. In each of the approaches, it is assumed that an active consumer C_a with expectation E_C is requesting an assessment of a service S and that the only points being used in assessment are those referring to S. That is, that the QDB has been searched for all relevant points and the e and d space populated with data.

4.5 Pattern Passive Approaches

The first set of approaches (Approach (a) and Approach (b) from Table 4.1) can be considered as being *pattern passive* in that the specification of potential service classes is not based on a detailed examination of points within spaces at runtime. That is, the pattern of service class layout is defined at design-time and remains constant across all assessments carried out by the approach. In this sense, the approach is considered ”passive” since it does not have the ability to adapt over time without manual reconfiguration of parameters..

The potential service classes identified by pattern passive approaches are to a great extent decided at design time which in many cases will generate a poor match to the actual service classes. However, in these instances validity will be correspondingly low and should serve as an indicator to the consumer that the assessment result may be incorrect. Thus, pattern passive approaches should still outperform traditional D-space assessment approaches in terms of the degree to which they support consumers in choosing the best services. This is because although both may perform poorly in terms of predictive accuracy pattern passive approaches will indicate when a poor match to actual service classes has occurred with an associated low validity measure. This is in contrast to traditional approaches which do not acknowledge the existence of service classes and will therefore be unable to determine the accuracy of their predictions in this respect.

Two pattern passive approaches are described below. The first, corresponding to combination (a) from Table 4.1 assumes that all parameters are fixed at design time. The second, corresponding to combination (b) allows for the definition of the range centroid at run time with the other two parameters remaining fixed.

4.5.1 Pre-defined Fixed Boundaries - Approach (a))

In the case where the number, placement and size of potential service classes are fixed the expectation space is split into n equally sized ranges R={R_1, ... , R_n} where each range R_i = ($E_{i(low)}$, $E_{i(high)}$) is delimited by an upper and lower expectation value such that $E_{i(high)} < E_{i+1(low)} \forall R_i$ and such that $E_{0(low)}$=0 and $E_{n(high)}$=1. That is, that the ranges are non overlapping and cover the entire E-Space range.

On receiving a query from an active consumer *c*, in terms of E_c , the following assessment process takes place:

(a) **Range Selection:** The target range R_T is determined such that $E_{T(low)} < E_c <= E_{T(high)}$. R_T is the range into which the active consumer's expectation falls.

(b) **Point Selection:**All e-space points e_i in the QDB are selected such that $e_i \in R_T$.

(c) **E-D Space Mapping:** The d-space points corresponding to the e-space points from the previous step are selected.

(d) **D-Space Aggregation:** The d-space points selected in the previous step

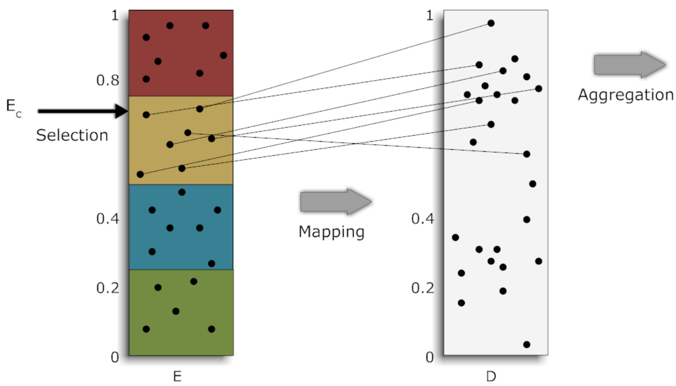


Figure 4.7: Pre-defined Fixed Boundaries: Selection, Mapping and Aggregation

are aggregated using the projection method described in Section 4.2.2. This results in a prediction P_S of the level of delivered service.

(e) **Validity Calculation:** The validity of the result is calculated based on a combination of the two validity measures for projection and service class quality.

The effectiveness of the approach is dependent upon the relationship between potential and actual service classes. In the case where the potential service classes relevant to E_c is subsumed by a single actual service class d-space points used in the final aggregation will be from a single service class and will therefore yield high accuracy and high validity predictions. In contrast, in the case where the potential service class relevant to E_c overlaps two or more actual service classes d-space points will be drawn from both. This leads to a prediction that is likely to be of low accuracy. In this case, the validity measure should be correspondingly low.

The accuracy of predictions of approaches is entirely determined by the arrangement of actual service classes which are not under the control of the approach. An approach relies on the validity measure to maintain effectiveness. However, it should still support consumers in discovering higher quality services than traditional approaches which will generally yield low accuracy predictions but have no measure to indicate this.

4.5.2 Dynamic Fixed Boundaries - - Approach (b)

The major problem with the fully-fixed approach outlined in Section 4.5.1 is that the choice of each parameter is essentially arbitrary - the nature of the ranges being decided wholly at design time. A partial solution to this problem which maintains fixed parameters whilst allowing some dynamicity is to keep the number of potential classes fixed whilst allowing the other characteristics of the potential classes to be established at run-time. This constitutes the second approach type defined by Table 4.1(b). Taking this approach, when a consumer request with associated expectation E_c is received the potential service class relevant to E_c is generated dynamically by defining a class, in which E_c forms the mid-point and which is delimited by a threshold value δ (Figure 4.8). An e-space point e_i is deemed as being within the potential service class iff $E_c - \delta < e_i < E_c + \delta$. The approach effectively divides the e-space into three potential service classes: a class into which E_c falls, a class which is greater than E_c and a class less than E_c. In such an approach, the following assessment process takes place:

(a) **Range and Point Selection:** All points in e-space e_i, are selected such that $E_c - \delta < e_i < E_c + \delta$

(b) **E-D Space Mapping:** The selected e-space points are mapped to corresponding d-space points.

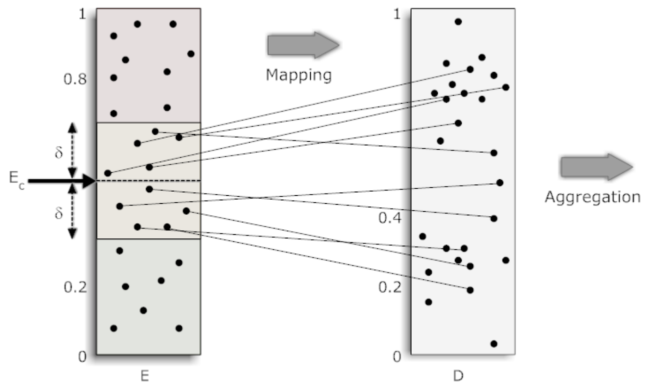


Figure 4.8: Dynamic Fixed Boundary: Selection, Mapping and Aggregation

(c) **D-Space Aggregation:** The selected d-space points are aggregated using the projection method described in Section 4.2.2. This results in a prediction P_S of the level of delivered service.

(d) **Validity Calculation:** The validity of the result is calculated based on a combination of the two validity measures for projection and service class quality.

Such an approach is still inherently arbitrary since the threshold (δ) which defines the size of the potential service class into which E_c falls is specified at design time. As for Approach (a) when E_c falls at the centre of an actual service class the predictive accuracy of the approach will be high. Also, the accuracy of the prediction generated by the approach is dependent upon the correlation of the target potential service class and the actual service class of the provider.

4.5.3 Pattern Passive Validity Measure

The potential service classes identified in the two pattern passive approaches will vary in the degree to which they match actual service classes. In terms of the choice of appropriate validity measure, two issues are relevant.

Firstly, the validity of an approach's prediction will be dependent upon E_c. The validity measure reflects the degree to which the final d-space points are from a single actual service class and this will differ based on the potential service class into which E_c falls. As illustrated in Figure 4.5, if E_c is 0.1, it falls into a potential service class that corresponds to a single actual service class and hence predictions will be of high accuracy. In contrast, for the same example if E_c is 0.6 it will fall into a potential service class that overlaps with two actual service classes yielding a prediction of lower accuracy. Thus, validity should be calculated for only the points used in aggregation (a local measure of validity) rather than as a global measure across all points since this will reflect the overall validity for all consumers. This will ensure that the measure of validity is useful to the consumer making the assessment.

Secondly, it is possible that two potential service classes might cover points from a single actual service class (Figure 4.9). In this case, the points used in the final aggregation for each potential service class will be similar since they will correspond to the same actual service class. Furthermore, the accuracy of the resulting prediction is likely to remain high for the same reason and it is desirable that the associated validity measure should remain high in such cases. Hence, the validity measures of separation (see Section 4.3.1.2) and silhouette coefficient (see Section 4.3.1.3) are inappropriate since they will result in low scores for such cases.

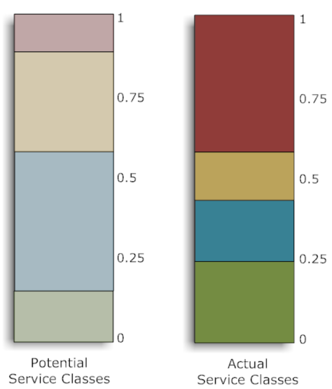


Figure 4.9: Two potential service classes relating to a single actual service class

The most appropriate measure of validity is therefore cohesion (see Section 4.3.1.2). This is a local measure which will return high validity when points used in aggregation are similar whilst ignoring separation of clusters. In the case where the final set of points used in aggregation are drawn from two or more actual service classes the density of points within the cluster will be high and therefore a cohesion-based measure will return low validity. Conversely, when the final set of points are drawn from a single actual service class there will be relatively high density and therefore high validity.

The measure of cohesion is affected by the size of clusters. Thus, a measure of "normalised cohesion" is defined and used here in which the result is weighted based on the size of the cluster in terms of the difference between minimum and maximum points.

$$\frac{\sum_{x \in C, y \in C}(x - y)}{max(x) - min(x)} \tag{4.9}$$

Using this measure will allow objective comparisons using cohesion across assessment instances and comparison of the results for multiple providers.

4.6 Pattern Active Approaches

The pattern passive approaches considered above define attributes of the potential service classes at design time. In each case, a validity measure is used to indicate the degree to which data from within a potential service class belongs to a single actual service class. This supports a consumer's decision making in that they are able to establish a level of confidence that can be attributed to a prediction. Although in many instances the confidence that a consumer might have in a prediction will be low such an approach is always preferable to approaches that do not acknowledge the existence of service classes and that do not provide a validity indicator as part of the assessment result.

A validity measure is necessary as it allows consumers to establish the level of risk in choosing a particular service. However, the accuracy of the prediction is of even greater significance to the consumer. The selection of potential service classes within the pattern passive approaches are essentially arbitrary in that they do not consider the individual differences in policy that might be exhibited between different providers or recognise the changing nature of service class boundaries over time (see Section 3.4). Thus, whilst always being preferable to existing non-conformance centric quality assessment methods, pattern passive approaches are naïve, and use validity indicators as a form of post-analysis to mitigate the negative effects of the arbitrary choice of potential service class attributes. In order to produce accurate predictions of provider performance non-naïve approaches which seek to identify the policies of individual providers and which recognise that such policies may change over time are required. Thus, in contrast to pattern passive approaches, "pattern active" approaches are capable of determining the pattern of service class layout at run-time. The pattern identified will not remain constant across all assessments carried out by these approaches since different patterns will be identified and used for assessment of different services and changes in service behaviour over time will lead to different patterns of service class layout being identified and used. The approaches are considered as “active” since they have the ability to adapt over time.

D-space data recorded as a result of monitoring service provision can be considered as a sample of the behaviour of a provider where each point has been generated from one of the provider’s service classes, but is unlabelled. The policy identification problem is therefore reduced to re-classifying the d-space points into groups each containing d-space points that were provided to consumers from a single service class. Having determined the class labels for all d-space points the corresponding e-space points may be identified to establish the segmentation policy of the provider with respect to the expectation of consumers. Changes in a provider’s policy to defining actual service class boundaries over time, mean that it is necessary to consider only recent d-space data as being relevant when identifying the d-space groups.

The goal of research within the domain of cluster analysis[151] is to partition data into meaningful groups when other information about their composition is unknown [43] whereas for the domains of constraint based reasoning and probability theory pattern classification is considered to be a special application of the underlying theories of the domain. Most significantly, cluster analysis is conceptually consistent with the concept of quality spaces which treat the assessment problem as one of manipulation and classification of points. In contrast, any probabilistic approach would require the problem to be re-framed in terms of probability distributions which whilst being a valid method of representation would be unlikely to offer any improvements in representation of the problem.

The overall assessment process, which supports the methods of clustering developed in this section is broadly similar to that of the pattern passive approaches, but with two significant differences. Firstly, an initial stage (Step 1) is carried out in which d-space points are examined and clusters identified. These are then mapped to e-space to identify potential service class boundaries (Steps 2 and 3). Secondly, during the e-space selection stage the possibility that E_C might fall into multiple potential service class ranges must be considered since outliers and other data characteristics might mean that the potential service classes are a poor match to the actual service classes. For completeness, the overall assessment process including the additional stages is as follows:

(a) **Identify Point Clusters in D-space:** D-Space points are classified using cluster analysis techniques. Ideally, points in a single cluster will correspond to all instances of provision within a single actual service class. This is a non-trivial step in which appropriate existing clustering techniques are applied (Figure 4.10(i)).

(b) **D-E Space Mapping:** The e-space points for each of the d-space points are identified.

(c) **Identify Service Class Boundaries in E-Space:** Using the clusters identified in d-space the boundaries of potential service classes are identified in e-space (Figure 4.10(ii)).

(d) **E-Space Selection**: All points in e-space within the same potential service class as E_c are selected. Since the potential service classes may be overlapping, E_c may fall into multiple classes. This step is non-trivial requiring the algorithm to address the multiple class problem.

(e) **E-D Space Mapping**: The identified e-space points are mapped to d-space to determine the level of service delivered by the service in each of these cases. These points will correspond to one of the clusters identified in Step 1 (Figure 4.10(iii)).

(f) **D-Space Aggregation**: the identified d-space points are aggregated using techniques described in Section 4.2.2, to predict likely future behaviour of the service for consumers with expectations within this service class.

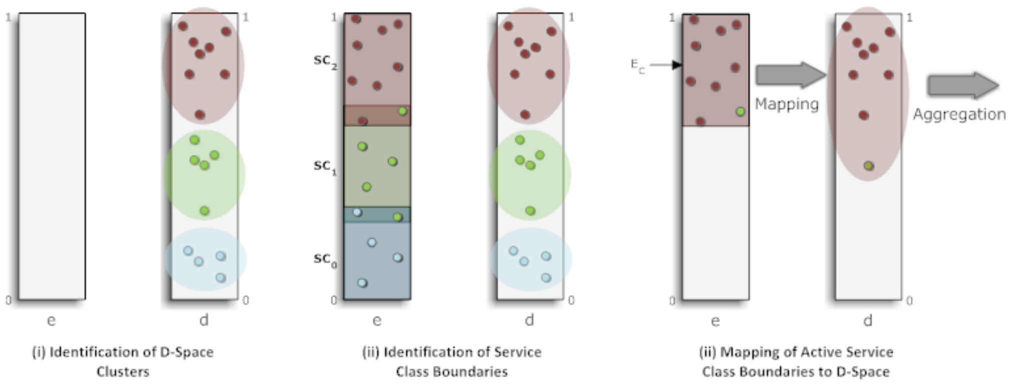


Figure 4.10: Pattern Active Assessment Process

The aim of describing the pattern active approaches developed in this section is to effectively demonstrate the improvements in predictive accuracy that are obtained by actively recognising service classes at run-time. An assessment approach incorporating a k-means clustering algorithm to classify d-space points is now developed. Firstly, an overview of clustering is provided whereby the specific clustering problem in d-space is defined. This is followed by a description of a k-means algorithm for clustering d-space points including consideration of a number of issues relevant to the k-means algorithm. Next, the remaining assessment stages are discussed and a means of addressing the multiple classes problem, when E_c falls into multiple potential service classes, is presented. Finally, a summary of the approach and a specification of the full algorithm is presented.

4.6.1 Fixed k Clustering - Approach (c)

The set of unsupervised learning techniques known as clustering seek to partition a set of data points D into a set of classes C, based on the establishment of proximity (or similarity) between the data points such that all data points within a single class are deemed to share a common characteristic in terms of their features. In its most general sense clustering can be considered to be an optimisation problem [31] which seeks a partitioning of the data that maximises its likelihood [11]. For many applications the problems of outlier detection, choice of initial centroids and choice of cluster numbers may need to be addressed or mitigated based on the type of clustering method used.

4.6.1.1 Clustering Method: The K-Means Clustering Algorithm

The K-Means clustering algorithm [83] is a partitional algorithm, in which data points are distributed into non-overlapping subsets such that each item is in exactly one subset forming a one-level partitioning of the data [146].

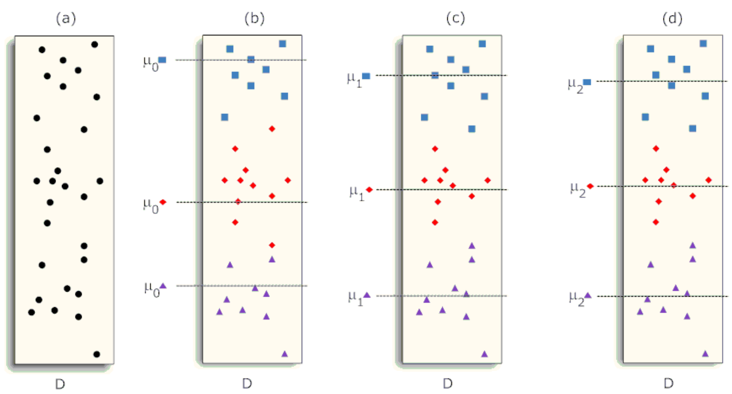


Figure 4.11: K-Means Clustering of d-space Points

The algorithm is an iterative approach to clustering in which a set of k centroids are initially selected. An initial set of clusters are defined by assigning each point in the space to the closest (or most similar) centroid (Figure 4.11 (a)). Following assignment the centroids of each cluster are recalculated as the mean of all points in the cluster (Figure 4.11 (b)). The k-means process then repeats, redistributing the points to the closest cluster defined in terms of the new centroids (Figure 4.11 (c)). Over multiple iterations the value of the objective function exhibited by the clusters increases until there is little or no change in cluster membership (Figure 4.11 (d)), at which point the clustering process terminates. In some cases the algorithm may reach a stage where a small number of points continue to oscillate between two clusters, meaning that the termination condition will never be met. Thus, a second termination condition is usually defined in terms of a maximum number of iteration is usually adopted.

The major advantage of using k-means as an alternative to other clustering algorithms is its efficiency: storage complexity of k-means is O((m+K)n) and time complexity is O(I*K*m*n) [151, 64] (where K is the number of clusters, I is the number of iterations, m is the number of attributes of each point and n is the number of points). This means that k-means is linear in the number of points. The standard K-means method[40] is appropriate for cluster analysis of d-space points for several reasons. Firstly, it is known that each d-space point was generated from a single service class. Thus, a partitioning mechanism which assigns each point to a single cluster is appropriate. Secondly, as discussed in Section 3.4.2 the distribution of data within each of the service classes defined by a provider can be assumed to be gaussian in the short term. This is significant as it can be shown that the standard measures of similarity and objective function used by k-means lead to clusters which are of high quality, if the underlying data is a multi-normal distribution as may be assumed for d-space data.

The k-means algorithm requires the specification of a value for k in order for clustering to take place. It is expected that in particular situations the value for k may be inferred at run-time based on existing information about the characteristics of service providers. For example, a provider may publish three advertisments relating to the same service to a directory service without providing details of the service class boundaries that distinguish the performance that a consumer might receive from each. The impact of this assumption in situations where k is a true representation of the actual number of service classes of a provider and where it is not are evaluated in Chapter 6.

4.6.1.2 Outlier Detection

In the context of clustering d-space points outliers may occur due to significant deviation of delivery from the target level of service of a service class. Such points may prove problematic if a k-means algorithm (which is especially sensitive to outliers) is used. The problem of outliers is illustrated in Figure 4.12(a). Here, a service has two classes, SC_0 and SC_1, represented by the blue and red points respectively. In the example, SC_0 has generated a collection of d-space points at around 0.3, and SC_1 a collection of d-space points at around 0.8. However, SC_1 has a single outlier, which is represented as point *p'* in Figure 4.12. During the initial d-space clustering phase the outlier is incorrectly identified as being part of the cluster corresponding to SC_0. Thus, when the D-E space mapping takes place overlapping potential service classes are identified in e-space (indicated by the shaded region in Figure 4.12). Whilst the technique described in Section somewhat mitigates the problem of overlapping service classes it is desirable that such situations should occur as infrequently as possible by removing outliers prior to clustering. Figure 4.12(b) illustrates the service classes identified by a clustering algorithm if the outlier point p' is removed prior to clustering.

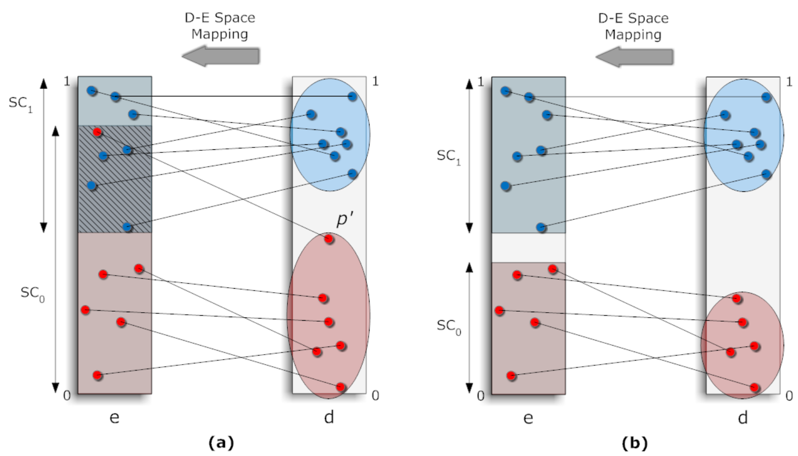


Figure 4.12: The Problem of Outliers

Two general options exist to deal with outliers. Firstly, they can be considered after clustering has taken place [151]. Algorithms of this type will judge the "outlierness" of a point based on its distance from the points within its labeled class. This is not appropriate for k-means as the existence of outliers might direct the clustering process and cause poor quality clusters to be developed due to the hill-climbing nature of the algorithm. Post-clustering detection might allow the validity measure of any prediction to be adjusted but would not allow the accuracy of predictions to be increased. Secondly, outliers may be considered before clustering. This type of approach tends to be developed with the aim of identifying the outliers for further analysis. However, such approaches can also be used to avoid clustering with points that will not cluster well [151]. It is generally accepted that for the purposes of outlier removal even crude approaches can lead to significant improvements on data where classes are known to exist [8]. As such, the grid-based method will be used to detect and remove outliers as a pre-clustering step. More specifically, the pre-clustering algorithm RemoveOutliers (see Algorithm 2) will be applied to d-space data prior to the clustering algorithm being run on the data.

Algorithm 2: Outlier Removal Algorithm - adapted from [151]

Data: $D = \{x_1, ..., x_n\}$

Result: $D' \subset D$

`/* determine the average density by dividing the space by the number of points */`

1 $density = 1/ \mid D \mid$;

2 $binsize = density * b$;

`/* create an array of bins, containing points within each range */`

3 Bins = split(D, binsize);

`/* remove any bins that do not exceed the threshold number of points (t) */`

4 $ValidBins = \{bin \mid bin \in Bins, \mid bin \mid \geq t\}$;

`/* return the set of points in the larger bins for use in later processing */`

5 $D' = \{x \mid x \in bin_j, bin_j \in ValidBins\}$;

The algorithm first splits the space into a set of fixed size bins, based on density (lines 1-3), assigning each point to a single bin. Then a threshold value (t) is used to remove any bins that do not contain a sufficient number of points (line 5) which are assumed to be outliers. Finally, the set of points contained in the remaining bins are returned to be used in the main clustering algorithm (line 6). By removing outliers prior to clustering, the risk of cases with large overlapping potential service classes being identified in e-space following the mapping from d-space clusters, is minimised.

4.6.1.3 Choice of Initial Centroids

The k-means algorithm is deterministic following the selection of initial centroids [11], and so in many cases the k-means algorithm will not converge to the optimal set of clusters [83], instead resulting clusters will represent local maxima [11] in terms of optimising the objective function. It is therefore necessary to choose initial centroids which are likely to give the best possible clustering quality rather than converge to local optima. Thus, the initial choice of centroids is a deciding factor in the quality of the clustering achieved by the algorithm and is therefore an important consideration in the design of any k-means algorithm.

Multiple existing approaches to choosing initial centroids for the k-means clustering algorithm exist though as with the clustering methods themselves there is no generally accepted "best" approach. Traditionally, centroids are chosen randomly [34] though this approach is likely to lead to the local convergence described above resulting in poor clusters [151]. The approach described by Tan et al [151] in which a hierarchical clustering algorithm is applied to a sample of the data set, in order to determine a set of good quality centroids is suitable for application in clustering d-space data. Such an approach is described as being practical for cases where the sample is relatively small and here k is relatively small compared to the number of samples. Since both of these criteria are met by the data within d-space this method will be used to choose the initial centroids for use by the incremental k-means clustering algorithm. This approach is used in Algorithm 3.

Algorithm 3: Initial Centroid Selection Algorithm - adapted from [151]

Data: $D = \{x_1, ..., x_n\}$
Result: $M = \{\mu_1, ..., \mu_k\}$, a set of k means
`/* put all of the points into a single (initial) cluster */`
1 $C_1 \longleftarrow D$;
2 $i \longleftarrow 1$;
`/* carry out the divisive process k times, to identify k clusters */`
3 **while** $i < k$ **do**
4     $i \longleftarrow i + 1$;
    `/* get the largest cluster and the two farthest points within */`
5     $(x_a, x_b, C_{par}) \longleftarrow distant(D)$;
    `/* split the cluster in two and reassign points */`
6     $C_i = \{x_j \mid x_j \in C_{par}, x_j < x_a\}$;
7     $C_{par} \longleftarrow C_{par} \setminus C_i$;
8 **end**
`/* return the median of each of k clusters as the initial centroids */`
9 $M = \{\mu_i \mid \mu_i = M(C_i)\}$;

Where *distant(D)* is a function which identifies the largest cluster (line 5) (defined in terms of the cluster containing the two points with maximum distance), returning the two points x_a and x_b and their parent cluster, C_{par}. $M(C_i)$ is a function which returns the mean of the points in the cluster C_i (line 6). The algorithm begins by taking a sample of the d-spaced data points from D (line 1). A basic divisive hierarchical clustering algorithm is then applied to the points (line 3), in which i clusters are iteratively broken down into i+1 clusters by splitting the largest cluster. The process terminates when the algorithm has split the data into k clusters and the points closest to the means of the final clusters are chosen as the initial centroids for the k-means clustering (line 9).

By carrying out a hierarchical clustering pre-analysis stage the initial centroids should provide a better correspondence to the clusters within the data than using a random or pseudo-random approach [34].

4.6.1.4 A K-Means Clustering Algorithm for Identifying Service Classes in D-Space

Based on the above description an algorithm for clustering d-space points into k clusters with the initial centroids chosen based on the provider's advertisement may be defined (see Algorithm 4). The application of Algorithm 4 to a set of D-Space data points results in a set of clusters of d-space points. The points within each cluster are deemed to refer to all the points generated from a single service class.

Algorithm 4: K-Means Clustering Algorithm - based on [83]

Data: $D = \{x_1, ..., x_n\}$
Result: $C = \{c_0, ..., c_k\}$

begin
  `/* remove outliers and select initial centroids using previous algorithms */`
  RemoveOutliers(D);
  M = SelectCentroids(D,k);
  `/* keep iterating until there are no point movements */`
  while *Points move between clusters* **do**
    `/* for every point in the data, associate with the closest cluster median */`
    for $x_i \in D$ **do**
      Reassign x_i to c$_j$ such that $\mid x_i - \mu_i \mid$ is minimised;
    end
    `/* following point reassignments, recalculate cluster means */`
    for $\mu_i \in M$ **do**
      $\mu_i = \frac{\sum_{x \in c_i}}{|c_i|}$;
    end
  end
end

4.6.1.5 Measuring Cluster Validity

A k-means clustering algorithm is able to partition any data set into any given number of clusters, regardless of the actual number of classes in the underlying data. As such, the validity of any model derived from such clustering is questionable. Having identified a set of non-overlapping clusters in d-space the validity of the clustering must therefore be established in order to allow the consumer to gain confidence in the resulting prediction.

In contrast to those identified by pattern-passive assessment approaches (see Section 4.6) the clusters of points identified in d-space by Approach (c) will not overlap due to the nature of the partitioning k-means clustering algorithm and measure of similarity used. As such, separation-based validity measures (which were previously discounted) may be considered as being appropriate for application in judging the distinctness of individual clusters. For the k-means clustering algorithm described above the silhouette coefficient is therefore chosen as the most appropriate validity measure due to the fact that it incorporates both the cohesion and separation of clusters in its calculation and has been effectively applied in a wide range of pattern classification problems (see Section 4.3.1.3).

4.6.2 D-E Space Mapping and Range Identification

Having identified clusters in D-space via the k-means algorithm the corresponding points of each cluster are identified in E-space. The purpose of this mapping is to identify the upper and lower expectation limits for each of the potential service classes by determining the lowest and highest expectation in each of the sets of points. The result of such identification is a set of potential service class ranges R={R$_1$, ... , R$_n$} where each range R$_i$ = (E$_{i(low)}$, E$_{i(high)}$) is delimited by an upper and lower expectation value (see Figure 4.13). Ranges may overlap such that $E_{i(high)} > E_{i+1(low)}$ and dependent on the nature of the data available to assessment, the ranges may not cover the entire expectation range. In an ideal

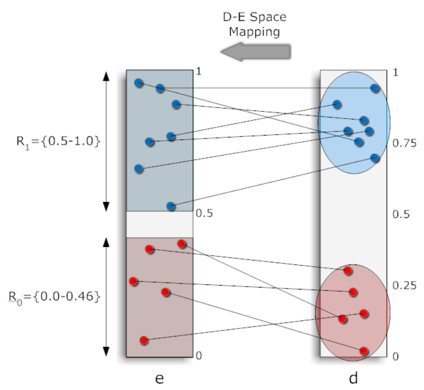


Figure 4.13: D-E Space Mapping and Range Identification

case the ranges identified in E-space should be non-overlapping. However, due to outliers in D-space data actual service level ranges with high rates of variance and the fact that T for each service level may change with time, it is possible that there will be a degree of overlap between the ranges identified in e-space. The definition of overlapping potential service class ranges in e-space has implications during the E-D Space Mapping stage of the assessment process, by which the range into which the current consumer's expectation E_c is determined. As such, these implications are dealt with below.

4.6.3 E-D Space Mapping and Aggregation

The purpose of the final stages of assessment in this approach is to ascertain the level of service that a consumer with an expectation E_c would receive if they were to use the service. As in previous approaches to E-D space assessment E_c is used to determine the potential service class into which the consumer falls, all e-space points within the class are identified, mapped to d-space and finally aggregated.

In previously described E-D space assessment approaches the potential service classes were defined such that they were non-overlapping and covered the entire e-space range. Thus, two assumptions were reasonable when selecting e-space points. Firstly, that E_c would fall into a single potential service class. Secondly, that each e-space point would be within the range of a single potential service class.

In contrast, in the case of dynamically recognising potential service classes through clustering, the e-space ranges may be overlapping. Thus, there are three possible relationships between E_c and the service ranges that affect the way in which E-D space mapping and aggregation may be carried out. The implications for each of these relationships on the assessment process is discussed below.

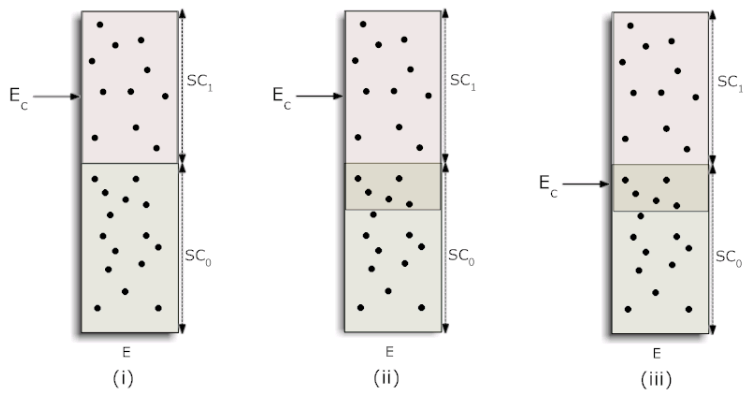


Figure 4.14: Relationships between E_c and Service Classes

In the simplest case the ranges defined in e-space are non-overlapping (Figure 4.14 (i)). Thus, the assumptions defined above regarding E_c and e-space point class membership hold. In this instance the final stages of assessment may be carried out as normal. Namely:

(a) Determine the potential service class into which E_c falls.

(b) Select all e-space points in the identified service class

(c) Map the selected e-space points to their d-space counterparts.

(d) Aggregate the d-space points and calculate validity

In the second potential case, there is an overlap between the ranges but E_c does not fall inside the overlap. In this case, it is possible to identity which service class E_c belongs to though the membership of some of the points in the range may be incorrect. In such cases, the effect of including the points from a second service class is two-fold. Firstly, the points used in the final aggregation will affect the point distribution used in determining the uncertainty in the data. Since these points will be outliers they will affect the validity of the prediction since cohesion and separation of the clusters will be decreased. Secondly, when determining the line of best fit through the data in order to make a projection the line will be biased towards the outliers. This will lead to a prediction which is less accurate though the correlation coefficient will show an increase in the uncertainty of the data. The overall effect of inclusion of such points will be a decrease in the predictive accuracy of the assessment result with an associated decrease in validity.

In the third case, where there is overlap between service class ranges in e-space and E_c falls into the overlapping region a second problem arises. Namely, that since both ranges cover E_c there is no immediate means of determining which service class is correct. In such a case, the assessment process may rely on the measure of uncertainty used in aggregation. By carrying out multiple aggregations, once for each service class in an overlapping region, it is possible to determine the prediction which is most likely to be correct based on the validity measure.

4.6.4 Hierarchical Clustering for Hidden Service Classes - Approach (d)

This section describes an assessment approach which employs hierarchical clustering to determine clusters in d-space where k is unknown a priori. At a high level the role of each stage of the quality assessment approach is identical to that within the k-means clustering approach described in Section 4.6.1. The difference occurs within the initial d-space selection stage where the hierarchical clustering approach is used to determine the optimal number of clusters (and hence the number of service classes) from the d-space data. Thus, the discussion in this section is limited to describing the implementation of the divisive clustering algorithm in terms of addressing the issues identified in the previous section when applied within the context of clustering d-space data.

This section begins with a description of a divisive hierarchical clustering method which uses an heuristic partitional algorithm to reduce the computational overhead inherent in the approach. Justification of the appropriateness of the method is provided by considering the characteristics of the data which is the subject of the analysis. Next, a method to establish the optimal number of clusters by computing the turning point of a chart series of k against an objective function is described. Finally, a description of the complete divisive clustering algorithm along with a pseudo-code illustration are provided.

Hierarchical clustering [101] is an approach to cluster analysis which presupposes very limited a priori knowledge and is suitable for situations where k (the number of actual clusters or optimal number of clusters) is unknown. The hierarchical clustering process involves producing a nested set of clusters with a single cluster at the top and single point clusters at the bottom. The value of k may then be inferred by examining the results of clustering at each level. Hierarchical clustering algorithms are generally more computationally complex than their partitional counterparts [151] (e.g. K-means). However, they tend to produce higher quality clusters and are less prone to the initial centroid selection problem which affects the k-means approach. By using a hierarchical clustering algorithm to cluster d-space points during the initial stages an assessment approach may be developed which is capable of assessing services where k is a priori unknown whilst maintaining high degrees of predictive accuracy and confidence.

One common application of hierarchical clustering is to classify data points into the number of classes which produce a model that best fits the underlying data in terms of an objective function. This is achieved through the iterative combination or division of clusters, a process which terminates when either a threshold objective value is met, when the fitness of models produced through further combination is insignificant, or when no further clustering is possible. The hierarchy of clusters formed by the clustering process consists of a single all-inclusive cluster at the root to n singleton clusters at the bottom [71] each containing a single point and is commonly recorded as a tree-like structure called a dendrogram [151] in which each level corresponds to a particular set of clusters over the data (see Figure 4.16). Sibling clusters in the tree partition the points covered by their common parent [8].

Two distinct approaches may be applied in achieving a hierarchical clustering of data. Taking an *agglomerative* approach [163] the algorithm begins by creating n clusters each containing a single data point (Figure 4.16 (a)). Clusters are combined iteratively based on their similarity until a single cluster containing all data points is formed. In contrast when taking a *divisive* approach [] the algorithm begins with a single cluster containing all data points then iteratively splits clusters

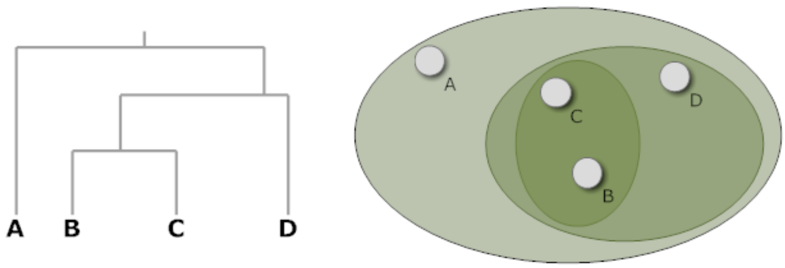


Figure 4.15: Representations of Hierarchical Clustering Results (Adapted from [151])

until n singleton clusters are formed (Figure 4.16 (b)).

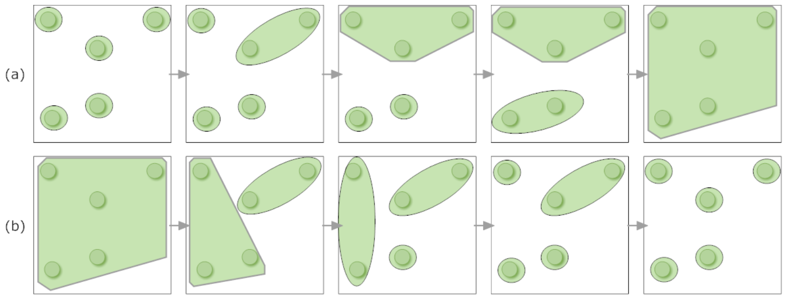


Figure 4.16: Agglomerative and Divisive Hierarchical Clustering

The resulting cluster models created at each iteration of the hierarchical clustering process are likely to differ depending on whether an agglomerative or divisive approach is adopted. The choice between a divisive and agglomerative approach strongly influences the quality of clustering achieved by the algorithm the most appropriate method of clustering being dependent upon the specific clustering task and the characteristics of the model which generated the data.

Compared to *partitional* algorithms (such as k-means) hierarchical clustering algorithms are relatively computationally expensive. In the worst cases agglomerative clustering has time complexity of $O(m^2 logm)$ [151]. Divisive methods are slower due to the significant computational overhead involved in determining how to best split large clusters in the early stages and have time complexity in the worst case of $O(2^n)$ making the problem of optimising the objective function using such a division NP-complete. Most divisive approaches employ heuristic techniques during the splitting phase reducing the average case computational complexity of the approach. However, in most instances divisive hierarchical approaches are not practical unless the number of divisions required can be restricted [43]. In practice agglomerative approaches to hierarchical clustering are most commonly used since the merging process is conceptually simpler and generally less computationally complex than divisive methods. However, divisive hierarchical clustering approaches tend to give a better model of the data at higher levels of the hierarchy since they benefit from complete information about the global distribution when making top-level partitioning decisions [85]. That is, if k is sufficiently low compared to the number of points then divisive hierarchical clustering is likely to yield higher quality clusters than by using agglomerative methods. Divisive clustering is less efficient than agglomerative clustering only if a complete tree is generated but is more efficient than agglomerative clustering if a complete hierarchy is not generated [85].

Unsupervised, a hierarchical clustering algorithm may iterate until no more clusters may be created. However, neither a model of a single cluster containing n data points or which consisted of n singleton clusters would give a good model of the underlying data. Specifying the criteria for use in identifying the "best" number of clusters is therefore a key consideration in any application of hierarchical clustering. An alternative perspective on this problem is to consider at what point the hierarchical clustering algorithm should no longer divide or combine clusters. In order to address this issue after each iteration the fitness of the model to the underlying data may be evaluated using an objective function which incorporates known desirable characteristics of the data. The clustering process might then terminate when a particular threshold objective value is exceeded or when the change in fitness from consecutive iterations is insignificant. Figure 4.17 illustrates an agglomerative hierarchical clustering approach to a data set generated from three actual data classes along with associated fitness scores for each iteration. The fitness objective function used in this example is based upon a measure of the "average normality" of the data distribution within the identified clusters.

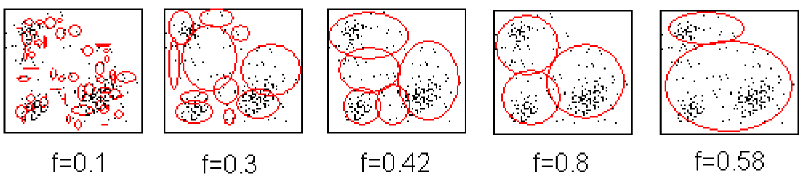


Figure 4.17: Objective Values of Iterative Hierarchical Clustering Models

4.6.4.1 Divisive Hierarchical Clustering of D-Space Points

In considering the most appropriate means of implementing a hierarchical clustering algorithm for the purposes of clustering d-space points it is useful to consider the characteristics of the model which led to the data generation and those required in the assessment process itself. The data model from which the d-space points are generated can be considered as the provider's policy with respect to the service classes and target levels of service within each class (see Section 3.4). It has been established that in the short term d-space data within each of theses service classes would approximate a normal distribution. In addition, a useful property of such policies is that the number of service classes is likely to be very small compared to the number of points generated. That is, the number of clusters is likely to be very small compared to the number of d-space points subject to clustering. In terms of required assessment characteristics the approach should be able to determine the number of service classes k, at runtime. In doing so, the quality of the identified potential service classes and hence the predictive accuracy and confidence of assessment results should be increased.

Divisive approaches to hierarchical clustering are computationally more complex than their agglomerative counterparts due to the intensive work required in determining a good set of subclusters to form from a parent. However, they are attractive for the purposes of quality assessment since at a high level in the dendrogram they produce clusters which are of high quality. This is significant as it is likely that the number of service classes will be small and hence the clusters of interest are likely to be close to the root of the generated dendrogram. Furthermore, since the goal of using hierarchical clustering is to be able to establish k, it is likely that the number of required splits of clusters will be very small. As such, the computational expense is likely to be far less than the worst case.

However, even if only a few cluster divisions are required finding an optimal split of clusters is computationally unfeasible, being of order $O(2^n)$. In order to make the solution more tractable therefore, a suitable heuristic partitional subroutine [85] must be applied. It has been observed that in order to reduce the complexity the repeated application of a fast partitional scheme can be used [10]. In employing a partitional subroutine the overall operation of the hierarchical clustering algorithm can be described as illustrated in Algorithm 5.

Algorithm 5: Divisive Hierarchical Clustering Algorithm - adapted from [154]

```
Data: D = {x_1, ..., x_n}
Result: C = {c_1, .., c_k}
begin
    while termination condition not met do
        Identify the best cluster to split
        Partition the cluster using the partitioning subroutine
        Evaluate the termination condition
    end
    Determine the optimal number of clusters;
end
```

The best cluster to split is the one which is least likely to correspond to data from a single actual service class. In determining this, the cluster with the highest sum of squares error (SSE) is selected. SSE is an appropriate measure as it can be shown that for clusters with Gaussian distribution, SSE when used in combination with Euclidean measures of distance are a good optimisation criteria. Furthermore, SSE can be easily calculated using the centroids from the k-means algorithm in linear time.

The partitioning subroutine may be enacted by using a partitional algorithm such as k-means or via some other measures of distance between the points in the cluster [8]). A generally-applicable method of divisive hierarchical clustering is the bisecting k-means [146] algorithm which has a time complexity that is linear in the number of items. In this approach k-means is used to bisect the identified cluster (a k-means clustering with k=2). As such, the bisecting k-means is adopted for use as the partitioning subroutine for the divisive hierarchical clustering algorithm described in this chapter.

The *BisectingPartition* algorithm (Algorithm 6, below) provides details of the operation of the partition subroutine which is used by the high level hierarchical clustering process described in Algorithm 5. The algorithm is identical to the k-means algorithm defined in Algorithm 4 with the exception that since the algorithm is carrying out a bisection, k is fixed as 2.

Algorithm 6: BisectingPartition Algorithm - adapted from [146]

Data: $D = \{x_1, ..., x_n\}$
Result: $C = \{c_1, c_2\}$
begin
    `/* remove outliers and select two initial centroids using previous algorithms */`
    RemoveOutliers(D);
    M = SelectCentroids(D,2);
    `/* repeat until points "settle" into final clusters */`
    while *Points move between clusters* **do**
        `/* for each point in the data, reassign to cluster with nearest mean */`
        for $x_i \in D$ **do**
            Reassign x_i to {c1, c2} such that $\mid x_i - \mu_i \mid$ is minimised;
        end
        `/* following reallocation of points between clusters recalculate the means of each */`
        $\mu_1 = \frac{\sum_{x \in c_1}}{|c_1|}$;
        $\mu_2 = \frac{\sum_{x \in c_2}}{|c_2|}$;
    end
end

The time complexity of the overall hierarchical clustering algorithm is $O(NlogN)$, assuming that the dendrogram is well balanced [10]. Such an increase in complexity over the standard k-means is acceptable given the likely increase in speed.

4.6.4.2 Identifying the Optimal Number of Clusters

A divisive hierarchical clustering algorithm will continue to split clusters until a terminating condition is met. These conditions include reaching a specified number of clusters (k), reaching a level of singleton clusters such that no further division is possible, or exceeding the threshold of some objective function. Of these the first can immediately be discounted since the purpose of applying a hierarchical clustering algorithm in this case is to determine k from the clustering outcome. Allowing the hierarchical clustering algorithm to continue until a complete dendrogram is formed is also unfeasible as the choice for a termination criteria since it is likely to only require the levels of a dendrogram closest to its root (as the number of service classes will generally be small compared to the total number of points used in making an assessment). The final criteria is therefore most appropriate. In this section, an appropriate objective function is identified and its implementation in determining the optimal number of clusters is described.

4.6.4.3 An Objective Function for Hierarchical Clustering of D-Space Points

Cluster validity measures have been considered in Section 4.3.1. In these cases, the value attributed was used during a post-clustering phase to determine the quality of the clusters and to infer confidence in the assessment result. As previously, the criteria for any objective function are that it should produce a value proportional to the average normality of the clusters and to the similarity of points to the cluster centroids. An additional criterion is that the function should not be computationally expensive. Since the function will be evaluated at each partitioning of the clusters it is desirable that its complexity should be minimised if the overall complexity of the hierarchical clustering algorithm is not to be affected. As previously (see Section 4.6.1) the silhouette coefficient is used here as the measure of cluster validity and as the objective function which is used to guide the hierarchical clustering algorithm in finding the number of clusters which is most likely given the data.

4.6.4.4 Hierarchical Clustering Termination Condition

Having identified an appropriate cluster validity measure a means of determining when it is at its optimal and hence the number of clusters in the data is needed. Typically, this is achieved through the calculation of the objective function value for each k, either after the clustering algorithm has completed or dynamically after each iteration of the partitioning subroutine (as is the case for divisive hierarchical clustering). From the resulting objectivity graph of the number of clusters (k) against the value of the objective function (the objectivity) it can be observed that objectivity increases up until a particular k at which point it flattens [157], peaks and begins to decrease [151]. In order to identify the correct k therefore, a method of detecting significant peaks in the objectivity graph is required.

A variety of methods exist to achieve this goal. The approach developed by Chiu et al [22] produces an estimate for the number of clusters by identifying the point in the graph where a big decrease in objectivity is seen, since this implies that two clusters have been merged which should remain separate [22]. Similarly, for a divisive approach it indicates that a cluster which should have remained whole has been bisected. In order to recognise when the correct number of clusters has been found an adapted version of the method is used. Whilst being basic it is likely that the approach should be sufficient to detect the first graph turning point to identify the correct number of clusters. This confidence being due to the observation that clusters in d-space will be generally well-formed and isotropic. The impact of not detecting the correct k at the point in the iteration of hierarchical clustering where it occurs is that the algorithm might potentially proceed until the singleton-cluster stage is reached. As described above, this is undesirable and is computationally complex. Thus, in order to ensure that this does not happen a termination criteria is specified for the hierarchical clustering algorithm to cease following a specific number of iterations. At this point the level in the tree for which the objective function was highest is chosen as the correct number of k. It may be observed that the algorithm DetectCorrectClusters (Algorithm 7) algorithm requires the objective function values for the previous three iterations (O_1,O_2,O_3) since it determines when k has been reached by comparing the three points (line 1). Thus, the algorithm is unable to detect the case where k=1. In this case, the fact that k=1 will not be identified until the maximum number of iterations is reached and the objective function evaluated for each k. As discussed previously, from the perspective of increasing accuracy of predictions, choosing to bisect a cluster that refers to an actual service class will probably have little impact on accuracy since the points within each child will still refer to points from a single actual service class. As such, being able to detect the case where k=1 is ideal for completeness but realistically the performance of the algorithm will not be affected.

Algorithm 7: DetectCorrectClusters Algorithm - adapted from [22]

```
Data: O1, O2, O3
Result: B
/* return true if O2 is a peak (i.e.  greater than the first
   point and the second point)                                  */
if O1 < O2 and O2 > O3 then
    B = true;
end
```

4.6.4.5 Hierarchical Clustering Approach Summary

The hierarchical clustering algorithm presented in this chapter may be used within an assessment approach for detecting points which belong to service classes, where the number of service classes is unknown a priori. The algorithm utilises a divisive k-means algorithm based on the observation that the number of service classes (k) is likely to be small compared to the number of points. Using bisecting k-means as a partitioning subroutine, at each stage the cluster with the highest SSE is split. The optimal number of clusters is identified by evaluating the objective function (the silhouette coefficient) after each bisection and then detecting the first turn in the graph of k against objectivity.

Based on the assumption that the cluster tree is well-balanced the average case complexity of the divisive hierarchical clustering algorithm is O(N log N). This is more complex than the k-means algorithm described in Section 4.6.1 though significantly less complex than the equivalent agglomerative algorithm. The improvement in performance is further enhanced by the fact that the hierarchical clustering algorithm will only complete a small proportion of the maximum number of iterations in finding the correct value of k – since when assessing established services with large amounts of performance data available, k will be small compared to the number of points. In contrast, an agglomerative approach would not identify a case with small k until nearing completion of the entire clustering dendrogram.

In terms of effectiveness the hierarchical clustering algorithm should yield higher quality clusters than those obtained through k-means where k is not known a priori - this being due to the fact that a k-means algorithm will use a value of k that does not reflect the number of actual service classes and demonstrates the latter algorithms's reliance on knowing the number of actual service classes. Indeed, it is likely that the performance of the hierarchical algorithm may be comparable with that of k-means where k is known a priori (this conjecture is evaluated in Section 6.2.4.3). Furthermore, as discussed in Section 4.6.4.1 the use of a divisive hierarchical approach will lead to higher quality clusters at the root of the dendrogram than for an agglomerative equivalent reinforcing the justification for choosing a divisive method. Improving the quality of clusters identified in d-space for unknown k has two direct benefits to consumers. Firstly, higher accuracy predictions (with associated higher validity) will be provided thus supporting consumers in locating better services. Secondly, consumers are allowed to obtain reliable predictions for the behaviour of services where the number of service classes offered is not known a priori.

4.7 Summary

This chapter introduced a novel high-level approach to assessment, which utilises patterns between E and D Space. Such patterns are exhibited due to the behaviour of providers in differentiating the level of service provided to consumers based on their expectations. The high level approach described returns a prediction of the level of service that a consumer might expect from a service along with a validity measure. Such information fulfills the requirement set down in Section 3.6 in enabling consumers to weigh evidence in order to combine with other sources. These approaches may fulfill the role of QAS in the conformance-centric SOA defined in Chapter 3.

The high level approach described was used as the basis for defining four specific E-D Space Assessment approaches. These specific approaches differ in terms of the degree to which they actively recognise the characteristics of service classes at runtime. Pattern passive approaches which assume a majority of the characteristics at design time, rely on uncertainty as an indicator of the degree of quality of the prediction made. Although naive in some respects pattern passive E-D Space assessment approaches should still prove more useful in supporting a consumer's decisions than existing D-Space approaches, which assume that no patterns exist in the data.

The second set of approaches defined seek to actively determine the characteristics of service classes at runtime by analysing patterns in d-space data. By actively recognising the boundaries of service classes the predictive accuracy of assessment results should be significantly increased over pattern passive approaches. This in turn will lead to consumers receiving higher quality from services.

Chapter 5

Implementations of a Conformance-Centric SOA

In order to illustrate the contribution of the approaches to quality assessment described in the previous chapter, it is necessary to evaluate them in terms of the degree to which they improve the quality of service experienced by consumers. This improvement is gained through the provision of high accuracy, high confidence predictions by a QAS which leads to an increase in the effectiveness of a consumer's decision making process.

The approaches developed in Chapter 4 make assumptions about the behaviour of consumers and providers in an SOA in relation to QoS. Importantly the focus of all approaches is on acknowledging the existence of service classes which lead to the segmentation of consumers and to the emergence of patterns within historic performance data. Indeed, the major distinction between the four approaches described is the way in which they detect or otherwise acknowledge these patterns.

In terms of consumer behaviour a QAS assumes that a consumer is able to express their ideals in terms of QoS attributes and that a consumer will perceive the highest QoS when their ideals are being met. Furthermore, a QAS assumes that providers will consider customers in this way and will adjust their policies accordingly. The model of a conformance-centric SOA developed in Chapter 3 embodies the behavior above and describes further details in terms of how entities will behave and interact under such assumptions. As such, an implementation of this conformance-centric SOA should be used in order to evaluate the approaches to quality assessment described in Chapter 4.

The implementation of a conformance-centric SOA for the evaluation of quality assessment approaches across a live network-based environment is unsuitable for a number of reasons, namely:

- **Limited Control of Environment:** it is difficult to control the behaviour of aspects of the live environment (such as network speeds) to the degree that would be required to support repeatability of experiments.
- **Effort Required for Adaption to a Conformance-Centric SOA:** existing families of standards formalise the interfaces and protocols by which agents within SOA are able to exchange information. However, actual implementations of entities and models of their behaviour are not typically provided. As such, there is considerable effort involved in adapting such environments, particularly the provision of full implementations of consumers, providers, services, assessment and performance monitoring services.
- **Monitoring of Evaluation Variables:** It is possible to monitor the performance of services delivered to consumers across a live environment (e.g. the delivered level of service in terms of QoS attributes). However, in order to usefully evaluate quality assessment approaches it is necessary to be able to evaluate such variables as: the level of quality experienced by users, the predictive accuracy of approaches and the difference between the level of services provided by a service chosen by the consumer compared to the other candidate services which were rejected. Acquisition and coordination of such monitoring information would be difficult in a live environment.

In order to overcome the limitations inherent in a live empirical evaluation of assessment approaches, a software application which is a simulated implementation of the conformance-centric SOA described in Chapter 3 has been developed. This chapter describes two implementations of a conformance-centric SOA which are used to demonstrate the quality assessment approaches developed in this thesis. The first of these - an agent-based architecture for Virtual Organisation (VO) formation in grid environments - is a physical implementation of a conformance-centric SOA in which the quality assessment approach is demonstrated to provide a crucial role in the formation and successful operation of VOs. The second - a conformance-centric SOA implemented as a simulation application - may be used to define and conduct experiments through the fine-grained initialisation, control and monitoring of scenario parameters. The application described is used in the next chapter to conduct empirical evaluation of the approaches to quality assessment developed in this thesis.

5.1 QoS Assessment for Virtual Organisations

Virtual organisations (VO) are formal, yet dynamic and generally temporary collaborations between service providers that emerge in order to meet requirements which could not be met by the services of individual providers. They are formed to exploit a market opportunity by combining a subset of resources and so spread the risk inherent in a venture. VOs are a key enabler in the development of grid computing based infrastructures which rely on the flexible and dynamic interaction of diverse services across a massive service oriented environment. Whilst the activities required to form and sustain such structures are complex, the structure and operation of VOs are hidden from the consumers of their services who are unaware whether the services are being provided by a single provider or by a VO.

Quality assessment plays a necessary and key role in the formation and maintenance of VOs by providing a means to assess the performance of potential and active VO services. The prototype software developed in the CONOISE-G project is a real world implementation of the principles and components of the conformance-centric architecture presented in Section 3.2. Significantly, the architecture supports the capture of consumer expectations and provides quality assessment and quality monitoring components that fulfill the roles previously described. The software was developed and demonstrated at a number of UK and International e-Science conferences (see for example [108]) and included an implementation of early versions of the assessment approaches defined in the previous chapter (details of the specific approach implemented in the final version of the CONOISE-G demonstrator are described in [135]).

In this section the CONOISE-G project and software is described to provide a practical demonstration of the approaches to quality assessment developed in the previous chapter. Whilst the specific approaches defined in the previous chapter were not implemented in the software itself, the example provided is explained in terms of the hierarchical clustering-based approach (Approach (d), Section 4.6.4). This serves to demonstrate the contribution and effectiveness of the approach in a realistic VO management scenario.

5.1.1 CONOISE-G Background

The CONOISE-G project sought to develop research and techniques to support next generation grid infrastructures through the exploration of research areas relevant to the formation, operation and perturbation of VOs. Specifically, the CONOISE-G project focussed on three major research components:

(a) **Trust**: Developing techniques for modeling, using and maintaining trust in virtual organisations where partial knowledge is pervasive.

(b) **Policing**: Developing theory and mechanisms for defining and enforcing social laws in virtual organisations including the policing of stakeholder behaviour.

(c) **Quality**: Developing techniques for representing and processing quality policies that support accreditation of service-providers and maintenance of relationships between service-providers and consumers.

In supporting the creation and operation of VOs the role of quality assessment is specifically relevant at three stages of a VO lifecycle:

(a) **VO Formation**: During the formation of VOs providers will attempt to locate other providers with whom to partner. Being able to assess the capabilities of such providers in terms of their likely quality of service is essential to provide assurance prior to collaboration.

(b) **VO Discovery**: When discovering services it is desirable that consumers be able to determine the likely quality of service in order to make an effective selection of the best service to use.

(c) **VO Reformation (Perturbation)**: VOs are dynamic in that the membership may change with time. This may be due to changing consumer requirements, poor performance of a provider leading to expulsion, or a provider intentionally withdrawing their resources. In each of these perturbations the VO must be able to reform by identifying other services to meet the shortfall.

As part of the CONOISE-G project an architecture for supporting the dynamic and autonomic operation of VOs was implemented. This architecture is described and an example presented which demonstrates the role of the quality assessment approaches developed in this thesis in the context of VO formation and reformation.

5.1.2 CONOISE-G Architecture

The architecture developed in the CONOISE-G project (see Figure 5.1) uses software agents to represent and act on behalf of consumers and providers of services. Each agent is responsible for the management and allocation of a set of resources with which it attempts to achieve the goals of its owner. Within the architecture a set of supporting third party agents act as information sources regarding the performance of services in the environment. Agents are endowed with the capability to communicate and negotiate with their peers and to reason and act upon information which they obtain from third parties in the process of discovering and selecting services. Agents within the CONOISE-G architecture have one of a number of roles, namely:

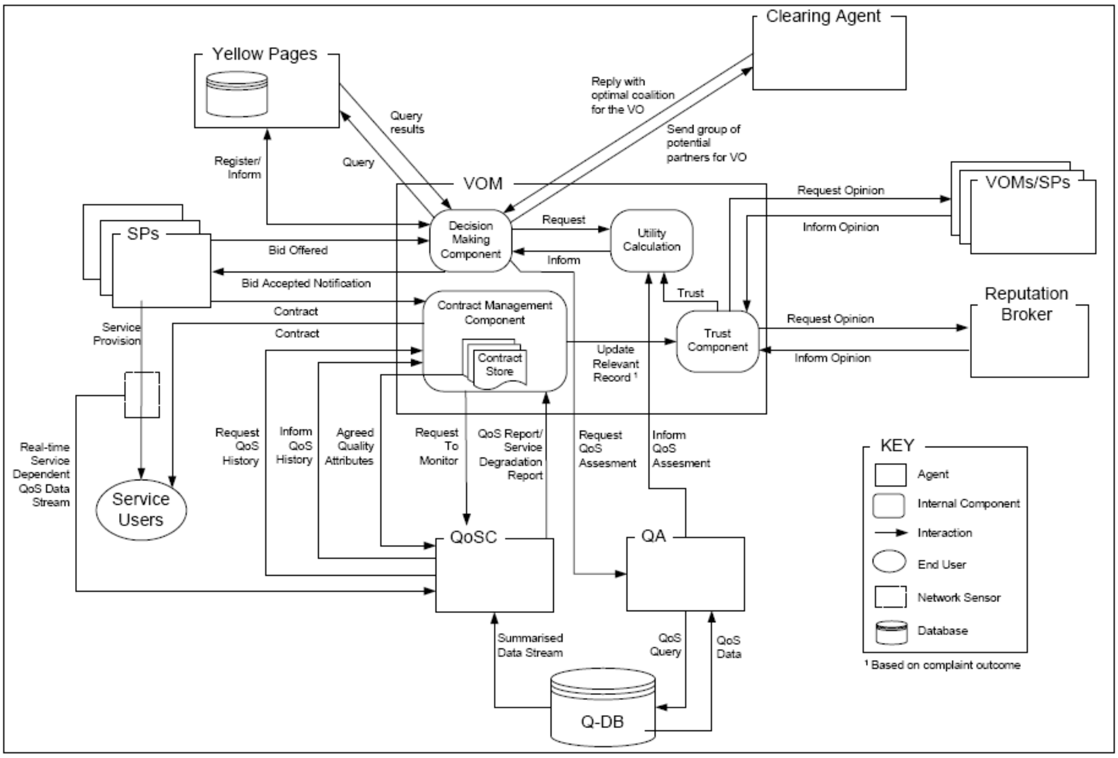


Figure 5.1: CONOISE-G System Architecture

- **Yellow Pages (YP):** A YP agent provides a directory service which allows services to be advertised by providers and discovered by consumers. Each advertisement provides details of a service's stated functional and non-functional capabilities.

- **Quality Assessment (QA):** A QA agent uses a set of historic performance information about services to predict the likely future behaviour of the service. This agent may be used by other agents within the environment as a source of evidence during the process of selecting a service which best meets their needs.

- **QoS Consultant (QoSC):** A QoSC agent collects performance information during the provision of services to consumers. This information a VO's supports performance and can be used to provide evidence in contract disputes and to provide information to support the QA process. This summarised performance information is stored within a Quality Database (QDB) which is accessible by the QA agent.

- **Reputation Broker (RB):** A RB agent maintains information regarding the history of service providers in successful contract fulfilment. An RB may be used by agents during service selection.

- **Service User (SU):** A SU agent represents an end consumer of services. It is aware of a consumer's functional and non-functional requirements and can negotiate on the consumer's behalf with the VOM and SP agents.
- **Service Provider (SP):** A SP agent provides services within an environment and can be offering these through a VO or as an individual agent. SPs are capable of negotiating with other agents who wish to use their services or combine them with their own to form a VO. Negotiation is enabled by the SP providing a "bid" to another agent to supply a particular service at a specified cost and with specified non-functional characteristics.
- **Clearing Agent (CA):** A CA agent is able to determine an optimal set of service providers to form a virtual organisation based on the bid information obtained from each provider.
- **VO Manager (VOM):** A VOM agent is responsible for the formation and maintenance of a virtual organisation. The agent negotiates directly with SP agents in order to gather "bids" for VO membership. Using information obtained from QA and Reputation Broker agents the VOM uses an internal utility component and external CA in order to decide upon the best set of services to use. During the operation of a VO the VOM is responsible for acting upon reported violations of contract by the QoSC and if necessary reforming the VO to ensure that its onward contracts are maintained.

These components interact to enable the dynamic formation, operation and reformation of VOs.

5.1.3 The Role of Quality Assessment in VO Formation and Perturbation

The formation of a VO is likely to be initiated when a consumer attempts to discover services to meet their requirements. The stages of this process and their relevance to the quality and QA aspects of the VO are as follows:

(a) **Requirements Specification:** A consumer specifies their functional and non-functional requirements to an SU agent. These requirements include a list of quality attributes for the services.

(b) **Service Discovery:** The agent uses a YP agent to obtain a list of all SPs who claim to be able to satisfy the consumer's requirements.

(c) **Service Negotiation:** The SU contacts each of the service providers individually and receives a set of bids from each. This results in a reduced set of service providers who are able to meet the consumer's terms.

(d) **Service Assessment:** The SU contacts third party agents to obtain information about each of the service providers. A QA is consulted which uses the requirements to determine a prediction of the likely behaviour of each service in the future.

(e) **Service Selection:** The SU selects the service which best meets the consumer's requirements.

The service providers on the list returned at Stage 2 may be SP or VOM agents. When an agent is contacted to negotiate in Stage 3 it may decide to form a VO to meet the consumer's requirements. The process of forming a VO is similar to that of service selection by a consumer, in that potential partners are discovered and assessed using the YP and QA agents. Additionally, following identification of the set of potential partners, the clearing agent is consulted in order to determine the optimal subset of providers based on resources and prices offered by each. Assuming that the consumer's requirements can be met and that service selection results in a VO being created by a VOM and chosen by the consumer the VO is formed and begins to operate, providing its component services to the consumer.

In collaboration with the QoSC the performance of the VO members is monitored to ensure that the consumer requirements (and the VOM-derived requirements) are being met. If a partner of the VO begins to fail or intentionally leaves the VOM may decide to reform the VO to replace the member concerned, enacting a process similar to the formation process but concerned with a single service provider only.

At each of the assessment stages described in the processes above the QA is responsible for producing predictions of the likely behaviour of services given the stated requirements along with an associated confidence score. Using a Quality Assessment agent to determine the performance of services in each of the cases above, has benefits for multiple stakeholders in the environment since:

- **Service Consumers:** are more likely to receive services from providers or VOs which offer them the highest quality of service available.

- **VO Managers:** as VO creators are likely to choose better, more reliable and higher quality services. This means that the VOs which they create are less likely to fail and may avoid contract violation and associated penalties.

- **Service Providers:** as VO members will profit from being members of VOs that are successful and long-lived. Quality Assessment provides an incentive and reward for offering high quality services through a VO or as an individual.

Service Type	**Customer Requirements**
TextUpdateService	4 updates/day, 100% availability
StreamAudioService	600mins/month, 40kbit/s bitrate, 22KHz sample rate
StreamVideoService	200mins/month, 100kbit/s bitrate, 6ms latency

Table 5.1: Mobile Multimedia Scenario - User Requirements

5.1.4 Assessing Service Providers with Private Service Classes in the CONOISE-G Architecture

This scenario considers a user of a mobile device who seeks to acquire a set of multimedia services as a contract package. The user's requirements are stated in terms of three services: a text updates service, a streaming audio service and a streaming video service; for which the user has stated functional and quality requirements. These requirements are summarised in Table 5.1. There are currently five service providers in the environment (SP_1 to SP_5) offering the services defined in Table 5.2. These services have previously been registered with the YP agent.

Type	SP_1	SP_2	SP_3	SP_4	SP_5
TextUpdateService	X		X		
StreamAudioService		X		X	X
StreamVideoService	X		X		

Table 5.2: Mobile Multimedia Scenario - List of Services

The user states the requirements to an SU which discovers a set of candidate SPs that claim to be able to meet the requirements.

5.1.4.1 VO Formation

A particular service provider, SP_3 decides that it would be feasible to form a VO to meet the requirements and begins the formation process. SP_3 is able to provide the TextUpdateService and StreamVideoService portions of the required package and decides to form a VO by engaging the services of other providers in the environment to provide the StreamAudioService. In this process SP_3 assumes the role of VO Manager (VOM).

By consulting the YP the VOM discovers three service providers (SP_2, SP_4 and SP_5) who state that they are capable of providing a StreamAudioService.

5.1.4.2 Quality Assessment

The VOM then contacts the QA agent, to assess each service in terms of the degree to which they are likely to meet the consumer's requirements. Figure 5.2 illustrates the actual service behaviour of these SPs over time, the historic performance information recorded in the QDB and the e and d components of the QDB data are illustrated in q-space notation for each of the providers. From

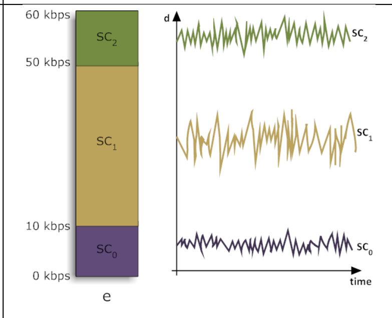
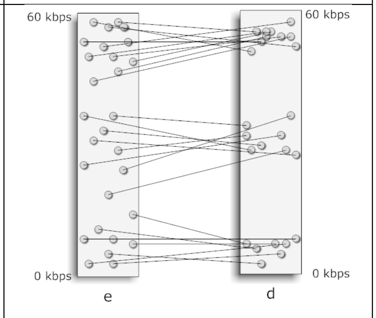
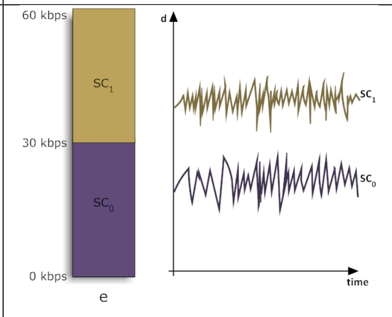
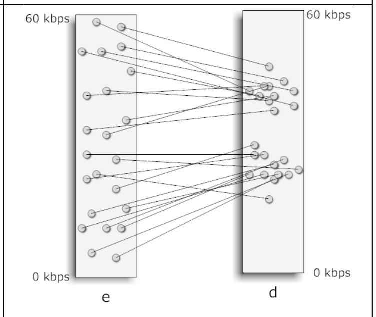
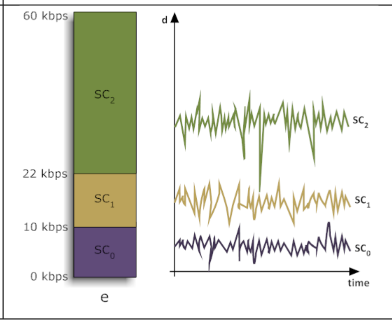
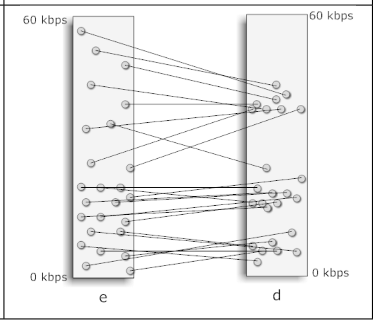
Service Provider	Behaviour	QDB Contents
SP_2		
SP_4		
SP_5		

Figure 5.2: Mobile Multimedia Scenario - Service Provider Behaviour

Figure 5.2, which shows the service behaviour, it can be seen that:

- SP_2 offers three service classes: SC_0 for consumers with expectation of 0kbps to 10kbps with an average delivered level of service of around 7kbps, SC_1 to consumers with expectation of 10kbps to 50kbps with an average delivered level of service of around 35kbps and SC_2 to consumers with expectation of 50kbps to 60kkbps with an average delivered level of service of around 55kbps.

- SP_4 offers two service classes: SC_0 for consumers with expectation of 0kbps to 30kbps with an average delivered level of service of around 25kbps and SC_1 to consumers with expectation of 30kbps to 60kbps with an average delivered level of service of around 40kbps.

- SP_5 offers three service classes: SC_0 for consumers with expectation of 0kbps to 10kbps with an average delivered level of service of around 7kbps, SC_1 to consumers with expectation of 10kbps to 22kbps with an average delivered level of service of around 15kbps and SC_2 to consumers with expectation of 22kbps to 60kkbps with an average delivered level of service of around 40kbps.

It is important to note that the service class definitions described are provided for the purposes of this example. These are hidden from the QA agent which only has access to the contents of the QDB when assessing each service.

Hierarchical Clustering The QA agent assesses each of the providers in turn using the hierarchical clustering approach described in Section 4.6.4. Figure 5.3 illustrates the initial clustering of d-space points for SP_2. It shows the first four iterations of the divisive hierarchical clustering algorithm in which the bisecting k-means algorithm is applied iteratively to the lowest validity cluster. Following

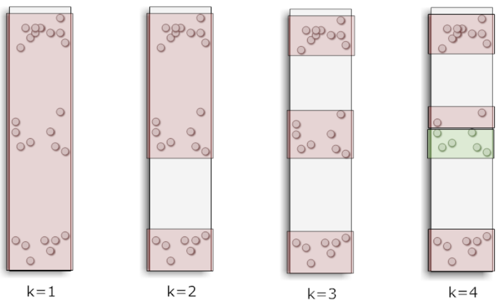


Figure 5.3: Mobile Multimedia Scenario - Hierarchical Clustering of D-Space Points

the fourth iteration the overall validity of the clustering decreases significantly following the incorrect bisection of the middle cluster. At this point the DetectCorrectClusters algorithm identifies the knee in the curve of cluster validity and correctly identifies that the peak of the curve occurred for k=3.

Mapping and Aggregation Process The overall assessment process, which follows from the initial clustering of d-space points is illustrated in Figure 5.4. Firstly, points are clustered in d-space (Stage 1). This is followed by a mapping to e-space to identify the range in which the consumer's requirement falls (Stage 2). Next, a mapping from e-space to d-space is carried out (Stage 3). Finally, the relevant points in d-space are aggregated (Stage 4) leading to the production of a prediction and confidence score.

5.1.4.3 Partner Selection

Having assessed all providers the VOM sends all "bid" information to the Clearing Agent to determine the optimal set of providers to meet the consumer require-

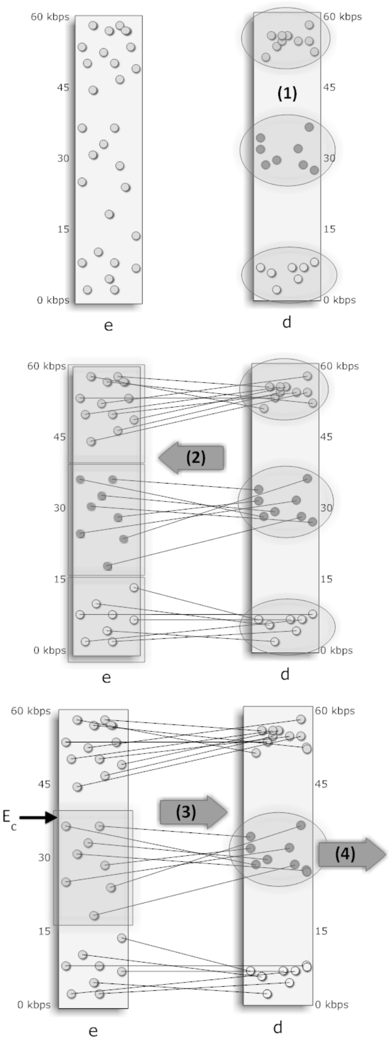


Figure 5.4: Mobile Multimedia Scenario - VO Formation Assessment Process

ments. The CA decides that in meeting the StreamAudioService component requirement, the full 600 mins/month should be provided by SP_2. Having formed a VO based on the set of providers chosen by the CA the VOM responds to the user agent with an offer to provide the services, which is accepted.

5.1.4.4 VO Perturbation

Following VO formation, the consumer begins to use the component services of the VO. This usage is continuously monitored by the QoSC and VOM. After a period of time it becomes evident that the bitrate of SP_2's StreamAudioService is deteriorating towards a level which risks violating the requirements of the consumer. At this point the VOM begins the process of reformation to replace the failing provider. In this case, the VOM is able to predict the likely behaviour of SP_2 and this means this provider is excluded from the assessment process. Consulting the YP the VOM determines that SP_4 and SP_5 are still able to provide StreamAudioServices and the QA is invoked to assess these services. Figure 5.5 illustrates the assessment process for the optimal service provider, SP_5.

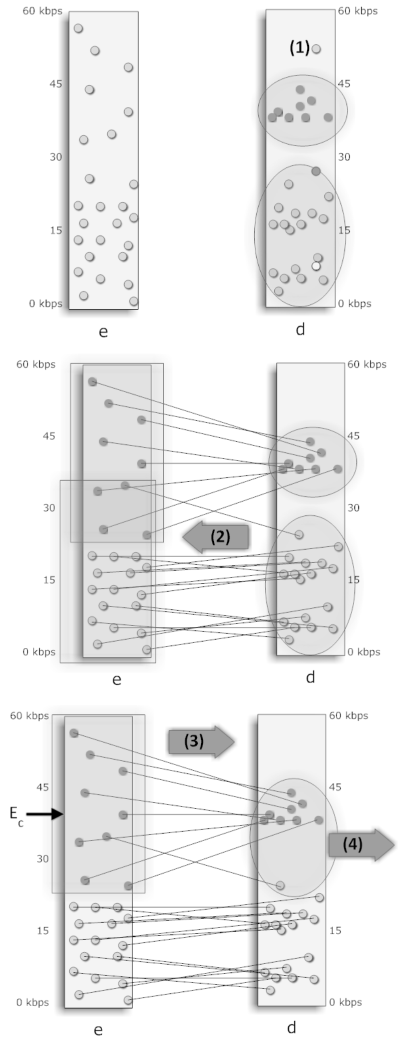


Figure 5.5: Mobile Multimedia Scenario - VO Perturbation Assessment Process

The assessment continues as for the initial assessment for SP_2. However, in this case, an outlier exists in the d-space data for SP_4, with the effect that the clustering algorithm assigns it to an incorrect service class (Stage 1). Following the mapping of the d-space clusters to e-space the bounds of the two service classes identified are overlapping with the expectation bounds of the service class to the lower end of the e-space range covering points that are from the second actual service class. However, since the expectation of the consumer (40kbps) is within the upper potential service class range in e-space the final e-d space mapping (Stage 3) and aggregation (Stage 4) are not affected. Following assessment the VOM compares the prediction for SP_4 and SP_5 with its own prediction for SP_2 gained from monitoring and decides to replace the StreamAudioService currently provided by SP_2 with that offered by SP_4.

5.2 Simulation Application

Evaluation of the effectiveness of assessment approaches requires fine grained control over many aspects of an SOA. Parameters which must be controlled include the behaviour of services, SLA creation and management, service monitoring, and the decision models of consumers used to services. Furthermore, to compare the performance of approaches it is necessary to accurately monitor multiple aspects of the environment e.g. changes in the quality received by consumers, the behaviour of services and detailed control of aspects affecting the characteristics of the data used in and produced by assessment approaches (such as noise). The experiments must also be repeatable.

In supporting the above requirements, a simulation application has been constructed as an implementation of the conformance-centric SOA model developed in Chapter 3. The application allows fine-grained control over the aspects identified above as well as providing tools for the testing and evaluation of quality assessment approaches. The entire application including simulation, model and visualisation aspects was developed as part of the work represented by this thesis.

The application is flexible and extensible in its design and implementation and so could be applied to the evaluation of many current or future approaches to assessment, which share the same evaluation criteria and have a compatible architecture such as the SOA described in Chapter 3 or a more basic SOA. This includes existing non-conformance centric SOAs. In order to demonstrate the suitability of the application as a general means of evaluating quality assessment approaches and hence to justify its use in evaluating the approaches presented in this thesis this section describes its architecture, underlying data model and behavioural aspects.

5.2.1 Architectural Overview

The simulation application consists of four main components, illustrated as the four containers in Figure 5.6:

- **Environment:** the core architecture component is an implementation of the conformance-centric SOA defined in Chapter 3. Services and Consumers are defined in terms of their behaviour and infrastructure components (quality assessment and directory service) are provided to support discovery and engagement between the two. During service provision, information relating to the quality of service delivered is logged - denoted by the circle on the

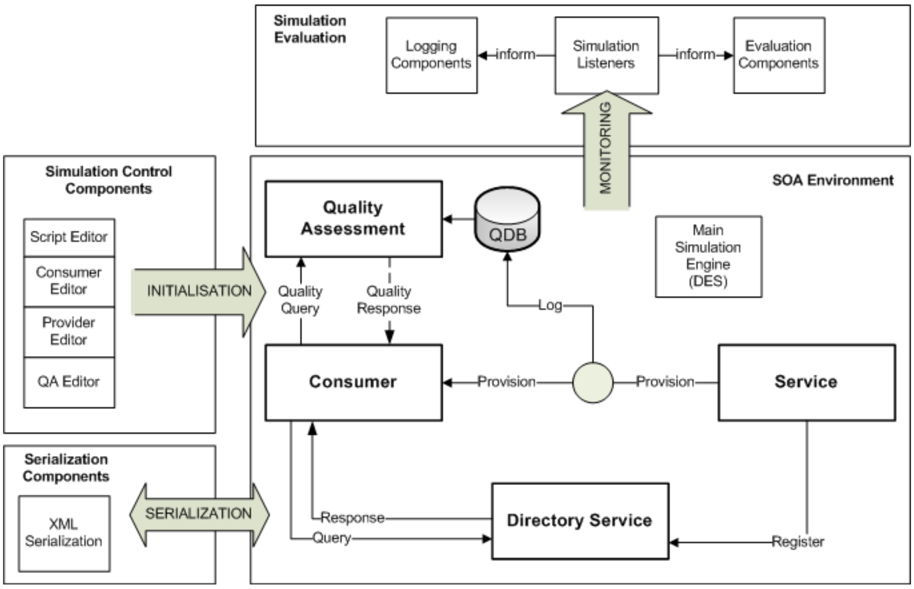


Figure 5.6: High Level Architectural Overview

provision arrow between Service and Consumer. The interaction between entities is defined and controlled by a Discrete Event Simulator (discussed in more detail in Section 5.2.3).

- **Initialisation Components:** API and GUI tools for the initialisation and control of SOA members and simulation parameters are provided. These allow the specfication of providers, consumers and assessment approaches, in terms of their behaviour.

- **Simulation Evaluation:** API and GUI tools for flexible monitoring and visualisation of multiple aspects of the SOA, including components for visualising quality of service experienced by users and level of service delivered.

- **Serialization:** tools to allow the saving and subsequent retrieval of simulation state to and from disk.

The entire architecture is supported by a rich API that allows customisation of many aspects of the simulation and SOA environment e.g. the development of new Quality Assessment approaches, consumer decision models and additional GUI evaluation components.

5.2.2 SOA Environment

The central component of the architecture is the SOA Environment subsystem which serves three purposes:

- **Coordination:** The environment acts as a container for SOA members, maintaining the sets of consumers, providers algorithms, events and other data that interact within the simulation.

- **Communication:** The environment provides a means of allowing the various parts of the system to discover each other and communicate (as would be provided by the messaging protocols in a real-world SOA).

- **Simulation:** The environment utilises a discrete event simulator (DES) to manage the actions of entities within the environment.

The four main elements within the environment correspond to aspects of the conformance-centric SOA defined in Section 3.2:

- **Service:** models and methods for specifying and controlling the behaviour of services in the environment. Services are able to register themselves with the directory service and respond to requests for service from consumers.

- **Consumer:** models and methods for specifying and controlling the actions of consumers in the environment. Consumers are able to consult directory services and quality assessment; and consume services. During service consumption the behaviour of the service is recorded in a QDB.

- **Directory Service:** maintains a table of all services currently active in the environment.

- **Quality Assessment** - models and methods for specifying and controlling the approaches to quality assessment available to consumers in the environment. Each QA in the environment has its own QDB containing information relating to services that it has recommended which have subsequently been used by consumers.

Multiple entities of each type may exist within the environment each with varying behaviour and scripted actions. The means by which each of the entities is initialised along with set of parameters that may be controlled for each is described in Section 5.2.4.

5.2.3 Discrete Event Simulation

At the core of the environment is a discrete event simulation (DES). Actions that have been defined for each of the SOA members are sorted in chronological order and are processed by the main event thread at the appropriate time in the simulation. A simulation clock is maintained and advanced at each iteration

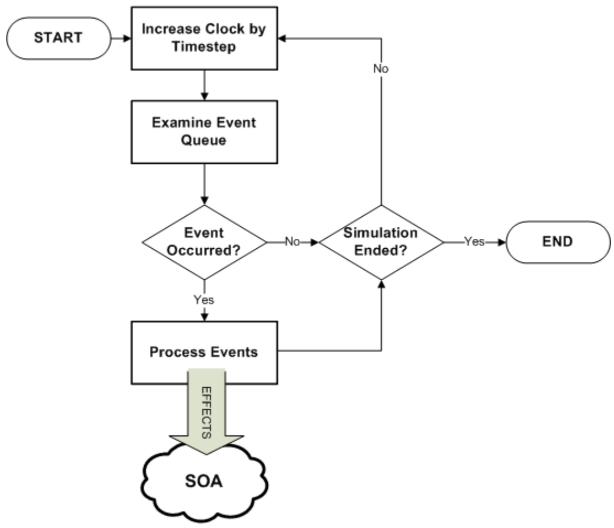


Figure 5.7: Simulation Application - DES Main Event Thread

of the loop (Figure 5.7). Three general types of events may occur within the environment:

- **Service Enter Event:** A service enters the environment and registers with the directory service.
- **Assessment Event:** A consumer requests an assessment providing an expectation and service. Quality assessment takes place and dependent upon the outcome the consumer may decide to use a service.
- **Provision Event:** A consumer uses a service. Following provision the actual behaviour of the service is recorded in the QDB along with the consumer's initial expectation and a rating of their experience.

The simulation terminates when all events have been processed or the simulation clock reaches a pre-defined end point. Although the environment supports the generation of events when a simulation is in progress, for the experiments conducted in Chapter **??** all events are defined during an experiment setup phase before the simulation is started. The distribution (and therefore frequency) of events is defined in the setup phase using the Script Initialisation user interface component which creates a series of individual events based on defined parameters using the Script Initialisation UI component (described in Section 5.2.4.4). AS such, the event queue is initially populated with a full set of events that will occur over the simulation with individual events leaving the queue when the point in time at which they have been defined as occurring is reached.

At each timestep of the simulation all events that are marked as occurring since the last timestep are processed, with the related action being taken (e.g. recording an instance of service provision, adding a service to the list held by the directory service, etc) before control is returned to the main event handling thread and the simulation timestep is advanced again. It should be noted that since the simulation processes all events that have occurred in the last timestep before incrementing the simulation time and returning control, the application cannot be used to evaluate the robustness or time efficiency of assessment approaches in terms of the volume, load or frequency of service requests(or other events).

5.2.4 Simulation Initialisation

A set of rich GUI components are provided to initialise and control the various SOA members and their parameters; and to initialise and control the simulation itself. For each member all aspects defined in the conformance-centric SOA model are adaptable. In each case the application provides standard and flexible components for setting these behaviours. As with other aspects of the application, these components may be customised to support extra functionality offered by new features. For each entity this section describes the parameters that are adaptable and the components provided by the application for their initialisation and control.

5.2.4.1 Consumer Initialisation

The tools provided for consumer initialisation allow for the provision of a consumer title and a set of behaviours which represent consumer decision models. Three decision models are defined for each consumer:

- **Predicted Delivery Model:** this describes the conformance curve for the results of assessment that return a prediction of the delivered level of service.
- **Predicted Rating Model:** this describes the conformance curve for the results of assessment that return a prediction of the consumer rating of a service.
- **Delivered Utility Model:** this describes a utility curve which is used to map the level of service delivered to a consumer to a utility value, that will be used by evaluation components to determine the actual level of quality received by a consumer.

The first two models are used by the consumer to decide which service to choose following assessment with the third being used by evaluation components. In many instances the two models relating to delivered values may be the same though the models are defined separately in the application to allow differentiation between quality perceived and quality experienced by consumers.

The application API defines base classes for each of the models. It may be extended to allow the creation of specific decision models that can be imported into the application. Figure 5.8 (right), illustrates the specification of a basic Delivered Utility Model. In addition to specifying models for consumer behaviour a

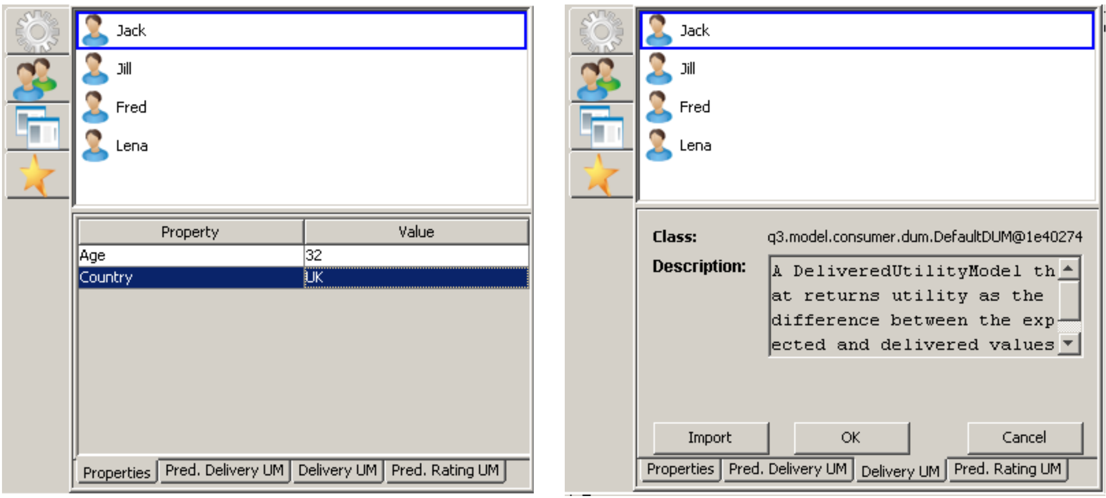


Figure 5.8: Simulation Application - Consumer Initialisation

customisable set of properties are definable for each consumer (Figure 5.8, left). This allows support for quality assessment approaches which require knowledge of consumer context in making an assessment such as collaborative filtering approaches.

5.2.4.2 Service Initialisation

In addition to a title and colour (which may be used to differentiate services on graphical evaluation components) the behaviour of individual services may be defined in terms of the delivered level of service provided throughout the simulation. As defined in the conformance-centric SOA such behavioural models allow the specification of service classes defined by expectation range. Figure 5.9 illustrates the behaviour of a provider with three service classes. The graph in the illustration is fully interactive, allowing users to plot points and lines and provides a range of graph filters for applying functions such as transposition to each series as well as the addition of normal noise. As for consumers, services may have a set of custom properties defined to support quality assessment approaches which require this information.

Figure 5.9: Simulation Application - Service Initialisation

5.2.4.3 Quality Assessment Initialisation

Parameters which define the operation of Quality Assessment members are specific to the individual approach type. Thus, the QA standard initialisation feature is minimal but provides a flexible parameter-driven interface which uses reflection to determine the parameters required by each QA in the environment and their data types, displaying input fields using a hierarchy of widget components. Figure

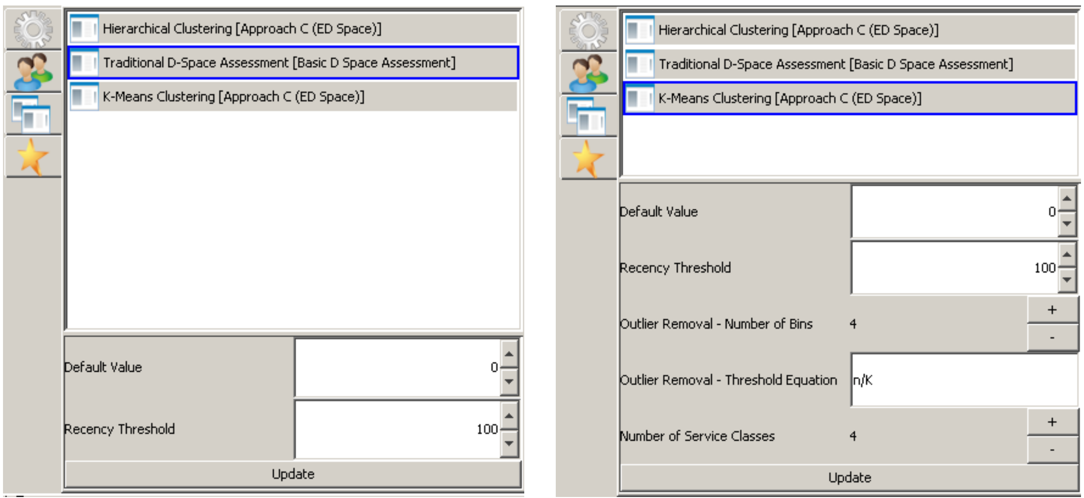


Figure 5.10: Simulation Application - Quality Assessment Initialisation

5.10 illustrates the parameter specification interfaces for two different quality assessment approaches with different initialisation parameters.

5.2.4.4 Script Initialisation

The script initialisation component provides tools to add individual events of the types described above to occur at specific times during the simulation. As described previously, in the case of all experiments, the event queue is filled at the start of the simulation and to facilitate increased speed of experimental setup an event sequencing tool is provided to allow the specification of patterns of events to occur periodically throughout the simulation. For example, Figure 5.11 illustrates the creation of a sequence of assessment events for an individual consumer with varying expectation at intervals of about 3 seconds for the first 100 seconds of the simulation.

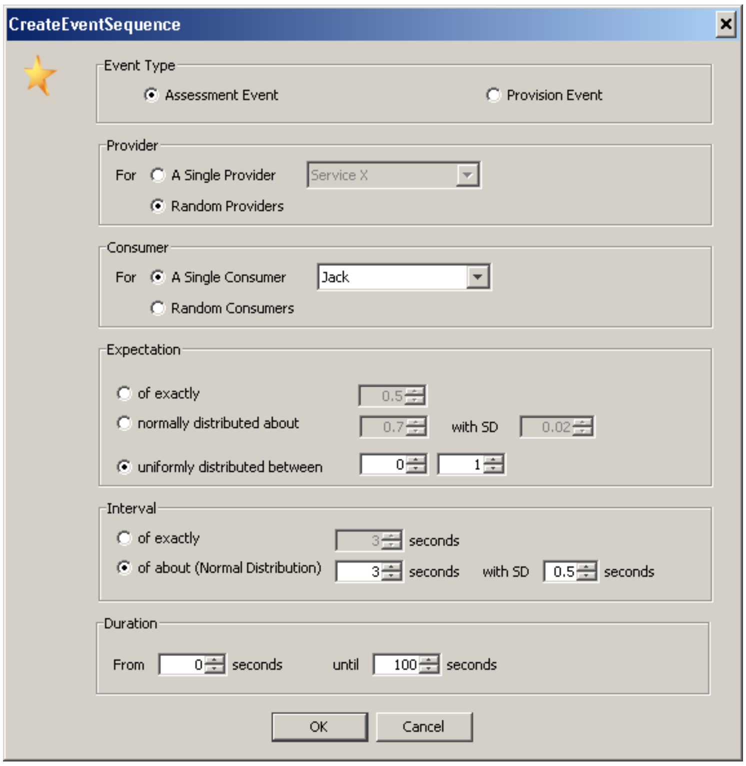


Figure 5.11: Simulation Application - Script Initialisation

The tool provided allows specification of series of events that occur with a particular distribution across time, giving the user the ability to define uniform (i.e. every n seconds) and normal (i.e. about every n seconds) distributions. Poisson distributions are sometimes considered as suitable models for representing the number of requests to a web server over time. Whilst the simulation environment allows for the specification of events of any distribution (via an Events API), the script initialisation tool provided does not include such a capability for two reasons. Firstly, Poisson distributions are used largely for "stress-testing" applications under worst case scenarios (e.g. exponential growth and decline in traffic), which can be used to optimise queueing to ensure server throughput and reliability. As previously discussed, the nature of the DES used by the simulation (and specifically the queueing mechanism) means that all events that have occurred in a timestep are processed prior to the simulation advancing, so the simulation cannot be used for load testing. Nor is time efficiency testing of assessment approaches a goal of this thesis. Secondly, although adopting a Poisson distribution rather than a Gaussian distribution will lead to of performance information being gathered about a service at different times, when analysing data to identify clusters the assessment approaches developed do not consider time as part of the clustering process. As such, the d-space data available in making assessments (and therefore assessment results) would be largely unchanged when specifying a different distribution of events in the simulation.

5.2.5 Simulation Evaluation

As part of the main DES thread, following the processing of any event the simulation alerts any component that has previously registered as an observer to the environment. Such observers may register interest in being informed when a range of triggers occur (Table 5.3). The use of observer components provides the basis

Trigger
Time Changed
Simulation Started
Simulation Stopped
Simulation Settings Changed
Service Entered/Exited
Assessment Occurred
Service Provision Occurred

Table 5.3: Observable Simulation Events

for two important capabilities in the application, namely:

- **Logging and Debugging:** components may be used to observe the actions of particular entities to ensure that they are behaving as intended. This is a critical feature in supporting the use of simulation to develop and debug new approaches to quality assessment.

- **Performance Evaluation:** a flexible and rich hierarchy of graphical components is provided to evaluate the behaviour and performance of entities in the environment.

Details of a range of evaluation components are discussed below. In addition to the specific features offered by each individual component all components in the application support the export of data in a range of formats including Comma Separated Value (CSV) and XML, as well as supporting copying and saving of graphical results as bitmap images.

5.2.5.1 Quality Evaluation

The role of quality assessment is to enable the consumer to identify services that lead to them experiencing higher quality of service. Thus, so that the application supports the evaluation of quality assessment approaches it is necessary to provide a means of determining the change in quality experienced by consumers in various configurations. The *quality evaluation component* provides a graphical representation of the quality experienced by consumers throughout the simulation. The component can be configured to display results by individual consumer and

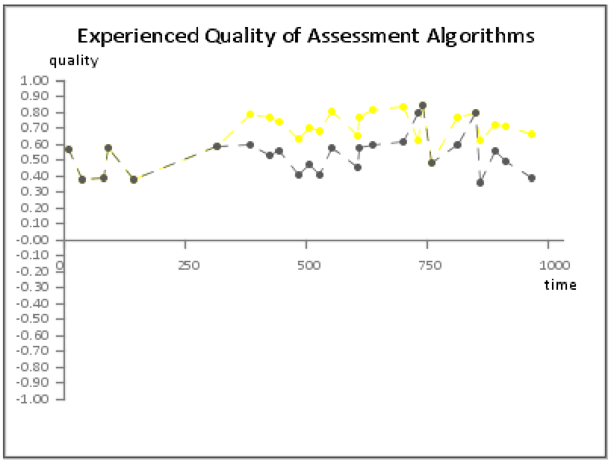


Figure 5.12: Simulation Application - Quality Evaluation Component

provider and displays individual series for each QA selected. To allow fine-grained analysis of algorithms the component can also be configured to display only instances where the consumer's expectation was in a specified range. Figure 5.12 illustrates the quality evaluation component configured to show the experienced quality of a single consumer for expectation in the range 0.4 to 0.5, when using two different quality assessment algorithms, indicated as the two series in the chart.

As discussed in Chapter 2, approaches from the field of Trust and Reputation and collaborative filtering tend to use ratings information in making service assessments. In contrast, other approaches utilise objective performance information. By supporting the recording of both ratings and performance information and providing tools for evaluating the quality experienced by consumers it is possible to use the application to fairly compare the performance of approaches from these distinct fields.

5.2.5.2 Prediction Evaluation

The quality evaluation component is a useful general purpose component for evaluating the quality experienced by consumers. However, in order to evaluate the cause of changes in quality it is useful to be able to directly monitor the performance of individual quality assessment algorithms and their prediction results. The *prediction evaluation component* observes assessment events allowing comparison of the level of service predicted in the assessment result and the level of service actually delivered by the service. The latter value is obtained by consulting the behaviour model of the service. As with the quality evaluation component

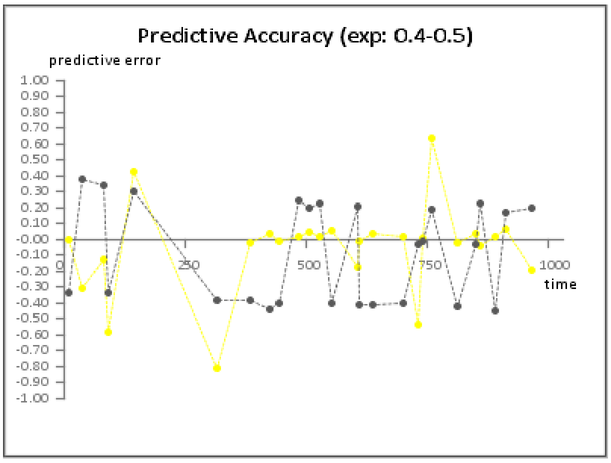


Figure 5.13: Simulation Application - Prediction Evaluation Component

the prediction evaluation component can be configured in terms of the consumers, providers and quality assessment approaches displayed as well as only displaying results for assessments within a particular expectation range. Figure 5.13 illustrates a prediction evaluation component monitoring the predictive accuracy of two quality assessment approaches for a single consumer, when the expectation of the consumer is within the range 0.4 to 0.5. The prediction evaluation component provides a means for establishing and comparing the predictive accuracy of different quality assessment algorithms.

5.2.5.3 Quality Assessment Observation Components

In contrast to the quality evaluation and prediction evaluation components, which provide a graphical representation of QA performance as a chart, a hierarchy of evaluation components are provided for detailed analysis and evaluation of the assessment process of each QA. These components are supported by two principles, illustrated in Figure 5.14. Firstly, each implemented QA approach may produce a custom AssessmentReport object that is attached to each assessment

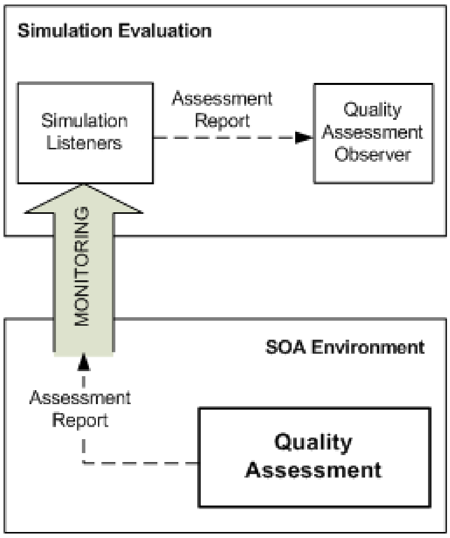


Figure 5.14: Simulation Application - Quality Assessment Observation Principles

result and may contain details of the process of each assessment instance. No restrictions are placed on the complexity or structure of this object which may be used to reconstruct the assessment process at a later date. Secondly, *quality assessment observation evaluation* components may be provided that are defined as being associated with a particular type of quality assessment approach. When an assessment event is passed to one of these components it examines the AssessmentReport object and displays the results appropriately. The simulation application provides a menu that gives the user access to the quality assessment observation component for each of the QA in the environment. Figure 5.15 illustrates a quality assessment observation component displaying a list of assessment results, services assessed in a specific assessment instance along with logging information recorded by the algorithm and a graph showing the distribution of data used in assessment.

5.2.6 Serialization

A key criteria for evaluation is that simulations should be repeatable. Hence, the application provides a serialization feature which saves all the simulation settings, SOA entities and configuration to an XML file. Figure 5.16 is an excerpt of the XML generated by experiment serialization. In addition, the user may define an experiment title and description to provide an explanation for the configuration and details of its use.

5.2.7 Simulation Process

The setup of the simulation environment when used to conduct experiments to evaluate the performance of quality assessment approaches typically comprises the

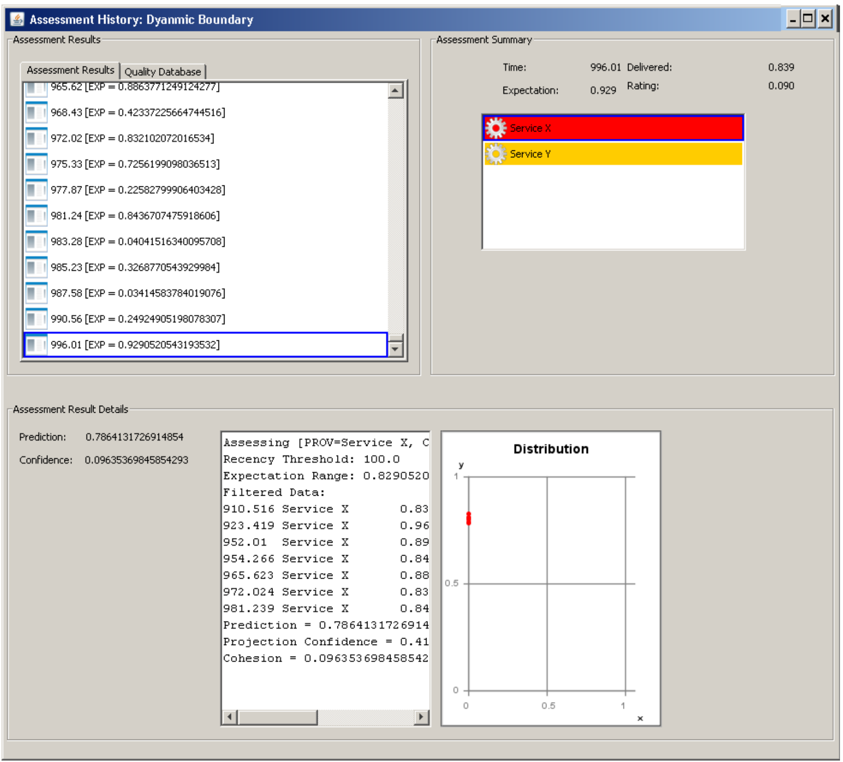


Figure 5.15: Simulation Application - Quality Assessment Observation Component


```
- <Experiment>
  - <Details>
      <Title>Unnamed Experiment</Title>
      <Author>Unknown</Author>
      <About> <H2>Unnamed Experiment</H2> </About>
    </Details>
  - <Settings>
      <DT>0.25</DT>
      <Delay>50</Delay>
      <Seed>0</Seed>
      <Start>0.0</Start>
      <End>1000.0</End>
    </Settings>
  - <ServiceSet>
    - <Service>
        <Id>c492df77-5af6-4807-b6d8-def46159189e</Id>
        <Name>Service X</Name>
        <Colour>255,0,0</Colour>
      - <ServiceClass>
          <Low>0.0</Low>
          <High>0.45</High>
        - <Point>
            <X>983.9370078740158</X>
```


Figure 5.16: Simulation Application - XML Serialization

following stages:

(a) **Simulation Setup:** Behavioural aspects relating to the simulation are defined including the length of the simulation, and the speed of simulation.

(b) **Service Setup:** One or more services are created within the environment and their behaviour specified.

(c) **Consumer Setup:** One or more consumers are created. Their decision models are specified

(d) **Assessment Algorithm:** One or more types of algorithm are added to provide assessment results. As discussed in the previous section these can be considered as running in parallel. The simulation will be run as if the consumer consulted each algorithm independently and the results were kept separately.

(e) **Scenario Specification:** A series of events of the types described in Section 5.2.3 are specified. In general two phases of simulation are specified:

   (i) **Population Phase:** A series of ProvisionEvents are defined in the initial stage of the simulation in order to provide information for use by the assessment algorithms.

   (ii) **Evaluation Phase:** A series of AssessmentEvents are defined in which assessment algorithms are asked to provide predictions for each of the available services, given a consumer's expectation. The consumer will decide which of the services to use and an associated ProvisionEvent will be generated dynamically.

(f) **Evaluation Component Setup:** A set of evaluation components are added to the environment interface to monitor specific aspects of the simulation.

In each of the experiments defined in the next chapter, details of their setup is described in terms of the above stages. When each simulation is run the evaluation components are used to capture results. These may be combined with other information about the simulation in order to generate the graphs included in the next chapter.

5.3 Summary

This chapter presented two implementations of a conformance-centric SOA. The first, an agent-based framework for the formation and management of VOs in a grid environment illustrates the application of the quality assessment approaches developed in this thesis in a realistic scenario and demonstrates interaction with a range of other components in order to support the decision making of consumers and VO managers. Whilst providing a means of demonstrating the functioning of the approach in a practical setting, the nature of the CONOISE-G architecture (as of any distributed open architecture) means that detailed monitoring and control over the behaviour of entities is not sufficient to evaluate the performance of the approach in a range of appropriate scenarios.

The simulation application presents a solution to the limitations of the CONOISE-G architecture in respect of QA evaluation. Through the provision of a virtual SOA which is an implementation of a conformance-centric SOA model, detailed control and monitoring is possible. The event simulation, evaluation components and serialization capability will support the detailed and repeatable analysis and evaluation of assessment approaches within a wide range of specific scenarios, as described in the next chapter.

Chapter 6

Results

In this chapter, an empirical evaluation of the quality assessment approaches developed in Chapter 4 is described. The chapter begins with an overview of the evaluation process in terms of the criteria used for evaluation and the range of scenarios used. The simulation application described in Section 5.2 is then used to carry out a range of experiments which evaluate the performance of approaches in the scenarios. Finally, a summary of the findings of the empirical evaluation is presented.

6.1 Evaluation Methodology

In order to demonstrate the contribution of the approaches developed in this thesis it is necessary to demonstrate their effectiveness in improving the quality of the services experienced by consumers. In this section, three criteria for evaluating such effectiveness are described and a justification is provided of their appropriateness for this purpose. This is followed by a discussion of a set of scenarios which test the approaches under a range of parameters and assumptions in order to demonstrate the conditions under which they are effective and to establish any limitations.

6.1.1 Evaluation Criteria

In Section 3.3.3, a model of consumer decision making, in terms of the calculation of expected utility of a candidate service was defined as:

$$EU(CS_i) = (C * P) + [(1 - C) * U_R] \tag{6.1}$$

where EU(CS_i) is the expected utility by the consumer using a candidate service CS_i, P is the predicted level of service offered by CS_i given a consumer's expectation and C is the consumer's subjective confidence that the prediction is accurate. The second component encapsulates the consumer's expected utility if CS_i is not chosen and is based on the individual's risk function, U_R. Prediction and confidence measures may be derived by the consumer from information obtained from a range of sources that may include a quality assessment service. The approaches defined in Chapter 4 define various techniques for providing information that may be used when calculating these values.

Quality experienced by a consumer in using services is directly dependent upon that consumer's ability to accurately evaluate the expected utility of candidate services and this in turn is wholly dependent upon the reliability of the information obtained from quality assessment services. Thus, in order to demonstrate their effectiveness in supporting consumers to discover higher quality services two aspects of the approaches must be evaluated.

- **Evaluation Criteria 1 - Predictive Accuracy:** Based on a given set of historic performance information, how close is the prediction of the assessment approach to the actual delivered level of service received by the consumer from the service following the assessment. When predictive accuracy is consistently higher consumers are likely to be able to reliably determine the degree to which services are likely to meet their needs.

- **Evaluation Criteria 2 - Prediction Validity:** To what degree is the assessment approach able to recognise the accuracy of its predictions? The validity of an assessment result should be proportional to its predictive accuracy. Overall validity of predictions may be affected by the degree to which potential service classes match actual service classes but may also be affected by other properties of the data available to the assessment or by the behaviour of services.

By demonstrating the accuracy of predictions and the reliability of prediction validity improvement in a consumer's ability to ascertain expected utility may be confirmed. Increasing the accuracy of predictions will inevitably lead to improved accuracy in calculating expected utility, though the degree to which this will lead to consumers receiving higher quality services cannot be determined without empirical evaluation. That is, the degree of improvement in prediction and confidence that will cause a particular increase in actual quality experienced by a consumer is not known. It is likely that the actual increase will depend upon the context in which the assessment is made and will be affected by the behaviour of candidate services and the expectation of individual consumers. Hence, a third criteria must be defined and used in evaluating approaches.

- **Evaluation Criteria 3 - Experienced Quality:** High predictive accuracy and effective measures of validity should lead to a consumer being able to determine the expected utility of using a particular service. However, tangible benefits will only be realised by a consumer if the results of an assessment lead to a higher quality service being chosen by the consumer than would be chosen when using another assessment approach. Assessment approaches should be evaluated in terms of the change in quality experienced by consumers through consuming services recommended by the approach over those recommended by other approaches.

Criteria 3 may be used to identify situations in which increases in predictive accuracy and validity lead to higher quality being experienced by the consumer. It is important to recognise that this relationship applies generally regardless of the individual assessment approach used to obtain prediction and confidence. By establishing the nature of this relationship and the scenarios in which increases in expected utility might lead to improved quality the general rule that improvements in prediction and validity of individual approaches will be of direct benefit to consumers can be proven.

6.1.1.1 A Note on Confidence

In Equation 6.1, C represents the subjective probability which the consumer has in the prediction P. In contrast the validity measures generated by assessment approaches are not a direct measure of confidence but are rather an indicator of the "goodness" of a model. Such measures are widely used in pattern classification applications, where they are applied heuristically rather than being based on theoretical statistical foundations.

In converting between measures of data validity and confidence using statistical techniques it is usual to identify the probability distribution function (PDF) for the variable concerned. However, the application of any PDF derived is limited to a specific pattern classification problem since the function is dependent upon assumed features of the data. Extensive Monte Carlo simulation might be carried out to develop PDFs for the validity measures as applied to the specific problem however such an activity would require substantial effort, and without adding substantial weight to the central arguments of the thesis stated in Section 1.2.

This leads in this evaluation to an assumption being made that a direct correlation exists between the validity measures generated by the assessment approaches and the subjective probability which measures a consumer's confidence in the prediction. Since both have the same range - and there is no evidence to suggest that another solution would be any more or less arbitrary - their equivalence is assumed. It is noted that changing the relationship between the variables would change the result of any expected utility calculation. However, since it is reasonable to assume that an individual consumer would apply the same principles across all candidate services, changing it would not alter the findings of the evaluation in terms of effectiveness of assessment approaches.

6.1.2 Evaluation Scenarios

The three identified criteria provide a set of objective measures with which the effectiveness of assessment approaches may be compared. Specifically, it is possible to compare approaches in terms of the *accuracy* and *validity* of their predictions and the degree to which the *quality experienced* by the consumer is improved.

In order to support the argument that the approaches developed lead to general increases in the quality experienced by consumers it is necessary to demonstrate the performance of the approaches in terms of these criteria in a range of selected scenarios. Since the range of scenarios from which to choose is vast, a subset of these must be considered as being sufficient to evaluate the approaches. The eleven scenarios described below have been selected to demonstrate the contributions of this thesis (stated in Section 1.2) in the remainder of this chapter. In general, these scenarios have been selected to evaluate and compare the performance of each of the four approaches developed in Chapter 4, in their best and worst case. An initial group of scenarios is also selected to provide a comparison with an existing non-conformance centric assessment approach and hence to define a baseline for comparison; and to demonstrate the significance of predictive accuracy and validity measure in allowing consumers to experience higher quality of service. In the remainder of this section the scenarios are divided into four groups and further justification provided of the purpose of the scenarios and their appropriateness.

6.1.2.1 Baseline Evaluation

In the first set of scenarios the key principles and components underlying ED space assessment approaches are evaluated by comparing them against a traditional D space assessment approach (see Section 3.7.1). This is intended to establish the weaknesses that are inherent in applying traditional approaches in a conformance-centric SOA. The scenarios also seek to demonstrate the relationship between increased accuracy predictions, validity measures and changes in quality experienced by the consumer. In each scenario the basic pattern passive approaches to assessment described in Section 4.5.1 and Section 4.5.2 are used.

- **Scenario 1 - Comparison of Basic ED Space with the Traditional Approach:** The Basic ED Space Approaches (Approach (a) and (b) from

Sections 4.5.1 and 4.5.1 respectively) will be compared with a generic D space assessment algorithm in terms of accuracy of predictions.

- **Scenario 2 - Effectiveness of Pattern Passive Validity Measure:** The normalised cohesion statistic employed by Approach (a) and Approach (b) as a validity measure will be evaluated in terms of its correlation with predictive accuracy.

- **Scenario 3 - Effect of Improved Predictions on Experienced Quality:** Experiments will be conducted to show that an increase in predictive accuracy leads to increases in experienced quality for consumers under realistic conditions.

- **Scenario 4 - Effect of Validity Measure on Experienced Quality:** Experiments will be conducted to show that the use of an effective validity measure leads to increases in experienced quality for consumers under realistic conditions.

The above scenarios seek to establish a baseline evaluation for ED space assessment. Specifically, by establishing the general improvements in performance of the basic ED space approaches over traditional approaches in terms of predictive accuracy the remaining scenarios may focus on evaluating the relative improvements offered by the variants of ED space assessment. Similarly, in demonstrating quantitatively the increase in experienced quality of consumers due to increased predictive accuracy and effective validity measures, the remaining scenarios may focus on evaluating approaches in terms of the first two criteria whilst assuring that such improvements will lead to improvements in experienced quality.

6.1.2.2 Effect of Boundary Choice Policy on Pattern Passive Approaches

Approach (a) (see Section 4.5.1) and Approach (b) (see Section 4.5.2) exhibit a number of similarities. Firstly, they may both be classed as pattern passive in the sense that they do not seek to identify service classes based on observed data. The complexity of the approaches is also comparable with both approaches being O(n) in worst case. Finally, both approaches utilise the same validity measure. However, the approaches are conceptually quite different in the ways that they select boundaries for potential service classes and it is difficult to determine which approach is generally preferable based on a solely conceptual analysis. Indeed, whether one of the approaches can be considered as generally preferable at all remains uncertain without conducting empirical analysis under a range of appropriate scenarios.

In the pair of scenarios described below Approach (a) and Approach (b) are compared in terms of their performance in two extreme cases, relating to the worst case and best case distribution of actual service classes for Approach (a) in a number of scenarios.

- **Scenario 5 - Uniformly Distributed Service Class Boundaries:** The case where the boundaries of actual service classes are aligned perfectly with the service class boundaries predicted by Approach (a) will be evaluated with a number of experiments conducted for different numbers of service classes.
- **Scenario 6 - Inverted Service Class Boundaries:** The case where the boundaries of actual service classes are poorly aligned to the service class boundaries predicted by Approach (a) will be evaluated with a number of experiments conducted for different numbers of service classes.

Best and worst case for Approach (b) are likely to be largely independent of service class distribution but sensitive to the expectation of the consumer in each scenario. Hence, as well as evaluating the approaches under a number of scenarios with varying numbers of service classes the results are presented in terms of improvements to consumers with a range of expectations.

By observing the relative performance of the approaches conclusions may be drawn regarding the effect of boundary choice policy (the distinguishing feature between the approaches) on approach effectiveness, and the degree to which either of the approaches may be considered as generally preferable.

6.1.2.3 Comparison of Pattern Passive and Pattern Active Approach under Semi-Private Policies

Whilst the pattern passive approaches evaluated in the previous section are similar in a number of ways the pattern active approach, described as Approach (c) in Section 4.6.1, represents a different class of approach to assessment. However, the three approaches are all designed to operate in an SOA where the number of service classes (N) offered by the service is known a priori and are therefore objectively comparable in terms of their effectiveness in supporting consumers in such environments. Whilst the pattern passive approaches are likely to be significantly affected by the distribution of service classes and the expectation of consumers, pattern active approaches should be more robust in this respect since they do not rely on assumptions regarding the segmentation of E space. Instead, they seek to establish the set of potential service classes at runtime by examining historic performance data. Hence, these approaches are susceptible to data quality issues in d-space data including separation and shape of clusters and noise.

- **Scenario 7 - Effect of Boundary Choice Policy on a Pattern Active Approach:** The experiments conducted in Scenario 6 will be used to establish the change in predictive accuracy demonstrated by Approach (c) in the worst case for the pattern passive approaches. This will suggest whether Approach (c) is generally preferable (over Approach (a) and Approach (b)) in a conformance-centric environment in terms of producing higher accuracy predictions.

- **Scenario 8 - Effect of Noise on Accuracy and Validity on a Pattern Active Approach:** A set of experiments will be conducted in which varying levels of noise are applied to d-space data in order to establish the robustness of the clustering algorithm employed by Approach (c) and to demonstrate the effectiveness of the silhouette coefficient in providing graceful degradation as noise increases.

These two scenarios seek to establish Approach (c) as generally preferable to the pattern passive approaches previously described by direct comparison of their predictive accuracy and demonstration of the robustness of the approach in the presence of noisy data.

6.1.2.4 Comparison of Pattern Active Approaches under Fully-Private Policies

In instances where information regarding the number of service classes offered by a provider is not known a priori, the performance of approaches previously evaluated is likely to be degraded. Whilst there is a general negative effect on the predictive accuracy of approaches the validity measure used by each should be able to recognise the poor match of service classes and be reduced accordingly. Hence, these approaches should still be able to operate in such environments, albeit in a diminished capacity. The degree to which performance is affected is difficult to ascertain through qualitative analysis.

In contrast, the pattern active approach described in Section 4.6.4 which employs a hierarchical clustering algorithm – referred to as Approach (d) – does not rely on a priori knowledge of the number of service classes when making an assessment rather inferring the value from the d-space data. As for Approach (c) the performance of Approach (d) is therefore likely to be susceptible to characteristics of d-space data such as noise and cluster separation.

- **Scenario 9 - Performance of Approaches under Fully-Private Policies:** The performance of all four approaches in a scenario where the number of service classes (N) is unknown a prior will be evaluated to establish the robustness of each approach in coping with such situations and to demonstrate the achievement of high predictive accuracy by Approach (d).

- **Scenario 10 - Detecting Various Fully-Private Policies:** Approach (d) will be evaluated in terms of the predictive accuracy across experiments with varying numbers of service policies. Approach (c) will be evaluated in the same setting (where it has knowledge of N a priori) to establish the appropriateness of Approach (d) for use in Semi-Private policy scenarios.

- **Scenario 11 - Effect of Noise on Accuracy and Confidence:** A set of experiments will be conducted in which varying levels of noise are applied to d-space data in order to establish the robustness of the clustering algorithm employed by Approach (d) and to demonstrate the effectiveness of the silhouette coefficient in providing graceful degradation as noise increases.

The three scenarios described above will enable quantitative comparisons of the approaches in terms of two criteria. Firstly, the limitations in the three approaches (Approach (a), Approach (b) and Approach (c)) when applied in fully-private policy settings may be ascertained. Secondly, the performance of Approach (d) in recognising a range of service classes may be objectively compared to the existing best algorithm. This may prove significant since a favourable comparison may establish Approach (d) as a generally preferable approach regardless of whether semi-private or fully-private service classes are adopted by providers.

6.1.2.5 Summary

The choice of scenarios evaluates the fitness for purpose of the approaches by identifying the likely change in utility experienced by the consumer (enabled through high confidence, high accuracy predictions) as well as the limitations of the individual approaches and the circumstances in which these occur.

6.1.3 Verisimilitude

Whilst the performance of the approaches will be evaluated in terms of the two criteria defined in Section, it is important to also identify those aspects of experiments that may affect confidence in any conclusions drawn. Foremost is the identification of the set of variables that might affect the performance of each approach, so that experiments may be designed to isolate the cause of any particular result obtained. Furthermore, when evaluating the effect of any individual variable on approach performance, it is necessary to choose values for that variable that are realistic. In addressing these points, the list below provides a brief overview of the variables of importance along with a brief description of their influence on experimental setup.

- **Service Behaviour:** The ability to provide accurate assessment for providers with a range of behaviour is arguably the most critical characteristic of any approach to assessment. In simulations, the behaviour of services is defined in terms of service classes and target levels of service. The majority of experiments are chosen to evaluate individual approaches performance in assessing providers with a range of behaviours, including varying the number and distribution of service classes. Particular experiments are chosen in order to evaluate the performance of approaches with behaviour that is likely to cause decreases in predictive accuracy in order to draw out the weaknesses of individual approaches and identify their significance.

- **Noise:** Noise within the historic service performance information recorded in the QDB, will undoubtedly be subject to noise. Whilst there will inevitably be limits on the ability of any approach to assessment to deal with noise, it is important to be able to demonstrate the degree to which noise occuring in data affects the reliability of assessment results. This is especially important for the approaches defined in this thesis, which include a validity measure in order to mitigate against the risk of inevitability noisy data. In order to isolate and evaluate the effect of noise on the approaches defined in this thesis (e.g. as separate from changes in the other variables defined here) Scenario 8 and Scenario 11 are designed solely to evaluate the performance of approaches under varying levels of noise. In other scenarios (which are designed to evaluate the effect of variables other than noise), noise is left intentionally low, so as to not overly interfere with the results.

- **Consumer Behaviour:** The behaviour of consumers in respect of requirements and experienced quality will vary widely. Confidence that the results obtained in this chapter are unaffected by differing consumer behaviours is assured by approaches being primarily evaluated in terms of the predictive accuracy and validity measure, rather than level of quality experienced. Justification for the choice of criteria as being sufficient in this respect is provided in Section 6.

- **Assessment Distribution/Frequency:** As discussed in Section **??**, whilst care has been taken to select individual techniques for approaches that will scale well (e.g. in terms of dataset size, number of dimensions, etc) the evaluation of approaches in this thesis focusses on their *effectiveness* in detecting patterns in historic information, for services with different types of behaviour. As such, the frequency and distribution of assessment requests in these experiments is chosen simply to provide enough information to the QA in making an assessment of a service. It is acknowledged that in order to robustly evaluate the *efficiency* of approaches it would be necessary to demonstrate their performance under various experiments in which both average and worst case request loads and distributions were used.

The discussion of variables above is intended to provide confidence that the simulated results obtained below would be observed in practice if deployed in a real world setting and to clearly define the limitations in scope of the evaluation.

6.2 Experimental Results

This section presents and analyses the experimental results arranged in the order of the scenarios described in Section 6.1.2.

6.2.1 Baseline Assessment Evaluation

In this section, the pattern passive approaches described as Approach (a) and Approach (b) in Sections 4.5.1 and Section 4.5.2 respectively are compared against a traditional assessment approach, Approach (x) (see 3.7.1), using three scenarios. The aim of the experiments conducted in the first set of scenarios is to show that the use of pattern passive approaches lead to improvements over traditional approaches in terms of the three criteria defined in Section 6.1.1. Three experiments are conducted. The first, provides a basic scenario in which a provider who is offering multiple service classes is assessed by a conformance-centric approach, Approach (a), and a non-conformance centric approach, Approach (x). This scenario serves to demonstrate the inherent weakness of traditional approaches in assessing providers that adopt a policy based on service classes. The second experiment provides a scenario that causes Approach (a) to yield low accuracy predictions. In this instance the experiment seeks to ascertain the degree to which the validity measure is a reliable indicator of predictive accuracy. The final two experiments seek to demonstrate the relationship between high accuracy predictions, high validity predictions and higher quality experienced by consumers. Proving the general correlation between the former two factors and the latter, will allow experiments in later sections to focus on evaluating predictive accuracy and validity with the subsequent increases in quality experienced by consumer being assured by the initial observation.

6.2.1.1 Comparison of Basic ED Space with Traditional Approach

Figure 6.1 illustrates the behaviour of a service that has two service classes defined, offering different levels of service over the simulation period. The first service class (SC_1) offers a level of service of around 0.2, to consumers whose expectation falls into the range 0.0 to 0.5 The second service class (SC_2) offers a level of service of around 0.8 to consumers whose expectation falls into the range 0.5 to 1.0, as illustrated in Figure 6.1. Over the duration of the simulation predictions of the likely

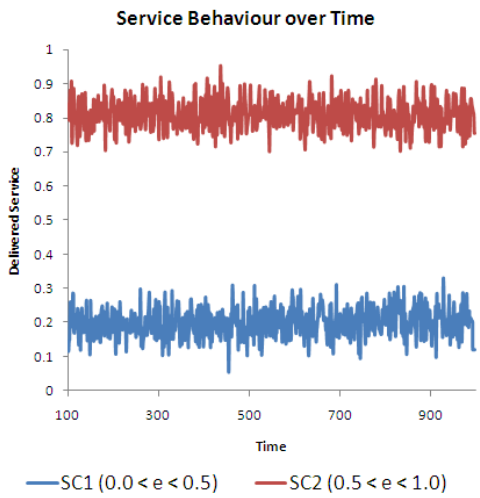


Figure 6.1: Behaviour over time plot for a Service with Two Service Classes

behaviour of the service are made by the pattern passive approach and a traditional approach. For the duration of the simulation provision events are generated for consumers with random expectation. This ensures that there is sufficient data available to make an assessment. The results of each assessment in terms of their predictive accuracy are recorded and illustrated in Figure 6.2. As with all experiments in this chapter, assessment events are generated for consumers with random expectation at 1 second intervals.

Figure 6.2 illustrates the prediction made by each of the approaches against the actual level of service delivered by the provider in each case. The two series of scattered points at the bottom and top of each graph correspond to provision from the two service classes illustrated in Figure 6.1. The third series (in green) illustrates the predictions made by each of the approaches. The results show a marked difference in the accuracy of predictions made by the two approaches. Approach (x) (see Figure 6.2, left) consistently provides predictions that represent the arithmetic average of the data recorded from the two service classes since the data used in making calculations is drawn equally from both. In contrast the

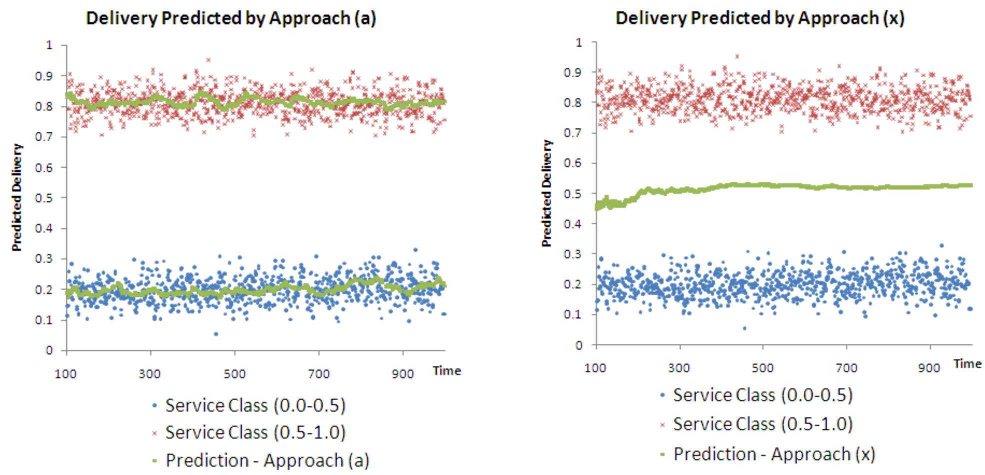


Figure 6.2: Comparison of Predictions of Approach (a) and Approach (x) for a Single Service with Two Service Classes

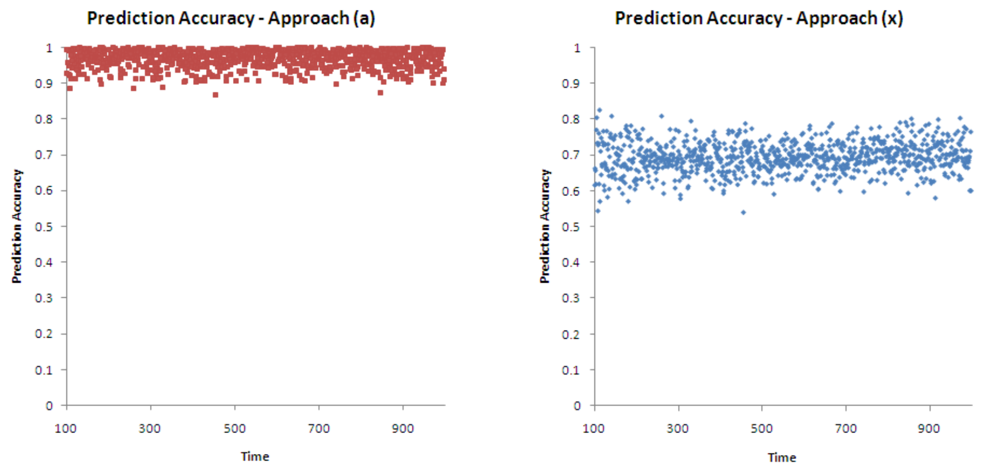


Figure 6.3: Comparison of Accuracy of Approach (a) and Approach (x) for a Single Service with Two Service Classes

predictions of Approach (a) (see Figure 6.2, left) are more closely correlated to the level of service delivered by the appropriate service class since the predicted service boundaries match the actual service classes well in this case. This is visible as the emergence of two green lines correlating to the level of service delivered in each service class.

Figure 6.3 illustrates the accuracy of the predictions made by each of the approaches where accuracy of 1 indivates that there was no difference between the level of service predicted by the approach and the level of service received by the consumer. The overall trend indicates that Approach (x) produced results that were on average significantly less accurate than those of Approach (a). It should be noted that the high predictive accuracy achieved by Approach (a) represents a best case for the algorithm since the service classes are a perfect match to those predicted by the approach. The effect of imperfect matches between predicted and actual service classes on the accuracy and validity measures produced by this approach are described in Section 6.2.2.

6.2.1.2 Effectiveness of Pattern Passive Validity Measure

The effect of the naïvety of the pattern passive approaches is mitigated through the calculation of validity measures for each assessment which seek to determine the "goodness-of-fit" of the service classes predicted by the approach to the actual service classes defined by the service provider. As described in Section 6.1.1.1 the validity measure cannot be considered as a measure of confidence though for the purposes of evaluation their equivalence is assumed. Thus, it is necessary to ensure that the validity measure is a good indicator of the degree to which the predicted and actual service classes match. The experiment described in this section aims to demonstrate the behaviour of the measure in a realistic scenario where the match of predicted service classes to actual service classes varies. Essentially, the experiment seeks to evaluate whether the validity measure is able to reliably identify cases where the predictive accuracy of the approach is low.

One possible way of ensuring that low accuracy predictions are generated by the approach is to define a service where the number of actual service classes differs from the number assumed by the approach. Although this scenario strictly falls outside the range of those implied by the assumptions of the approach it is useful to consider it, in that for certain expectations it constitutes a theoretical worst case for the predictive accuracy of the approach and when considered in the overall evaluation of the approach it should be interpreted as such.

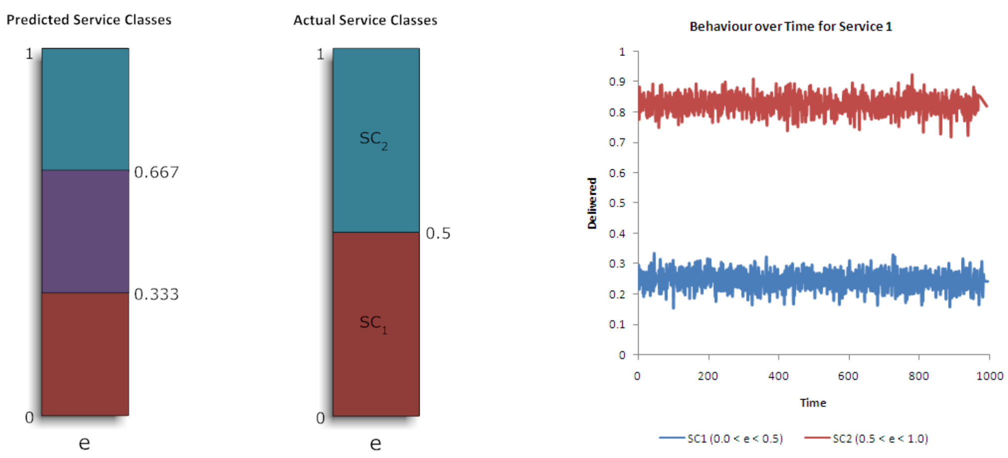


Figure 6.4: Predicted Service Classes for n=3 and Behaviour of a Service with 2 Service Classes

As with the previous experiment since the focus of evaluation is the performance of the validity measure a single service (as defined by the behaviour in Figure 6.4) is used in the simulation. This service has two service classes which partially overlap with the three service classes that are assumed by the approach in order to force the approach to produce predictions which are of low predictive accuracy for a range of expectation. The creation of provision events and assessment events remains the same as the previous experiment with the intention of recording the accuracy and validity measure for each prediction in order to evaluate the correlation between the two variables.

The accuracy (and therefore validity) of the predictions will be dependent upon the predicted service class into which the consumer's expectation falls. For expectations in the range 0.0-0.333, the assessment will consider only points from SC1 as relevant and therefore there should be high accuracy, high validity results. This also applies to the range 0.667 to 1, where only points from SC2 will be used. However, for expectation within the middle predicted service class in the range 0.333 to 0.666 points will be used from both SC1 and SC2 leading to poor accuracy of predictions.

In order to illustrate the differing performance of the approach the accuracy and validity of predictions described in the results below are considered across three ranges, that fall into the ranges identified above. Specifically, the ranges 0.0-0.3, 0.4-0.6 and 0.7-1.0.

Figure 6.5 illustrates the prediction and accuracy of assessments with expectation in the ranges 0.0-0.3 and 0.7-1.0. As previously, the series in green indicates predictions, and the other two series represent the delivered level of service from each of the two providers. When making assessment for expectation in these ranges the approach will draw points from a single actual service class. This is evident in the high accuracy results for these two ranges. In contrast, the predictive accuracy of the approach for expectations in the range 0.4-0.6 is far lower (Figure 6.6). In these cases the prediction was made based on data from two adjacent actual service classes and is erratic. Referring to Figure 6.1, it can be seen that the area of overlap of actual and potential service classes occurs for points within the range 0.333 to 0.667 which encapsulates the entire range of expectation whose results are displayed in Figure 6.6. The degree to which the prediction tends towards one of these service classes is dependent upon the proportion of points taken from each.

Results relating to the value of the validity measure returned for the predictions in the expectation ranges are illustrated in Figure 6.7. These results are consistent with the desirable qualities of the validity measure previously defined in that the

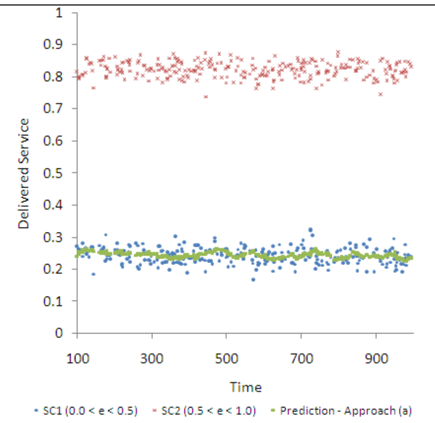
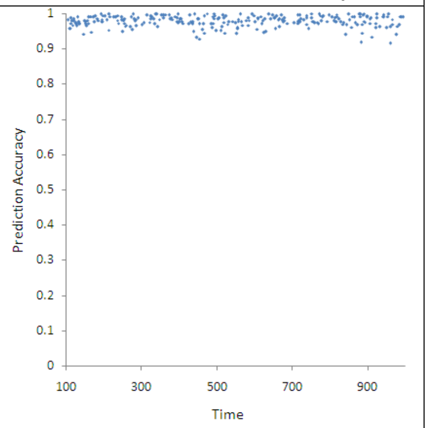
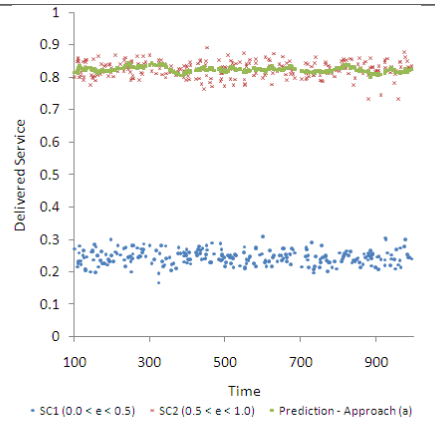
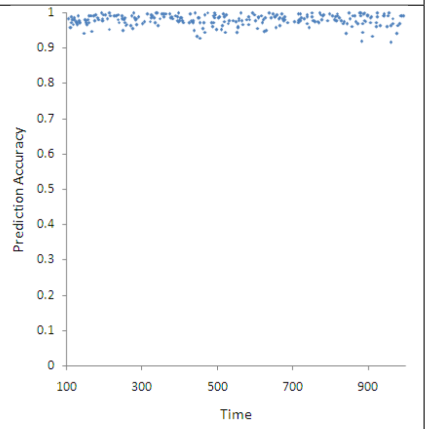
<table>
<tr><th>Exp.</th><th>Predicted Delivery</th><th>Predictive Accuracy</th></tr>
<tr><td>0.0-0.3</td><td>
</td><td>
</td></tr>
<tr><td>0.7-1.0</td><td>
</td><td>
</td></tr>
</table>

Figure 6.5: Effect of Overlapping Service Classes (No Visible Effects)

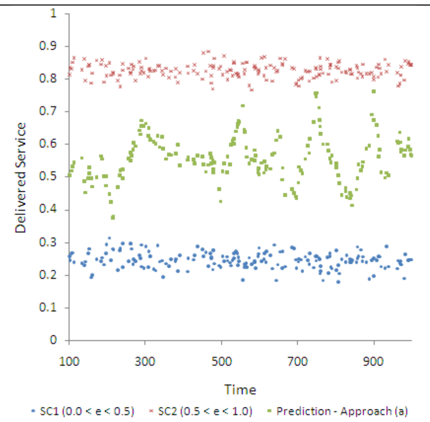
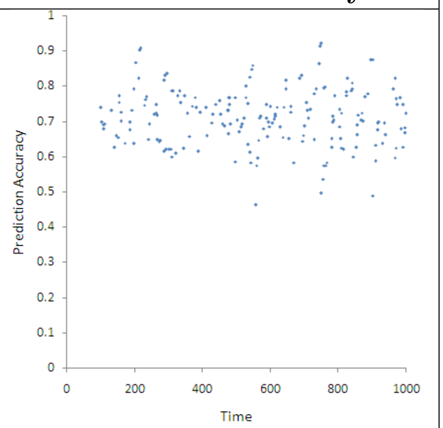
<table>
<tr><th>Exp.</th><th>Predicted Delivery</th><th>Predictive Accuracy</th></tr>
<tr><td>0.4-0.6</td><td>
</td><td>
</td></tr>
</table>

Figure 6.6: Effect of Overlapping Service Classes (Visible Effects)

validity measure is assigned a high value in the cases where a good match between actual and predicted service classes is identified and low in the case in which there is substantial inconsistencies between the two.

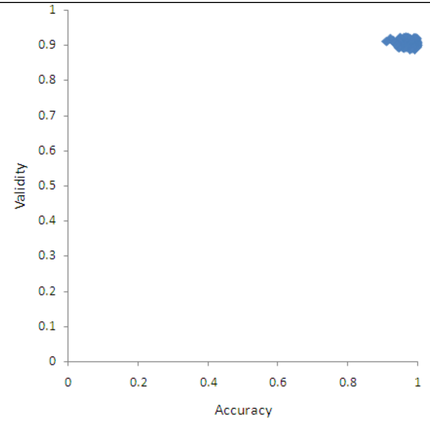
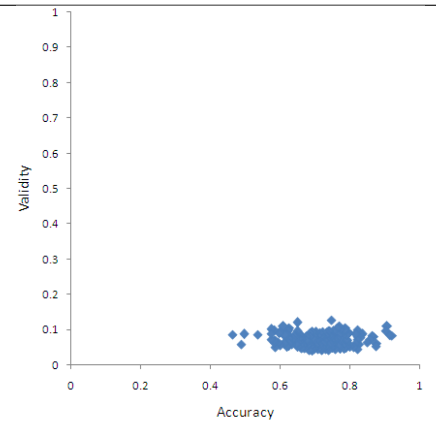
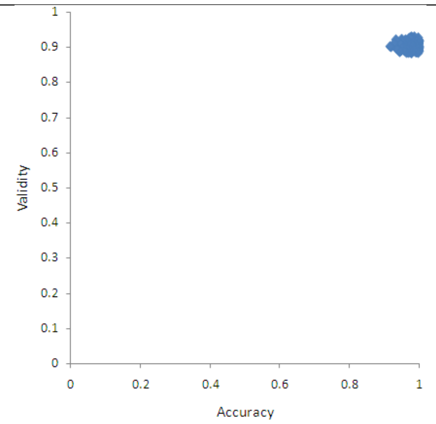
Exp.	Predictive Accuracy vs. Validity
0.0-0.3	
0.4-0.6	
0.7-1.0	

Figure 6.7: The Relationship between Accuracy and Validity

6.2.1.3 Effect of Improved Predictions on Experienced Quality

The previous experiments have established that the use of a pattern passive approach may lead to significant improvements in predictive accuracy over traditional approaches,in a conformance centric SOA and the validity measure has been demonstrated to be an effective means of identifying cases where the naivety of the approach has led to a poor prediction. However, in Section 3.3 it was considered that improvements in predictive accuracy and a consumer's subjective confidence may only be considered as significant,if they lead to a consumer engaging services that mean they will experience higher quality. To illustrate this, the following two experiments use scenarios in which multiple providers exist within the environment and the differing results of approaches lead to alternative services being chosen and therefore changes in experienced quality. The purpose of these experiments is to demonstrate the general relationship between improving the predictive accuracy and validity of predictions,and the subsequent higher quality experienced by consumers. Demonstrating this relationship will mean that later experiments may be evaluated in terms of the former two,whilst assuring that the latter (which is the factor of most significance) is achieved. That is, it will be appropriate for later experiments to be evaluated based purely on the changes in predictive accuracy and validity of various approaches.

In this section,a scenario is considered to demonstrate improvements in quality that are due to high predictive accuracy, comparing Approach (a) with Approach (x). In the next section, a scenario is considered to illustrate the same for the validity of predictions.

The two services whose behaviour is defined in Figure 6.8 both offer three service classes with boundaries defined at 0.0, 0.333, 0.666 and 1.0. However, the actual level of service offered by each of the services differs in such a way that dependent upon the consumer expectation either of the two services may offer higher quality service. For example, for an expectation of 0.1, Service 1 would offer a level of service of between 0.075 and 0.125, whereas Service 2 would offer a level of service of around 0.25, meaning that Service 1 would be the best choice for the consumer with this expectation. In contrast, for an expectation of 0.9, Service 1 would offer a level of service of about 0.8 whereas Service 2 would offer a level of service of around 0.9. In this case, Service 2 would be the best choice. An effective approach should be able to distinguish between Service 1 and Service 2 in providing better service for consumers with expectations across the range and subsequently lead to higher quality being experienced by consumers in general.

Further to the example described above the expectation range may be partitioned based on which of Service 1 and Service 2 will offer the highest quality of service to the consumer. Table 6.1 illustrates a selection of these partitions (the "∼" character indicates approximation). For example, for consumers with expectation in the range 0.0-0.1, Service 1 will provide a delivered level of service of approximately 0.1 and Service 2 will provide a delivered level of service of approximately 0.25 - these values being determined using the behaviour over time graphs for the services in Figure 6.8 (for each of the services the expectation range 0.0-0.1 falls into the lower service class with bounds of 0.0-0.333 in each case). Results from these subranges may be used to evaluate the performance of assessment in

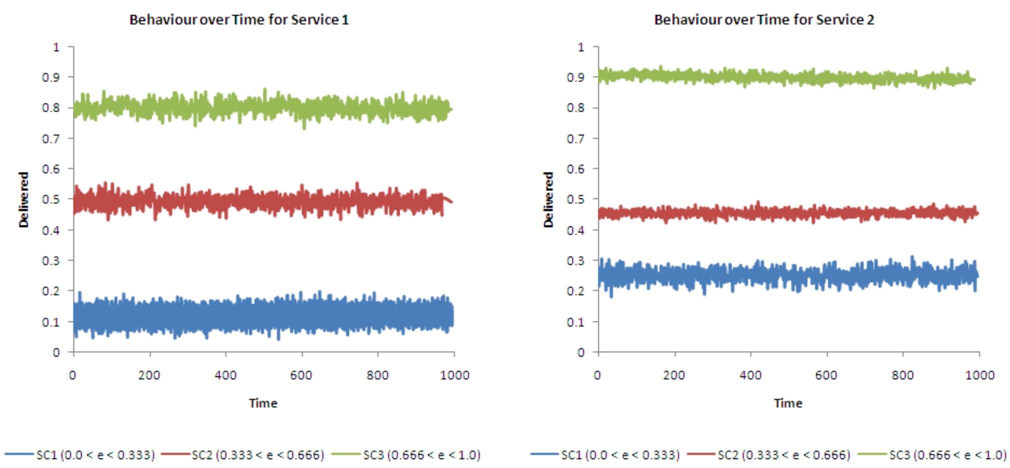


Figure 6.8: Behaviour over Time Plot for Two Services with Different Policies

recognising the "best" provider in each case.

Range	Service 1	Service 2	Best Service
0.0-0.1	~ 0.1	~ 0.25	S1
0.2-0.3	~ 0.1	~ 0.25	S2
0.7-0.8	~ 0.8	~ 0.9	S1
0.9-1.0	~ 0.8	~ 0.9	S2

Table 6.1: Partitions of e-space and Associated Service Behaviour

The purpose of this experiment is to demonstrate that higher accuracy predictions lead to improvements in quality experienced by the consumer. Hence, an assessment approach in which predictions are generated randomly will also be evaluated in order to provide a baseline for establishing the significance of the difference between Approach (a) and Approach (x).
The two series represented by the red and green lines in each of the graphs of Figure 6.9 show the service that would have been delivered by each service for each of the expectation ranges 0.0-0.1 and 0.2-0.3. In each case, the series composed of bluepoints superimposed over the top of the lines indicates the service that was chosen by consumers using each approach. The results indicate that predictions from Approach (x) lead to consumers choosing Service 1, regardless of a consumer's expectation. The prediction of Service 1 will tend towards the arithmetic average and will so predict around the 0.5 level of service. In contrast, Approach (a) correctly identifies the changing optimality of the service based on different consumer expectation - selecting S1 for consumers with expectation in the range 0.0-0.1 and S2 for those in the range 0.2-0.3, matching the best services for those ranges as defined in Table 6.1. A similar contrast in service selection is exhibited for the selection of services in the two higher expectation ranges (Figure

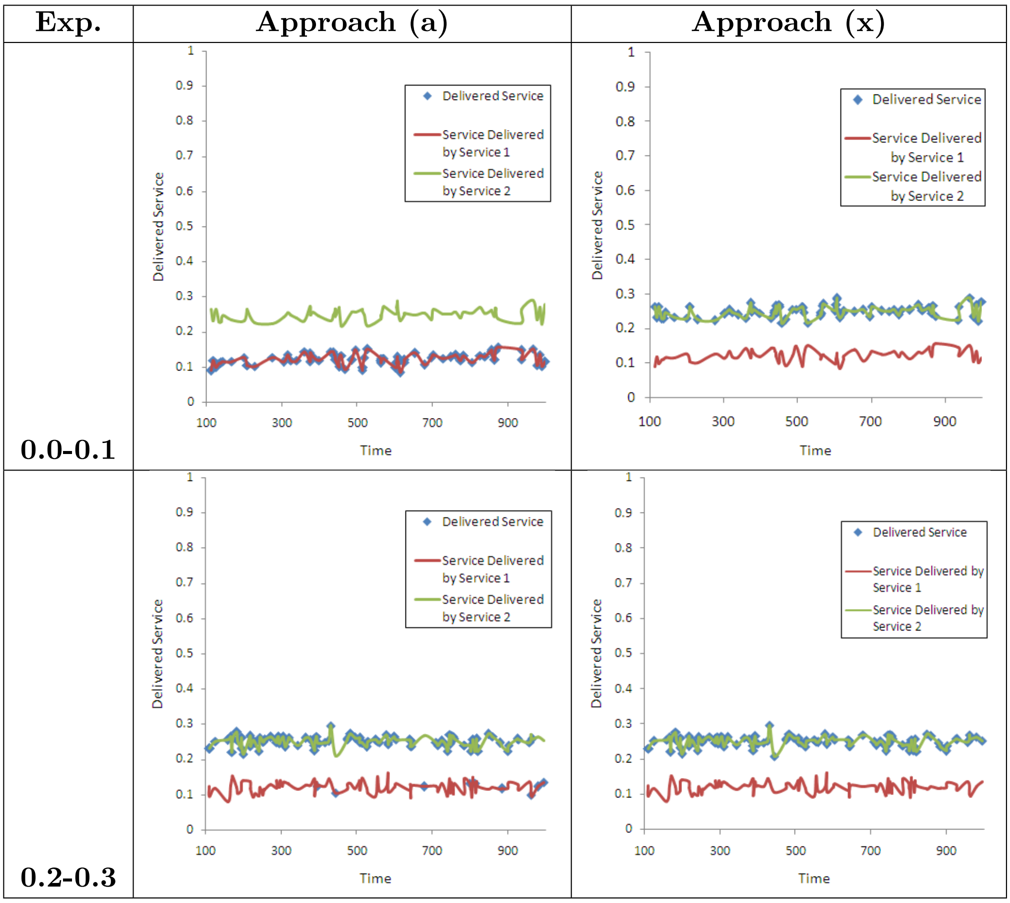


Figure 6.9: Comparison of Delivered Service received by Consumers using Approach (a) and Approach (x) (lower expectations)

6.10). These results illustrate the different selections made by consumers when using the results from the two different assessment approaches. The ranges of expectation have been chosen in order to emphasise instances where the approach led to different services being selected. As a further indicator of the significance of using the different approaches for selecting services, Figure 6.11 illustrates the accuracy of predictions for the two approaches over all consumer expectations. The graph for Approach (x) illustrates a clear banding which corresponds to the differences between the prediction and the level of service in each of the actual service classes. However, predictive accuracy when using the results of Approach (a) in selecting services is significantly higher due to the fact that the approach is able to consider the existence of service classes. The significance of the different services that are selected by the consumer after using Approach (x) and Approach (a) is evident in the level of quality experienced by the consumer over the duration of the simulation (Figure 6.12). In the case where Approach (x) is used, quality experienced by the consumer has a mean value of 0.80 and standard deviation of 0.21. Conversely, in the case where Approach (a) is used, quality experienced has

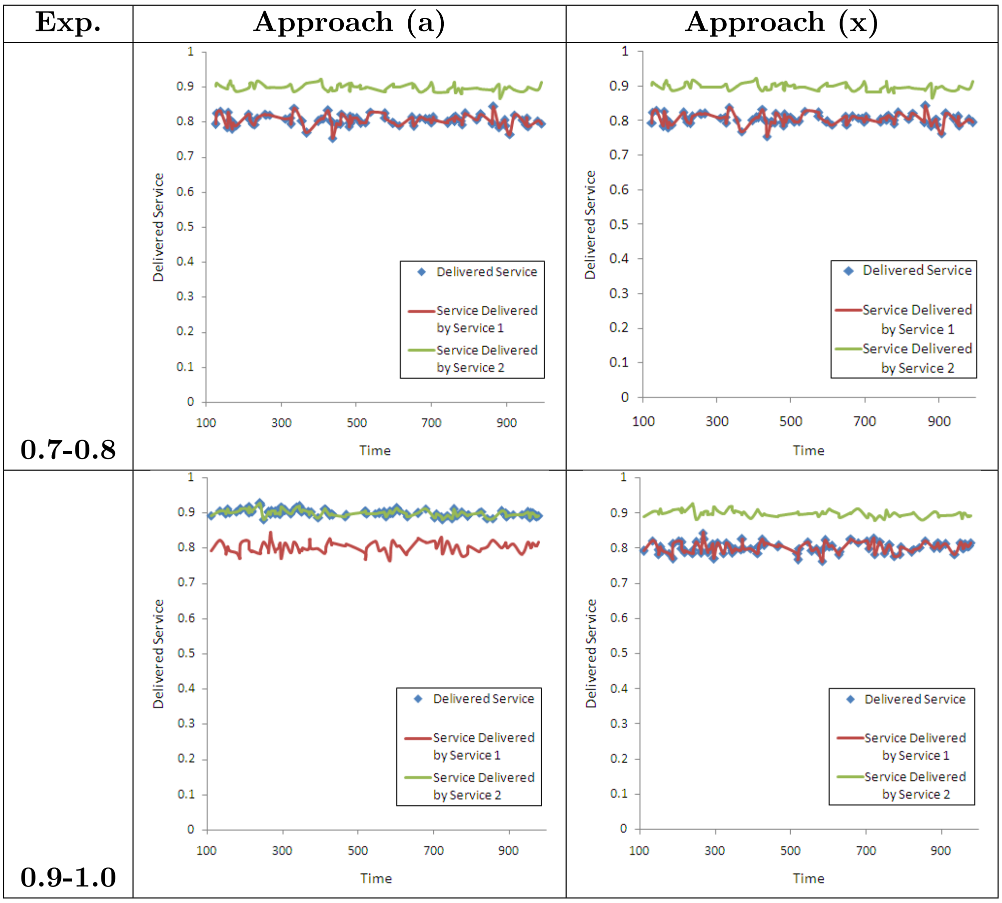


Figure 6.10: Comparison of Delivered Service received by Consumers using Approach (a) and Approach (x) (higher expectations)

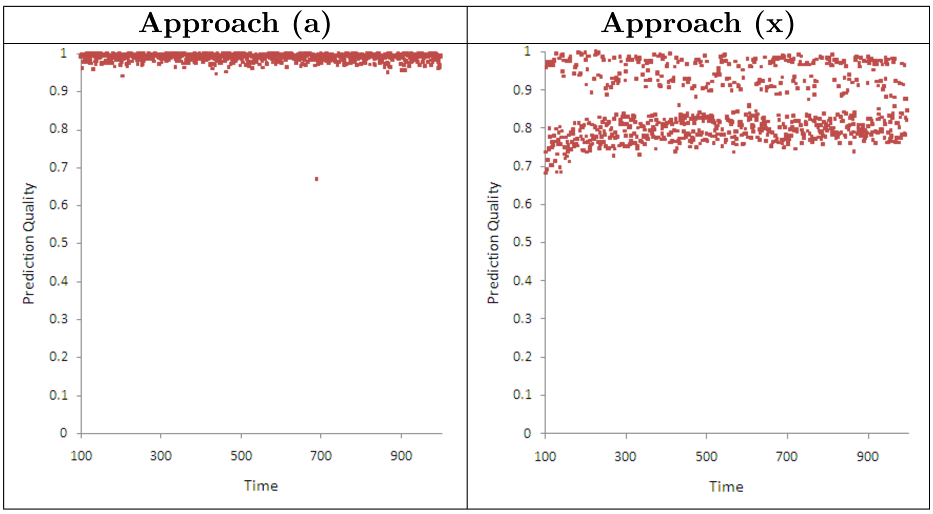


Figure 6.11: Comparison of Prediction Quality for Approach (a) and Approach (x)

a mean of 0.85 and standard deviation of 0.16.

When measured on an absolute scale the difference between levels of quality

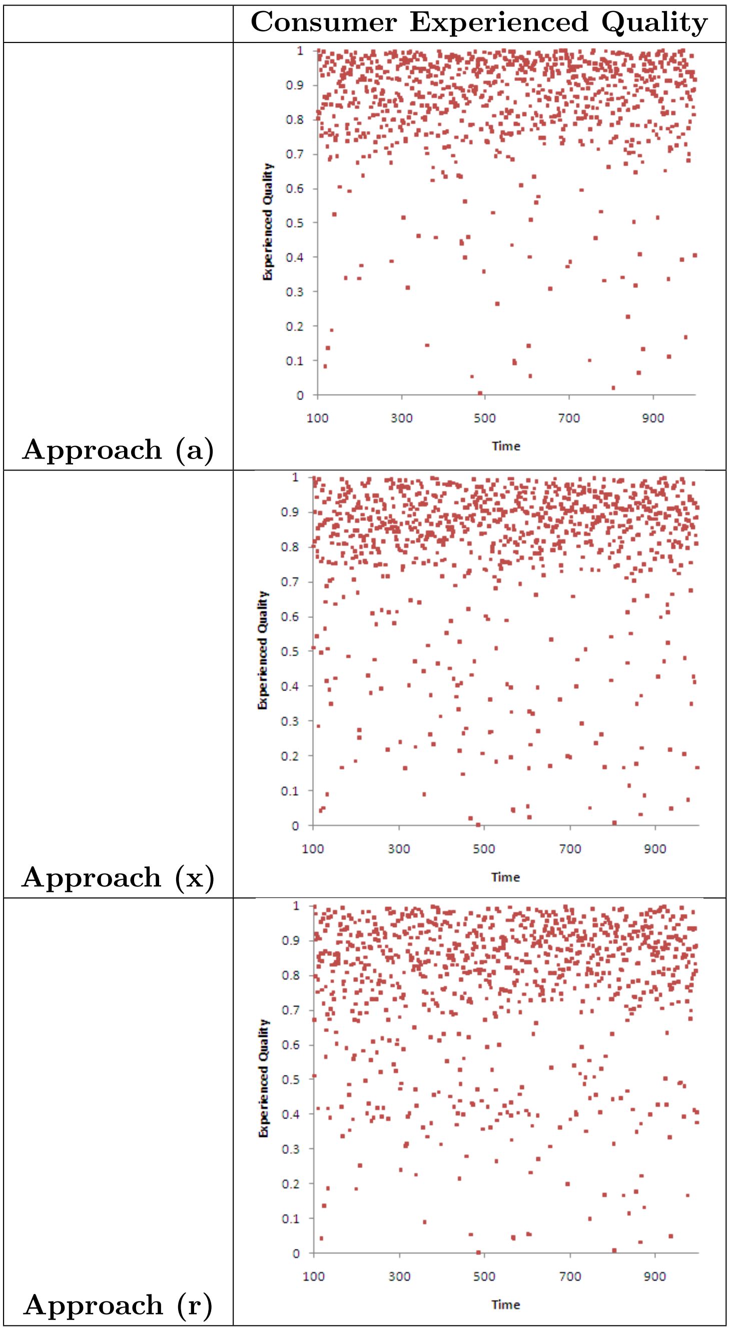


Figure 6.12: Comparison of Experienced Quality for Approach (a) and Approach (x) and Approach (r) (Random Prediction)

experienced by consumers when using each of the approaches appears to constitute a relatively small improvement. However, as illustrated in Figure 6.12, the random selection approach yields a mean of 0.78 with standard deviation of 0.21. This constitutes a difference of just 0.02 with the mean of Approach (x) and with the same standard deviation. When considered in this context, a more favourable comparison may be made of the difference in mean quality experienced whilst using Approach (a) to Approach (x).

6.2.1.4 Effect of Validity Measure on Experienced Quality

The validity measure produced by assessment approaches is used by the consumer in determining their subjective confidence in the assessment prediction. Thus, as with the prediction itself the quality of the measure may be judged in terms of the degree to which it supports consumers in experiencing higher quality from services. In this experiment, a scenario is used in which the validity measure acts to differentiate between two services such that its use leads to selection of a different service than would be chosen without the measure. In order to ensure that any observed changes are as a direct result of the effects of the validity measure the experiment evaluates two versions of Approach (a), in which one is adapted such that the validity measure calculation is set to return a constant value of 1.0 (making the validity measure redundant as a differentiating factor). The adapted approach is compared against the standard Approach (a) and against the random selection approach introduced in the previous scenario in order to establish the significance of the results.

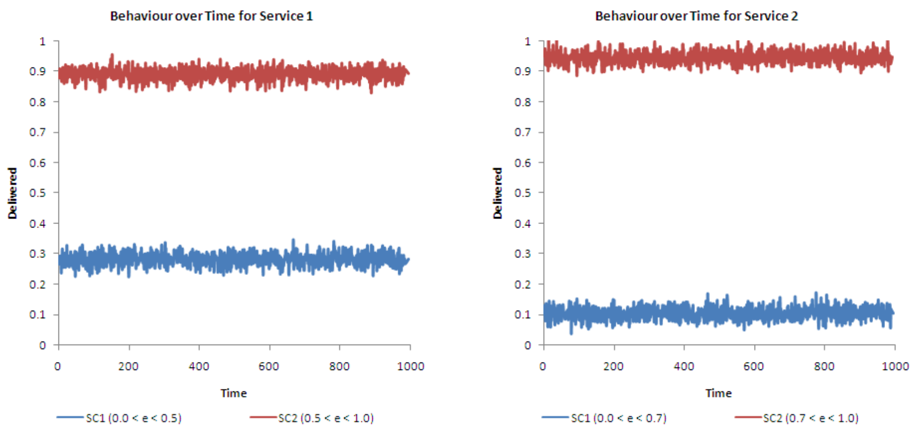


Figure 6.13: Behaviour of Two Service Providers

Two services were used in the simulation both with two service classes but different service class boundaries defined by each service (Figure 6.13). Service 1 defines two uniform service classes, with the boundary at 0.5. The boundary of the two service classes for Service 2 is 0.7.

It is important to note that the aim of this scenario is not to identify the approach that makes the most accurate predictions since the service class distribution of Service 2 makes it inevitable that predictive accuracy will be low for the pattern passive approach and therefore consumers are not likely to receive high quality services using any of the approaches. Rather, this experiment seeks to evaluate the degree to which the validity measure can support the consumer in recognising where a prediction may be inaccurate. Hence, the key measure for comparison is the relative quality experienced by consumers in using the approaches. As in the previous experiment (see Figure 6.2.1.3) the use of the random approach to assessment denoted Approach (r), will provide a baseline in order to provide context for this comparison.

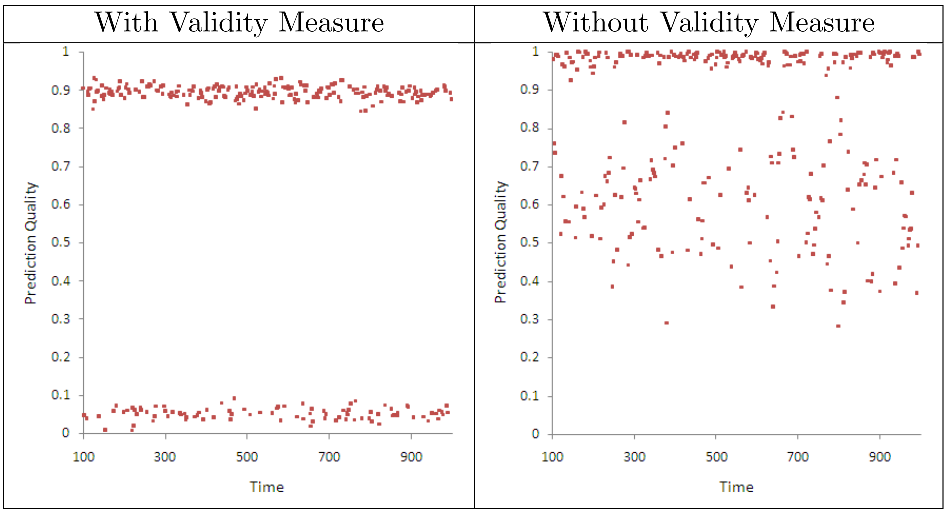


Figure 6.14: Comparison of Prediction Quality for Approach (a) - with and without validity measure

Figure 6.14 compares the predictive quality of both approaches across the entire expectation range. A clear banding of predictive accuracy is evident in the assessment results of the standard Approach (a) with the effect that the predictive quality is generally lower across the whole simulation than for the Approach without the validity measure. Whilst this result may appear to suggest that the use of the validity measure presents a hindrance to predictive accuracy it should not be evaluated in isolation from the experienced quality which is the measure of interest for this experiment. When considering the level of quality experienced by consumers using each of the assessment approaches the results indicate (Figure 6.15) that the quality experienced by consumers using the approach with validity measure experienced an average quality of 0.73 (across the range of expectation) with the results for the approach without validity being an average of 0.66. Whilst this is a small absolute difference and there appears to be visually no clear distinction between Approach (a) with and without quality the result must be interpreted in the context of the baseline result (Approach (r)). In this experiment, the average quality experienced when using the random approach was 0.63. This baseline value can be explained by considering the behaviour of the two service providers used in the experiment (see 6.13). For each assessment, in the worst case a consumer's expectation would fall at one limit of an e-space service class boundary. For SP1, the consumer expectation that would yield the worst case is a value of just over 0.5 (e.g. 0.51). In this case, a consumer will be delivered a level of

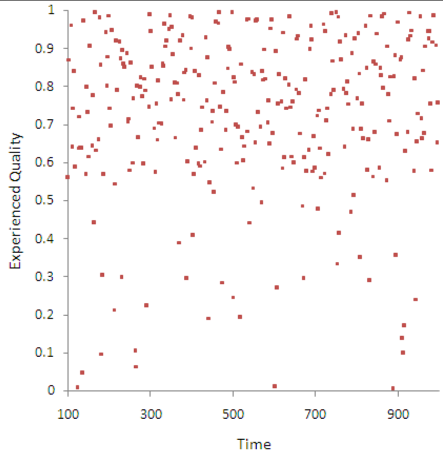
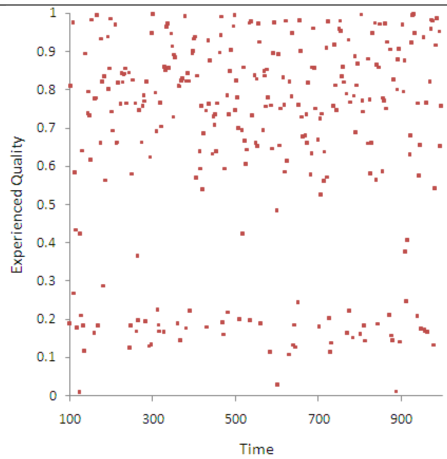
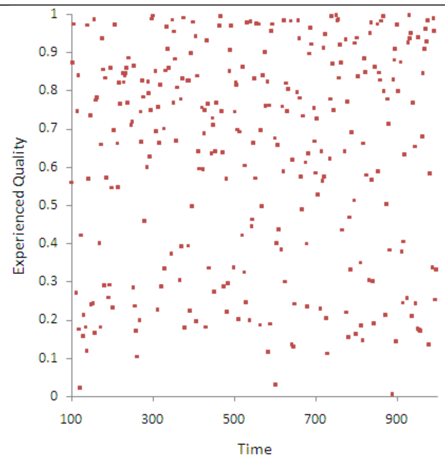
	Consumer Experienced Quality
With Validity	
Without Validity	
Random Selection	

Figure 6.15: Comparison of Experienced Quality for Approach (a) with validity, Approach (a) without validity and Approach (r)

service of 0.93 (from SC2) resulting in an overall experienced quality of 0.58 (i.e. 1-(0.93-0.51)). For SP2, the expectation that yields the worst case is just under 0.7 (e.g. 0.69). In this case, a consumer will be delivered a level of service of 0.12, resulting in an overall experienced quality of 0.43 (i.e. 1-(0.93-0.51). These values are lower than the quality experienced under the random approach (the baseline).

However, they represent the worst case and therefore all other assessments (for consumers with other expectations) will result in higher quality experienced. In such context, the seemingly high baseline value of 0.63 is explained.

In comparison with this figure the result for Approach (a) without validity constitutes a less than 5% improvement in experienced quality over selecting a provider at random. For Approach (a) with validity measure this figure is increased by a further 10% which constitutes a significant increase in experienced quality. This result reinforces the importance of using a combination of predictive accuracy and validity in supporting consumer decision making. Whilst in this scenario the approach without validity measure achieved higher accuracy predictions in general this did not lead to an associated increase in experienced quality.

6.2.1.5 Summary

The results presented in this section highlight key deficiencies inherent in traditional approaches to quality assessment and provide a quantitative analysis of the performance improvements that may be realised by using a basic pattern passive ED space approach. The specific algorithm used in the traditional approach is technically simple in its implementation using the arithmetic mean of past performance to determine future performance. However, the results in this section demonstrate a conceptual limitation in this *type* of approach due to the lack of consideration of service classes within the performance data. The technical complexity of the algorithm used by the traditional approach would have little bearing on the prediction accuracy since the limitation is inherent to any algorithm in this class of approaches.

Similarly, the basic ED space approach may be considered as being naïve since it does not seek to actively identify service classes but rather acknowledges their presence and uses the validity measure as an indicator of certainty in its predictions. Such an observation only serves to further reinforce the argument above. By comparing the performance of basic implementations of both types of approach it has been possible to identify the conceptual advantage of one over the other and clearly attribute the improvement to the conformance-centric nature of Approach (a). Furthermore, the technical simplicity of Approach (a) might prove to be an attractive quality in an approach to assessment.

The results demonstrate that increases in the accuracy of predictions and the associated validity measure can lead to consumers being able to select services that enable them to experience higher qualities of service. This is a key observation since this allows experiments in the remainder of this section to focus on the comparative differences in predictive accuracy of varying approaches whilst having confidence that such increases are able to offer tangible increases in experienced quality for consumers in a range of scenarios.

6.2.2 Effect of Boundary Choice Policy in Pattern Passive Approaches

As noted in the previous section, the best case performance of Approach (a) occurs when the actual service classes defined by a provider are an exact match to the potential service classes defined by the approach. In contrast, the worst case for the approach occurs when the boundaries of the actual service classes align with the centroids of the potential service classes. In this section, the performance of the two pattern passive approaches is evaluated in terms of the accuracy of predictions generated by each. The two scenarios are chosen to reflect the cases above and seek to establish whether the "dynamic" nature of Approach (b) is generally preferable to the "static" method of Approach (a).

In order to ensure a fair comparison of the two approaches the independent variable in the experiments that comprise each scenario is chosen as the size of the defined potential service classes since this may be directly controlled by adjusting the design time parameters of each approach. As a general observation it is likely that overall results for Approach (b) will not be significantly affected by specific service class policies since it has no inherent preference for particular distributions of service classes. In contrast, the performance of Approach (a) will be significantly different based on the arrangement of service classes.

6.2.2.1 Uniformly Distributed Service Class Boundaries

The case where the service classes defined by providers are evenly defined across E space (Figure 6.16, left) constitutes a best case for Approach (a) since these will correspond to the service class boundaries defined by the approach. In contrast, in cases where Approach (b) has a threshold value defined such that the predicted service class size is the same as for Approach (a) a match with actual service classes will only occur when the expectation of the consumer falls at the centroid of an actual service class (Figure 6.16, right). Approach (b) will generate predicted service classes that present a worst possible match with actual service classes when the expectation of the consumer falls at the boundary of an actual service class. In this case the approach will select an equal number of E space points from the actual service classes lying on either side of the boundary. Significantly, this suggests that the performance of Approach (b) will depend upon the expectation of the consumer requesting the assessment.

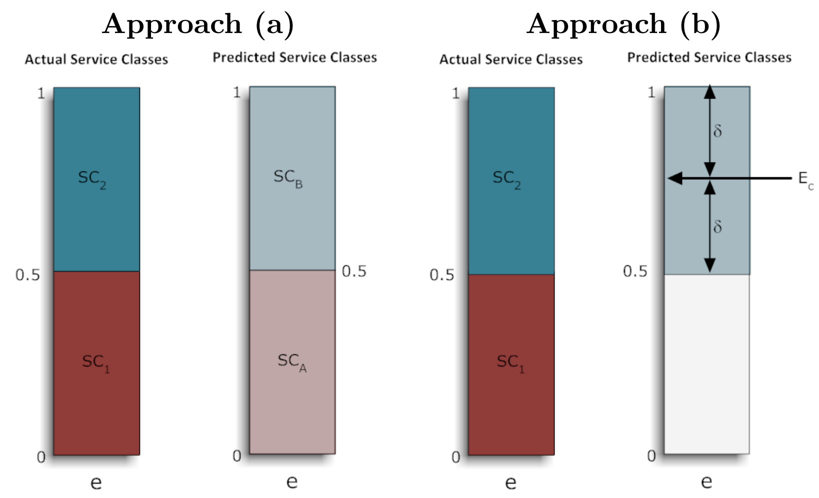


Figure 6.16: Uniformly Distributed Service Class Boundaries - relationship with Approach boundaries

SC	Approach A (n)	Approach B (δ)
2	2	0.250
3	3	0.167
4	4	0.125
5	5	0.100
6	6	0.083

Table 6.2: Relationship between service class numbers (Approach (a)) and threshold (Approach (b))

In this set of experiments a single service is used with behaviour defined such that the service classes are distributed evenly across the expectation range. The target level of service is specified at the centre of each service class with a small level of normal noise (standard deviation, σ^2=0.02) being introduced to each.

The number of service classes assumed by Approach (a) are aligned with the number of service classes offered by the provider in each experiment with the threshold value (which defines service class size) of Approach (b) being defined in order to ensure the same size ranges are used by each each approach and class size is therefore factored out of the experiment. The range of experiments in terms of number of service classes and threshold values are summarised in Table 6.2.

In each experiment, the predictive accuracy of both approaches is recorded in terms of the predictive error. That is, the difference between the level of service predicted by the approach and the level delivered to the consumer is recorded. For ease of representation the results are grouped into 0.1 width segments of expectation.

The results illustrate a clear difference in the accuracy of predictions of Approach (a) and Approach (b) in this scenario (see Figure 6.17). Significantly, the results

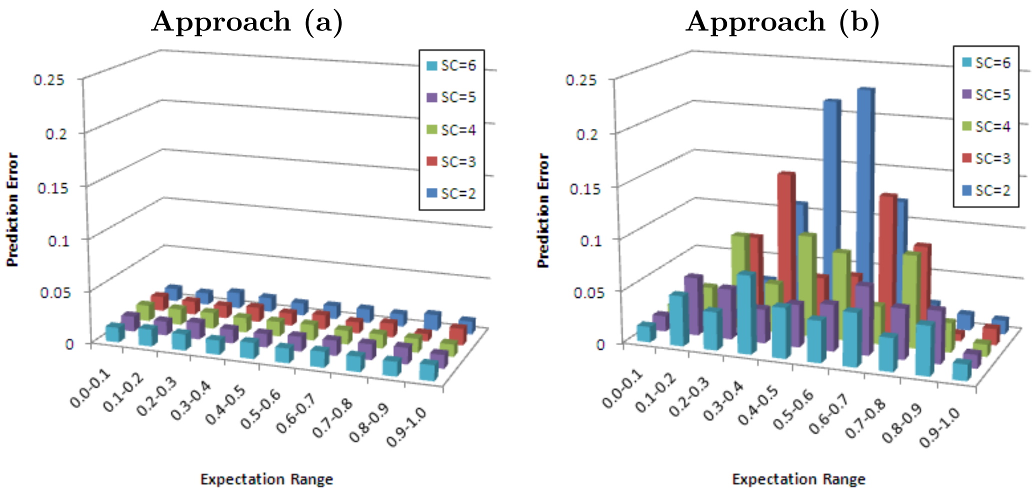


Figure 6.17: Comparison of Prediction Errors for Approach (a) and Approach (b) - Best Case

obtained for Approach (b) confirm that the accuracy of predictions will deteriorate as the consumer expectation moves closer to service class boundaries. In the experiment where SC=3, two peaks in predictive error are clearly evident in the 0.3-0.4 and 0.6-0.7 expectation ranges. This is unsurprising, since the actual service class boundaries of the service in this experiment fall at 0.333 and 0.666, meaning that assessment will draw a similar amount of points from each of the service classes in these cases.

The results for Approach (b) also indicate that the absolute error of prediction reduces significantly as the number of service classes increases. This is misleading since as the number of service classes increases the magnitude of potential error is equivalently reduced. Figure 6.18 illustrates the results for Approach (b) normalised against the size of actual service classes for each experiment showing that there is no significant decrease in the magnitude of prediction error between service classes.

Overall, the results obtained under this scenario illustrate a significant increase in accuracy of predictions of Approach (a), when compared to Approach (b). However, confidence in the overall effectiveness of the approach is not assured based on this observation since the scenario represents a contrived best case for Approach (a), that whilst being realistic is not representative of the range of behaviours that may be adopted by a service.

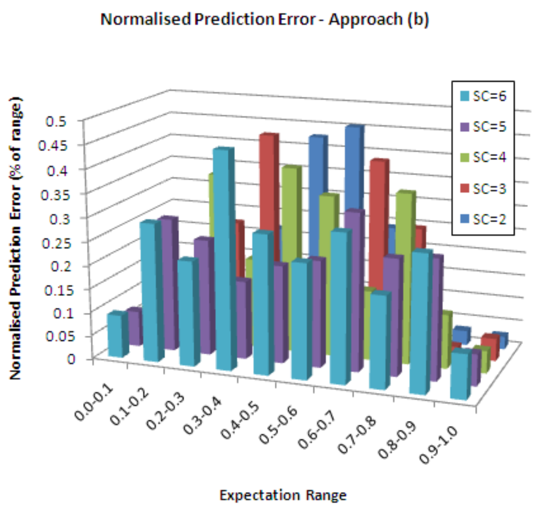


Figure 6.18: Normalised Prediction Error for Approach (b) - Best Case

6.2.2.2 Inverted Service Class Boundaries

The worst case performance of Approach (a) occurs where the service classes predicted by the approach are the "inverse" of the actual service classes defined by the provider (see Figure 6.19, left). In this case, when making an assessment the approach will gather points from the two service classes on either sides of the boundary leading to predictions that are not an accurate reflection of the behaviour in either of the classes. A useful way to simulate such a situation is to specify the number of service classes assumed by the approach as one greater than the actual number of service classes. It should be noted that the inconsistency between the assumed number of service classes and the actual service classes is made in this case in order to simulate a poor match. In this case, the threshold parameter (δ) used to define the size of potential service classes by Approach (b) is adjusted to ensure that same size of potential service classes by Approach (a) (see Figure 6.19, right).

As with the previous scenario multiple experiments will be performed for a single provider and a number of service classes with the service class range being the same for each of the two Approaches by the adjustment of parameters.

The results indicated in Figure 6.20 show that Approach (b) provides generally more accurate predictions across the entire expectation range for the various numbers of service classes used in the experiments. A general improvement in absolute predictive accuracy with increasing number of service classes – which is similar to that exhibited by Approach (b) in the previous experiment. This is due to the potential maximum error being reduced as described previously.

Approach (b) exhibits a general trend in having small predictive errors at the

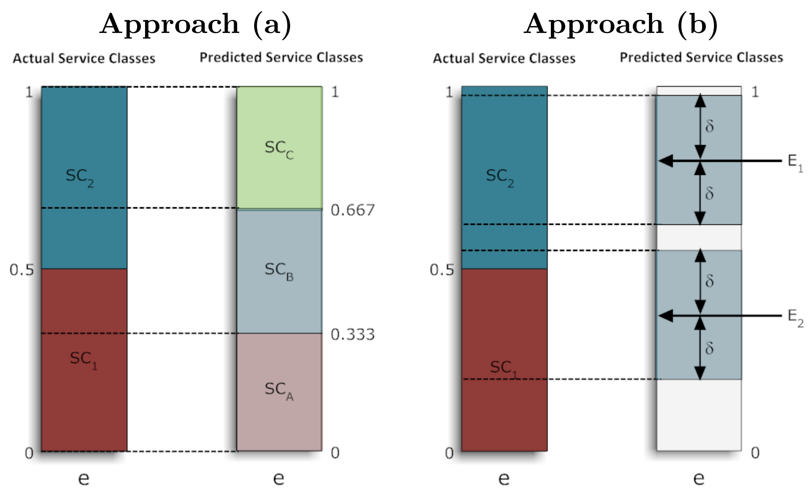


Figure 6.19: Inverse Distribution of Service Class Boundaries - relationship with Approach boundaries

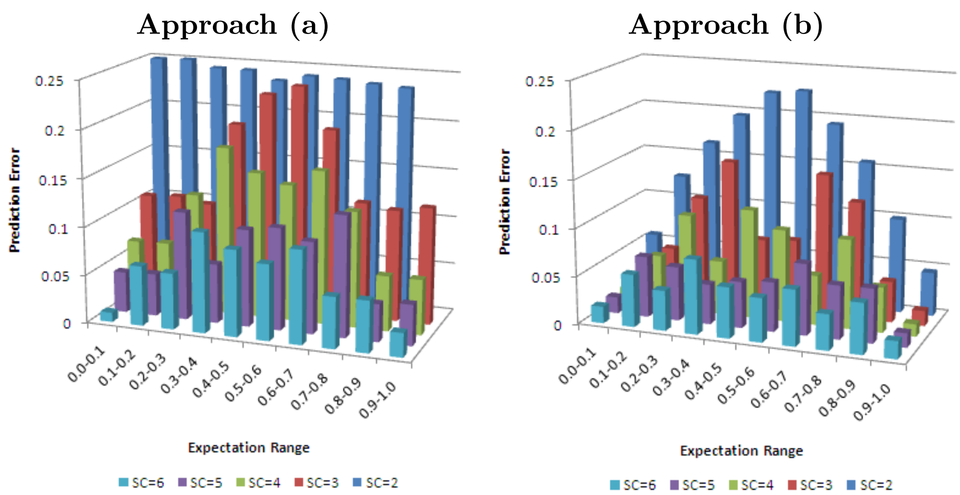


Figure 6.20: Comparison of Prediction Errors for Approach (a) and Approach (b) - Worst Case

boundaries of the expectation range (the subranges 0.0-0.1 and 0.9-1.0). This is a direct result of the dynamic expectation range created by the approach using the expectation of the current consumer and threshold value to define the width of the range. In cases where the consumers expectation falls close to upper and lower expectation limits the range defined by the approach is bordered by only one other service class and the likelihood of including points from a second range in the assessment is therefore reduced. The overall predictive error of Approach (b), normalised by service class size across all of the above experiments was 0.263. This compares very favourably against the same value for Approach (a) of 0.430.

6.2.2.3 Summary

The two scenarios evaluated in this section constitute the best and worst case performance of Approach (a), that is wholly dependent upon the relationship of its fixed potential service classes to the actual service classes of providers. Although the predictive error of Approach (a) in its best case is minimal and shows significantly higher accuracy than Approach (b) this occurs only when the service class boundaries predicted by the approach exactly match those of the service. It is very unlikely that this best case will occur frequently in reality since it relies on a very specific and regular arrangement of service classes.

In contrast, the worst case for Approach (a) occurring with an "inverse" match between the two sets of service classes where the favourable performance of Approach (b) constituted a large improvement over Approach (a). Although contrived in order to demonstrate the worst case for Approach (a) the policy adopted by the service in this scenario can be considered more indicative of that which might be typically expected from services.

Comparison of the performance of Approach (b) across the two scenarios suggests that the approach is generally insensitive to the policy adopted by providers when defined in terms of a uniform service class size. Since the size of each range is essentially determined by choosing boundary points on the expectation range it is reasonable to extrapolate that the performance of Approach (b) in the two scenarios above can be considered as being an average case across all provider service class policies.

6.2.3 Performance of Pattern Active Approaches

Pattern passive approaches are naïve and dependent parameters that are defined at design time. Predictive accuracy in pattern passive approaches is dependent upon the match between potential and actual service classes. They do not examine provider performance data in establishing behaviour. This implies two assumptions for the approach to be effective: that all providers define the same service class boundaries and that the boundaries do not change with time. The first assumption means that the approach cannot effectively identify individual differences and will lead to inaccurate predictions in many cases; the second that if a provider changes its service policy with respect to service classes the performance of pattern passive approaches may degrade unpredictably.

Pattern active approaches should be more robust under such changing behaviour since they seek to recognise patterns in the data specific to the provider in question and use this to infer service classes. Two pattern active approaches were defined in Sections 4.6.1 and 4.6.4. The first of these, which is intended to be used where the number of service classes is known and uses a k-means clustering algorithm is evaluated here and referred to as Approach (c).

First, the ability of pattern active approaches to overcome the problems experienced in the worst case for Approach (a) and Approach (b) is demonstrated. Pattern active approaches are not affected substantially by the boundaries of service classes as defined in e-space while pattern passive approaches are. However, they are influenced by variation in the performance data which may be related to noise introduced during monitoring or due to a poor service. The second scenario in this section seeks to identify how the performance of a pattern active approach is affected by noisy data.

6.2.3.1 Comparison of Pattern Active and Pattern Passive Approaches

It is unlikely that Approach (c) will be able to match the predictive accuracy of Approach (a) in its best case since this would require the k-means clustering algorithm to consistently recognise perfect clusters. Due to the clustering issues described in Section 4.6.1.1 it is unlikely that this will happen. However, in all other cases Approach (c) should significantly outperform Approach (a) and Approach (b) in terms of predictive accuracy. This performance is aided by the fact that the service delivery data used in the scenario has low level noise and there are clear separation between the target service levels for each service class. As with the scenarios above the assumption is made that Approach (c) is aware of the number of service classes a priori - this value being used as the k for the k-means clustering algorithm (see Section 4.6.1.4).

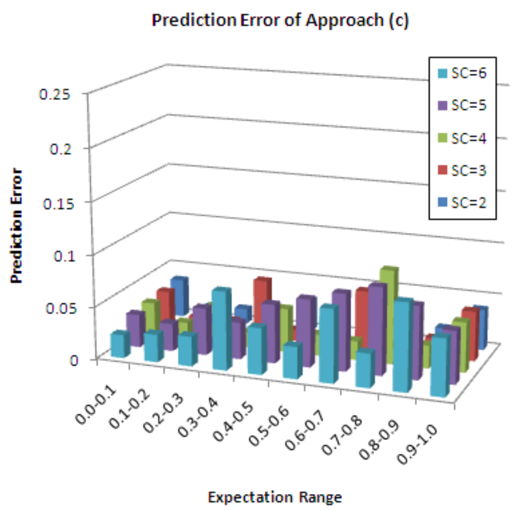


Figure 6.21: Prediction Error for Approach (c) - Worst Case

The performance of Approach (c) (see Figure 6.21) clearly demonstrates improvement over Approach (a) in all experiments other than the best case for Approach (a), which was expected. In the majority of cases Approach (c) also outperforms Approach (b) in terms of predictive accuracy. Significantly, the degree of improvements that can be gained by using Approach (c) are particularly evident when a lower number of service classes are defined by the provider.

Whilst Approach (c) outperforms Approach (b) in general there are two observations that may be made. Firstly, the performance of Approach (b) betters that of Approach (c) in cases where the expectation is close to the end of the range. This is linked to the observation made in the last section that for expectation near the range limits there is less potential for Approach (b) to consider incorrect data in the result. Secondly, although there is a large difference between the performance of the approaches for a lower number of service classes the difference is reduced when SC is larger. This is likely inherent to the nature of the clustering used by Approach (c) - as the number of service classes increases the distinction between the target levels of service for each also decreases meaning that the clusters identified by the algorithm are less distinct. In the case of SC=6, where the service class size is 0.167 the noise (with σ^2=0.02) added to the monitored data may be enough to lead to the incorrect identification of clusters.

As a general approach, when considered across the range of experiments which make up the two scenarios Approach (c) is generally preferable since it remains largely unaffected by the number of service classes and their boundaries whereas Approach (a) and Approach (b) performance varies depending upon the expectation of the consumer and the service class boundaries. An approach which consistently provides low error predictions is generally preferable to one whose performance is so heavily tied to the characteristics of the underlying data.

6.2.3.2 Effect of Noise on Accuracy and Confidence of Pattern Active Approaches

Approach (c) introduces the silhouette coefficient (see Section 4.3.1.3) as a validity measure rather than the normalised cohesion used by the pattern passive approaches. It serves the same purpose, in acting as a confidence measure that the prediction is correct. The accuracy should not be affected by boundaries of classes defined by the provider since these are determined based on analysis of the data. However, prediction will depend upon the degree to which the clustering algorithm is able to recognise these classes from the data. Clustering may be adversely affected by decreasing the distance between the service levels of each class and increasing the level of intra-class noise in the data. The effect of the former was evident in the results in the previous section. The effect of decreasing this distance is considered in the next section where the approach is evaluated for providers with various numbers of service classes.

In this section, a set of experiments are conducted to determine the effect of noise on the predictive accuracy of Approach (c). In the series of experiments, a single provider with four uniformly distributed service classes is added to the environment. The target level of service for each of the service classes is specified at the centre of each and a level of noise sampled from a normal distribution is added to the delivered level of service. The standard deviation of the normal distribution from which the noise is generated is adjusted across the experiments. The levels of noise used are selected in order to provide a range of situations from those where there is clear separation to those where there is substantial overlap between the levels of service delivered by individual service classes (Figure 6.22). Carrying out the experiments in this scenario serves two purposes. Firstly, by determining the effect of the degree of noise on the predictive accuracy of Approach (c), it becomes possible to establish important limits on its effectiveness and the situations in which these limits occur. Secondly, by establishing the accuracy of predictions in each case and comparing this with the attached validity measure, it is possible to gain confidence in the use of the silhouette coefficient as a reasonable measure of uncertainty. As with previous experiments the performance of the approach will be evaluated in terms of the accuracy and validity of predictions to consumers with expectation across the e-space range.

The results, recorded in Figure 6.23, Figure 6.24 and Figure 6.25 shows the predictive accuracy and validity of the approach with three levels of noise and illustrate a clear and continuous decrease in accuracy of predictions as the amount of noise added to the data increases (see Figure 6.23,Figure 6.24 and Figure 6.25). Significantly, the predictive accuracy of the approach remains substantially high (0.9-1.0) with low variance in levels of noise up to σ^2=0.04. The relatively substantial degradation in accuracy of prediction which occurs at σ^2=0.07 (see Figure 6.25) may be attributed to the fact that at this point the data from service classes begins to exhibit a high degree of overlap yielding clusters that are a poor match to actual service classes. As a general observation this is entirely expected, since clustering algorithms are unlikely to be able to detect clusters whose boundaries overlap substantially. It is evident that the approach copes reasonably well with the degree of overlap exhibited by a σ^2=0.04 level of noise.

The correlation between the two assessment measures is further illustrated in Figure 6.26. in which the accuracy of each prediction is plotted against the validity measure.

The generally positively linear relationship between prediction accuracy and validity measure is a reasonable indicator that the silhouette coefficient is a suitable

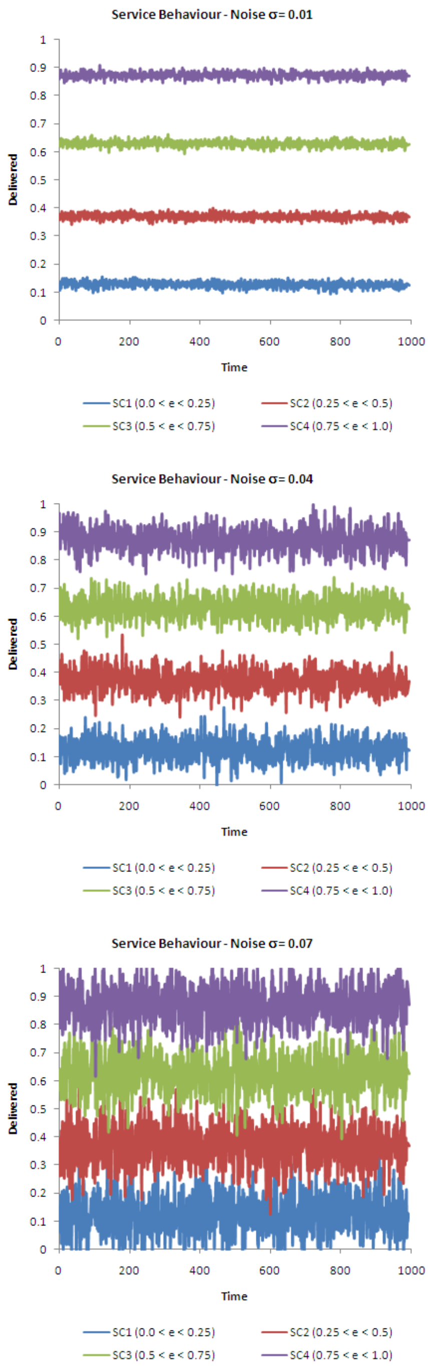


Figure 6.22: Service provision affected by three different levels of noise

measure of cluster validity.

6.2.3.3 Summary

Conceptually, a pattern active approach offers a technical advantage over pattern passive approaches since it examines the historic performance data of each service in order to dynamically establish potential service class boundaries. Compared

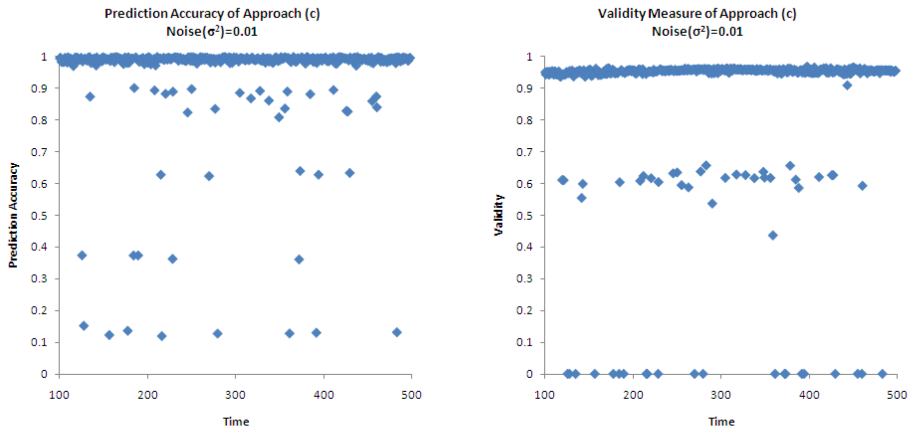


Figure 6.23: The Effect of Noise on Prediction Accuracy and Validity Measure of Approach (c) - Noise = 0.01

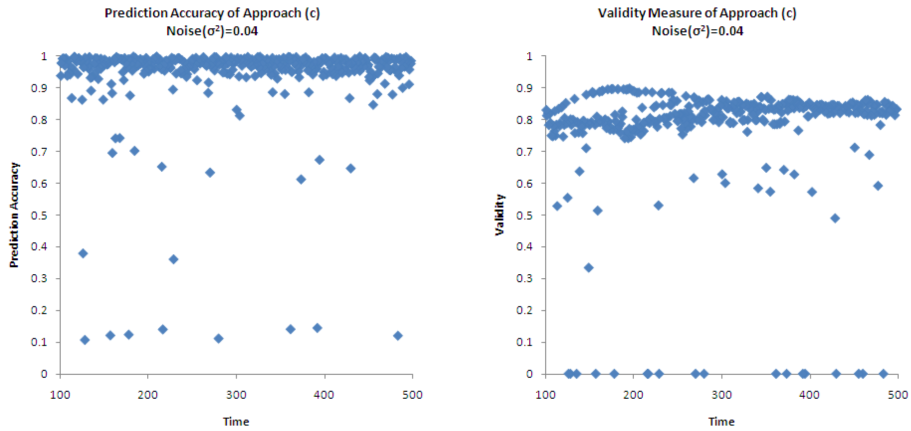


Figure 6.24: The Effect of Noise on Prediction Accuracy and Validity Measure of Approach (c) - Noise = 0.04

to the pattern passive approaches, Approach (c) performs favourably in terms of the accuracy of its predictions. The approach is generally more robust in that it exhibits consistently lower predictive error. Whilst being comparable to Approach (c) in some instances performance of Approach (b) degrades for a lower number of service classes and for expectations in potential classes near the boundaries of actual service classes.

As expected, the addition of noise to assessment data led to a general decrease in accuracy of predictions proportional to the amount of noise. Such an effect is inevitable and it is realistic to expect that in some cases high levels of noise will be present in assessment data. However, the results indicate two positive aspects of Approach (c) in this respect. Firstly, the approach shows a tolerance to noise which constitutes a partial overlap in the delivered level of service of neighbouring

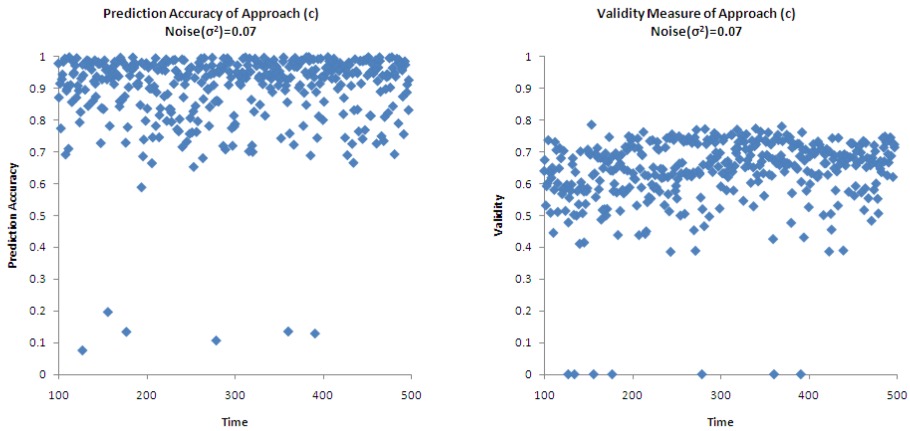


Figure 6.25: The Effect of Noise on Prediction Accuracy and Validity Measure of Approach (c) - Noise = 0.07

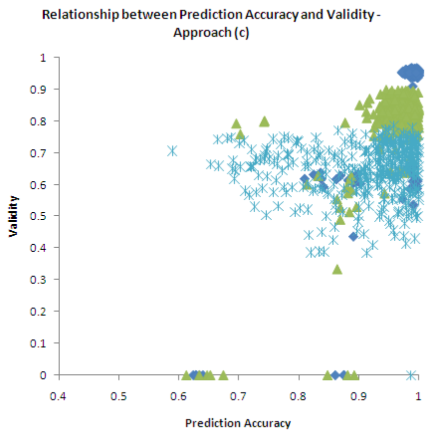


Figure 6.26: Relationship between Prediction Accuracy and Validity for Approach (c)

service classes. Secondly, the silhouette coefficient measure adds robustness to the approach in that the accuracy of the prediction degrades gracefully with the validity measure in the presence of increasing noise. As demonstrated in Section 6.2.1.4 the effectiveness of a validity measure is a key aspect in supporting consumers in selecting between candidate services.

6.2.4 Comparison of Pattern Active Approaches under Fully-Private Policies

The approaches evaluated as Approach (a), Approach (b) and Approach (c) in the previous sections are designed to operate in an SOA where the number of service classes (N) is known a priori. Such information provides the basis for establishing parameters for the assessment algorithms in order to increase the quality of the match between the potential service classes identified by the approach and the actual service classes defined by the provider. In the case of Approach (a) and Approach (c) the number of service classes is used directly in choosing the number of segments into which to divide the expectation space. For Approach (b) the number of service classes is used indirectly to establish the size of potential service classes. Without knowledge of the number of service classes each of these approaches would produce lower accuracy predictions in most cases.

In contrast, the hierarchical clustering approach (see Section 4.6.4) to assessment is designed to operate in environments where providers have fully private service policies - that is, where the assumption of a priori knowledge of N is not made. The approach described here as Approach (d), which utilises an unsupervised hierarchical clustering algorithm determines the most likely number of actual service classes based solely on analysis of the historic performance data. Whilst constituting a more generally applicable approach, as with Approach (c), Approach (d) is inevitably susceptible to issues of data quality such as noise.

In this section, Approach (d) is evaluated in three scenarios to establish its performance in terms of its ability to assess services with fully-private policies. Firstly, a scenario is provided to illustrate the effect of relaxing the assumption of a priori knowledge of N on existing approaches in terms of predictive accuracy. Secondly, a set of experiments are used to show the ability of Approach (d) in detecting a range of numbers and arrangement of service classes. Finally, the effect of noise on the assessment approach is evaluated using a similar scenario to the one used in Section 6.2.3.2 to evaluate Approach (c).

6.2.4.1 Performance of Approaches under Fully-Private Policies

The previous section established Approach (c) as the most generally applicable approach to assessment of providers where N is known a priori. In this experiment a basic scenario in which the number of service classes assumed by Approach (c) is different from the actual number of service classes is presented and evaluated. The performance of Approach (c) will be compared to that of Approach (d) in terms of the predictive accuracy of assessment results. Approach (a) and Approach (b) will also be evaluated in the scenario though the emphasis in evaluation will be upon Approach (c) and Approach (d).

In the experiment, a single provider with five uniformly distributed service classes each with a constant target level of service at the middle of the class (Figure 6.27) is added to the environment. In order to ensure that the difference in performance of the approaches is solely due to the mismatch between k (the number of clusters assumed by the k-means approach) and N, and not due to a difference in the

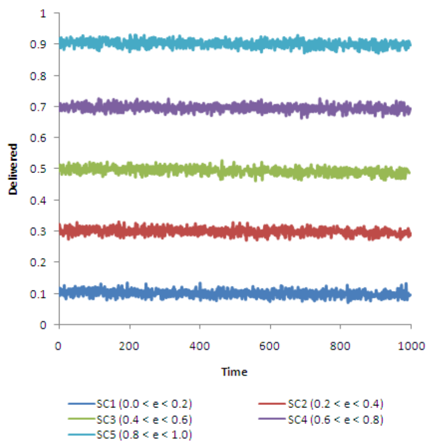


Figure 6.27: Behaviour of a Service with 5 Service Classes

approaches tolerance to poor quality data, a minimal amount of noise (σ^2=0.01) will be added to the data. In Section 6.2.3.2 it was demonstrated that this level of noise does not have a significantly adverse effect on the ability of Approach (c) to recognise service classes.

For the three approaches evaluated, k is assumed to be two which will lead to substantial inconsistency between the potential and actual service classes. It is pertinent to note that it is likely that the technical complexity of Approach (c) when compared to Approach (a) and Approach (b) may actually lead to the former under-performing the latter two in terms of predictive accuracy in the scenario described. This is due to the number of actual service classes that may be covered by the potential service classes identified by each approach. For Approach (a) and Approach (b) the number of actual service classes used in making any assessment is at least two and at most four. However, for Approach (c) it is possible that points from five actual service classes might be identified as part of one of the two clusters identified by the algorithm. Figure 6.28 illustrates the worst case for each approach (as just described) in terms of the relationship between predicted and actual service classes.

The results illustrated in Figure 6.29 show the predictive accuracy of the four approaches for the duration of the simulation. Approach (b) achieves significantly better performance than either Approach (a) or Approach (c) with lower variance and generally higher accuracy though still deviates from the optimal. The two most poorly performing approaches are Approach (a) and Approach (c) which exhibit clearly and consistently lower accuracy predictions and higher variance across the duration of the simulation.

The results for Approach (d) show consistently high accuracy with few variations. However, in instances where variation does occur the predictive accuracy is

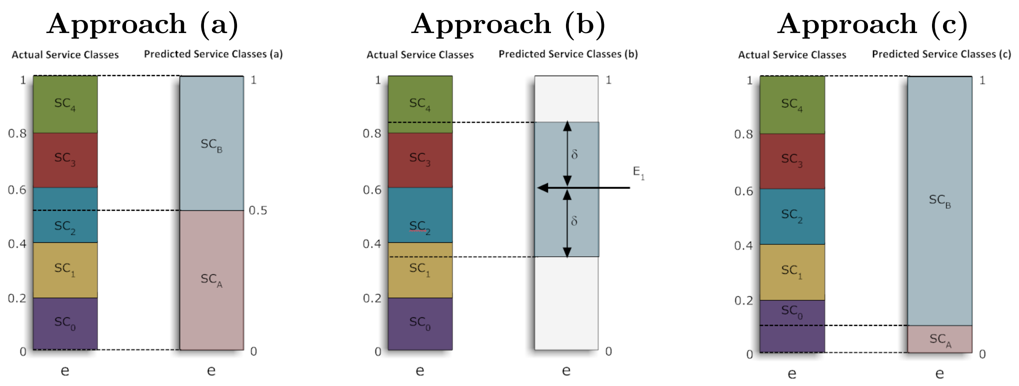


Figure 6.28: Poor Mappings for Approach (a), Approach (b) and Approach (c)

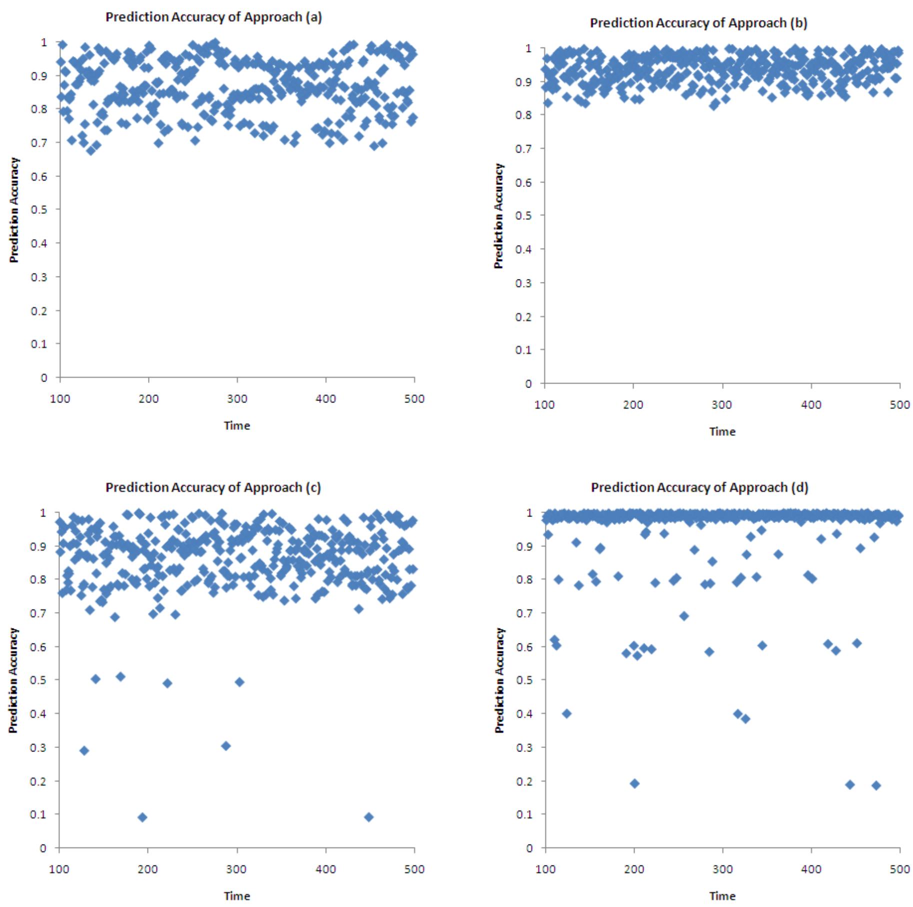


Figure 6.29: Effect of Poorly Matching Service Classes on Accuracy (Good Match)

generally lower than for any of the predictions by the other approaches. This reflects cases where the stopping criterion employed by the algorithm to determine the number of service classes is ineffective either leading to the wrong number of service classes being identified or reaching the maximum number of divisive iterations of the algorithm. Overall, the results show a clear increase in performance of Approach (d) over the others in cases where N is unknown and where approaches assume a different number of potential service classes to the number that are defined by the provider.

6.2.4.2 Detecting Various Fully-Private Service Policies

The results presented in the previous section demonstrate the weaknesses of approaches to assessment that assume knowledge of N. This is perhaps an unfair comparison since the assumption of the approach is reasonable in certain situations. Evaluation of the previous section results might conclude that a different approach to assessment should be used depending upon whether the assumption of N is made. That is, Approach (c) which provides generally more accurate predictions than Approach (a) and Approach (b) should be used where the number of service classes is known and Approach (d) should be used where the number of service classes is unknown. However, such a statement is not justifiable without first comparing the performance of Approach (c) and Approach (d) in a range of scenarios where the number of service classes is known to Approach (c) a priori.

This scenario serves to achieve two purposes. Firstly, to demonstrate the performance of Approach (d) in accurately predicting the performance of services with a range of service class policies through effectively identifying service class boundaries. Secondly, to compare the performance of Approach (d) with Approach (c) where the latter is aware of the number of service classes. The latter comparison will provide substantial evidence to justify the statement that the preferable approach is different depending on the availability of information relating to the number of service classes.

This scenario consists of a number of individual experiments in which a single provider is added to the environment with a varying number and arrangement of service classes in each experiment. As previously, the target level of service is approximately constant and a level of noise is added to the data which is small relative to the size of the service classes. Figure 6.30 shows the behaviour of the service in each experiment (the experiments are identified as (i) through (v)).
The number of service classes in each experiment is made available to Approach (c) though Approach (d) does not use this information to improve the assessment process. That is, it relies entirely on the stopping criterion to decide when an appropriate number of potential service classes has been reached. Since the scenario seeks to compare the general performance of the approaches across the range of consumer expectation the predictive accuracy of all assessments will be recorded. That is, the performance of the approaches will not be compared in terms of the predictive accuracy of assessments for consumers in specific ranges of expectation.

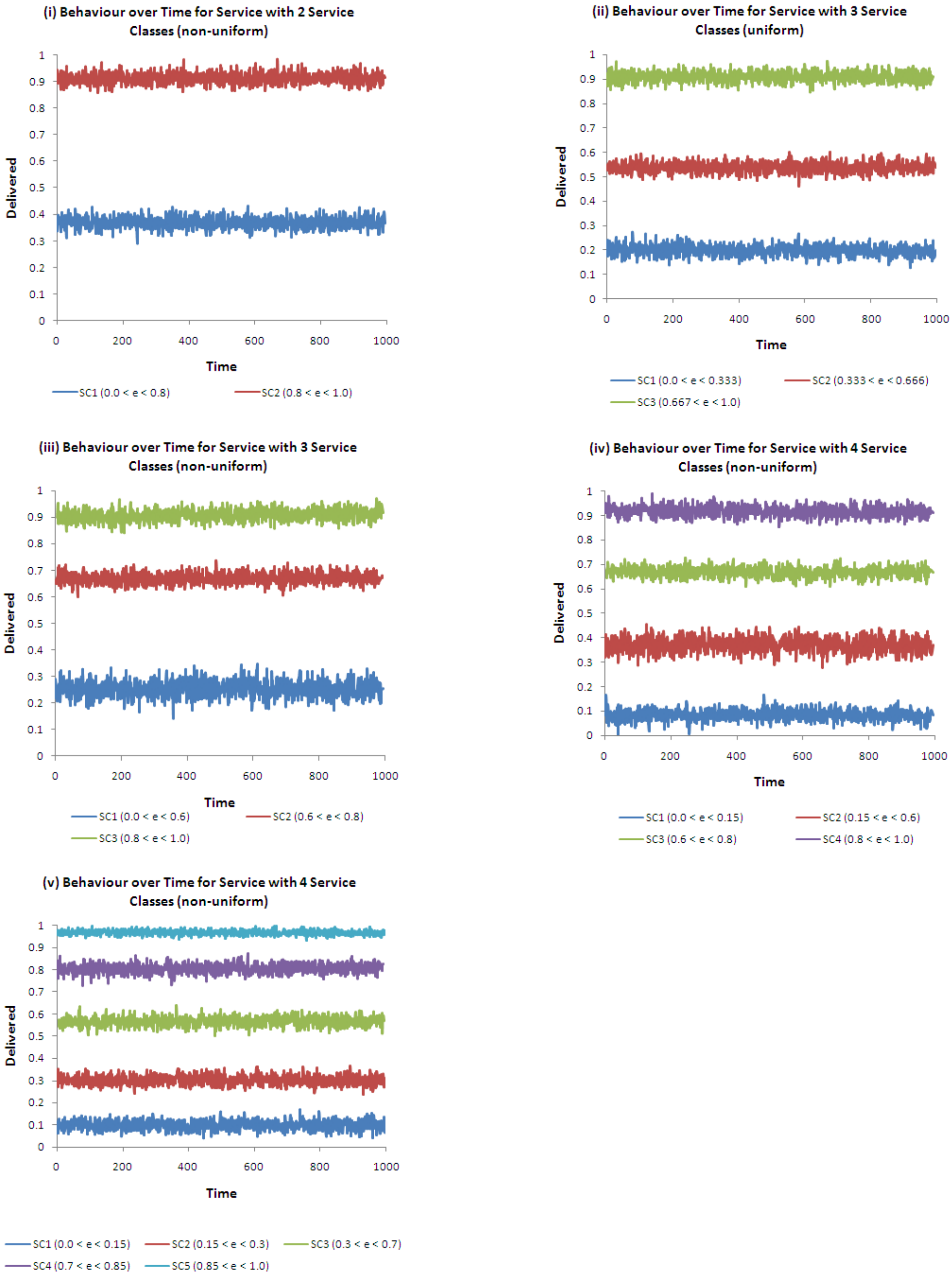


Figure 6.30: Behaviour over time graphs for Services with various Service Class Distribution

The results (Figure 6.31 and Figure 6.32) show little significant difference in the performance of each of the approaches for the majority of assessment results across the range of expectation. In the cases where there are two service classes (Figure 6.31 (i)), three uniformly distributed service classes classes (Figure 6.31 (iii)) and four service classes (Figure 6.32 (iv)), Approach (d) shows a significantly lower variance in predictive accuracy than Approach (c). In the other two cases

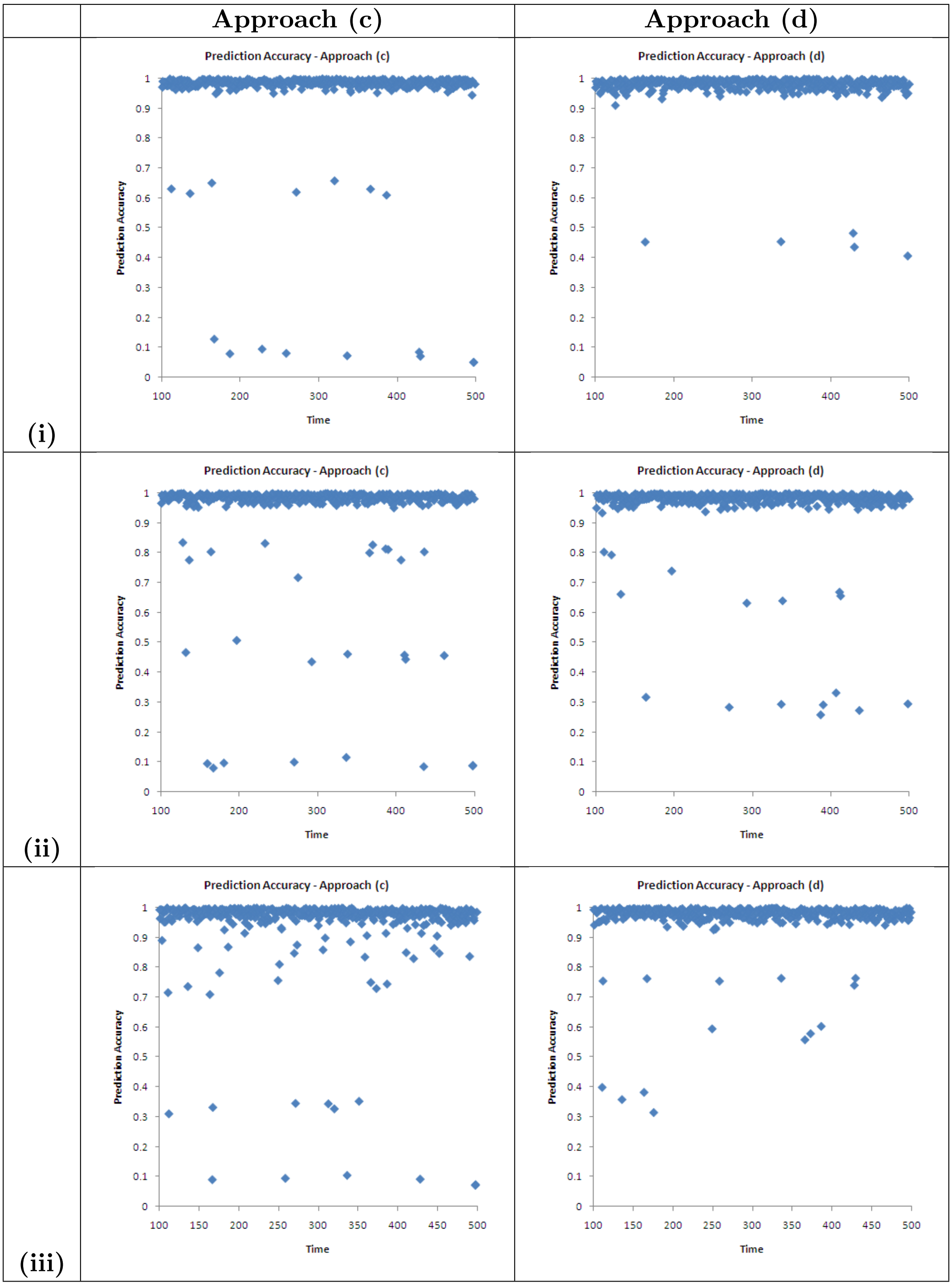


Figure 6.31: Comparison of Prediction Accuracy of Approach (c) and Approach (d) for Services with 2 and 3 Service Classes

where there are three non-uniformly distributed service classes (Figure 6.31 (ii)) and five service classes (Figure 6.31 (v)) Approach (d) outperforms Approach (c) to a lesser extent. The overall set of results demonstrates that the approach is more consistent in producing higher accuracy predictions in these cases. Table 6.4 compares the average predictive accuracy and variance of accuracy across the five

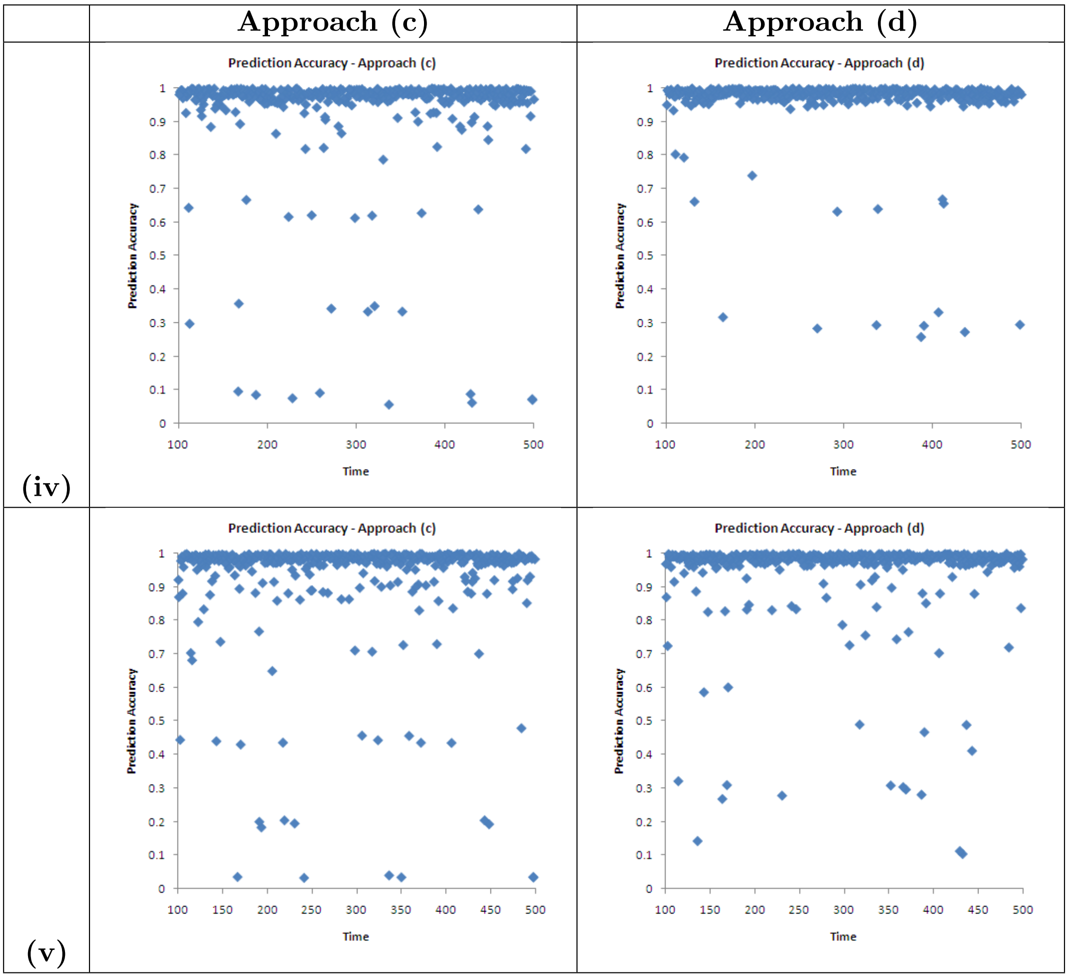


Figure 6.32: Comparison of Prediction Accuracy of Approach (c) and Approach (d) for Services with 4 and 5 Service Classes

experiments. The results suggest that Approach (d) is the approach of preference

SC	$\mu_{(a)}$	$\sigma_{(a)}$	$\mu_{(b)}$	$\sigma_{(b)}$
SC=2	0.961	0.134	0.970	0.078
SC=3	0.958	0.136	0.967	0.095
SC=3	0.946	0.144	0.967	0.081
SC=4	0.935	0.158	0.967	0.095
SC=5	0.933	0.164	0.949	0.141

Table 6.3: Accuracy and Variance of Approach (c) and Approach (d)

regardless of whether N is known a priori or not. This statement is based purely on comparison of performance of the scenarios described and differences in variance of predictive accuracy and may not be generalisable. However, the results do suggest that there is not a significant difference between the two approaches in terms of predictive accuracy of assessment results.

6.2.4.3 Effect of Noise on Accuracy and Confidence

As with any clustering technique the effectiveness of the hierarchical clustering algorithm employed by Approach (d) is susceptible to noise in the data sample as well as to general properties of the data including cluster shape, separation and cohesion. In the previous scenario the performance of the approach was evaluated using a range of behaviours which affected the number and distribution of clusters with noise maintained at a tolerable level. In this scenario, the effect of noise on the performance of Approach (d) is evaluated. Specifically, this scenario seeks to demonstrate how the predictive accuracy of Approach (d) is affected by the range of levels of noise in the data used previously to evaluate Approach (c) (see Section 6.2.3.2). As previously, the inevitable decrease in predictive accuracy of the approach should coincide with a relative decrease in the measured validity of the result.

This scenario seeks to demonstrate three aspects of the approach. Firstly, by establishing the tolerance of the hierarchical clustering algorithm to noise in the sample data the limitations of the approach are established. Secondly, comparing the predictive accuracy and prediction validity will provide further support for the robustness of the approach as it recognises cases where there are data quality issues. Finally, using the same simulation setup that was used to evaluate Approach (c) in Section 6.2.3.2 allows observations to be made in terms of the comparative ability of the two approaches to deal with noise.

As in previous experiments Approach (c) is provided with the value of N prior to the experiment in order to allow an objective comparison. The scenario setup in terms of provider behaviour and assessment events are maintained exactly as described in Section 6.2.3.2.

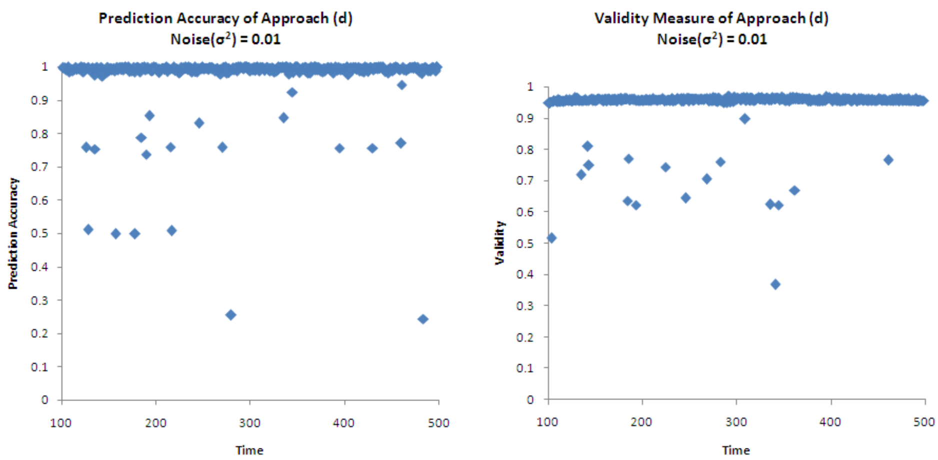


Figure 6.33: The Effect of Noise on Prediction Accuracy of Approach (d) - Noise = 0.01

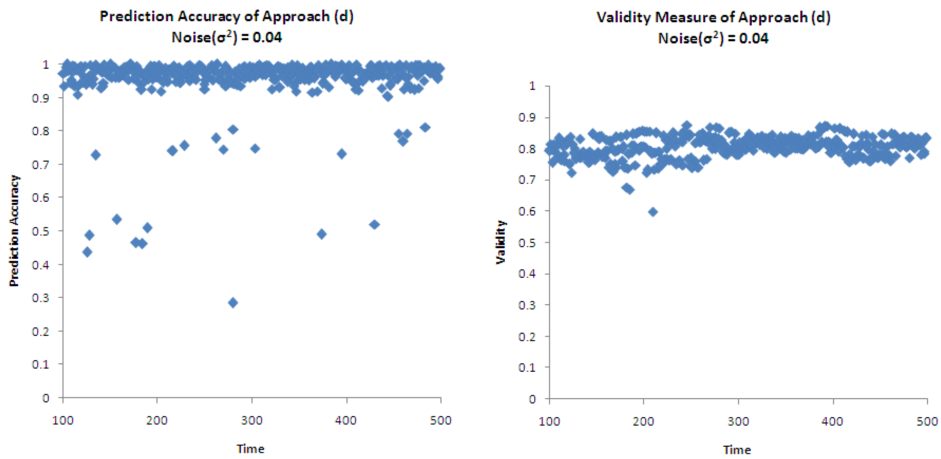


Figure 6.34: The Effect of Noise on Prediction Accuracy and Validity Measure of Approach (d) - Noise = 0.04

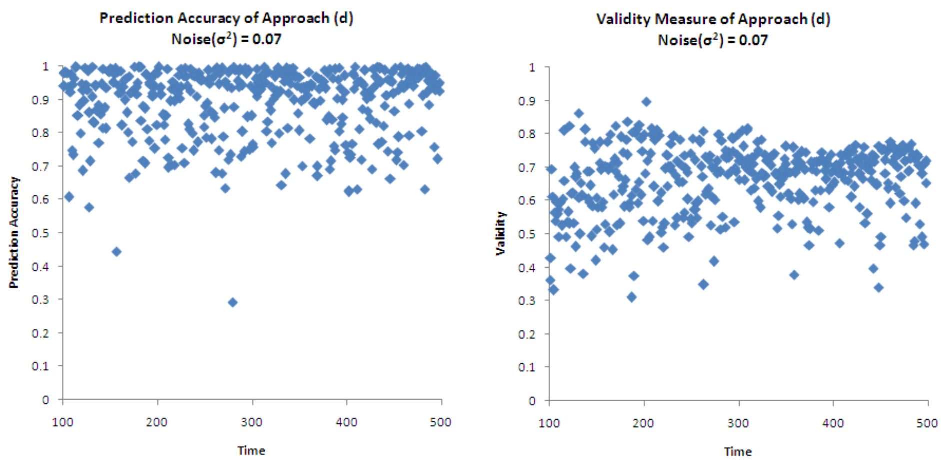


Figure 6.35: The Effect of Noise on Prediction Accuracy and Validity Measure of Approach (d) - Noise = 0.07

The experimental results (see Figures 6.33, 6.34, 6.35) appear visually to be very similar to those obtained for Approach (c) (see Figure 6.23, Figure 6.24 and Figure 6.25). Quantitatively, the results show that a small improvement in predictive accuracy is realised by Approach (d) which also has variance of between 20% and 45% less than Approach (c) suggesting that the results of the former are a more reliable indicator of general performance. Table 6.4 provides an overall comparison of Approach (c) and Approach (d) in terms of accuracy and validity with various levels of noise.

In terms of establishing the relationship between predictive accuracy and validity of assessment results, Figure 6.36 illustrates the mean accuracy against mean validity for each of the experiments. The second series on the graph denoting the equivalent correlation for Approach (c) shows the consistency between the two,

Noise	$\mu_{(c)}$	$\mu_{(d)}$	$\sigma_{(c)}$	$\sigma_{(d)}$
0.01	0.956	0.978	0.145	0.080
0.02	0.953	0.963	0.142	0.102
0.04	0.943	0.954	0.135	0.086
0.05	0.924	0.943	0.126	0.097
0.07	0.898	0.898	0.128	0.104
0.10	0.846	0.873	0.136	0.110

Table 6.4: Accuracy and Variance of Approach (c) and Approach (d)

which is understandable since the approaches both use the silhouette coefficient in evaluating the quality of clusters. The results illustrate a desirable quality of

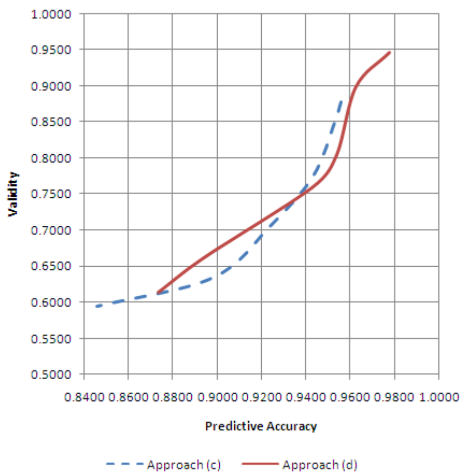


Figure 6.36: The Relationship between Prediction Accuracy and Validity for Approach (c) and Approach (d)

both approaches – their graceful degradation in the face of poorer quality data which will support a consumer in using the assessment result as evidence in their decision making. Furthermore, the approach has been shown to be comparable to Approach (c) in terms of the level of accuracy achieved in prediction even when allowing Approach (c) to know N a priori. The approaches are also comparable in terms of the degree to which noise affects predictive accuracy and validity.

6.2.4.4 Summary

The predictive accuracy of Approaches (a), (b) and (c) is significantly diminished when operating in environments where their assumption of N is incorrect reinforcing the need for an approach that is able to infer the number of service classes from the data when providers adopt a fully private policy. In the first two sets of experiments in this section Approach (d) was demonstrated to be effective in this role across a range of arrangements and numbers of service classes indicating that it can be generally applied to problems of this type.

When Approach (c) and Approach (d) are compared in terms of predictive accuracy Approach (d) outperforms Approach(c) in a range of scenarios. Significantly, a small increase in performance is evident even when Approach (c) is provided with the number of service classes a priori. This increase in performance may be attributed to the quality of the clusters identified by the approaches suggesting that for problems of this type hierarchical clustering is generally better at identifying clusters than a k-means algorithm. However, as is often the case with clustering there is a tradeoff between quality of clusters and computational complexity. Based on the analysis of the complexity, it is likely that Approach (c) is more scalable than Approach (d) though this has not been evaluated empirically.

When evaluated in terms of the effects of noise Approach (d) performs comparably to Approach (c). The set of experiments and their results provide confidence in the robustness of the approach in terms of withstanding a degree of noise in the data and the appropriateness of the silhouette coefficient in reflecting any decreases in predictive accuracy due to poor data quality in the validity measure.

The results presented in this section support the conclusion that Approach (c) may be appropriately applied in environments where providers adopt fully private service policies.

6.3 Evaluation Summary

This chapter presented the results of an evaluation of the four approaches developed as part of this thesis. The three criteria presented at the start of the chapter (namely accuracy, validity and experienced quality) define the outputs and outcomes of any assessment approach that must be measured for any approach in order to demonstrate its effectiveness. The set of scenarios which followed constitute a collection of experiments that justified in Section 6.1.2 as being sufficient to demonstrate the general effectiveness of the approaches (or otherwise). Specifically, the scenarios sought to provide an incremental approach to the evaluation process demonstrating the contribution of fundamental aspects of the general ED space approaches before isolating and evaluating particular factors and then evaluating the performance of approaches under a range of service behaviours often relating to the worst case and best case for particular approaches.

The experiments carried out within the scenarios demonstrated the effectiveness of each approach as well as identifying aspects of the environment that affect their performance. In evaluating the results comparisons were made against one of three baselines: a traditional D space assessment approach, a random selection approach and in group comparisons where the individual approaches developed were compared against each other. Specifically, the results observed across the range of experiments led to the following observations being made:

- Traditional D Space approaches to assessment are wholly inappropriate for use in a conformance-centric SOA where providers adopt a differentiated services approach by segmenting consumers into service classses. Even the most naive ED space approaches showed significant increases in performance over a traditional D space approach.

- Increasing the predictive accuracy of approaches may lead to the consumer experiencing higher quality services. However, in isolation high predictive accuracy will not necessarily lead to this outcome. Consumers' decision making is optimised when consumers are able to consider the predictions of approaches in the context of their associated validity measures.

- The pattern passive approaches (a) and (b) are both sensitive to properties relating to the expectation space. For Approach (a) the provider-defined service class boundaries and their relationship with the predicted service classes has a significant effect. Approach (b) is less affected by service class boundaries but its performance is dependent upon the expectation of the consumer requesting the assessment. Since in each case this leads to substantial inconsistency in quality of results neither approach can be considered ideal. However, Approach (b) may be considered as generally preferable based on its favourable performance in later scenarios.

- Approach (c) which used a k-means clustering algorithm to identify service class boundaries was able to overcome weaknesses inherent in the more naïve pattern passive approaches. The approach appears to be less sensitive to properties of E space than its counterparts since it makes less prior assumptions about the distribution of points in this space instead inferring them from D space data.

- Approach (d) is capable of operating in environments where no prior assumptions are made regarding the providers service class policy. The approach was demonstrated as being capable of effectively identifying service classes of varying number and arrangement.

- The *normalised cohesion statistic* was established as being appropriate as measures of prediction validity in that their values exhibited a desirable positive correlation with accuracy of predictions.

- Similarly, although based upon clustering algorithms that are inherently sensitive to the quality of d-space data the silhouette coefficient was able to reliably recognise situations where the predictive accuracy of Approach (c) and Approach (d) was affected and provide graceful degradation through the lowering of the validity measure.

- In comparing the two pattern active approaches it was found that Approach (d) is generally more effective than Approach (c) at identifying service classes and even performed slightly better when the latter was provided with N a priori. This suggests a strength inherent in the hierarchical clustering algorithm when applied to data with features specific to the assessment problem though this observation should be considered taking into account the substantial increase in complexity of Approach (d).

Through conducting the above evaluation it has been shown that the range of ED space assessment approaches offer consumers in environments where conformance-aware providers adopt service classes as a provision policy an effective tool for supporting decision making to find the services which best meet their needs. The choice of approach employed will depend upon the specific scenario in which it will operate though all approaches offer significant improvement over the current approaches to assessment which cannot operate effectively in such conformance-centric environments.

Chapter 7

Conclusion

A Service Oriented Architecture can be considered as a marketplace in which consumers and providers of services interact through a process of discovery, negotiation and consumption. Both consumers and providers are necessary components and both must gain benefit from such interaction if they are to be persuaded to continue their membership of the marketplace. Incentivising providers and consumers is of particular importance to SOA which is still in its infancy as a marketplace architecture and must therefore seek to attract the business of existing and new providers and consumers who are also able to operate in traditional spaces.

Although both are necessary components of an SOA the relationship between provider and consumer is such that the onus is typically on the former to attract the latter into using their products or services. Again, this is particularly evident in electronic marketplaces which may be global and where providers may offer functionally similar services that can be substituted for one another easily by a consumer. In such an environment competition is high and providers will seek to understand and meet the needs of their consumers to gain and maintain their custom and eventually profit from the relationship. Providers of services will vary their policy of serving consumers but will rationally seek to ensure that the consumers with whom they choose to engage have their needs satisfied. Management of multiple consumers with diverse requirements of a functional service is problematic and providers will likely adopt a differentiated services approach to provision through service classes.

As in any competitive marketplace consumers are faced with a dilemma of which product or service to choose to meet their needs. In traditional, well-established marketplaces consumers are able to use a number of sources to inform their decision of which provider to choose. However, in an SOA marketplace such channels of communication are not well established and the risks that consumers take in consuming services are increased substantially. The lack of widely available, reliable information regarding the quality of services in SOA leading to high risk, is one of the key reasons why the successful commercialisation of SOA has been slower than expected.

The approaches to assessment developed in this thesis sought to address the gap identified above by providing information relating to past performance of services. By increasing the reliability of such information the decision making capability of consumers is improved leading to consumers being able to more consistently identify and consume those services which are most likely to meet their needs. Specifically, consumer decision making is best improved through the provision of predictions of future performance with associated measures of validity. By providing such information consumers are able to determine the highest expected utility offered by a set of candidate services. Furthermore, the accuracy of predictions produced by Quality Assessment may be increased substantially by considering the nature of the service classes offered by individual providers.

In addition to the direct benefits realised by consumers in terms of experienced quality which have been described and evaluated in this thesis improvements due to the improved assessment approaches extend to other areas of an SOA. A significant effect of improving abilities of consumers to choose services which best meet their needs is that providers that offer consistently high quality services are rewarded through more frequent custom. In addition to the obvious direct financial benefits experienced by good providers, malicious providers or those who are generally poor may be dis-incentivised to continue operation in the environment and choose to withdraw. This will lead to the average quality of providers in the environment increasing, reinforcing the improvements in quality experienced by consumers. Furthermore, a portion of the financial rewards realised by good providers will likely be spent on improving services. Again, this effect is self-reinforcing - over time leading to likely substantial increases in the overall quality of service experienced by consumers.

Most importantly all of the above benefits realised from improved quality assessment approaches will with time increase the reputation of Service Oriented Architectures as a viable platform for offering marketplaces in which both consumers and providers may flourish. Increases in reputation of the platform will lead to wide-spread commercial adoption of SOA by both consumers and providers which will in turn lead to increases in competition, diversity of services, investment and maturity of the platform.

7.1 Contributions

Below the main contributions of this work (see Section 1.2) are restated and concepts, approaches and evaluation are discussed to demonstrate that they have been achieved within this thesis:

7.1.1 A Comprehensive Model of a Conformance-Centric SOA

The consideration of quality as conformance whilst being widely accepted has typically been ignored in the development of specific SOA models with quality instead being equated to performance and its subjective nature being unacknowledged. Moreover, where a reasonable model of quality as conformance has been considered its application has been limited to one area of SOA such as consumer provided ratings, requirements specification or objective quality assessment. Considering the behaviour of entities which take a conformance centric view of quality in isolation ignores more complex behaviours that emerge where all entities adopt the view and more importantly, where entities are also aware that all other entities also adopt the view. The model of a SOA presented in Chapter 3 considers holistically the behaviours of providers, consumers and assessment services where such an assumption is made. Significantly, the behaviour of service consumers and service providers who assume a conformance-centric view of quality have implications for the design and function of quality assessment approaches.

7.1.2 A Conformance-Centric Approach to Quality Assessment

A significant observation made from using the conceptual model is that the perception of quality by consumers is subjective and context dependence. The role of a quality assessment service should therefore not be one of recommendation or selection since doing so makes an assumption that individual differences do not exist or that consumers are able to objectively and accurately specify their requirements. Rather, a quality assessment service should aim to serve as a provider of accurate information relating to provider performance. Specifically, the consumer decision model described in Section 3.3.3 defines predictions of likely future provider behaviour confidence in predictions and a consumer risk function as information requirements for making such decisions. To usefully support consumer decision making any approach to quality assessment should therefore provide both a prediction of the likely future performance of a service and an indicator of confidence in this figure.

Section 4.1 described a generic approach to quality assessment which returns a prediction of the likely future behaviour of a service by identifying relevant points and using linear regression; and a validity measure based on statistical analysis of the data which may be used to infer confidence. A range of validity measures were identified for use across the range of specific approaches to assessment described in Chapter 4. In the evaluation, the aspects of consumer decision making were evaluated in three ways. Firstly, each specific approach was evaluated in terms of the accuracy of its predictions. Secondly, each validity measure was evaluated in terms of its reliability when indicating predictive accuracy. Finally, the link between high accuracy, high confidence predictions and improved consumer decision making in terms of receiving higher quality services was evaluated. In general it was established that the approaches to assessment described in this thesis which assume a role of information provider are more effective in supporting consumer decision making than traditional approaches.

The conceptual model (see Chapter 3) describes provider behaviour as being defined by a set of service classes which are a rational response to dealing with numerous consumers with varying expectations. Providers will seek to keep details of the service classes private though they may expose limited details of their policies through advertisements. Since providers will decide the level of service provided to consumers based on their requirement approaches to assessment will only be effective if they first seek to identify the service class which is relevant to the expectation of the current consumer. Multiple specific techniques for identifying service classes are defined in Section . These can be classified in terms of the degree to which they assume characteristics of the service classes and range from naive approaches which assume all characteristics at design time to approaches which utilise clustering techniques to discover characteristics of service classes at runtime. In Chapter 6, each of the approaches was evaluated in terms of predictive accuracy and effectiveness of the validity measure. The evaluation showed that traditional approaches to quality assessment are misleading when applied in conformance-centric SOA and that an approach to quality assessment which utilised hierarchical clustering is generally preferable to the other types described in Chapters 4 and 5.

7.1.3 Environments for the Empirical Evaluation of Quality Assessment Approaches

The approaches to quality assessment defined in Chapter 4 are designed to operate in a conformance-centric SOA and overcome some of the weaknesses inherent in traditional approaches when operating in these environments. Therefore, any evaluation of the approaches should take place within an implementation of such a conformance-centric SOA. This is challenging because fully defined conformance-centric SOA do not yet exist as their non-conformance-centric counterparts do. Furthermore, in order to conduct effective evaluation any experiments must allow fine-grained control over scenario parameters and support the evaluation of multiple aspects of simulation including the predictive accuracy of approaches and changes in the quality experienced by consumers.

The model of a conformance-centric SOA defined in Chapter 3 provided the basis for the development of two environments for the evaluation of the conformance-centric approaches to quality assessment developed in Chapter 4. The first of these, the CONOISE-G architecture (see Section 5.1) served to provide a demonstration of the approaches in a real-world scenario though lacked some of the detailed control required for full evaluation and did not allow the direct comparison of approaches. This environment was used to provide a qualitative demonstration of the context in which a QAS operates and to demonstrate its value in the formation and perturbation of virtual organisations.

The simulation application (see Section 5.2) provided a full implementation of the conformance-centric SOA described in Chapter 3 and includes features for the control of simulation parameters and the visualisation of results. The application is extensible and can be used to test and evaluate a range of approaches to quality assessment in addition to those developed in this thesis. The environment was used to objectively and effectively evaluate the four approaches to assessment defined in 4 including simulation of traditional and other approaches to assessment in order to provide baselines for comparison.

7.2 Future Work

The application of the conformance-centric view of quality to a SOA taken in this thesis is novel and this work has therefore focussed on developing approaches to assessment which made particular assumptions in order to support the explanation and evaluation of its core principles. Whilst these assumptions were necessary in order to achieve the stated aims of this thesis the work may be furthered by reconsidering these in order to develop more robust, scalable and general approaches to the problem of quality assessment in conformance-centric SOA. Specifically, the following three topics provide viable and potentially valuable avenues for future research, each being a direct and logical progression from the work described in this thesis.

7.2.1 Multiple Attributes

The central contribution of this thesis surrounds the concept of quality as conformance which has implications across an SOA. In order to focus on this contribution and its implications the assumption was taken that a single QoS attribute would form the basis of assessment (see Section 3.1.1). Choices in terms of the design of each approach were taken based on an assumption that the approach should be easily adaptable for use with multiple attributes. This is particularly evident in choosing clustering methods and statistical validity measures which may be easily applied to multi-dimensional problems.

The approaches described in this thesis would require insubstantial adaptation at a algorithmic level in order to support multiple attribute assessment. However, consideration of multiple attributes will have implications for the way in which providers provide their services since in a given service provision instance the level of service delivered for each attribute may not be independent. For example, there is a direct and general correlation between the *bandwidth* and *frame rate* of a service. Approaches to assessment would be improved by further considering the patterns between the quality spaces of individual attributes due to the relationship between QoS attributes. Identification and generalisation of such rules would be particularly useful in situations where there is limited information available to conduct an assessment of a service.

7.2.2 Time Series Clustering

The effect of time on the behaviour of services has been recognised as being significant in terms of both a provider's ability to adjust the delivered level of service within each service class and changing the boundaries of the service classes themselves (see Section 3.4.2). In the assessment approaches developed these assumptions about behaviour were acknowledged through the inclusion of a "decay function" (see Section 4.2.1) that selects only recent performance information for use in assessment. This mitigates against the risk of a provider changing their behaviour and in doing so makes an implicit assumption that recent performance data is more relevant to assessment than older data. Removal of time as a parameter also aids the use of clustering to identify service classes since more recent data is likely to exhibit less variation and noise therefore leading to crisper, more distinct clusters in d-space (see Section 4.6).

However, it is possible that considering time within the service class identification process might allow patterns in d-space resulting from service classes to be identified more effectively. This might be achieved by treating historic performance data as being generated from a number of *time series* with each time series corresponding to performance data from a single service class. Recent research into *time series analysis*[32] particularly *time series clustering* might be applied to approaching the problem of recognising service classes in a way that uses the concept of *time* to aid the recognition process.

7.2.3 Model Based Approaches

The approaches to assessment developed in this thesis are based on the observation that the behaviour of providers is determined by the expectation of individual consumers leading to the formation of clusters of points in d-space that may be subsequently identified through various techniques. The specific methods of identifying clusters described and evaluated carry out analysis of the points each time an assessment is requested by a particular consumer. For the k-means and hierarchical clustering approaches this involves a degree of computational complexity that is higher than linear with respect to time (see Section 4.6.1.1 and Section 4.6.4). This problem is compounded by increases in the number of attributes as discussed above. It is likely that the clustering of time series will also significantly increase the computational complexity of assessment approaches.

One way in which the complexity of the most effective approaches might be reduced involves taking a model-based approach to assessment. In such an approach rather than taking all historic provider performance data as points a mathematical model of behaviour may be created and maintained for each provider. Such a model would be consulted during each instance of assessment without requiring the re-assessment of performance data through re-clustering. In such an approach there will be tradeoffs between the age of models (which will need to be refreshed periodically) and the accuracy of assessments. A form of unsupervised learning might be developed to enable model refresh in which the assessment approach monitors the accuracy of its predictions and refreshes the model when a particular threshold of accuracy is reached.

Bibliography

[1]

[2] The American Society for Quality. http://www.asq.org/, 2007.

[3] A. Abdul-Rahman and S. Hailes. Supporting Trust in Virtual Communities. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, page 6007. IEEE Computer Society, January 2000.

[4] R. J. Al-Ali, O. F. Rana, and D. W. Walker. G-QoSM: Grid Service Discovery Using QoS Properties. *Journal of Computing and Informatics: Special Issue on Grid Computing*, 21(4):363–382, 2002.

[5] J. Anselmi, D. Ardagna, and P. Cremonesi. A QoS-Based Selection Approach of Autonomic Grid Services. In *Proceedings of the 2007 Workshop on Service-oriented Computing Performance*, pages 1–8. ACM, 2007.

[6] F. Barbon, P. Traverso, M. Pistore, and M. Trainotti. Run-Time Monitoring of Instances and Classes of Web Service Compositions. In *ICWS '06: Proceedings of the IEEE International Conference on Web Services*, pages 63–71. IEEE Computer Society, 2006.

[7] M.N. Bennani and D.A. Menasce. Resource Allocation for Autonomic Data Centers using Analytic Performance Models. pages 229–240. IEEE Computer Society, 2005.

[8] P. Berkhin. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, Inc., 2002.

[9] H. Billhardt, R. Hermoso, S. Ossowski, and R. Centeno. Trust-Based Service Provider Selection in Open Environments. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, pages 1375–1380. ACM, 2007.

[10] B.Larsen and C. Aone. Fast and Effective Text Mining using Linear-Time Document Clustering. In *KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22, New York, NY, USA, 1999. ACM.

[11] P. S. Bradley and U. M. Fayyad. Refining Initial Points for K-Means Clustering. In *Proceedings of the 15th International Conference on Machine Learning*, pages 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

[12] J. Breese, D. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52. Morgan Kaufman, San Francisco, CA, 1998.

[13] J. Brooke, D. Fellows, and J. MacLaren. Resource Brokering: The EUROGRID/GRIP Approach. In *Proceedings (CD-ROM) of the UK eScience All Hands Meetings (AHM 2004)*, September 2004. URL: http://www.allhands.org.uk/2004/submissions/papers/178.pdf.

[14] M. Bruno, G. Canfora, M. Di Penta, G. Esposito, and V. Mazza. Using Test Cases as Contract to Ensure Service Compliance Across Releases. In *Proceedings of the International Conference on Service Oriented Computing (ICSOC 2005)*, pages 87–100. Springer Verlag, Heidelberg, Germany, November 2005.

[15] G. Canfora and M. Di Penta. Testing Services and Service-Centric Systems: Challenges and Opportunities. *IT Professional*, 8(2):10–17, 2006.

[16] H. Cao, X. Feng, Y. Sun, Z. Zhang, and Q. Wu. A Service Selection Model with Multiple QoS Constraints on the MMKP. In *Proceedings of the FIP International Conference on Network and Parallel Computing Workshops, 2007. NPC Workshops. I*, pages 584–589, September 2007.

[17] V. Cardellini, E. Casalicchio, V. Grassi, and F. Lo Presti. Efficient Provisioning of Service Level Agreements for Service Oriented Applications. In *Proceedings of the 2nd International Workshop on Service Oriented Software Engineering*, pages 29–35. ACM, 2007.

[18] J. Cardoso, J. Miller, A. Sheth, and J. Arnold. Modeling Quality of Service for Workflows and Web Service Processes. Technical Report UGACS-TR-02-002, LSDIS Lab, Computer Science Department, University of Georgia, May 2002.

[19] J. Carter, E. Bitting, and A. A. Ghorbani. Reputation Formalization for an Information-Sharing Multi-Agent System. *Computational Intelligence*, 18(4):515–534, November 2002.

[20] D. Chalmers and M. Sloman. A Survey of Quality of Service in Mobile Computing Environments. *IEEE Communications Surveys and Tutorials*, 2(2):2–10, 1999.

[21] M Chen and J.P. Singh. Computing and Using Reputations for Internet Ratings. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 154–162. ACM, 2001.

[22] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pages 263–268. ACM, 2001.

[23] M. Claypool, P. Le, M. Waseda, and D. Brown. Implicit Interest Indicators. In *Proceedings of the ACM Intelligent User Interfaces Conference*, pages 33–40, January 2001.

[24] B. Yan D. Qian Z. Luan D. Yang, Z. Wuand. Agent-Based MADM Approach to the Dynamic Web Service Selection. *The 2nd IEEE Asia-Pacific Service Computing Conference*, pages 260–265, 11-14 Dec. 2007.

[25] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, P. Samarati, and Fabio Violante. A Reputation-Based Approach for Choosing Reliable Resources in Peer-to-Peer Networks. In *Proceedings of the 9th ACM conference on Computer and communications security*, pages 207–216. ACM, 2002.

[26] A. Dan, D. Davis, R. Kearney, R. King, A. Keller, D. Kuebler, H. Ludwig, M. Polan, M. Spreitzer, and A. Youssef. Web Services on Demand: WSLA-Driven Automated Management. *IBM Systems Journal, Special Issue on Utility Computing*, 43(1):136–158, March 2004.

[27] C. Dellarocas. Immunizing Online Reputation Reporting Systems against Unfair Ratings and Discriminatory Behavior. In *EC '00: Proceedings of the 2nd ACM Conference on Electronic Commerce*, pages 150–157, New York, NY, USA, 2000. ACM Press.

[28] G. N. Demir, A. S. Uyar, and S.G. Oguducu.

[29] V. Deora, J. Shao, G. Shercliff, P.J. Stockreisser, W.A. Gray, and N.J. Fiddian. Incorporating QoS Specifications in Service Discovery. In *Web Information Systems WISE 2004 Workshops*, volume 3307 of *Lecture Notes in Computer Science*, pages 252–263. Springer Berlin / Heidelberg, 2004.

[30] Vikas Deora, Jianhua Shao, W. Alex Gray, and Nick J. Fiddian. Expectation Based Quality of Service Assessment. *International Journal on Digital Libraries*, 6(3):260–269, March 2006.

[31] C. Ding and X. He. Cluster Merging and Splitting in Hierarchical Clustering Algorithms. In *Proceedings of The Second IEEE International Conference on Data Mining (ICDM'02)*, pages 139–147. IEEE Computer Society, 2002.

[32] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. *Proceedings VLDB Endowment*, 1(2):1542–1552, 2008.

[33] J. R. Douceur and J. S. Donath. The Sybil Attack. In *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS '02)*, pages 251–260. Springer, March 2002.

[34] R.O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, USA, 1973.

[35] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. 2001.

[36] M. Ekstrom, A. C. Bicharra Garcia, and H. Bjornsson. Rewarding Honest Ratings through Personalised Recommendations in Electronic Commerce. *International Journal of Electronic Business (IJEB)*, 3(3/4):392–410, 2005.

[37] R. Elmasri and S. Navathe. *Fundamentals of Database Systems (3rd Edition)*. Addison Wesley, Reading, Massachusetts, USA, 1999.

[38] F. Emekci, O. D. Sahin, D. Agrawal, and A. El Abbadi. A Peer-to-Peer Framework for Web Services Discovery with Ranking. In *Proceedings of the IEEE International Conference on Web Services*, pages 192–199, July 2004.

[39] A. Farag and M. Muthucumaru. Evolving and Managing Trust in Grid Computing Systems. In *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, pages 1424–1429, 2002.

[40] E.W. Forgy. Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics*, (21):768–769.

[41] I. Foster, C. Kesselman, J. Nick, and S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Global Grid Forum: Open Grid Service Infrastructure Working Group, June 2002. URL: http://globus.org/alliance/publications/papers/ogsa.pdf.

[42] I. Foster, C. Kesselman, and S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Applications*, 15(3):200–222, 2001.

[43] C. Fraley and A. E. Raftery. How many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *The Computer Journal*, 41:578–588, 1998.

[44] S. Frolund and J. Koistinen. Quality of Service Specification in Distributed Object System Design. In *Proceedings of the 4th USENIX Conference on Object-Oriented Technologies and Systems (COOTS)*, April 1998.

[45] Z. Gao and G. Wu. Combining QoS-based Service Selection with Performance Prediction. *Proceedings of the IEEE International Conference on e-Business Engineering, 2005. ICEBE 2005*, pages 611–614, October 2005.

[46] J. Garofalakis, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis. Web Service Discovery Mechanisms: Looking for a Needle in a Haystack? In *Proceedings of International Workshop on Web Engineering 2004*, 2004.

[47] J. Golbeck and J. Hendler. Inferring Reputation on the Semantic Web. In *Proceedings of the 13th International World Wide Web Conference*. ACM, 2004. URL: http://www.mindswap.org/papers/GolbeckWWW04.pdf.

[48] F. A. Grabowski. *Plato, Metaphysics and the Forms (Continuum Studies in Ancient Philosophy)*. Continuum International Publishing Group, London, UK, May 2008.

[49] G. Grimmett and D. Stirzaker. *Probability and Random Processes (Third Edition)*. Oxford University Press, Oxford, UK, 2001.

[50] J. Gross. *Linear Regression*. Springer-Verlag, Berlin, Germany, 2003.

[51] S. Guan, X. Dong, W. Wu, Y. Mei, and S. Liao. Trust Management and Service Selection in Pervasive Computing Environments. *Computational Intelligence and Security Workshops, 2007. CISW 2007. International Conference on*, pages 620–623, 2007.

[52] A. Hafid and S. Fischer. A Multi-Agent Architecture for Cooperative Quality of Service Management. In *Proceedings (CD-ROM) of the IFIP/IEEE Conference on Management of Multimedia Networks and Services (MMNS'97)*, pages 41–54, 1997.

[53] J. Hendler and D. L. McGuinness. The DARPA Agent Markup Language. *IEEE Intelligent Systems*, 15(6):67–73, November 2000.

[54] M. Hosomani, H. Narayanappa, and H. Rajan. Monitoring the Monitor: An Approach towards Trustworthiness in Service Oriented Architecture. In *Proceedings of the 2nd International Workshop on Service Oriented Engineering*, pages 42–46, September 2007.

[55] M. S. Hossain, A. Alamri, and El Saddik A. A Framework for QoS-Aware Multimedia Service Selection for Wireless Clients. In *WMuNeP '07: Proceedings of the 3rd ACM workshop on Wireless Multimedia Networking and Performance Modeling*, pages 16–22, New York, NY, USA, 2007. ACM.

[56] J. Hu, C. Guo, H. Wang, and P. Zou. Quality Driven Web Services Selection. In *ICEBE '05: Proceedings of the IEEE International Conference on e-Business Engineering*, pages 681–688, Washington, DC, USA, 2005. IEEE Computer Society.

[57] N. Hu, P. A. Pavlou, and J. Zhang. Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. In *EC '06: Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 324–330, New York, NY, USA, 2006. ACM.

[58] M. N. Huhns and M. P. Singh. Service-Oriented Computing: Key Concepts and Principles. *IEEE Internet Computing*, 9(1):75–81, January 2005.

[59] I. Im and A. Hars. Does a One-Size Recommendation System Fit All? The Effectiveness of Collaborative Filtering Based Recommendation Systems Across Different Domains and Search Modes. *ACM Trans. Inf. Syst.*, 26(1):4, 2007.

[60] T. Irwin. *Plato's Ethics.* Oxford University Press (USA), New York, USA, 1995.

[61] R. Ismail and A. Josang. The Beta Reputation System. In *Proceedings (CD-ROM) of the 15th Bled Conference on Electronic Commerce*, 2002. http://sky.fit.qut.edu.au/ josang/papers/JI2002-BI.

[62] ISO. ISO9000 Quality Management Systems – Fundamentals and Vocabulary. Technical report, International Standards Organisation (ISO TC 176/SC 1), 2005.

[63] Y. Ito and S. Tasaka. Quantitative Assessment of User-Level QoS and its Mapping. *IEEE Transactions on Multimedia*, 7(3):572–584, June 2005.

[64] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[65] N. R. Jennings. An Agent-Based Approach for Building Complex Software Systems. *Communications of the ACM*, 44(4):35–41, April 2001.

[66] A. Josang, R. Ismail, and C. Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43(2):618–644, 2007.

[67] P. Juhua, X. Zhang, and W. Zhenxing. The Research on QoS for Grid Computing. In *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, volume 2, pages 1711–171, 2003.

[68] R. Jurca and B. Faltings. An Incentive Compatible Reputation Mechanism. In *Proceedings of the International Conference on E-Commerce Technology (CEC'03)*, pages 285–292. IEEE, June 2003.

[69] S. Kalepu, S. Krishnaswamy, and S. W. Loke. Reputation = f (User Ranking, Compliance, Verity). In *Proceedings of the IEEE International Conference on Web Services (ICWS' 2004)*, pages 200–207, California, USA, July 2004. IEEE Press.

[70] I. Kant. *Foundations of the Metaphysics of Morals (translated by L.W. Beck)*. Prentice Hall, Saddle River, New Jersey, USA, 1989.

[71] G. Karypis, E-H. (Sam) Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modelling. *IEEE Computer*, 32:68–75, August 1999.

[72] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Interscience, New York, USA, March 1990.

[73] A. Keller, G. Kar, H. Ludwig, A. Dan, and J. L. Hellerstein. Managing Dynamic Services: A Contract Based Approach to a Conceptual Architecture. In *Proceedings of the 8th IEEE/IFIP Network Operations and Management Symposium (NOMS 2002)*, pages 513–528, 2002.

[74] J. O. Kephart and D. M. Chess. The Vision of Autonomic Computing. *Computer*, 36(1):41–50, 2003.

[75] J. O. Kephart and R. Das. Achieving Self-Management via Utility Functions. *IEEE Internet Computing*, 11(1):40–48, 2007.

[76] S. Lamparter, A. Ankolekar, R. Studer, and S. Grimm. Preference-Based Selection of Highly Configurable Web Services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1013–1022, New York, NY, USA, 2007. ACM.

[77] S. Lamrous and M. Taileb. Divisive Hierarchical K-Means. In *Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, page 18. IEEE Computer Society, 2006.

[78] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[79] J. Liu, Q. Zhang, B. Li, W. Zhu, and J. Zhang. A Unified Framework for Resource Discovery and QoS-Aware Provider Selection in Ad Hoc Networks. *SIGMOBILE Mob. Comput. Commun. Rev.*, 6(1):13–21, 2002.

[80] Y. Liu, A. H. Ngu, and L. Z. Zeng. QoS Computation and Policing in Dynamic Web Service Selection. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, pages 66–73. ACM Press, 2004.

[81] H. Ludwig. Web Services QoS: External SLAs and Internal Policies Or: How Do We Deliver what we Promise? In *Proceedings of the International Conference on Web Information Systems Engineering Workshops (WISEW'03)*, pages 1–4, Los Alamitos, CA, 2003. IEEE Computer Society.

[82] C. M. MacKenzie, K. Laskey, F. McCabe, P. Brown, and R. Metz. Reference Model for Service Oriented Architecture 1.0. Technical report, Organisation for the Advancement of Structured Information Standards, August 2006. URL: http://www.oasis-open.org/committees/download.php/19679/soa-rm-cs.pdf.

[83] J.B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, California, USA, 1967. University of California Press.

[84] U. S. Manikrao and T. V. Prabhakar. Dynamic Selection of Web Services with Recommendation System. In *NWESP '05: Proceedings of the International Conference on Next Generation Web Services Practices*, page 117, Washington, DC, USA, 2005. IEEE Computer Society.

[85] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*, chapter 17. Cambridge University Press, Cambridge UK, 2008.

[86] E. Martin, S. Basu, and T. Xie. Automated Robustness Testing of Web Services. In *Proceedings of the 4th International Workshop on SOA And Web Services Best Practices (SOAWS 2006)*, October 2006.

[87] E. Martin, S. Basu, and T. Xie. WebSob: A Tool for Robustness Testing of Web Services. In *Companion to the Proceedings of the 29th International Conference on Software Engineering*, pages 65–66, Washington, DC, USA, 2007. IEEE Computer Society.

[88] O. Martin-Diaz, A. Ruiz-Cortés, D. Benavides, A. Duran, and M. Toro. A Quality-Aware Approach to Web Services Procurement. *Lecture Notes in Computer Science*, 2819:42–53, September 2003.

[89] Y. Matsui, S. Kihara, A. Mitsuzawa, S. Moriai, and H. Tokuda. An Extensible Object Model for QoS Specification in Adaptive QoS Systems. In *Proceedings of the Second IEEE International Symposium on Object-Oriented Real-Time Distributed Computing*, page 129, 1999.

[90] E. M. Maximilien and M. Singh. Conceptual Model of Web Service Reputation. *ACM SIGMOD Record*, pages 36–41, 2002.

[91] E. M. Maximilien and M. P. Singh. Agent-Based Architecture for Autonomic Web Service Selection. In *Proceedings of the Workshops of the Autonomous Agents and Multi-Agent Systems Conference (AAMAS'03)*, 2003.

[92] E. M. Maximilien and M. P. Singh. Toward Autonomic Web Services Trust and Selection. In *ICSOC '04: Proceedings of the 2nd International Conference on Service Oriented Computing*, pages 212–221. ACM Press, 2004.

[93] E. M. Maximilien and M. P. Singh. Agent-Based Trust Model Involving Multiple Qualities. In *AAMAS '05: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 519–526, New York, NY, USA, 2005. ACM Press.

[94] E. M. Maximilien and M. P. Singh. Self-Adjusting Trust and Selection for Web Services. In *ICAC '05: Proceedings of the Second International Conference on Automatic Computing*, pages 385–386, Washington, DC, USA, 2005. IEEE Computer Society.

[95] E. Michael Maximilien and Munidar P. Singh. An Ontology for Web Service Ratings and Reputations. In *Proceedings of the Workshop on Ontologies in Agent Systems (OAS2003)*, pages 25–30. CEUR Publications, 2003.

[96] E.M. Maximilien and M.P. Singh. A Framework and Ontology for Dynamic Web Services Selection. *IEEE Internet Computing*, 8(5):84–93, 2004.

[97] D. A. Menasce. QoS Issues in Web Services. *IEEE Internet Computing*, 6(6):72–75, November 2002.

[98] D. A. Menasce and V. Dubey. Utility-Based QoS Brokering in Service Oriented Architectures. *IEEE International Conference on Web Services, 2007. ICWS 2007.*, pages 422–430, 9-13 July 2007.

[99] Nolan H. Miller, Paul Resnick, and Richard J. Zeckhauser. Eliciting Honest Feedback in Electronic Marketplaces. Technical report, August 2002. URL: http://web.hks.harvard.edu/publications/workingpapers/Index.aspx.

[100] L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, and A. Halberstadt. Ratings in Distributed Systems: A Bayesian Approach. In *Workshop on Information Technologies and Systems*, 2001.

[101] F. Murtagh. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *The Computer Journal*, 26:354–359, 1983.

[102] OASIS. Introduction to UDDI: Important Features and Functional Concepts. Technical report, 2004. URL: http://www.uddi.org/pubs/uddi-tech-wp.pdf.

[103] A. Padovitz, S. Krishnaswamy, and S. W. Loke. Towards Efficient Selection of Web Services. In *Proceedings of the Second International Joint Conferences on Autonomous Agents and Multi-agent Systems*. ACM Press, July 2003.

[104] M. P. Papazoglou and D. Georgakopoulos. Service-Oriented Computing. *Communications of the ACM*, 46(10):24–28, 2003.

[105] A. Parasuraman and V. A. Zeithaml. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, 64(1):12–40, 1988.

[106] A. Parasuraman, V. A. Zeithaml, and L. L. Berry. A Conceptual Model for Service Quality and its Implications for Future Research. *Journal of Marketing*, 49:41–50, 1985.

[107] J. Patel, W. T. Luke Teacy, N. R. Jennings, M. Luck, S. Chalmers, N. Oren, T. J. Norman, A. D. Preece, P. M. D. Gray, G. Shercliff, P. J. Stockreisser, J. Shao, W. A. Gray, N. J. Fiddian, and S. Thompson. Agent-based Virtual Organisations for the Grid. *Multiagent and Grid Systems*, 1(4):237–249, 2005.

[108] J. Patel, W.T.L. Teacy, N. R. Jennings, M. Luck, S. Chalmers, N. Oren, T.J. Norman, A. Preece, P.M.D. Gray, G. Shercliff, P.J. Stockreisser, J. Shao, W.A. Gray, N.J. Fiddian, and S. Thompson. CONOISE-G: Agent Based Virtual Organisations for the Grid. *International Journal of MultiAgent and Grid Systems: Special Issue on Smart Grid Technologies*, 1:237–249, 2005.

[109] M. Di Penta, G. Canfora, G. Esposito, V. Mazza, and M. Bruno. Search-Based Testing of Service Level Agreements. In *GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, pages 1090–1097, New York, NY, USA, 2007. ACM.

[110] R. Pirsig. *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values*. Bantam Books, April 1984.

[111] A. Preece, S. Chalmers, N. Oren, T. Norman, A. Preece, P. Gray, G Shercliff, P. J. Stockreisser, J. Shao, A. Gray, N. Fiddian, J. Patel, L. Teacy, N. Jennings, M. Luck, and S. Thompson. Managing Virtual Organisations on the Grid using Agent Technology. *AgentLink News*, pages 14–17, 2005.

[112] T. Pyzdek. *Six Sigma Handbook*. McGraw-Hill Publishing Company, 2000. URL: http://www.qaa.ac.uk/aboutus/annualReports/0405/annualreport/annrev0405.pdf.

[113] QAA. QAA Annual Review 2004-05. Technical report, Quality Assurance Agency for Higher Education (QAA), 2006.

[114] S. D. Ramchurn, T.D. Huynh, and N.R. Jennings. Trust in Multiagent Systems. *The Knowledge Engineering Review*, 19(1):1–25, July 2004.

[115] S. Ran. A Model for Web Services Discovery with QoS. *ACM SIGecom Exchange*, 4(1):1–10, 2003.

[116] T. Rea. *Engendering Trust in Electronic Environments: Roles for a Trusted Third Party*, pages 221–236. Kluwer Academic Publishers Group, 2001.

[117] P. Resnick and H.R. Varian. Recommender Systems. 40:56–58, March 1997.

[118] P. Resnick and R. Zeckhauser. Trust Amongst Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. *The Economics of the Internet and E-Commerce*, 11:127–157, 2002.

[119] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation Systems. *Communications of the ACM*, 43(12):45–48, December 2000.

[120] M. D. Rohr. Is Goodness Comparative? *The Journal of Philosophy*, 75(9):494–503, September 1978.

[121] P. J. Rousseuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, November 1987.

[122] J. Roy and A. Ramanujan. Understanding Web Services. *IT Professional*, 3:69–73, November 2001.

[123] J. Carbo Rubiera, J. M. Molina Lopez, and J. Davila Muro. A Fuzzy Model of Reputation in Multi-Agent Systems. In *AGENTS '01: Proceedings of the Fifth International Conference on Autonomous Agents*, pages 25–26, New York, NY, USA, 2001. ACM Press.

[124] J. Sabater and C. Sierra. Reputation and Social Network Analysis in Multi-Agent Systems. In *AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 475–482, New York, NY, USA, 2002. ACM.

[125] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.

[126] B. Savery. The Unique Quality Goodness – A Myth. *International Journal of Ethics*, 47(2):210–223, January 1937.

[127] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and Metrics for Cold-Start Recommendations. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, New York, NY, USA, 2002.

[128] J. Schneider, G. Kortuem, J. Jager, S. Fickas, and Z.Segall. Disseminating Trust Information in Wearable Communities. *Personal and Ubiquitous Computing*, 4(4):245–248, 2000.

[129] Jenifer M. Schopf, Mike D'Arcy, Neill Miller, Laura Pearlman, Ian Foster, and Carl Kesselman. Monitoring and Discovery in a Web Services Framework: Functionality and Performance of the Globus Toolkit's MDS4. Technical report, Argonne National Laboratory, 2005.

[130] Sandip Sen and Neelima Sajja. Robustness of Reputation-Based Trust: Boolean Case. In *AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 288–293, New York, NY, USA, 2002.

[131] M.A. Serhani, R. Dssouli, A. Hafid, and H. Sahraoui. A QoS Broker Based Architecture for Efficient Web Services Selection. In *Proceedings of the 2005 IEEE International Conference on Web Services, 2005. ICWS 2005.*, pages 113–120, 11-15 July 2005.

[132] A. ShaikhAli, O. F. Rana, R. Al-Ali, and D. W. Walker. UDDIe: An Extended Registry for Web Services. In *Proceedings of the 2003 Symposium on Applications and the Internet (Workshops)*, pages 85–89, 2003.

[133] U. Shardanand and P. Maes. Social Information Filtering: Algorithms for Automating Word of Mouth. In *Proceedings of: Human Factors in Computing Systems CHI 95*, pages 210–217, New York, NY, USA, 1995. ACM.

[134] G. Shercliff, P.J. Stockreisser, J. Shao, W.A. Gray, and N.J. Fiddian. Supporting QoS Assessment and Monitoring in Virtual Organisations. In *IEEE International Conference on Service Computing (SCC 05)*, volume 2, pages 249–250. Springer, 2005.

[135] Gareth Shercliff, Jianhua Shao, W. Alex Gray, and Nick J. Fiddian. A Multiple Quality-Space Mapping Approach to Qos Assessment. In *CIT 2006: Proceedings of the Sixth IEEE International Conference on Computer and Information Technology*. IEEE Computer Society, September 2006.

[136] Gareth Shercliff, Jianhua Shao, W. Alex Gray, and Nick J. Fiddian. QoS Assessment of Providers with Complex Behaviours: An Expectation-Based Approach with Confidence. In *Proceedings of the 4th International Conference on Service-Oriented Computing*, Lecture Notes in Computer Science, pages 378–389. Springer, 2006.

[137] G. Silver, J. A. Miller, A. P. Sheth, J. Myers, A. Maduko, and R. Jafri. Modeling and Simulation of Quality of Service for Composite Web Services. *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI'03)*, pages 1–10, July 2003.

[138] M. P. Singh. Peering at Peer-to-Peer Computing. *IEEE Internet Computing*, 5(1):4–5, 2001.

[139] M. P. Singh. Trustworthy Service Composition: Challenges and Research Questions. In *Proceedings of the Autonomous Agents and Multi-Agent Systems Workshop on Deception, Fraud and Trust in Agent Societies*, pages 39–52. Springer-Verlag, 2002.

[140] M. P. Singh, B. Yu, and M. Venkatraman. Community-Based Service Location. *Communications of the ACM*, 44(4):49–54, 2001.

[141] F. Siqueira and V. Cahill. Quartz: A qos architecture for open systems. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS 2000)*, pages 197–204, April 2000.

[142] W. Song and V. V. Phoha. Neural Network-Based Reputation Model in a Distributed System. In *CEC '04: Proceedings of the IEEE International Conference on E-Commerce Technology*, pages 321–324, Washington, DC, USA, 2004. IEEE Computer Society.

[143] J. D. Sonnek and J. B. WEissman. A Quantitative Comparison of Reputation Systems in the Grid. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, pages 242–249, November 2005.

[144] A. Soydan Bilgin and M.P. Singh. A DAML-Based Repository for QoS-Aware Semantic Web Service Selection. In *Proceedings of the IEEE International Conference on Web Services*, pages 368–375, 2004.

[145] C. Spearman. The Proof and Measurement of Association between Two Things. *American Journal of Psychology*, (15):72–101, 1904.

[146] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques (Technical Report 00-034). Technical report, 2000.

[147] Patrick J. Stockreisser, Gareth Shercliff, Jianhua Shao, W. A. Gray, and N. J. Fiddian. Managing Quality of Service in Virtual Organisations. In *iiWAS*, pages 413–424. Austrian Computer Society, 2005.

[148] L. Taher, H. El Khatib, and R. Basha. A Framework and QoS Matchmaking Algorithm for Dynamic Web Services. In *Proceedings of the 2nd International Conference on Innovations in Information Technology*, September 2005.

[149] D. Talia. The Open Grid Services Architecture: Where the Grid meets the Web. *IEEE Internet Computing*, 6:67–71, November 2002.

[150] A. Talwar, R. Jurca, and B. Faltings. Understanding User Behaviour in Online Feedback Reporting. pages 134–142, June 2007.

[151] P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, chapter 8. Addison-Wesley, 2006.

[152] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. TRAVOS: Trust and Reputation in the Context of Inaccurate Information Sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, March 2006.

[153] R. K. Teas. Expectations, Performance Evaluation and Consumers' Perceptions of Quality. *Journal of Marketing*, 57(4):18–34, 1993.

[154] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Fourth Edition.* Academic Press, 2009.

[155] M. Tian, A. Gramm, T. Naumowicz, H. Ritter, and J. Schiller. A Concept for QoS Integration in Web Services. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering Workshops (WISEW'03)*, pages 149–155, 2003.

[156] M. Tian, A. Gramm, H. Ritter, and J. Schiller. Efficient Selection and Monitoring of QoS-Aware Web Services with the WS-QoS Framework. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, pages 152–158, September 2004.

[157] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the Number of Clusters in a Dataset via the Gap Statistic. Technical report, Stanford University, 2000.

[158] Yathiraj B. Udupi, Pinar Yolum, and Munindar P. Singh. Trustworthy Service Caching: Cooperative Search in P2P Information Systems. In *Agent-Oriented Information Systems, 5th International Bi-Conference Workshop*, pages 32–44, 2003.

[159] S. Vinoski. Service Discovery 101. *IEEE Internet Computing*, 7(1):69–71, 2003.

[160] L-H. Vu, M. Hauswirth, and K. Aberer. Towards P2P-Based Semantic Web Service Discovery with QoS Support. In *Proceedings of the Workshops of the 3rd International Conference on Business Process Management (LNCS 3812)*, pages 18–31, 2006.

[161] J. Wang, J. Pouwelse, R. L. Lagendijk, and M.J.T. Reinders. Distributed Collaborative Filtering for Peer-to-Peer File Sharing Systems. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, pages 1026–1030, New York, NY, USA, 2006.

[162] Yao Wang and Julita Vassileva. A Review on Trust and Reputation for Web Service Selection. *Distributed Computing Systems Workshops, 2007. ICDCSW '07. 27th International Conference on*, page 25.

[163] J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[164] E.W. Weisstein. Correlation Coefficient, 2010. URL: http://mathworld.wolfram.com/CorrelationCoefficient.html.

[165] A. Whitby, A. Jsang, and J. Indulska. Filtering Out Unfair Ratings in Bayesian Reputation Systems. In *In Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 106–117, July 2004.

[166] M. J. Wooldridge and N. R. Jennings. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.

[167] G. Wu, J. Wei, X. Qiao, and L. Li. A Bayesian Network Based QoS Assessment Model for Web Services. *IEEE International Conference on Services Computing, 2007. SCC 2007.*, pages 498–505, 2007.

[168] D. Xu, K. Nahrstedt, and D. Wichadakul. QoS-Aware Discovery of Wide-Area Distributed Services. In *Proceedings of the 1st International Symposium on Cluster Computing and the Grid*, page 92. IEEE Computer Society, 2001.

[169] L. Yang and H. He. Grid Service Selection Using QoS Model. In *Proceedings of the Third International Conference on Semantics, Knowledge and Grid*, pages 576–577. IEEE Computer Society, 2007.

[170] P. Yolum and M. Singh. An Agent-Based Approach for Trustworthy Service Location. In *Proceedings of the First International Workshop on Agents and Peer-to-Peer Computing (AP2PC)*, pages 45–56, 2002.

[171] S. Yoon, D. Kim, and S. Han. WS-QDL Containing Static, Dynamic, and Statistical Factors of Web Services Quality. In *Proceedings of the IEEE Conference on Web Services (ICWS'04)*, pages 807–808. IEEE Computer Society, 2004.

[172] B. Yu and M. P. Singh. Detecting Deception in Reputation Management. In *AAMAS '03: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 73–80, New York, NY, USA, 2003. ACM.

[173] K. Yu, X. Xu, M. Ester, and H-P. Kriegel. Feature Weighting and Instance Selection for Collaborative Filtering: An Information-Theoretic Approach. *Knowledge and Information Systems*, 5(2), 2003.

[174] T. Yu and K. Lin. The Design of QoS Broker Algorithms for QoS-Capable Web Services. In *EEE '04: Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, pages 17–24, Washington, DC, USA, 2004. IEEE Computer Society.

[175] G. Zacharia and P. Maes. Trust Management through Reputation Mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, October 2000.

[176] G. Zacharia, A. Moukas, and P. Maes. Collaborative Reputation Mechanisms in Electronic Marketplaces. In *Proceedings of the IEEE 32nd International Conference on System Sciences*, page 8026. IEEE, 1999.

[177] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Z. Sheng. Quality Driven Web Services Composition. In *Proceedings of the Twelfth International Conference on World Wide Web*, pages 411–421. ACM Press, 2003.

[178] J. Zhang and R. Cohen. An Incentive Mechanism for Eliciting Fair Ratings of Sellers in e-Marketplaces. In *AAMAS '07: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent systems*, pages 1–3, New York, NY, USA, 2007. ACM.

[179] Y. Zhang and Y. Fang. A Fine-Grained Reputation System for Reliable Service Selection in Peer-to-Peer Networks. *IEEE Transactions on Parallel and Distributed Systems*, 18(8):1134–1145, 2007.

[180] C. Zhou, L. Chia, and B. Lee. DAML-QoS Ontology for Web Services. In *Proceedings of the IEEE Conference on Web Services (ICWS'04)*, pages 472–479. IEEE Computer Society, 2004.

[181] C. Zhou, L. Chia, and B. Lee. Service Discovery and Measurement based on DAML-QoS Ontology. In *Special Interest Tracks and Posters of 14th World Wide Web Conference*, pages 1070–1071, 2005.

[182] C. Zhou, L.T. Chia, B. Silverajan, and B. Lee. UX - An Architecture Providing QoS-Aware and Federated Support for UDDI. In *Proceedings of the International Conference on Web Services (ICWS 2003)*, pages 171–176, 2003.

[183] J. Zhou and E. Niemela. Toward Semantic QoS Aaware Web Services: Issues, Related Studies and Experience. In *IEEE/WIC/ACM International*

Conference on Web Intelligence, 2006 (WI 2006), pages 553–557, December 2006.