



AIS and Semantic Query¹

Rana Kashif Ali²
Digital Media Systems Laboratory
HP Laboratories Bristol
HPL-2004-233
December 23, 2004*

semantic web,
artificial immune
system, query
expansion,
information
retrieval

The semantic web has created various exciting opportunities to explore. Here we present a nature inspired solution to one such opportunity; that of semantic queries for information retrieval. We take our inspiration from the human immune system and develop an analogy between antibodies and queries. Successful antibodies are those that are activated by an infection. These antibodies are stimulated to clone, but imperfectly, giving rise to a multitude of similar antibodies that are better suited to tackle the infection. Analogously, queries producing relevant results can be cloned to give rise to various similar queries, each of which may be an improvement on the original query. The semantic web, being concept based, has a set of rules for creating expressive yet standardised queries with clear semantics guiding their modification. This paper discusses the implementation and evaluation of such an immune based information retrieval technique for the semantic web. Two query mutation operators, *RandomMutationOperator* and *ConstrainedMutationOperator* are proposed and compared in terms of their *precision*, *recall* and *convergence*. We have found the presented approach to be viable, and we discuss the potential for further improvements.

* Internal Accession Date Only

¹For more information please contact Steve.Cayzer@hp.com

²School of Computer Science, University of Birmingham, Birmingham, U.K.

Approved for External Publication

The University of Birmingham
School of Computer Science
MSc in Advanced Computer Science

Summer Project

AIS and Semantic Query

Rana Kashif Ali
msc43kar@cs.bham.ac.uk

Supervisors

Prof. Xin Yao
The University of Birmingham
x.yao@cs.bham.ac.uk

Dr. Steve Cayzer
HP Labs, Bristol
steve.cayzer@hp.com

September 9, 2004

Abstract

The emergence of semantic web on the one hand has created various exciting opportunities to explore but on the other hand has left many questions unanswered or vague. Here we present a nature inspired solution to one such vague question that deals with information retrieval on the semantic web. We take our inspiration from the human immune system and develop an analogy between antibodies and queries. Successful antibodies are those that are activated by an infection. These antibodies are stimulated to clone, but imperfectly, giving rise to a multitude of similar antibodies that are better suited to tackle the infection. Similarly, queries producing relevant results can be cloned to give rise to various similar queries which will be ideal for producing more relevant results and may even be an improvement on the original query. Also the semantic web, being concept based, has a set of rules for creating expressive yet standardised queries with clear semantics guiding their modification. The thesis discusses the implementation and evaluation of such an immune based information retrieval technique for the semantic web. Two query mutation operators namely, *RandomMutationOperator* and *ConstrainedMutationOperator* are proposed and compared in terms of their *precision*, *recall* and *convergence*. We have found the presented approach to be viable with potential for improvements.

Keywords: Semantic Web, Artificial Immune System, Query Expansion, Information Retrieval

Acknowledgements

I would like to take this opportunity to express my gratitude to Allah the Almighty for giving me the courage to complete the work and for making things easy for me.

Special thanks to Dr. Steve Cayzer for his invaluable advices at crucial points of the project, for keeping me interested and motivated, for proof reading the thesis and for always being there from the start to the very last minute of the project. Thanks to Prof. Xin Yao for agreeing to supervise me and for always taking time out from his busy schedule and providing an alternative view to the project.

I would also like to thank all my classmates for creating a wonderful atmosphere for studies and fun. Thanks to my mother for her endless prayers for my success and to my brother for partially funding the MSc and accommodating me with him. Thanks to my fiance, shafaq, for her love, kindness and patience.

Rana Kashif Ali

Contents

Abstract	ii
Acknowledgements	iii
Contents	vi
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Organisation of the report	2
2 Background Research	3
2.1 Human Immune System (HIS)	3
2.1.1 Acquired Immunity	5
2.2 Artificial Immune System (AIS)	6
2.3 Semantic Web	8
2.3.1 Resource Description Framework (RDF)	9
2.3.2 Schemas and Ontologies	10
2.3.3 Jena	11
2.4 Query Expansion	14
2.4.1 Types of Query Expansion	14
2.4.2 Query Expansion for the Semantic Web	15
3 AIS for Semantic Query Expansion	17
3.1 Mapping between the AIS and HIS	17
3.2 Overall Structure and Flow of Information	18
3.3 The Algorithm	18
3.3.1 Initial user feedback	18
3.3.2 Initialisation of AIS	20
3.3.3 Affinity Evaluation	21
3.3.4 Selection	24
3.3.5 Clonal Expansion	24
3.3.6 Mutation Operators	25

3.3.7	Replacement Strategy	29
3.3.8	Halting Criteria	29
3.4	Input Parameters and their Description	29
3.5	AIS vs GA	30
4	AIS based Semantic Search Utility	31
4.1	Implementation Details	31
4.2	Usability	33
5	Experiments and Results	39
5.1	Experimental Setup	39
5.1.1	Test Data	39
5.1.2	Plan	40
5.1.3	Automated Test Script	40
5.2	Results	42
5.2.1	First Phase	42
5.2.2	Second Phase	49
5.3	Summing it all up	53
6	Future Work	54
6.1	Critique	54
6.2	Future Directions	54
6.2.1	Testing	54
6.2.2	User Testing	55
6.2.3	Making it Suitable for Multiple Facets	55
6.2.4	New Mutation Operators	55
7	Conclusion	57
8	Online Resources	58
8.1	Poster	58
8.2	Source Code	58
8.3	Test Data and Configuration Files	58
8.4	Semantic Search Utility	58
8.5	Visualisation of the HIS	59
8.6	Graphs	59
A	Summer Project Declaration	63
B	Statement of Information Search Strategy	66
B.1	Semantic Web	66
B.2	HIS & AIS	67
B.3	Query Expansion	67

C	Data	68
C.1	wwite_tidy.n3	68
C.2	swed_org_type_skos.n3	71
C.3	wwite_index_skos.n3	72
C.4	Input Data Sets	73
C.4.1	set1	73
C.4.2	set2	73
C.4.3	set3	73
D	Source Code	74
D.1	RandomMutationOperator.java	74
D.2	ConstrainedMutationOperator.java	75

List of Figures

2.1	Organs of the Human Immune System [20]	4
2.2	Negative selection and clonal expansion [8]	7
2.3	Difference between the current web and the semantic web [27]	8
2.4	Structure of a triple	9
2.5	A simple RDF Model representing the author	9
2.6	RDF/XML representation of the model	10
2.7	N3 representation of the model	11
2.8	A slice of organisational ontology	12
3.1	AIS Infrastructure and flow of information	18
3.2	The AIS algorithm	19
3.3	Genotype and Phenotype of individuals	21
3.4	Types of Result Feedback	21
3.5	Affinity Evaluation Process	23
3.6	Clonal Expansion Process	25
3.7	Working of the mutation operators	28
4.1	Class Diagram	32
4.2	Organisation details	34
4.3	Results grouped by queries	35
4.4	Nonsel self results	36
4.5	Self results	37
4.6	Selected queries sorted by fitness	38
5.1	Precision for input set1	44
5.2	Precision for input set2	44
5.3	Precision for input set3	44
5.4	Recall for input set1	46
5.5	Recall for input set2	46
5.6	Recall for input set3	46
5.7	Convergence for RandomMutationOperator	48
5.8	Convergence for ConstrainedMutationOperator	48
5.9	Precision vs Mutation Rate for RandomMutationOperator	50
5.10	Precision vs Mutation Rate for ConstrainedMutationOperator	50
5.11	Recall vs Mutation Rate for RandomMutationOperator	51

5.12 Recall vs Mutation Rate for ConstrainedMutationOperator . . .	51
5.13 Minimum Iterations for Maximum Recall vs Mutation Rate . . .	52

List of Tables

2.1	semantic web vs. current web in the context of query expansion .	16
3.1	Mapping between the HIS and the presented model	17
3.2	Interpretation of the parameters	29
5.1	Input parameters common for both operators	42
5.2	Input parameters for <i>RandomMutationOperator</i>	43
5.3	Input parameters for <i>ConstrainedMutationOperator</i>	43

Chapter 1

Introduction

Semantic web is an extension to the current World Wide Web (WWW) in which resources are connected semantically rather than through hyperlinks. This semantic connectivity is achieved by meta-data about resources. The availability of such meta-data has opened new areas for researchers to explore. Once such area of research is information retrieval which also forms part of this thesis. The meta-data on the semantic web is represented in Resource Description Format (RDF) and is structured according to ontologies (discussed in Chapter 2) that are accessible to all. The access to the meta-data and ontology make them suitable for machine processing. Data in RDF can be queried using languages such as Resource Description Query Language (RDQL) [24]. In this thesis we present a nature inspired approach to query meta-data on the semantic web, our inspiration being the Human Immune System (HIS). The HIS is a complex adaptive system which learns to protect the body again harmful infections caused by germs. When the HIS encounters infections, it produces various antibodies, some of which are more suitable to overcome the infection. The suitable antibodies undergo mutation to produce various similar antibodies. Some of these newly produced antibodies might be an improvement over the original antibody and can better tackle the infection. We bring this idea in the realms of query expansion by establishing an analogy between the antibodies and queries. Mutation of a query may result in queries that are better suited to answer a particular search criteria. Query expansion is quite an old area of research that deals with modification of the original query for efficient information retrieval. We take the idea first proposed by Lee et al [17] and extend it with a concrete implementation and evaluation of an Artificial Immune System (AIS) for information retrieval on the semantic web. The presented approach is promising for two reasons. Firstly, computational models of HIS often called AIS have been shown to be useful for on line learning and document classification tasks. Secondly, query languages designed specifically to fetch information from the semantic web exist and are flexible enough to be modified using a set of rules. Two query expansion operators namely, *RandomMutationOperator* and *ConstrainedMutationOperator* have been proposed to work within the AIS and their performance is evaluated

in terms of their *precision, recall* and *convergence*.

1.1 Organisation of the report

The report is composed of eight chapters. Chapter 1 forms the introduction, provides the reader with some basic ideas and organisation of the report. It is followed by Chapter 2 which is a detailed discussion of the concepts that form the basis of the presented work. Chapter 3 holds explanation of the AIS based information retrieval technique and incorporates the description of the two query expansion operators. Chapter 4 discusses is an illustrated guide to the usage of the proposed immune based semantic search utility. In chapter 5 we talk about the experimental setup and the results obtained while comparing the two query mutation operators. Chapter 6 outlines the possible future directions of the project and also performs a critical analysis. We conclude in chapter 7, followed by a chapter on the on line resource associated with the project. Finally, there are four appendices, the first is the mini-project declaration, the second is the statement of information search strategy, the third is chunk of the sample data used and is meant give to the reader a feel of the data used in this project and the fourth appendix contains selected source code. Since the work is an extension of a mini-project some of the text from the mini-project is also incorporated in this thesis. Bulk of chapter two is copied from the mini-project since the background information is same for both the projects. All other chapters, however are written from scratch

Chapter 2

Background Research

The inherent nature of the project required diverse literature search in the areas of immunology and computing. Here we provide the details of research that forms the basis of the presented work. The first section gives a brief overview of the Human Immune System (HIS) followed by a survey of areas in which Artificial Immune System have been applied. In section 2.3, we have presented the notion of semantic web and its underlying technologies. Finally the area of Query Expansion is explored and distinction is drawn between the traditional approaches to Query Expansion and Query Expansion for the semantic web. We have also tried to explain how the query expansion process would work on the semantic web.

2.1 Human Immune System (HIS)

This section gives an overview of the HIS with the aim of providing the reader with necessary foundation to understand the presented work. Emphasis has been put on those aspects of HIS that we have tried to model in this work. The text presented here is not for the immunologically savvy readers who can whet their appetite with [28, 21]. For readers with no foundation of immunology, an extremely good introduction to HIS by Hofmeyr [13], that covers the HIS in sufficient depth as well as breadth, is recommended.

Immune system is the defence mechanism employed by the body against harmful organisms called *pathogens* for example viruses, bacteria and fungi. A similar term to pathogen is *antigen* which is any molecule that can stimulate the immune system [8]. Its important at this stage to note that both these terms have been used interchangeably in the literature related to AIS for example [6], and should be considered the same in the context of this work.

HIS is truly distributed with organs called *lymphoid organs* [26], situated at different parts in the body as shown in the Figure 2.1. Organs of the HIS are connected by *lymphatic vessels*, a transparent fluid called *lymph* flows in these vessels and finally get mixed in the blood. There are various bean-shaped *lym-*

phatic nodes where the lymph gets cleaned, it is this place where the pathogens are encountered by the immune system.

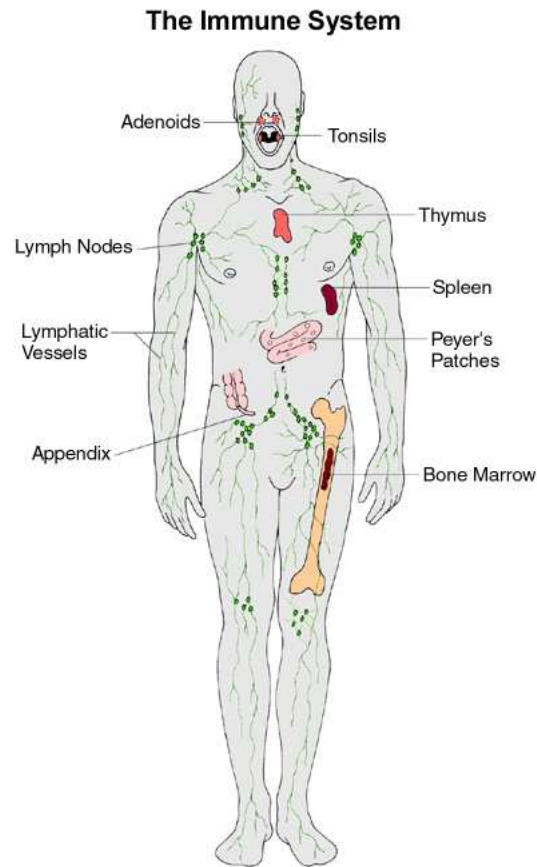


Figure 2.1: Organs of the Human Immune System [20]

Beside being distributed HIS is multi-layered and tackles pathogens at various levels of resistance to give the body maximum defence against diseases. The layers of HIS along with their composition are presented below

- *Physical Layer:* It is the skin that blocks the pathogens from entering the body.
- *Physiological Layer:* Consists of fluids secreted by the body for example saliva, tears and sweat. These body fluids have the enzymes to kill the germs and prevent them from entering the body for example tears wash away pathogens trying to enter through the eyes.

- *Innate Immunity*: This is the layer of immunity which a pathogen is faced with once it passes through the above mentioned layers. It is also called *non-specific* because it remains the same irrespective of the pathogen. The layer is characterised by the presence of *phagocytes* called *macrophages*.
- *Acquired Immunity*: Also known as *specific* because of its ability to tackle and adapt to different types of pathogens including ones never encountered before. It is composed of *lymphocytes* which are discussed later in this section

Of the above mentioned layers, acquired immunity require further explanation in the context of this work.

2.1.1 Acquired Immunity

Acquired immunity is characterised by the presence of cells called *lymphocytes* which can be broadly categorised in to *B-cells* and *T-cells*.

T-cells

They mature in the *thymus* with the primary function of signalling to regulate the immune system, for example, they propagate the type and location of antigen. When an antigen is destroyed, T-cells broadcast a message that ends the immune response and brings the immune system to its normal state [38].

B-cells

These are produced in the *bone-marrow* with receptors called *antibodies* attached to their surface to detect antigens. Similarly antigens are covered with molecules called *epitope* that allow them to be recognised by antibodies [5]. If a particular antibody matches with an epitope on the antigen, the B-cell bearing the antibody releases various similar antibodies to tackle the antigen. This process is explored further, later in this section.

Both the above mentioned cells are produced on the basis of a principle called *negative selection* which is discussed below.

Negative Selection

Lymphocytes of various affinities are produced in the thymus and bone-marrow. Although they are produced with the intention of protecting the body but some of the produced cells react against body's own cells which makes their survival dangerous for the body. To avoid the body being damaged by its own cells the immune system has a mechanism called *negative selection* to differentiate between body's own cells often called *self* with foreign cells called *non-self*. The negative selection or self/non-self discrimination ability enforces the deletion of cells that react against the self, as soon as they are produced.

Clonal Expansion

Once an antigen matches an antibody, the B-cell bearing the antibody undergoes the process of *clonal expansion*. Clonal expansion is a two step process namely *proliferation* and *differentiation*. In the first step the selected B-cell is cloned proportional to its affinity, to produce various similar B-cells. In the second step the clones with the highest affinity for the antigen gets converted into *memory cells* and *plasma cells*. The memory cells ensure a quick immune response for any similar infection in the future whereas the plasma cells generate bulk amount of antibodies that destroy the antigen.

Both negative selection and clonal expansion are shown in Figure 2.2. Of the above mentioned details, the ones we will incorporate in our AIS infrastructure are self/non-self, clonal expansion and affinity maturation.

2.2 Artificial Immune System (AIS)

Computational models and problem solving approaches inspired from the biological immune system are called artificial immune systems [5]. According to Morrison et al [18], *Artificial Immune Systems are adaptive search algorithms based on the biological immune system with the central task of pattern matching between antigens and antibodies*.

Dasgupta et al [4], have summarised the literature in the area of AIS in the past five years. The paper attempts to classify the prominent papers, from the year 1999 to the year 2003, on the basis of aspects of immune system modelled, representation used for modelling and their applications. It is recommended for readers who want to get themselves familiarised with the recent developments in this area.

Here we outline the literature related to AIS that was consulted while working on this project. Some of the literature is directly related to our work while the rest was consulted to better our understanding of AIS

- **Morrison et al [18]:** have applied the concept of immunity to the process of website recommendation. They have based their work on the principles of affinity maturation and Jerne's *idiotypic networks theory*, which holds that its not only antigens with which the antibodies interact, antibodies can interact with other antibodies as well.
- **de Castro et al [7]:** have based their work on the clonal selection principle and shown how an AIS could be used for pattern recognition and optimisation tasks. They also have presented an algorithm called CLON-ALG which is an evolutionary-like algorithm incorporating vocabulary from immunology instead of natural genetics and a shape-space formalism.
- **Lee et al [17]:** have applied the immune metaphor for information retrieval on the semantic web. They have presented a framework for exploratory information retrieval for multi-faceted datasets, for example,

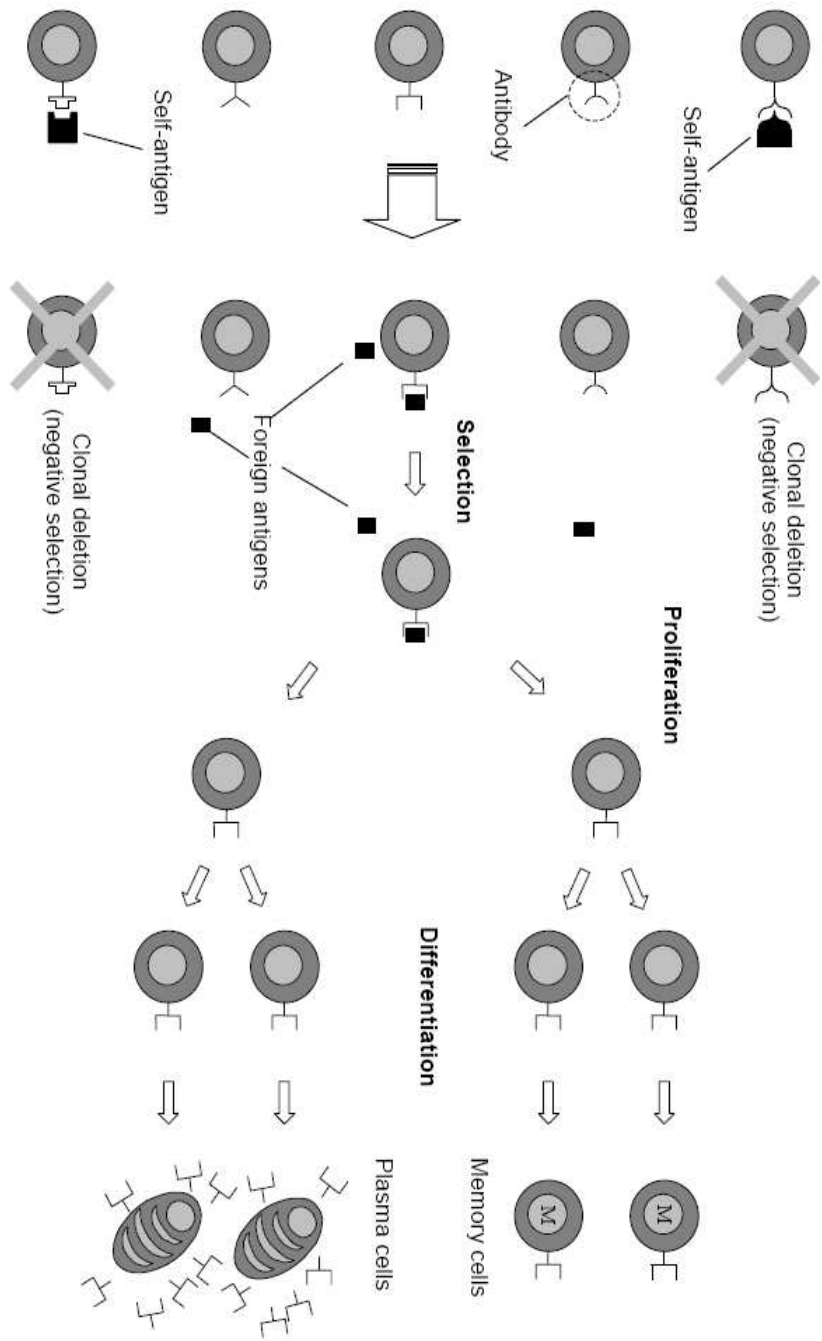


Figure 2.2: Negative selection and clonal expansion [8]

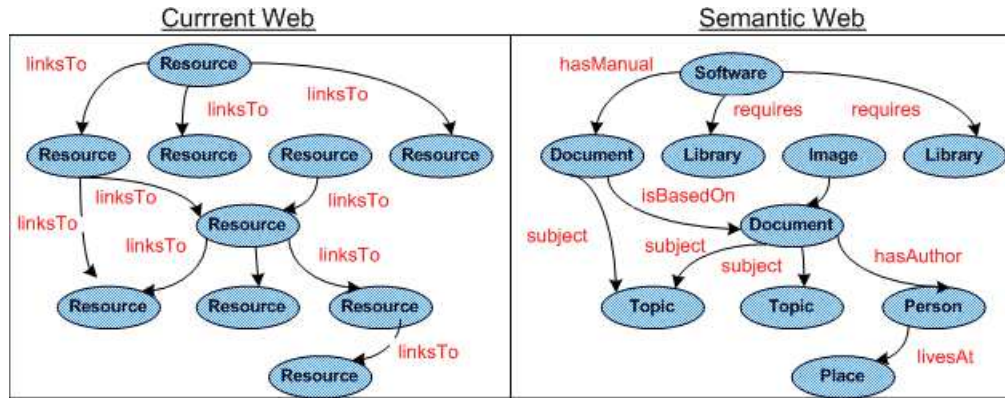


Figure 2.3: Difference between the current web and the semantic web [27]

Gene Ontology (GO) [11]. Their work is by far the most relevant work for this project and hence is of particular importance.

- **Twycross et al [37]:** have presented an immune based approach to document classification. It uses cooperative co-evolution to generate a classifier which consists of a set of detectors that can distinguish between good and bad samples.

Based on the literature searched in the area of AIS, some of the areas in which they have proved successful are pattern recognition, computer security, optimisation, dynamic learning and data mining. Some of the other resources that helped understanding AIS are [15, 1, 9]

2.3 Semantic Web

The current World Wide Web (WWW) is a collection of resources primarily web pages, linked to each other by means of hyperlinks only. Although the representation is useful for human consumption, it has definite drawbacks when the consumer is a machine and not a human.

Semantic web is an effort to make the data on the WWW machine understandable by creating well-defined relationships among various resources rather than a naive hyperlink. Putting it in the words of its founder, Tim Berners-Lee, *The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation* [2].

Figure 2.3 depicts the difference between the current WWW and the envisioned semantic web.

To materialise the idea of semantic web a key requirement is a standard way of adding semantics to the available data. In the realms of semantic web every



Figure 2.4: Structure of a triple

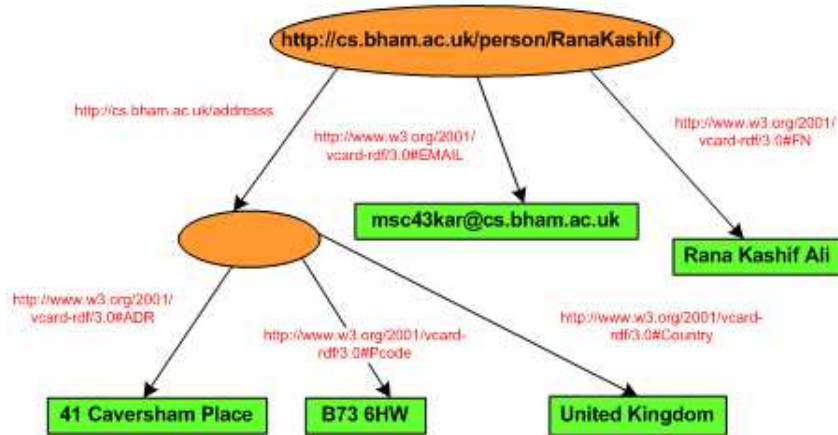


Figure 2.5: A simple RDF Model representing the author

thing is considered as a resource rather than a mere web page, the resource is uniquely identifiable by means of a Uniform Resource Identifier (URI). Every resource is conceptually related to various other resources as shown in the Figure. It is important to note that , it is not a requirement that a resource be directly accessible on the internet, for example, a mouse, keyboard or a monitor can be resources that are uniquely identifiable but their URI do not point to a specific web page on the internet.

2.3.1 Resource Description Framework (RDF)

RDF is a language for representing information about resources in the WWW [23]. It is a W3C specification with an abstract syntax that reflects a simple graph-based data model [22]. Any RDF model is a collection of assertions called *triples*. A triple has a very simple structure as shown in Figure 2.4.

Each triple represents a relation between two resources in the WWW that are identifiable via a URI. Combining these elementary triples we can obtain a graphical model as shown in Figure 2.5, which is a very simple model describing the person named Rana Kashif Ali.

For the purpose of automated reasoning these RDF models need to be serialised.

RDF/XML

When an RDF model is serialised to an XML it is called RDF/XML. The corresponding RDF/XML representation of the above model is shown in Figure 2.6.

```
-----  
<?xml version="1.0"?>  
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
  xmlns:dc="http://purl.org/dc/elements/1.1/"  
  xmlns:bham="http://cs.bham.ac.uk/"  
  xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#">  
  <rdf:Description rdf:about="http://cs.bham.ac.uk/RanaKashif">  
    <bham:address rdf:parseType="Resource">  
      <vcard:Pcode>B73 6HW</vcard:Pcode>  
      <vcard:Country>United Kingdom</vcard:Country>  
      <vcard:ADR>41 Caversham Place</vcard:ADR>  
    </bham:address>  
    <vcard:EMAIL>msc43kar@cs.bham.ac.uk</vcard:EMAIL>  
    <vcard:FN>Rana Kashif Ali</vcard:FN>  
  </rdf:Description>  
</rdf:RDF>  
-----
```

Figure 2.6: RDF/XML representation of the model

Notations 3 (N3)

Another textual representation of the RDF model is N3, invented by Tim Berners-Lee [19]. N3 version of the model is depicted in Figure 2.7.

RDF/XML is the recommended serialisation format by W3C. It is important to note that RDF/XML and N3 are only textual representations of the actual RDF graph and should not be misunderstood as being RDF.

2.3.2 Schemas and Ontologies

The RDF models presented above must follow standard vocabularies and relations which anyone can access to extract the meaning out of the RDF models. This is where schemas and ontologies come into play.

What is an ontology ?

We begin answering the question by quoting Gruber [12],

"an ontology is a specification of a conceptualisation"

```

-----
@prefix bham: <http://cs.bham.ac.uk/> .
@prefix vcard: <http://www.w3.org/2001/vcard-rdf/3.0#> .
@prefix : <http://cs.bham.ac.uk/#> .

:RanaKashif
  :address [ vcard:ADR "41 Caversham Place";
            vcard:Country "United Kingdom; vcard:Pcode "B73 6HW" ] ;
  vcard:FN "Rana Kashif Ali" ;
  vcard:EMAIL "msc43kar@cs.bham.ac.uk" .
-----

```

Figure 2.7: N3 representation of the model

According to the above definition of ontology, it provides a declarative formalism to the knowledge of a domain. For example an organisational ontology would describe all the relationships that exist among various concepts in the domain of organisations. Every domain has certain concepts which are linked to each other by relations, a declarative representation that captures all the concepts, relationships and inference rules is called the ontology for the particular domain. It is quite evident that ontology making is not a trivial task. It requires the author to be a very knowledgeable person in the domain. An ontology is usually very complex and has various facets for example Gene Ontology (GO) [11], as a result the data following a certain ontology is organised according to those facets. Figure 2.8 shows an organisational ontology with facets of topic, type, legal status and geography. For the purpose of simplification only two facets, that is, *type* and *topic* are further expanded in the figure. The red dotted lines represent an implicit relation between two different concepts within the same domain, that is, both these concepts can be used together when adding semantics to the data.

RDF Schema (RDFS)

RDF Schema is weaker form of ontology representation which provides an elementary level vocabulary and the facility to specify relationships between resources. More sophisticated ontologies can be produced using the RDF Schema.

2.3.3 Jena

Jena is an open source framework for building semantic web applications [16]. It is implemented in Java and is an open source initiative backed by HP. Jena includes

- API for RDF manipulation which facilitates the reading and writing of RDF in XML, N3 and N-triples

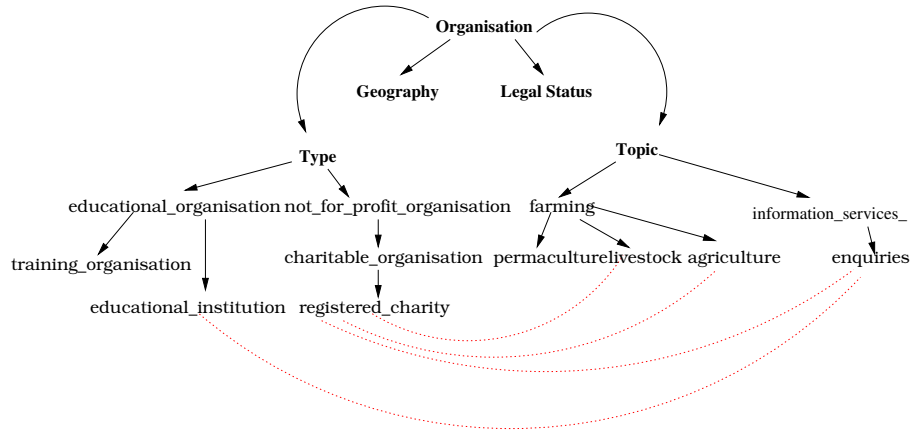


Figure 2.8: A slice of organisational ontology

- Ontology API
- A query language for RDF called RDQL which is a W3C Member submission [24]

Now we shall demonstrate the usage of Jena on our sample organisational data

RDF Manipulation

Jena, when provided with an RDF file in XML or N3 format creates an in memory representation of the model and allows reading and writing on that model. Since we only require the to read models and there is apparently no task of modifying the actual data, here we present how we can actually navigate the models using Jena. Below is that java code that uses Jena to read a file, *RDFTest* is the name of the class where the code resides.

```

-----
String fileName = "wwite_tidy.n3";
//create a default model
Model model = ModelFactory.createDefaultModel();

//establish an input stream for the data file
InputStream in =
RDFTest.class.getClassLoader().getResourceAsStream(fileName);

//create a reader based on the format of the data file
N3JenaReader reader = new N3JenaReader();

//read the data to create an in-memory representation.

```

```
reader.read(model, in, "");
```

```
//now the variable model is ready for manipulation
```

Once the Jena model is created it can be used for manipulation, for example, the following code demonstrates how we can get all those records which have the property EMAIL.

```
-----  
//acquire an iterator over the desired resources  
Iterator iter = model.listSubjectsWithProperty(VCARD.EMAIL);
```

```
//fetch the individual records  
while(iter.hasNext()) {  
  //fetch the resource  
  Resource r = (Resource)iter.next();  
}
```

Resource Description Query Language (RDQL)

The RDF manipulation that was shown above provides a basic level of querying which is not sufficient for our task, a more flexible approach is needed to query the models. RDQL provides a declarative approach to query a model. Its grammar resembles to that of Structured Query Language (SQL). RDQL can be used as a command line tool or from within the Java code. Here we demonstrate how we can query an organisational data to get all the organisations which are registered charities, deal with animal welfare and were established before the year 1900. The RDQL query would be of the form,

```
-----  
SELECT  ?resource, ?name
```

```
WHERE
```

```
(?resource swed:has_organisation_type swed_ot:registered_charity)  
(?resource swed:has_topic  
<http://jena.hp1.hp.com/2004/02/swed/wwrite_index#animal_welfare>)  
(?resource swed:has_primary_prorg_name ?name)  
(?resource swed:has_year_formed ?year)
```

```
AND
```

```
?year <= 1900
```

```
USING
```

```
swed FOR <http://jena.hp1.hp.com/2004/02/swed#>  
swed_ot FOR <http://jena.hp1.hp.com/2004/02/swed/org_type#>
```

Here *swed* and *swed.ot* are name spaces which define the relation *has_organisation_type* and concept *registered_charity* respectively. The USING clause here prevents the use of fully qualified names in the WHERE clause. Another mentionable thing is that all the variables are preceded by the question mark.

Assuming we have the query in a String variable in the Java code, the results can be obtained by invoking the Jena Query Engine as follows

`Query query = new Query(queryString);`
`query.setSource(model);`
`QueryExecution qe = new QueryEngine(query);`
`QueryResults results = qe.exec();`

2.4 Query Expansion

Query Expansion is the process of iteratively reformulating an initial query by appending closely related words to the original query. The intent of expansion is to increase the precision and/or recall by retrieving more and more relevant results.

2.4.1 Types of Query Expansion

Research in the area of query expansion can be traced back to 1972 and is still ongoing. Efthimiadis has summarised the vast area of research in [10]. According to Efthimiadis query expansion can be performed in one of the following three ways

Manual Query Expansion (MQE)

In MQE much of the task of formulating the query is left to the searcher. The searcher performs an initial key-word based search, based on the results and personal knowledge adds more terms to the initial query. The process is repeated over and over again to with both success and failure in each iteration unless the user achieves the desired results. This approach relies on the user a bit too much and generally a user with a better understanding of the area being searched is able to get more relevant results as compared to a novice user. Another drawback of this approach is the time and effort required to achieve the desired results.

Automatic Query Expansion (AQE)

AQE also known as pseudo-relevance feedback hides the process of expansion from the user. This type of expansion can be completely independent for example the initial 10 results might be used from the initial query and some statistics

applied to those results to extract terms that can be used for expansion. A user feedback in the form of yes-no, relevant-irrelevant can be also be taken to direct the expansion process. The user feedback is very subtle and the process has the capability of expanding on its own if no user feedback is available.

Interactive Query Expansion (IQE)

In IQE the expansion process proceeds with a constant interaction of the system and the user. The system is primarily responsible for letting the user know of possible expansion terms, the user interacts by specifying or selecting the terms that he/she wants to be used in the expansion process. As compared to AQE the user interaction level is very high in this process.

2.4.2 Query Expansion for the Semantic Web

The literature in the area of query expansion is useful to some extent but we have not come across any specific methods of query expansion for the semantic web except for [17]. In the approach presented in this report, the expansion process is automatic and relies on the *ontology* and *relevance-feedback*. The expansion process continues with or without user feedback. No user feedback is given a slightly positive weight as discussed in chapter 3 and the expansion process continues on its own.

Central to our approach is the concept of *relevance feedback*, first proposed in 1971 [25], has been shown to increase the quality of results obtained as a result of expansion [3]. User feedback is meant to remove results that are irrelevant and to create a pool of relevant results without the user explicitly changing the search strategy.

On the semantic web, where the data is kept in a standardised manner with relations between resources properly defined, the task of expansion becomes relatively easier as compared to the traditional world wide web. The possible expanding direction of the query is guided by the terms taken from the *ontology* i.e. there is no need to search for similar terms within the results retrieved or consult a thesaurus for synonyms. It is important to note that the queries on the semantic web are based on the document meta-data rather than content, in other words the search is not based on key words but concepts. Table 2.1 outlines the differences between the current web and the semantic web in the context of query expansion.

Efthimiadis has identified the elements that need to be taken under consideration for any form of query expansion

- Source for the expansion terms
- Method used to select the terms to be used in the expansion process

The source for the expansion terms in our case would be the ontology and the method used to select the terms would be the AIS infrastructure or more precisely the GA working inside it.

Table 2.1: semantic web vs. current web in the context of query expansion

Current Web	Semantic Web
keyword based	concept based
keywords are not linked to one another	concepts are linked to other concepts
thesaurus can be consulted to add synonyms to the original query	ontology is much richer than thesaurus
queries are content oriented	queries are meta-data oriented

Chapter 3

AIS for Semantic Query Expansion

In this chapter, first, an analogy is established between the AIS and the HIS. The algorithm at the core of our approach is then presented followed by explanation of its constituents and how the semantic query expansion process occurs within it. Finally, a distinction is drawn between an AIS and a Genetic Algorithm (GA).

3.1 Mapping between the AIS and HIS

It is important at this stage to map the AIS and HIS concepts so that terms from both can be used interchangeably. Table 3.1 shows the mapping.

The justification behind representing an *antigen* as a collection of all the *non-self* is pretty straight forward. Both *antigen* and *non-self* are something that is not part of the body and hence are considered same in our approach. An *antigen* can either be harmful or harmless. In biology harmful antigens are recognised (and destroyed) by the immune system whereas we are interested in recognising harmless antigens. It is to be noted that there are certain aspects of the HIS that are not modelled in the presented work primarily because their incorporation was not immediately apparent. Some of the aspects that are not

HIS	AIS
Self	Results marked as irrelevant by the user
Non self	Results marked as relevant by the user
Antibody	Semantic Queries
Antigens	All the relevant results (non self)
Mutation	Semantic query expansion

Table 3.1: Mapping between the HIS and the presented model

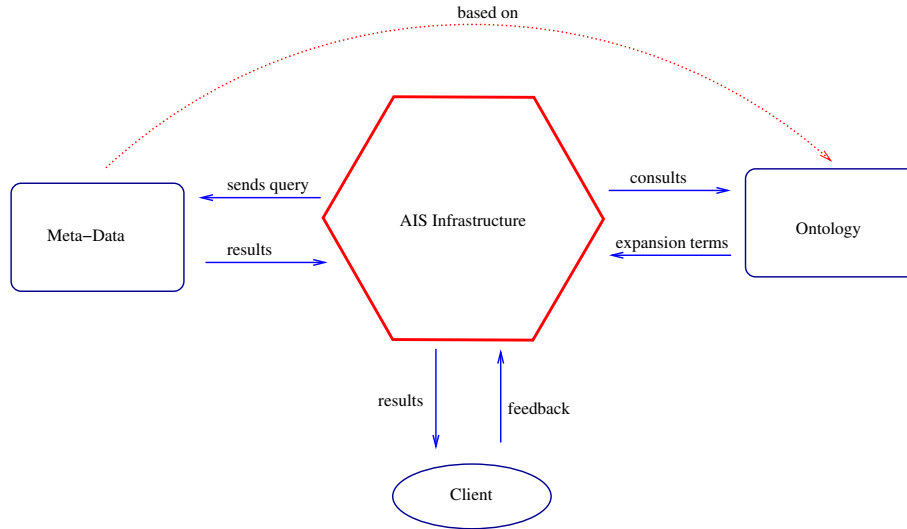


Figure 3.1: AIS Infrastructure and flow of information

modelled include immune memory and antibody-antibody interaction.

3.2 Overall Structure and Flow of Information

Before delving deep into the complexities of the AIS infrastructure a high-level view is presented in Figure 3.1. The figure shows the flow of information among the AIS, meta-data, ontology and the client. Its aimed at establishing a clear picture in the mind of the reader that how would the system actually works.

3.3 The Algorithm

The presented approach is based on the algorithm shown in Figure 3.2. Further explanation of the constituents of the algorithm, for example individuals, selection, mutation schemes and the halting criteria, is in the following sections.

3.3.1 Initial user feedback

Initial user feedback involves the user specifying one organisation of interest and saying that s/he wants to find similar organisations. The web interface displays names of all the organisations to the user. The user may click one organisation of interest to see its details. The organisational details page bears a link titled 'FIND SIMILAR', clicking on this link triggers the AIS initialisation and the adaptive search process. Due to the small size of organisations (100) this approach worked well for us but for huge data sets it would not be nice to

- **begin**
- take initial user feedback
- initialise \mathbf{Q} , query population based on feedback
- **while** (halting criteria not met)
 - **begin**
 - display the results of queries in \mathbf{Q} and take user feedback
 - add relevant results to *non-self* and irrelevant to *self*
 - construct an *antigen* representation from the *non-self*
 - evaluate affinities of queries in \mathbf{Q}
 - select queries with highest affinities, \mathbf{Q}_s using fitness proportionate selection
 - perform clonal expansion on the selected queries to form \mathbf{Q}_c
 - apply mutation operator to transform \mathbf{Q}_c to \mathbf{Q}_m
 - replace the the previously selected queries \mathbf{Q}_s with \mathbf{Q}_m
 - **end**
- **end**

Figure 3.2: The AIS algorithm

display the names of all organisations (probably thousands) for a user to select a particular organisation of interest (starting point).

3.3.2 Initialisation of AIS

When the user clicks the 'FIND SIMILAR' link the AIS is initialised with a query population of size equal to `INIT_POP_SIZE`. The initial query population is generated randomly using the *type* and *topic* ontologies. Queries with non-empty result set are regarded as valid queries whereas queries that do not produce any result are invalid. Throughout the AIS we ensure that only valid queries propagate. The pseudo code for initialisation of the AIS is give below

```
-----  
set generatedQueries = 0  
while(INIT_POP_SIZE > generatedQueries) {  
- randomly select a type and assign it to the new query  
- generate a random number, count, between 0 and MAX_TOPICS_IN_QUERY  
- select count number of topics randomly from the ontology  
- combine the type and topics to make a query  
- if(query produces some results) {  
    generatedQueries++;  
    add query to the AIS  
  }  
}  
-----
```

Once the AIS is initialised, the antibodies/queries within it are extracted and displayed along with their results. The user may give feedback by specifying whether a particular result is irrelevant/relevant (self/non-self). The results specified as relevant are added to the *self* part of the AIS whereas results marked as irrelevant are first given an antigen representation and then added to the AIS. In the source code, each result is represented by the java class *SemanticResult.java*. The antigen is merely a collection of these results with the relevance property equal to true. Now we talk about the genotype and phenotype of the two individuals in the AIS ie. antibody and antigen.

Genotype and Phenotype

In any nature inspired approach an essential requirement is the formal representation of the individuals and their genetic composition. This representation, in case of simple problems, can be a binary string. However as more problem specific information is incorporated the representation tends to be much more than a binary string. Representation of individuals should be appropriate as well as flexible ie. it should capture interesting features of the search space. We have used this as the guiding principle in representing the individuals and their genetic composition i.e. the phenotype and the genotype. The phenotype, in case of an antibody, is the RDQL query whereas the genotype is a java class

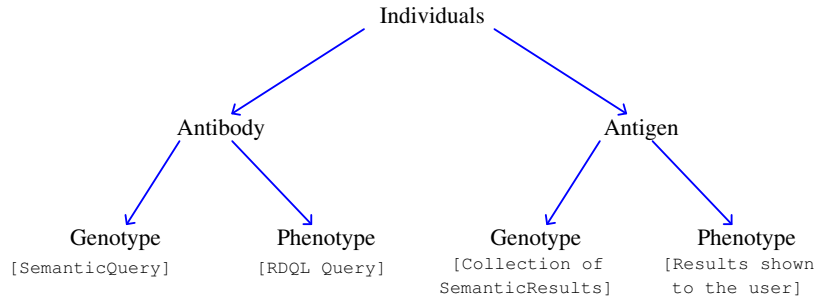


Figure 3.3: Genotype and Phenotype of individuals

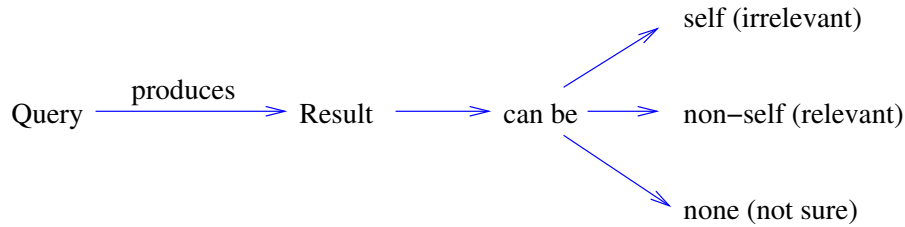


Figure 3.4: Types of Result Feedback

called *SemanticQuery* which has a basic template of an RDQL query and additional information for example type and topic information required for the query construction. It also holds the results produced by that particular query. On the other hand the genotype of the antigen is a collection of instances of the java class *SemanticResult*. The individuals within the AIS along with their genetic composition are shown in Figure 3.3. It is important to note that the *SemanticQuery* is a java class with various properties as mentioned above whereas the RDQL query is actual query string used to fetch data using Jena. Similarly the class *SemanticResult* is a formal representation of the results shown to the user.

3.3.3 Affinity Evaluation

Once the user has given some feedback affinities of antibodies/queries need to be evaluated. Affinity of an antibody in our case is measure of how well it binds to the non-self. For example, if a query binds to greater number of non-self and smaller of number of self then it should have a higher fitness. If on the other hand the query binds neither to self nor to non-self then it should have a medium affinity. Basically we had three different types of results to deal with while evaluating affinities of antibodies, shown in Figure 3.4.

We used the following formula to evaluate the affinities

$$affinity = \frac{non-self \times positive\ wt + self \times negative\ wt + none \times neutral\ wt}{total\ number\ of\ results}$$

Where the values used for different weights are, *positive wt* = 1, *negative wt* = 0 and *neutral wt* = 0.4. The choice of the values for different weights was empirical. Figure 3.5 demonstrates how the binding measure is combined with the formula to evaluate the affinity of antibodies.

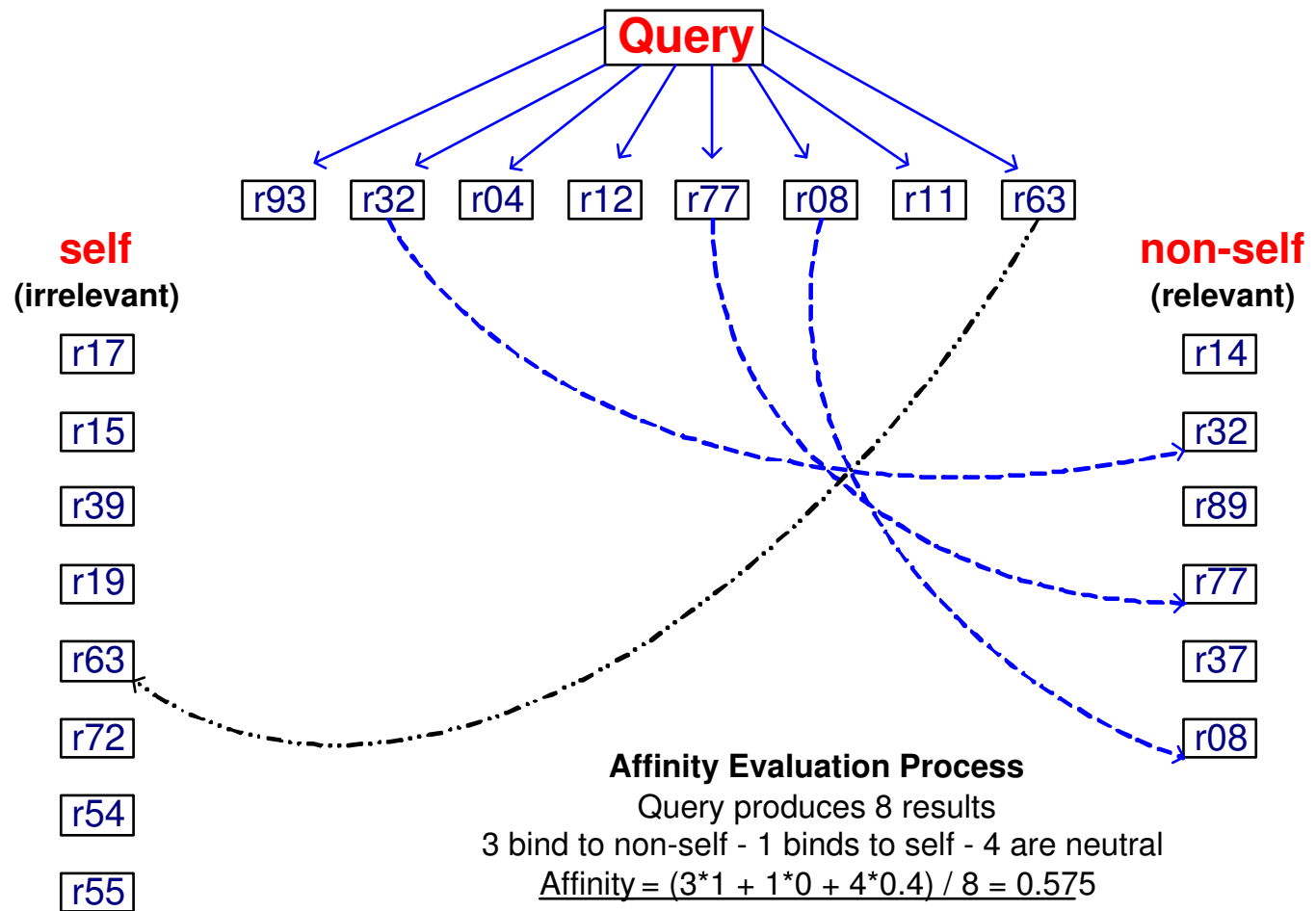


Figure 3.5: Affinity Evaluation Process

3.3.4 Selection

In a pure AIS individuals are selected so as to maximise the collective affinity against the antigen called *affinity maturation*. It is important to note that affinity maturation like roulette wheel selection is fitness proportionate and has a similar effect. For this reason we have incorporated roulette wheel selection process in the AIS. The input parameter in this phase of the algorithm is, SELECTION_COUNT , the number of individuals to be selected. The pseudo code of how the roulette wheel was setup is given below.

```
-----  
- evaluate the sum of affinities of all antibodies call it sum  
- assign selection probability to each antibody equal to affinity/sum  
- create an array of size equal to the input parameters WHEEL_SIZE  
- set toIndex = 0 and fromIndex = 0  
- for each antibody {  
    toIndex = fromIndex + selectionProbability * WHEEL_SIZE  
    fill the array with antibody between fromIndex and toIndex  
    set fromIndex = toIndex  
  }  
-----
```

Finally, pseudo code for spinning the roulette wheel and selecting an antibody is pasted below

```
-----  
- generate a random number between 0 and WHEEL_SIZE call it index  
- return the antibody at location index in the array  
-----
```

As mentioned above roulette wheel selection was used since it produces similar effect as affinity maturation, however other selection schemes such as tournament selection and stochastic uniform selection can also be used and are left as part of future work.

3.3.5 Clonal Expansion

Clonal expansion involves generation of fitness proportionate closely similar clones. The higher the fitness of an antibody the more number of clones would be produced. This process is tightly coupled with the mutation process where we mutate the clones to produce various similar individuals. Once again there are certain constraints which need to be fulfilled for the results of the query to be displayed to the user. The constraints are summarised below

$$(R - (S \cup N) \neq \phi) \quad (3.1)$$

Where R is the result set of a query, N represents the set of non-self (relevant results) and S represents a set of self (irrelevant results).

The clonal expansion process is shown in Figure 3.6.

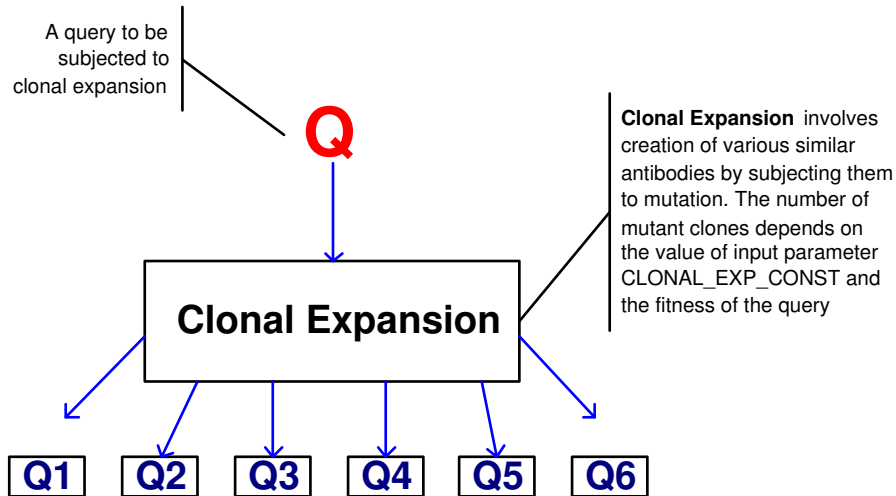


Figure 3.6: Clonal Expansion Process

3.3.6 Mutation Operators

As mentioned above, the various similar individuals produced as a result of clonal expansion undergo the process of mutation. Here we discuss the two query mutation operators that we developed and how mutation is carried out using them.

Constrained Mutation Operator

This mutation operator maintains most of the existing characteristics of the individual and only appends additional characteristics to it. When a query is subjected to this type of mutation either the same *type* of the query is retained or a *type* is randomly chosen from the type ontology. The change in type is governed by the input parameter TYPE_CHG_PROB. In terms of topics, each topic may have to undergo one or all of the following three operations, depending on the input parameter MUTATION_RATE.

- Append
- Delete
- Change

Pseudo code for the *ConstrainedMutationOperator* is given below and the complete java source code is included in the appendix titled *Source*.

- generate a random number between 0 and TYP_CHG_PROB

```

- if(random < TYP_CHG_PROB) {
    replace existing type with a one randomly chosen from the ontology
  }
- else {
    retain the old type
  }
- define three variable, append, delete and change
- initialise the variables with random numbers between 0 and 1
- for each topic in the query {
    if(append >= MUTATION_RATE) {
      append a random topic to the query
    }
    if(delete >= MUTATION_RATE) {
      delete the topic from the query
    }
    if(change >= MUTATION_RATE) {
      replace the topic with a randomly selected topic
    }
  }
}

```

Random Mutation Operator

In case of this operator the change in type is similar to the *ConstrainedMutationOperator* but it differs in how the topics are selected for the mutant query. The topics for the off spring are selected with probability equal to MUTATION_RATE from the topics of the parents and all the available topics in the ontology. There is also a constraint on the maximum number of topics to be selected. The algorithm is presented below and the source is included in the appendix.

```

- generate a random number between 0 and TYP_CHG_PROB
- if(random < TYP_CHG_PROB) {
    replace existing type with a one randomly chosen from the ontology
  }
- else {
    retain the old type
  }
- generate a random number between 1 and MAX_TOPICS_IN_QUERY, count
- while(count != 0) {
    generate a random number between 0 and MUTATION_RATE, rate
    if(rate <= MUTATION_RATE) {
      choose a topic randomly from the old query and add to the new one
    }
    else {

```

```
    choose a topic randomly from the topic ontology
    add the topic to the new query
  }
}
```

Diagrammatic representation of the mutation process using the above mentioned operators is shown in Figure 3.7. Other mutation operators namely *SpecialisingMutationOperator* and *GeneralisingMutationOperator* were also planned but were not implemented due to shortage of time. They are mentioned in the future work section.

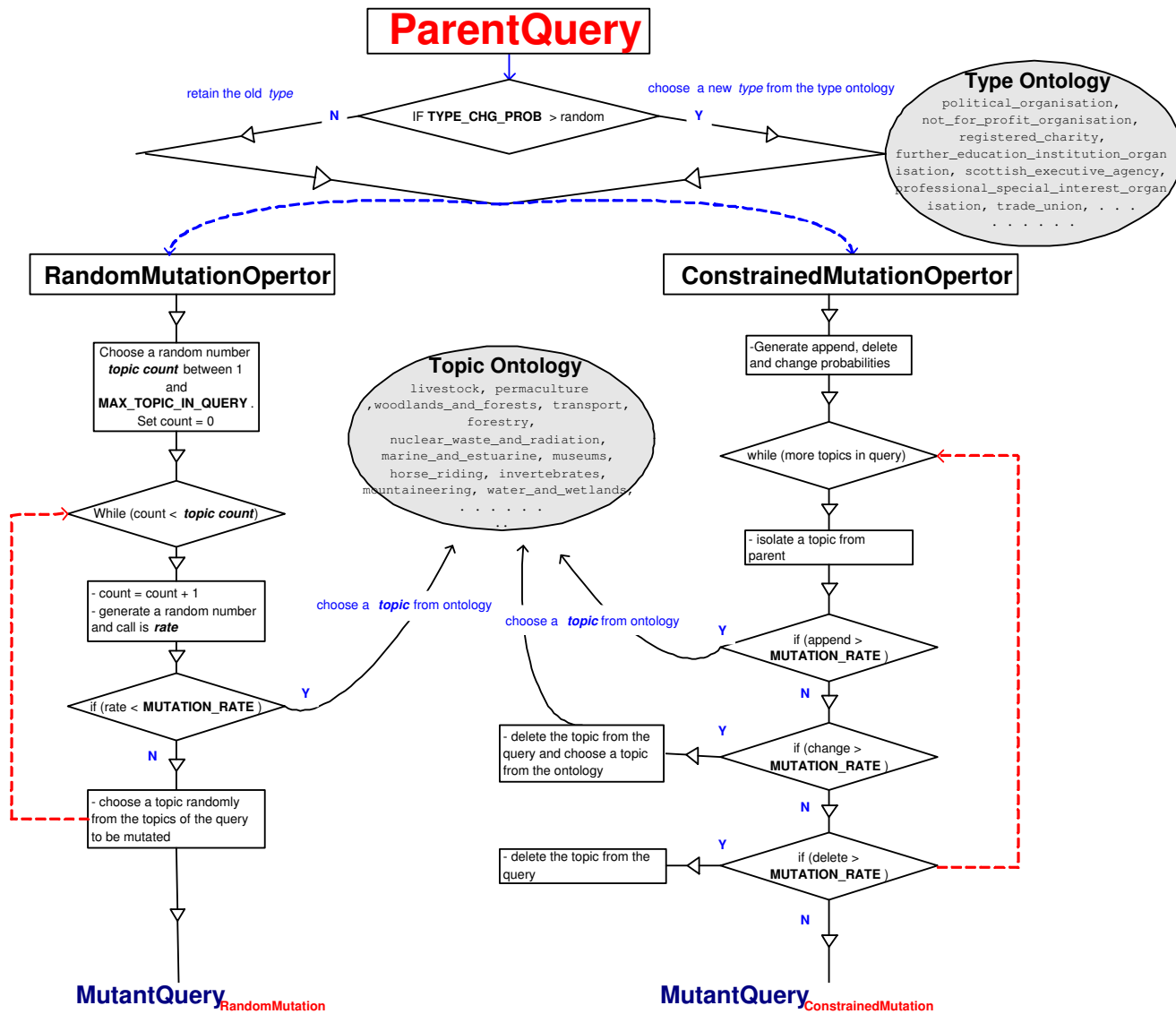


Figure 3.7: Working of the mutation operators

Parameters	Description
MAX_TOPICS_IN_QUERY	Maximum number of topics in a query, used only in case of <i>RandomMutationOperator</i>
INIT_POP_SIZE	The number antibodies to produce when the AIS initialises
SELECTION_COUNT	The count of antibodies to be selected
CLONAL_EXP_CONST	Clonal expansion constant
WHEEL_SIZE	Clonal expansion constant
MAX_QUERY_ATTEMPTS	Number of attempts to be made to generate a valid antibody
MUTATION_OPERATOR	The mutation operator used by the AIS, can either be <i>RandomMutationOperator</i> or <i>ConstrainedMutationOperator</i>
TYP_CHG_PROB	Probability of changing the type of a query
MUTATION_RATE	Rate of mutation

Table 3.2: Interpretation of the parameters

3.3.7 Replacement Strategy

All individuals that are selected during the *selection phase* are replaced with the offspring. The unselected individuals however, remain in the AIS to maintain diversity in the population.

3.3.8 Halting Criteria

There can be two possible halts to the search process. Firstly, the user can end the search if s/he has found all the desired results. Secondly, the process would come to an end if further query expansion is not possible. The scenario in which further expansion is not possible is when no query can be formed that satisfies the constraints in Equation 3.1. Although on a large data set this scenario is very unlikely but to prevent the algorithm from entering an infinite loop an upper limit to the number of attempts to be made to generate valid queries can be supplied as an input parameter to the AIS. The name of the input parameter to achieve this is MAX_QUERY_ATTEMPTS.

3.4 Input Parameters and their Description

There were various input parameters to the AIS. The parameters along with their description are mentioned in the Table 3.2

3.5 AIS vs GA

The AIS being a nature inspired approach resembles quite closely with a GA. However, there are certain definite differences between the two which are explained below

- AIS has a concept of *self* and *non-self* which helps it to learn and adapt to changes.
- In an AIS selection of individuals occurs so as to maximise the overall affinity, a process called affinity maturation. GAs on the other hand use selection schemes such as roulette wheel and tournament. However since affinity maturation is a fitness proportionate selection like roulette wheel it may have a similar effect.
- Another contrasting characteristic is the presence of the process called *clonal expansion* which is absent in GAs.
- In a pure AIS there is a concept of *concentration* of antibodies and the overall ability of an antibody to survive depends on two factors namely *concentration* and *affinity*. This is different from a GA where fitness is the only criteria for survival.
- Though not modelled in the presented work, an AIS has a concept of *immune memory*. Quick response to similar infections is possible only because of the presence of this feature.
- An AIS involves interaction among individuals for example antibody-antigen and antibody-antibody interactions. In a GA however, individuals do not interact directly with each other.
- Finally, a very subtle difference is the vocabulary used in each approach. GAs use vocabulary taken from natural evolution whereas AIS use vocabulary taken from immunology.

Chapter 4

AIS based Semantic Search Utility

The project involved the development of an AIS based semantic search utility. This chapter not only discusses the implementation details but also demonstrates the functionality of the developed application.

4.1 Implementation Details

Since the search utility was intended to be a web application the code naturally splits into two areas namely, *back-end* and *front-end*.

Back-End Code

The AIS infrastructure explained in the previous chapter forms the *back-end* of the application. It is responsible for various tasks some which are, query expansion, user feedback analysis and fitness evaluation etc. All the code was written in Java with extensive use of the *Jena* semantic web framework by HP. The *back-end* code was designed to be very loosely coupled so that any one wishing to extend the work can simply plug the new implementations of AIS components such as fitness function, mutation operators, genotype, phenotype, antibody, antigen, selection scheme and even the AIS. A simplified class diagram of the *back-end* code is shown in Figure 4.1.

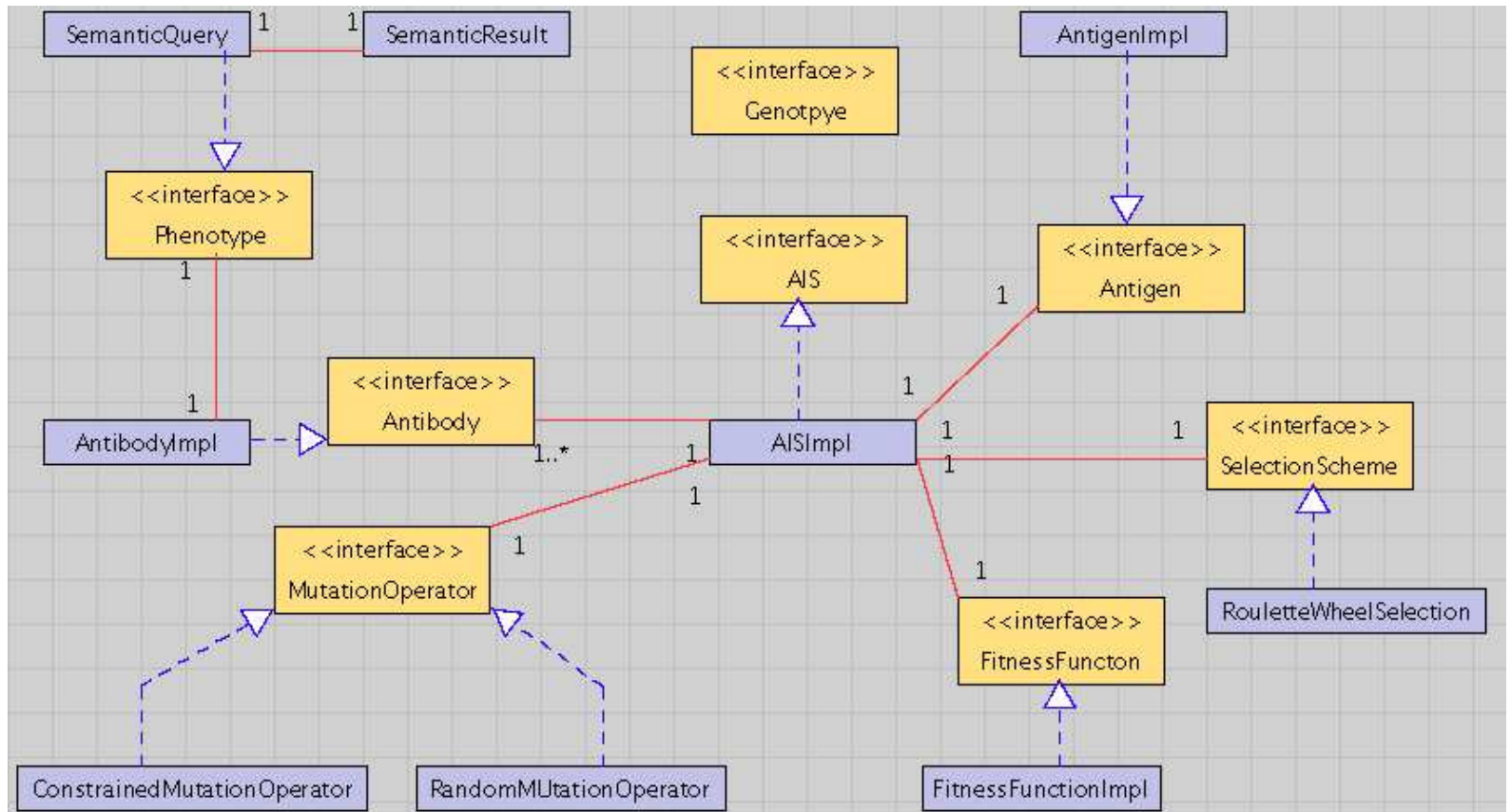


Figure 4.1: Class Diagram

The entities with the label *interface* are Java interfaces and the arrow sign indicates an interface implemented by a particular class. Consider the interface *SelectionScheme* as an example which is implemented by the Java class *RouletteWheelSelection*. If we want to test the AIS with a different selection scheme, lets say tournament selection, we would only need to create another Java class *TournamentSelection* that implements the *SelectionScheme* interface. In essence, a Java interface isolates common characteristics from similar entities and provides flexibility and seamless incorporation of various concrete implementations. Finally, the solid lines with numbers on both their ends represent the *association* relationship in Java. For example the *association* between the class *AISImpl* and *Antibody* is one-to-many implying there are one or more antibodies in an AIS.

Front-End Code

To present the dynamic content generated by the AIS *back-end*, Java Servlets and Java Server Pages (JSP) were used to develop the *front-end*. JSP combines the power of Java and Hyper Text Markup Language (HTML). *Tomcat* servlet container was used to serve the JSPs on the web.

4.2 Usability

This section is an illustration of how the developed application can be used for information retrieval.

Starting Point

At the very beginning the user is displayed with all the organisations in the data file. The user is expected to click on any organisation of interest, this brings up the organisational details page shown in Figure 4.2. From here there is a link named *Find Similar*, clicking on the link activates the AIS which generates relevant queries. The results of the generated queries are then displayed. These results are grouped by queries and sorted by fitness of queries, the higher the fitness of the query the upper position it occupies. Figure 4.3 shows the results grouped by query.

Search Process

The interface shown in Figure 4.3 facilitates user feedback by allowing the user to specify what is relevant and what is not. Once again *SELF* refers to irrelevant results and *NON-SELF* refers to relevant results. User feedback is then sent to the AIS once again which adapts and generates more queries which are presented in a similar way. At any point the user can view the rated results, which can either be relevant or irrelevant by clicking the links named *SELF* and *NON-SELF* on the top left and right of the main interface. This pops up the pages shown in Figure 4.5 and 4.4.

Organisation Details	
National Energy Action	
Organisation number	urn:x-hp-swed:prorg0064
Acronym	NEA
Year formed	1981
Description	NEA aims to confront, publicise and solve the social problem of fuel poverty. It tries to achieve this through informed representations to interested parties, whether central government, local government, fuel boards, etc. In addition, NEA promotes practical work to alleviate the problem in the form of work carried out to insulate and draught-proof the homes of low-income households, particularly for the elderly, disabled and one-parent families.urn:x-hp-swed:prorg0064
Type	registered_charity
Topics	information_services_enquiries education policy_and_policy_development education_training energy
Telephone	voice: http://www.w3.org/2001/vcard-rdf/3.0#fax fax: 0191-261 5677
Email	-
URL	http://www.nea.org.uk/
Primary Contact	Information Officer-name withheld-
Postal address	England Tyne and Wear Newcastle upon Tyne St Andrew's House NE1 6GG 90-92 Pilgrim Street
FIND SIMILAR	

Figure 4.2: Organisation details

These interfaces allow the user to remove any result that might have been chosen by mistake. There are some other links as well which were initially included for testing purposes but give a good insight into the whole process of query expansion and information retrieval. For example, clicking the *CURRENTLY SELECTED QUERY POPULATION* link brings up a window displaying the fitness, RDQL and description of all the selected queries. This is shown in Figure 4.6

End of search

The search process can be terminated by the user at any point when he/she feels that all the relevant results have been identified. The search also ends when all the result space has searched and there are no more results that are neither part of the self nor non-self. On a huge data set the latter would be very rare to observe.

RESULTS				
SELF	PREVIOUSLY SELECTED QUERY POPULATION	CURRENTLY SELECTED QUERY POPULATION	COMPLETE CURRENT POOL [selected+nonslected]	NON-SELF
1				
IRRELEVANT	QUERY SELECT organisations where TYPES = [private_limited_company] AND TOPICS = [recreation]			RELEVANT
<input type="checkbox"/>	Festival of the Countryside			<input type="checkbox"/>
2				
IRRELEVANT	QUERY SELECT organisations where TYPES = [registered_charity] AND TOPICS = [built_environment]			RELEVANT
<input type="checkbox"/>	Cathedral Camps			<input type="checkbox"/>
<input type="checkbox"/>	Campaign for the Protection of Rural Wales			<input type="checkbox"/>
<input type="checkbox"/>	National Trust for Scotland			<input type="checkbox"/>
<input type="checkbox"/>	Action with Communities in Rural England			<input type="checkbox"/>
<input type="checkbox"/>	Barn Owl Trust			<input type="checkbox"/>
3				
IRRELEVANT	QUERY SELECT organisations where TYPES = [registered_charity] AND TOPICS = [animal_welfare]			RELEVANT
<input type="checkbox"/>	Zoological Society of London, The			<input type="checkbox"/>
<input type="checkbox"/>	Barn Owl Trust			<input type="checkbox"/>
<input type="checkbox"/>	National Animal Welfare Trust			<input type="checkbox"/>
<input type="checkbox"/>	Humane Slaughter Association			<input type="checkbox"/>
4				
IRRELEVANT	QUERY SELECT organisations where TYPES = [private_limited_company] AND TOPICS = [business_and_commerce]			RELEVANT
<input type="checkbox"/>	Planning Exchange, The			<input type="checkbox"/>
5				
IRRELEVANT	QUERY SELECT organisations where TYPES = [registered_charity] AND TOPICS = [developing_world]			RELEVANT
<input type="checkbox"/>	Pesticides Trust, The			<input type="checkbox"/>
<input type="checkbox"/>	Trust for Education and Development			<input type="checkbox"/>

Figure 4.3: Results grouped by queries

NON - SELF

Vision 21 - Action for Sustainable Communities

Vision 21 Scotland is based on the conviction that our society faces major change; that sustainable development demands both new and integrated policies at all levels, and fundamental changes in communities and lifestyles.

The mission statement of Vision 21 is: 'To nurture a common vision of our society into the 21st Century; practical changes in lifestyle and local communities; and the development of integrated policies for a just, participatory, secure and sustainable society.' [\[REMOVE\]](#)

John Muir Award

The Award aims to promote the social and personal development of people through the conservation of wild places. It is non competitive and open to all.

The Award offers three levels: Discovery (introductory), Explorer (intermediate) and Conserver (advanced). Each level requires a greater effort and commitment but has the same four challenges:

- Discover a wild place;
- Explore its wildness;
- Take personal responsibility for its protection; and
- Share your discoveries.

The award is delivered in partnership with youth organisations, NGOs and local authorities. [\[REMOVE\]](#)

Northern Ireland Environment Link

NIEL acts as a forum for voluntary organisations concerned with conservation of the countryside, wildlife, and the environment of Northern Ireland. It acts as a link between members and the government, provides information on environmental matters and promotes understanding through increasing public awareness. [\[REMOVE\]](#)

Regenerative Technology

Regenerative Technology investigates methods of natural therapy appropriate to the needs of people, their habitat, and their environment. [\[REMOVE\]](#)

[close](#)

Figure 4.4: NonselF results

SELF
<p>International Bee Research Association</p> <p>IBRA seeks to increase people's awareness of the vital role of bees in agriculture and the natural environment. Promoting the study and conservation of all types of bees. [REMOVE]</p>
<p>Global Witness</p> <p>Global Witness focuses on areas where environmental exploitation causes human rights abuses and funds conflicts. Information obtained through research and field investigations is used to brief governments, intergovernmental organisations, NGOs and the media, in order to achieve positive change. Global Witness has no political affiliation. [REMOVE]</p>
<p>Scottish Society for the Prevention of Cruelty to Animals</p> <p>The objectives of the Scottish SPCA are to prevent cruelty to animals and to encourage kindness and humanity in their treatment. The Society attempts to achieve this in Scotland through its inspectors, by education and by assisting in the formation of new and improved legislation. Fourteen Animal Welfare Centres, which care for domestic, farm and wild animals, including cleaning oiled birds and rehabilitating seals, are run by the Society. [REMOVE]</p>
<p>Welsh Federation of Young Farmers' Clubs</p> <p>The YFC is a voluntary organisation for young people in the countryside. A democratic system is operated whereby members decide from a wide range of activities which they would like to participate in at club/county/welsh level. The organisation's principle objective is to facilitate the personal development of the members and to enable them to participate fully in the community.</p> <p>The Welsh YFC organises a Rural Life Project Scheme where clubs are invited to undertake a particular scheme within their communities based on an environmental, community or heritage theme. [REMOVE]</p>
<p>Royal Highland and Agricultural Society of Scotland</p> <p>The Society is involved in agricultural promotion and education. It is proprietor of the Royal Highland Centre, and organises the Royal Highland Show. [REMOVE]</p>
<p>Vincent Wildlife Trust</p> <p>The Vincent Wildlife Trust operates an otter rehabilitation centre for orphaned or injured otters from throughout the UK with reintroductions occurring in Northern Ireland and eastern England. [REMOVE]</p>
<p>close</p>

Figure 4.5: Self results

Selected Queries [sorted by fitness]	
1	<p>FITNESS: 0.7999999999999999</p> <p>DESCRIPTION: SELECT organisations where TYPES = [voluntary_sector_organisation] AND TOPICS = [ecology]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:voluntary_sector_organisation) (?resource swed:has_topic swed_toi:ecology) USING swed FOR swed_toi FOR swed_ot FOR</p>
2	<p>FITNESS: 0.7</p> <p>DESCRIPTION: SELECT organisations where TYPES = [non_departmental_public_body] AND TOPICS = [built_environment]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:non_departmental_public_body) (?resource swed:has_topic swed_toi:built_environment) USING swed FOR swed_toi FOR swed_ot FOR</p>
3	<p>FITNESS: 0.7</p> <p>DESCRIPTION: SELECT organisations where TYPES = [professional_special_interest_organisation] AND TOPICS = [ecology]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:professional_special_interest_organisation) (?resource swed:has_topic swed_toi:ecology) USING swed FOR swed_toi FOR swed_ot FOR</p>
4	<p>FITNESS: 0.5</p> <p>DESCRIPTION: SELECT organisations where TYPES = [voluntary_sector_organisation] AND TOPICS = [recreation]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:voluntary_sector_organisation) (?resource swed:has_topic swed_toi:recreation) USING swed FOR swed_toi FOR swed_ot FOR</p>
5	<p>FITNESS: 0.4</p> <p>DESCRIPTION: SELECT organisations where TYPES = [not_for_profit_organisation] AND TOPICS = [ecology]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:not_for_profit_organisation) (?resource swed:has_topic swed_toi:ecology) USING swed FOR swed_toi FOR swed_ot FOR</p>
6	<p>FITNESS: 0.4</p> <p>DESCRIPTION: SELECT organisations where TYPES = [trade_related_umbrella_organisation] AND TOPICS = [farming]</p> <p>RDQL: SELECT ?resource WHERE (?resource swed:has_organisation_type swed_ot:trade_related_umbrella_organisation) (?resource swed:has_topic swed_toi:farming) USING swed FOR swed_toi FOR swed_ot FOR</p>

[close](#)

Figure 4.6: Selected queries sorted by fitness

Chapter 5

Experiments and Results

The methodology adopted to test the viability of the presented approach forms the initial part of this chapter. Followed by the results obtained and their discussion.

5.1 Experimental Setup

This section outlines the data used for experimental purposes and draws a distinction between manual and automated testing of a web based application.

5.1.1 Test Data

The data used to carry out the experiments was related to organisations and included information such as the name, acronym, purpose, location, contact information, type, topics etc. for a particular organisation. The data further included type and topic ontologies for organisations. The names and descriptions of the data files used are written below and included with further detail and sample data as an appendix.

- **wwite_tidy.n3** The file with semantically enriched data of 100 organisations.
- **swed_org_type_skos.n3** Type ontology for organisations.
- **wwite_index_skos.n3** Topic ontology for the organisations.

The ontology files, namely *swed_org_type_skos.n3* and *wwite_index_skos.n3* form two different facets of the data and were heavily used to construct new queries and to modify the existing ones. The first file, *wwite_tidy.n3*, on the other hand contained semantic data of organisations enriched using the ontology files. All the data was in N3 format and was provided by HP Labs (Bristol).

5.1.2 Plan

Two query mutation/expansion operators were developed namely *RandomMutationOperator* and *ConstrainedMutationOperator* described in Chapter 3. The experiments were primarily focused on finding the difference between the two mutation operators. Specifically, we wanted to see

- Which operator leads to a better performance when the AIS is supplied with different input data sets and standard input parameters. Where performance is measured by the *precision*, *recall* and *convergence*.
- How does the *precision*, *recall* and *convergence* changes with the changing mutation rate.

Where *precision* signifies the accuracy of the results in every iteration and is evaluated as

$$precision = \frac{relevant\ retrieved\ results}{retrieved\ results}$$

recall is a measure of the relevant results obtained so far

$$recall = \frac{relevant\ retrieved\ results}{relevant\ results}$$

and *convergence* is the number of iterations required to find maximum results or reach maximum recall. It is important to note that with the increasing number of iterations the value of precision tends to wards 0, since it becomes increasingly difficult for the AIS to find a very small number of organisations from a large data set. Similarly the value of recall tends to wards 1, since more relevant results are found after every iteration of the AIS. Finally, we needed to determine how many times we would have to run the experiments to be able to draw credible conclusions on the performance of the AIS. We divided the experiments in two phases the first phase involving 180 runs and the second involving 300 runs, break up is given below

$$first\ phase \rightarrow 2(operators) \times 3(data\ sets) \times 30(runs) = 180$$

$$second\ phase \rightarrow 2(operators) \times 5(mutation\ rate\ values) \times 30(runs) = 300$$

It is important to note two main differences between these phases. Firstly, three different input data sets are used in the *first phase* compared to only one in the *second phase*. The input data set used in the *second phase* is the one at which both mutation operators performed well in the *first phase*. Secondly, in the *first phase* the mutation rate remains the same whereas in the *second phase* we test the AIS by changing the mutation rate from 0 to 1 (5 values in all). It is to be noted that input data sets are subsets of the data file *wwite_tidy.n3*, this is explained further in the following section.

5.1.3 Automated Test Script

Since the project involved development of web-based interface we had the option of either testing the application manually or by using some automated script.

By manual testing we mean using the actual web interface and noting down observations on the paper. We realised it quite early that performing the experiments using the developed web based application would be too time consuming considering the time frame and the number of planned runs. We therefore implemented an automated way of testing the AIS. The idea was to simulate a user interacting with the system rather than actually using the system. The simulation was based on the following core ideas

- A set of organisations be identified initially and regarded as relevant (organisation to be found by a user).
- Any organisation that is not relevant may be regarded as irrelevant.
- The simulation code should pick a random organisation from the input set and consider it as the starting point of the AIS.
- After every iteration of the AIS relevant and irrelevant items be selected equal to MAX-NONSELF and MAX-SELF, which are the input parameters to the algorithm.
- This selection of items should be considered as the user feedback.
- The simulation stops when all the organisations in the input set have been found or a maximum iteration limit of MAX_ITERATIONS is reached.

Pseudo-code for the User Simulation

```
-----  
- initialise the AIS  
- randomly pick an organisation from the input set of organisations  
  and add it to the non-self of AIS  
- remove the organisation from the input set  
- set count = 0 and allResultsFound = true  
- while( (MAX_ITERATIONS > count) && (allResultsFound == false) ) {  
- fetch antibodies/queries from the AIS  
- isolate all the results  
- generate a random number between 0 and MAX_NONSELF and  
  call it relevantCount  
- pick relevantCount number of results from all results  
  and add them to AIS as non-self  
- remove the recently added non-self from the input set  
- generate a random number between 0 and MAX_SELF call it  
  irrelevantCount  
- pick irrelevantCount number of results from all results  
  and add them to AIS as self  
- note down observations such as precision, recall, iterations  
  count, number of self and non-self selected, total number  
  of results produced etc.
```

Input Parameters	Values
SELECTION_COUNT	10
CLONAL_EXP_CONST	8
WHEEL_SIZE	100
MAX_SELF	5
MAX_NONSELF	5
MAX_QUERY_ATTEMPTS	250
MAX_ITERATIONS	40
TYPE_CHG_PROB	0.5

Table 5.1: Input parameters common for both operators

```
- if(input set = empty)
allResultsFound = true
-}
```

Selection of the input data sets

The input data sets were selected purely on the basis of human judgement. From the limited data set of 100 organisations three different input sets were extracted. Each data set contained organisations that were related in a broad sense. Types and topics of the organisations were not considered in the selection of the input data sets. The fact that the sets were chosen based on a user’s subjective judgement makes the test set more realistic. The selected input data sets for the experiments are listed in the appendix C (section C.4).

5.2 Results

Now we present the the obtained results. The results are divided in two phases as discussed above.

5.2.1 First Phase

In the first phase we noted the *precision*, *recall* and *convergence* using two mutation operators working on three different input data sets. The input parameters for the AIS (common for both operators) are shown in Table 5.2.1 whereas the parameters specific to *RandomMutationOperator* and *ConstrainedMutationOperator* are shown Tables 5.2.1 and 5.2.1 respectively. The explanation of the mutation operators is in Chapter 3 (section 3.3.6)

Input Parameters	Values
MAX_TOPICS_IN_QUERY	8
MUTATION_OPERATOR	RandomMutationOperator
MUTATION_RATE	0.7

Table 5.2: Input parameters for *RandomMutationOperator*

Input Parameters	Values
MUTATION_OPERATOR	ConstrainedMutationOperator
MUTATION_RATE	0.5

Table 5.3: Input parameters for *ConstrainedMutationOperator*

Precision

Objective To compare the *precision* of two different mutation operators and to note how it changes with different input data sets.

Discussion The comparison of *precision* for *RandomMutationOperator* and *ConstrainedMutationOperator* on different input data sets is shown in Figures 5.1, 5.2, 5.3. On average *ConstrainedMutationOperator* yields better precision in first 10 iterations but the precision drops after 20 iterations. This drop in precision is due to the fact that the higher precision value in the earlier iterations enables collection of greater number of non-self and the fewer number of unfounded non-self. However, for both the operators even the highest value of precision remains under 0.2. .

Student's t-test which is unpaired and yields two-tailed *p-values*, was performed to see the significance of the results. The analysis was performed using an on line tool [36]. The obtained values are shown in the table below. Two of the *p-values* are well below 0.05 but one is quite high (set2). Hence, the considering the average, we may reject the *null hypothesis* with reasonable confidence. Statistically, the difference between the two operators in terms of precision is slightly significant.

<i>precision</i>	set1 (iteration 4)	set2 (iteration 12)	set3 (iteration 3)	<i>average</i>
t-value	3.8636	1.6969	8.3238	4.59776
p-value	0.0003	0.0975	0.0001	0.03263

Outcome The approach being exploratory in nature fetches a large number of irrelevant results along with the relevant, in every iterations which brings the value of precision very down. We found slightly significant difference between the operators in terms of precision. If precision is critical this approach is may not be the best one.

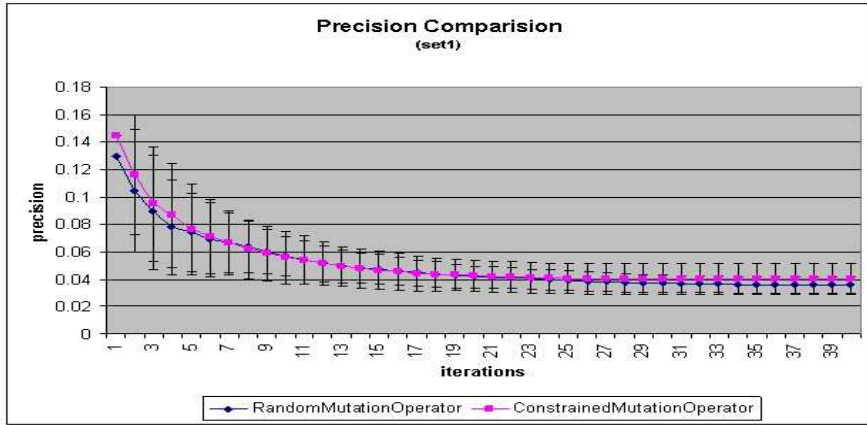


Figure 5.1: Precision for input set1

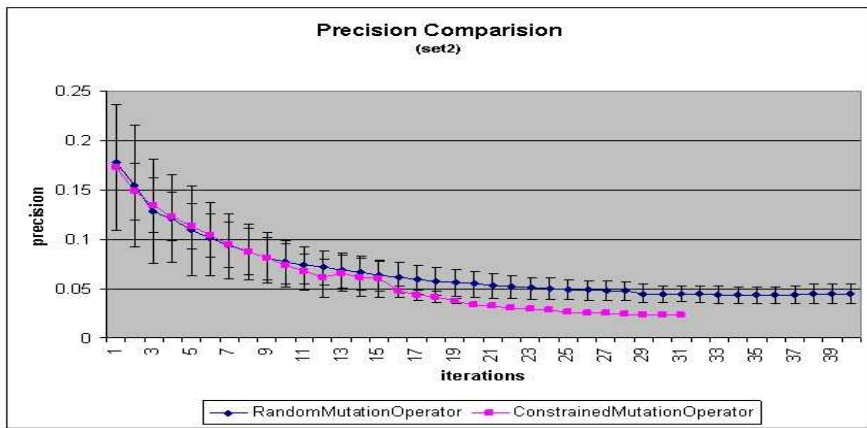


Figure 5.2: Precision for input set2

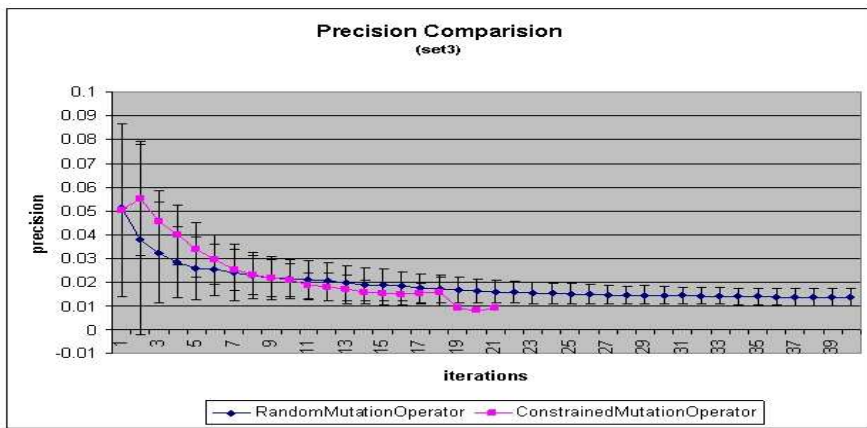


Figure 5.3: Precision for input set3

Recall

Objective To compare the *recall* of two different mutation operators and to note how it changes with different input data sets.

Discussion Figures 5.4, 5.5 and 5.6 show the recall comparison for input set1, set2 and set3 respectively. There are two very important observations in these results. First, the value of recall achieved and secondly, the number of iterations required to reach the maximum recall. It is evident the the *ConstrainedMutationOperator* achieves more recall than its counterpart and that too fairly quickly. The significance of the *ConstrainedMutationOperator* lies in the fact that it reaches approximately a recall of 0.9 within the first 10 iteration making it very suitable for a web application.

Student's t-test was performed to see the significance of the results using an on line tool [36]. The obtained values are shown in the table below. All of the *p-values* fall below 0.01, hence we can reject the *null hypothesis* with strong confidence. Statistically, the difference between the two operators in terms of recall is extremely significant.

<i>recall</i>	set1 (iteration 9)	set2 (iteration 11)	set3 (iteration 6)
t-value	-11.3	-14.1	8.6320
p-value	¡ 0.0001	¡ 0.0001	¡ 0.0001

Outcome We found that although both operators yield high recall values (between 0.8 and 0.9) but *ConstrainedMutationOperator* performs better since it not only achieves a higher value of recall but in a quick fashion. Also we found the difference between the operators to be statistically very significant.

Convergence

Objective To compare the number of iterations required to collect majority of relevant results and note the effect of different input data sets.

Discussion The *convergence*, on different input data sets, for the *RandomMutationOperator* is shown in Figure 5.7 and for the *ConstrainedMutationOperator* is shown in Figure 5.8. One very obvious observation is the behaviour of *ConstrainedMutationOperator* on the input set2 and set3 where the search process terminates very early compared to all the other instances. This indicates that *ConstrainedMutationOperator* takes less number of iterations to find maximum results then the *RandomMutationOperator*. On the whole *ConstrainedMutationOperator* produces large number of both relevant and irrelevant results compared to *RandomMutationOperator*. Another important difference between the two is, in case of *RandomMutationOperator* there is always a gradual rise in the number of non-self whereas in case of *ConstrainedMutationOperator* we can see a fall as well. This fall in non-self is due to the fact that the graph is an average of 30 runs of 40 iterations each. Most of the time maximum non-self was

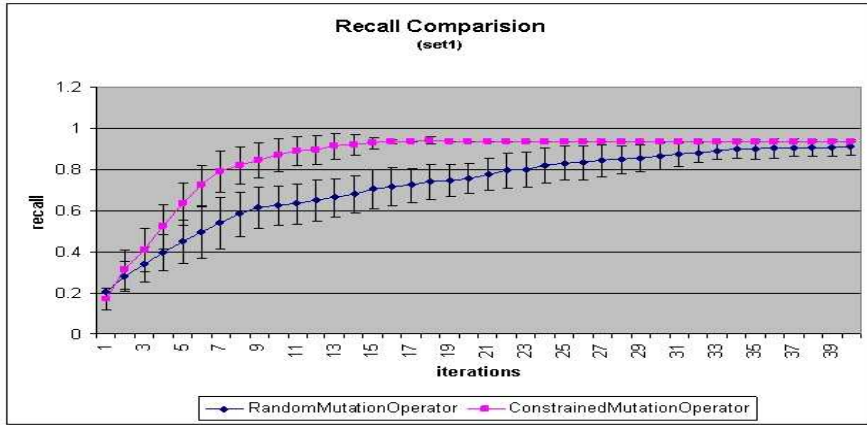


Figure 5.4: Recall for input set1

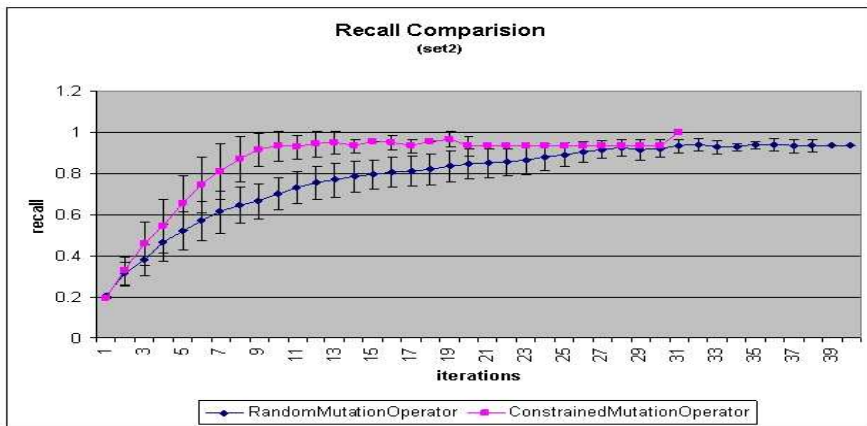


Figure 5.5: Recall for input set2

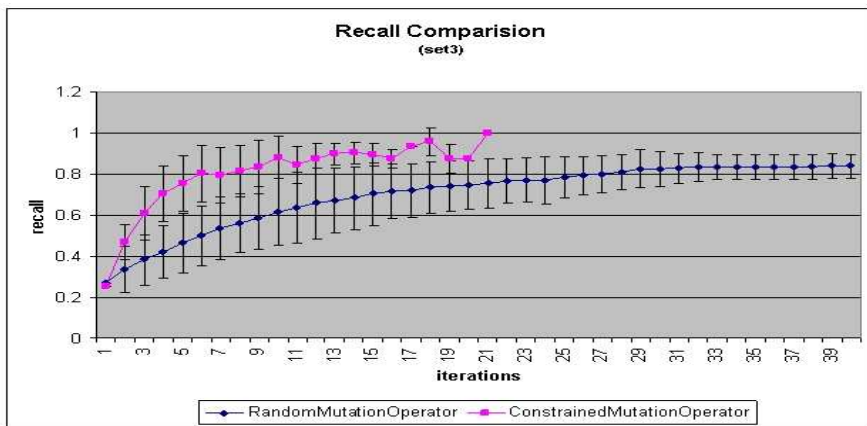


Figure 5.6: Recall for input set3

collected at the peak but there were some exceptions which took more number of iterations and were unable to collect high number of non-self. Hence we see a down fall in the number of non-self.

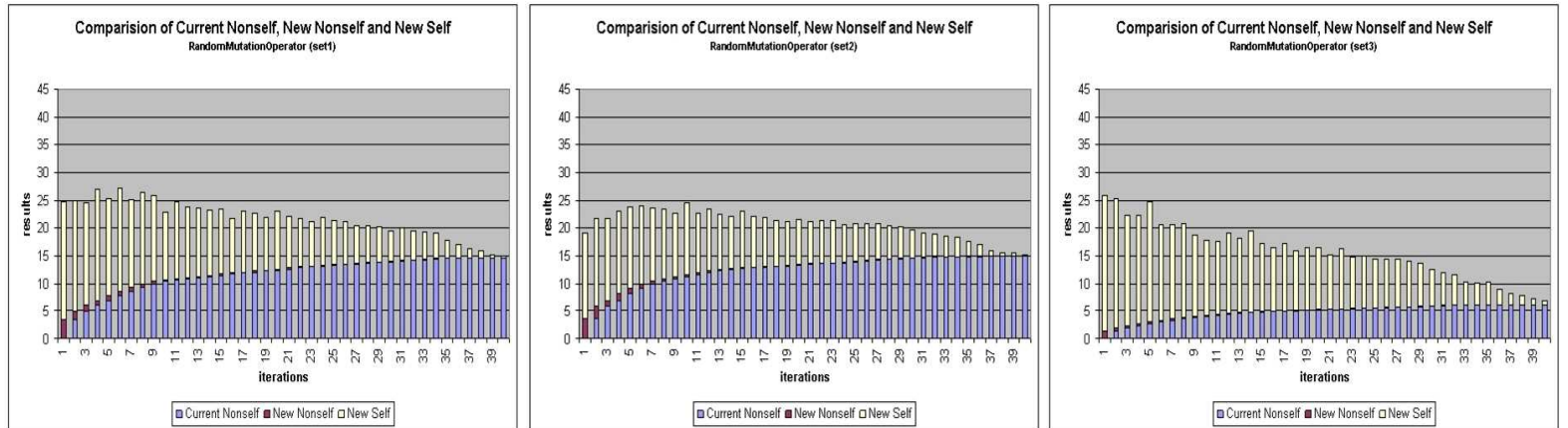


Figure 5.7: Convergence for RandomMutationOperator

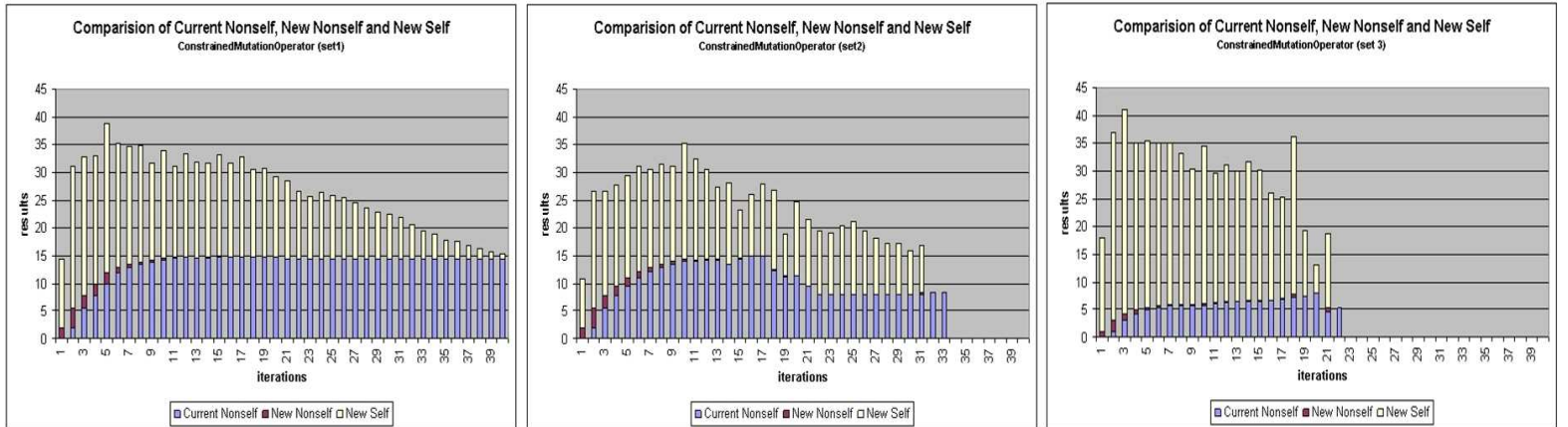


Figure 5.8: Convergence for ConstrainedMutationOperator

Outcome We found that *RandomMutationOperator* exhibits delayed convergence and in every iterations finds lesser number of both relevant and irrelevant results compared to the *ConstrainedMutationOperator*. It would not be wrong to say that the *ConstrainedMutationOperator* is much more aggressive, since it fetches greater number of both self and non-self in every iterations and *converges* rapidly.

5.2.2 Second Phase

The second phase involved selection of one input data and note the effect of changing the mutation rate. There was no definite rule of choosing the input data set for this phase. We opted for input data set2 since both the operators perform reasonably well on it.

Precision Comparison

Objective To observe how precision changes with changing mutation rate on both the operators.

Discussion Figure 5.9 and 5.10 respectively show the precisions for *RandomMutationOperator* and *ConstrainedMutationOperator*. As expected there was nothing significant about the precision of both the operators and it remained under 0.2. There was however an exceptional case for the mutation rate of 0 and *ConstrainedMutationOperator*, the precision remained relatively higher than all the other mutation rates but only at the expense of a low recall.

Outcome Precision remained quite low for both the operators.

Recall Comparison

Objective To observe, for both operators, how recall changes with the mutation rate and also change in recall between number of iterations.

Discussion The graphs related to these experiments are shown in Figures 5.11 and 5.12. Since every iterations means a separate user feedback to the AIS it is very crucial that the recall should increase as early as possible. In terms of the *RandomMutationOperator* the mutation rate of 0 and 0.25 give a better early recall but an overall lower recall value compared to the mutation rate of 0.5 and 0.75. The *ConstrainedMutationOperator* performs much better since it achieves a maximum recall value of 1 for all mutation rates except 0 where it remain rather low. Although a recall of 1 is achieved 4 times but the mutation rate of 1 results in the fewest number of iterations. Another very interesting observations is the change in recall between iteration 5 and iteration 10 in both the operators. Although, in both the operators recall changes significantly between iteration 5 and 10 but the *RandomMutationOperator* is outperformed by the *ConstrainedMutationOperator* which converges much faster.

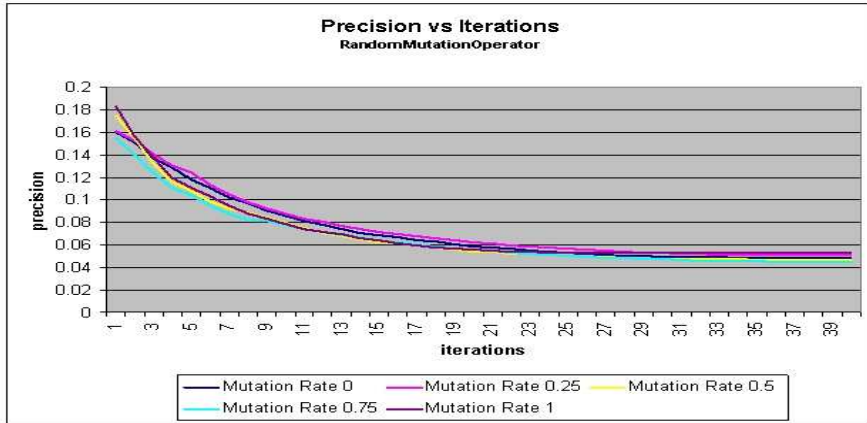


Figure 5.9: Precision vs Mutation Rate for RandomMutationOperator

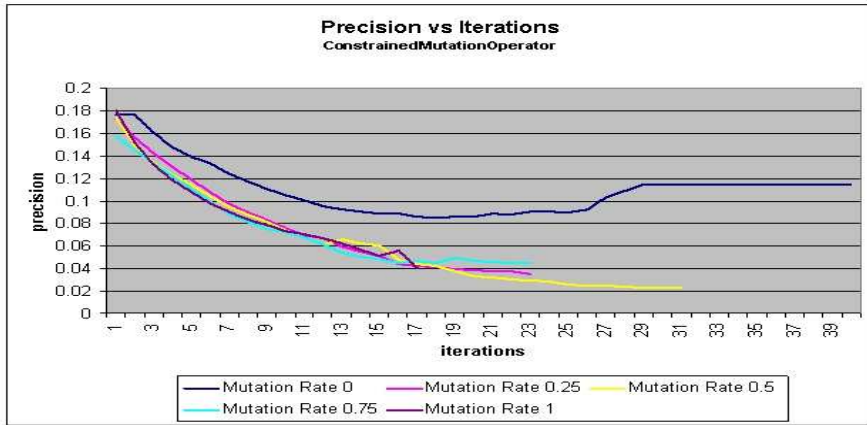


Figure 5.10: Precision vs Mutation Rate for ConstrainedMutationOperator

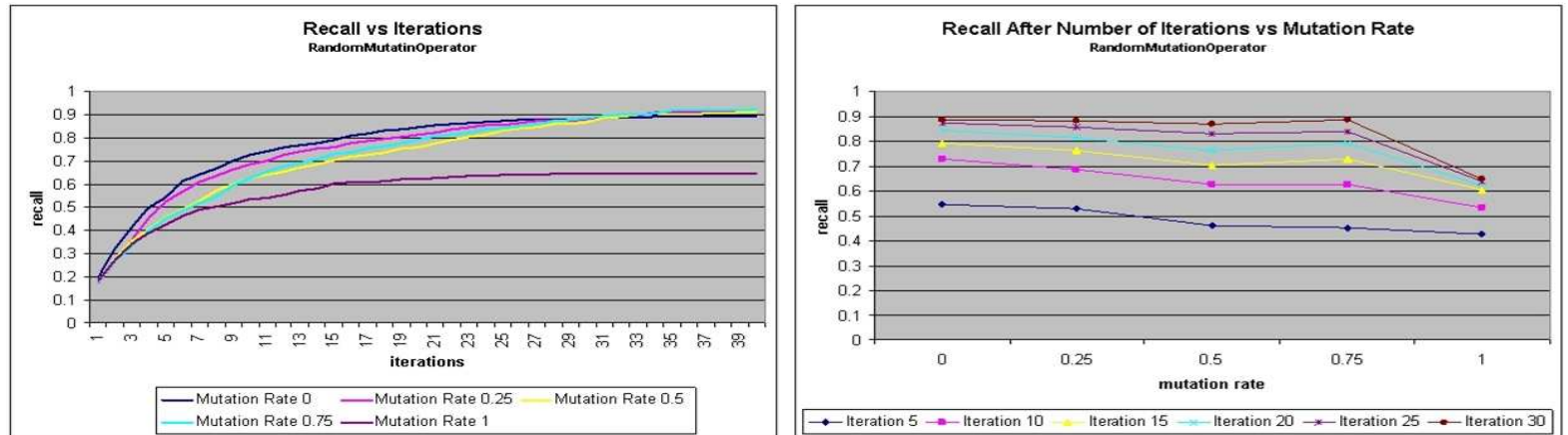


Figure 5.11: Recall vs Mutation Rate for RandomMutationOperator

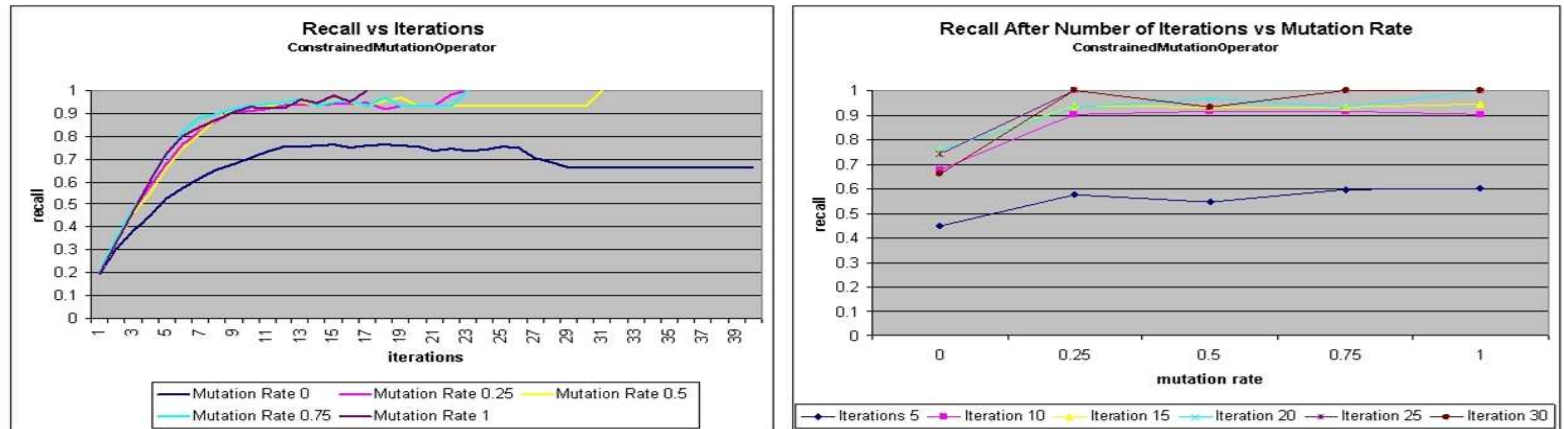


Figure 5.12: Recall vs Mutation Rate for ConstrainedMutationOperator

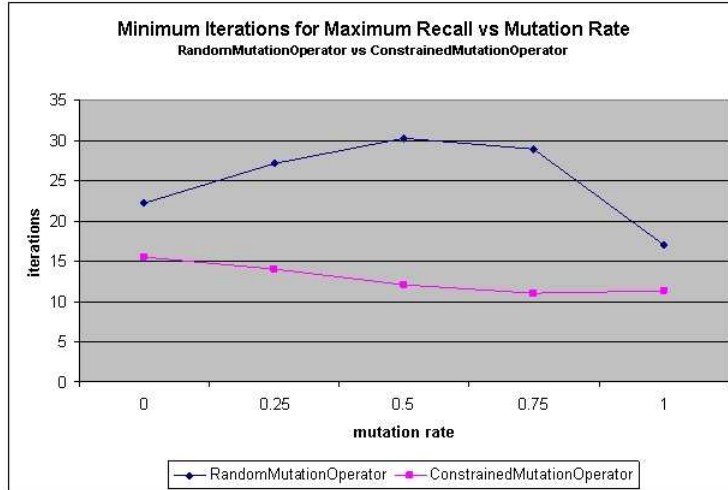


Figure 5.13: Minimum Iterations for Maximum Recall vs Mutation Rate

Outcome We found that the *ConstrainedMutationOperator* not only results in higher recall but very quickly. We also found that the maximum increase in recall is between iteration 5 and iteration 10. In case of the *RandomMutationOperator* increasing the mutation rate from 0 to 0.75 results in an overall higher recall which falls down between 0.6 and 0.7 between the mutation rate of 0.75 and 1. On the other hand the *ConstrainedMutationOperator* performs poorly for the mutation rate of 0, improves drastically at the mutation rate of 0.25, goes down once again at the rate of 0.5 and then performs at its optimal at mutation rate of 0.75 and 1.

Convergence Comparison

Objective To observe the number of iterations required to reach maximum recall with the changing mutation rate for both the operators.

Discussion The two operators behave in completely opposite ways with the changing mutation rates as shown in Figure 5.13. The *ConstrainedMutationOperator* clearly outperforms the *RandomMutationOperator* by taking lesser number of iterations to reach maximum recall on all the different mutation rates. The *ConstrainedMutationOperator* is optimal at the mutation rate of 0.75 and 1 whereas the *RandomMutationOperator* works well at the two extremes i.e 0 and 1.

Outcome We found that *ConstrainedMutationOperator* reaches a higher recall for all the different mutation rates in lesser number of iterations compared

to the *RandomMutationOperator*.

5.3 Summing it all up

Two query mutation operators were implemented and tested. We found slightly significant difference between the two in terms of precision, however the difference between the two in terms of recall was very significant. Recall of the *RandomMutationOperator* of different input set on average remained same, between 0.8 and 0.9. For *ConstrainedMutationOperator* the recall value was a higher, between 0.9 and 1. In terms of precision both the operators performed poorly on all the input sets. The worst performance, in terms of precision, for both operators was on input set3 which had only 8 organisations to be found. This low precision is because of the fact that the AIS retrieved similar number of irrelevant results as in other cases but lower number of relevant results. In terms of convergence we found *RandomMutationOperator* to be slow and steady and *ConstrainedMutationOperator* to be fast and aggressive making it suitable for use in a web based utility. By changing the mutation rate and observing the recall and precision values we found the two operators to behave in opposite manner. Recall for *RandomMutationOperator* decreased steadily with the increasing mutation rate whereas the recall for *ConstrainedMutationOperator* increased sharply with the increasing mutation rate. We also measured the increase of recall within iterations and found that for both operators recall increased sharply up to ten iterations, however the increase in recall for *ConstrainedMutationOperator* was more rapid. Maximum recall for *ConstrainedMutationOperator* at the 10th iterations was 0.9 and that of the *RandomMutationOperator* was 0.7.

Chapter 6

Future Work

Restricted by the time frame, we feel that there is a lot that can be implemented and tested. This chapter first critically analyses the project followed by recommendations for the future direction of the project. Finally mentioned are some of the lessons we learnt from this project. All in all this chapter is aimed at helping anyone who wishes to extend this work.

6.1 Critique

- The presented approach for information retrieval is tested only on a small data set of 100 organisations. Its sufficient to prove the presented idea but will need further rigorous testing on larger data sets.
- Our AIS currently handles only two ontologies namely *type* and *topic*. It is not designed to tackle semantic data with more than two facets.
- Behaviour of the AIS by changing various other input parameters is unknown. These parameters include the SELECTION_COUNT, TYP_CHG_PROB, CLONAL_EXP_CONST and MAX_TOPIC_IN_QUERY.
- The user interface needs to be modernised for ease of navigation.

6.2 Future Directions

Some of the areas that need to be worked are

6.2.1 Testing

One important issue in the evaluation of a semantic web related application is the availability of valid semantic data. The data can be quite overwhelming for a new starter but it is important to note that it follows certain guidelines. This semantically enriched data can be generated for rigorous testing. However, the

generated data might not be able to give a better qualitative feel of the search process. In short, generation of data is an option that can be further explored to test the information retrieval process on huge data sets.

6.2.2 User Testing

Currently we tested the AIS by simulating user interaction. It is planned to integrate the search utility with the semantic portal being developed at HP Labs (Bristol) and to see how it behaves with actual user testing.

6.2.3 Making it Suitable for Multiple Facets

The current code base was designed to work for only two facets (ontologies). It can be extended to incorporate n number of facets. It is recommended that any attempt to modify the code to be able to make it suitable for multiple facets should focus on the following.

Structure of the Query

Currently the structure of the query is contained in the java source file named *SemanticQuery.java*. Substantial changes would be required in this file to make it suitable to represent a query with n number of facets.

Mutation Operators

When we think of a multi-faceted data sets i.e. a data set that uses multiple ontologies, the current mutation operators automatically become invalid. The current query mutation operators namely *RandomMutationOperator* and *ConstrainedMutationOperator* need to be enhanced to be able to digest more ontologies. We feel that work in this area would produce a rich user experience.

6.2.4 New Mutation Operators

The *ConstrainedMutationOperator* can be extended to produce *SpecialisingMutationOperator* and *GeneralisingMutationOperator*. Both the operators use *inferencing* capabilities of the Jena API. Considering the ontology tree, these operators work simply by incorporating terms fetched by either moving up the ontology tree (general terms) or down (specific terms).

GeneralisingMutationOperator

As the name suggests the main task of this operator would be to produce general results starting from specific results. Jena's inferencing API can be used to retrieve general terms for various facets and those general terms could be combined to form new breed of queries that generate results that are more general.

SpecialisingMutationOperator

This operator is exactly the opposite to the *GeneralisingMutationOperator*. Inference is used to obtain specific terms from general terms. These specific terms can be combined to form queries that produce specific result.

Crossover

A crossover operator is also expected to produce interesting results. Type and topic information from two queries can be swapped randomly to achieve crossover.

Chapter 7

Conclusion

Sound theoretical foundations were established in diverse areas such as HIS, semantic web and query expansion. From the area of HIS, we isolated concepts such as antibody-antigen interaction, clonal expansion, self/non-self discrimination and affinity maturation. In terms of semantic web, we showed by practical examples, how meta-data is represented in RDF. An overview of ontologies, Jena semantic web framework and RDQL was also presented. Different query expansion techniques were identified and the suitable one selected for our work.

Based on the knowledge gained from the above mentioned diverse areas, an AIS information retrieval utility for the semantic web was designed and implemented with a web interface. We also designed and implemented two query expansion/mutation operators, *RandomMutationOperator* and *ConstrainedMutationOperator*, which sit at the core of the AIS algorithm. The developed utility was then tested with the mutation operators by simulating user interaction. Strategy for this automated test was also outlined.

We found the performance of the AIS, as a whole, to be promising, yielding high level of *recall* and *convergence*. The *precision*, however remained understandably low. It was primarily because of the exploratory nature of the approach.

The individual performances of both the operators were also compared. A two-phase test strategy was planned and executed. In the *first-phase* we compared both the operators using different input data sets. In the *second-phase* we tested them by changing the mutation rate.

The *ConstrainedMutationOperator* was found to outperform the *RandomMutationOperator*, in terms of *recall* and *convergence*. The *precision* for both the operators, however, remained relatively low and they performed quite similar.

We have found the presented approach to be viable and promising with potential for improvement. A critical analysis of the work was also presented along with the identification of further areas of research.

Chapter 8

Online Resources

Following are some of the online resources associated with this work.

8.1 Poster

<http://studentweb.cs.bham.ac.uk/~msc43kar/poster.jpg>

8.2 Source Code

The complete *back-end* code is accessible at
</home/students/msc/msc43kar/workspace/Project>

The Java Server Pages (JSP) used for the display, *front-end*, are accessible at
/home/students/msc/msc43kar/public_html/jsp

8.3 Test Data and Configuration Files

The following link points to all the data used for testing purposes as well as the the input parameters configuration file.
</home/students/msc/msc43kar/data+conf>

8.4 Semantic Search Utility

Live, immune based search utility configured to use the ConstrainedMutation-Operator. Accessible at the following location
<http://studentweb.cs.bham.ac.uk/~msc43kar/jsp/index.jsp>

8.5 Visualisation of the HIS

This is not directly related to the presented work but is an interesting visualisation of the HIS developed by the author. It visualises aspects of immune system such as, antibody-antigen interaction, immune memory. The applet is accessible at the following location and requires the presence of Java Runtime Environment (JRE) in the browser.

<http://studentweb.cs.bham.ac.uk/~msc43kar/ImmuneApplet.html>

8.6 Graphs

All the graphs in the report are published on the web and can be seen at the following two locations

<http://studentweb.cs.bham.ac.uk/~msc43kar/results.html>

<http://studentweb.cs.bham.ac.uk/~msc43kar/results2.html>

Bibliography

- [1] AIS as a recommender. http://www-uk.hpl.hp.com/people/steve_cayzer/downloads/Presentation020514hawaii.ppt.
- [2] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 2001.
- [3] Bodo Billerbeck, Falk Scholer, Hugh E. Williams, and Justin Zobel. Query expansion using associated queries. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 2–9, New York, 2003. ACM Press.
- [4] D. Dasgupta, Z. Ji, and F. Gonzalez. Artificial immune system (ais) research in the last five years. In *Proceedings of the international Conference on Evolutionary Computation (CEC)*, volume 1, Canbara, Australia, 2003.
- [5] Leandro N. de Castro and Jonathan Timmis. Artificial Immune Systems: A Novel Approach to Pattern Recognition. In *Artificial Neural Networks in Pattern Recognition*, pages 67–84. University of Paisley, January 2002.
- [6] Leandro N. de Castro and Fernando J. Von Zuben. An evolutionary immune network for data clustering. In *Proceedings of the IEEE Brazilian Symposium on Artificial Neural Networks (SBRN)*, pages 84–89, New York, 2000. IEEE.
- [7] Leandro N. de Castro and Fernando J. Von Zuben. Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, 6(3):239–251, June 2002.
- [8] Leandro Nunes de Castro. An introduction to the artificial immune systems. tutorial presented at ICANNGA 2001.
- [9] L.N. de Castro. An introduction to the artificial immune systems. <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/lnunes/tutorial01.pdf>. Tutorial at ICANNGA 2001.
- [10] Efthimis N. Efthimiadis. *Annual Review of Information Systems and Technology (ARIST)*, volume 31, chapter Query Expansion, pages 121–187. Information Today Inc, Medford, NJ, 1996.

- [11] Gene Ontology Consortium. <http://www.geneontology.org/>.
- [12] Tom R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.
- [13] Steven A. Hofmeyr. An interpretative introduction to the immune system. In *Design Principles for Immune Systems and Other Distributed Autonomous Systems*, pages 3–28. Oxford University Press, New York, 2001.
- [14] HP Labs, Bristol. <http://www.hpl.hp.com/>. accessed on February 26, 2004.
- [15] Immune systems - an evolutionary metaphor. http://www-uk.hpl.hp.com/people/steve_cayzer/downloads/030213_ais.ppt.
- [16] Jena. <http://jena.sourceforge.net/>.
- [17] Doheon Lee, Jungja Kim, Mina Jeong, Yonggwan Won, Seon Hee Park, and Kwang-Hyung Lee. Immune-based framework for exploratory bio-information retrieval from the semantic web. In *Artificial Immune Systems: Second International Conference, ICARIS 2003, Edinburgh, UK, September 1-3, 2003, Proceedings*, volume 2787/2003 of *Lecture Notes In Computer Science*, pages 128–135, Heidelberg, 2003. Springer.
- [18] T Morrison and U Aickelin. An artificial immune system as a recommender for web sites. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, volume 1, pages 161–169, University of Kent at Canterbury, September 2002. University of Kent at Canterbury Printing Unit.
- [19] Notation 3. <http://www.w3.org/DesignIssues/Notation3.html>. accessed on April 26, 2004.
- [20] What is the immune system ? <http://www.muschealth.com/infectious/immune.htm>. accessed on February 26, 2004.
- [21] J. Piel. Life, death and the immune system. *Scientific American*, 269(3):20–102, 1993.
- [22] Rdf: Concepts and abstract syntax. <http://www.w3.org/TR/rdf-concepts/>. accessed on April 26, 2004.
- [23] Rdf primer. <http://www.w3.org/TR/rdf-primer/>. accessed on April 26, 2004.
- [24] Rdbl - a query language for rdf. <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>. W3C Member Submission 9 January 2004.
- [25] J. J. Rocchio. *The Smart Retrieval System - Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval, pages 313–323. Prentice Hall, Englewood, Cliffs, NJ, 1971.

- [26] L. Schindler, D. Kerrigan, and J. Kelly. Understanding the immune system. <http://press2.nci.nih.gov/sciencebehind/immune/immune01.htm>. accessed on February 26, 2004.
- [27] An introduction to the semantic web. http://www.hpl.hp.com/personal/steve_cayzer/downloads/040312_semweb_public_version.pdf. accessed on April 26, 2004.
- [28] Lauren Sompayrac. *How the Immune System Works*. Blackwell Science, 1999.
- [29] Steve Cayzer. http://www-uk.hpl.hp.com/people/steve_cayzer/index.htm.
- [30] Dave Beckett's resource description framework (rdf) resource guide. <http://www.ilrt.bris.ac.uk/discovery/rdf/resources/>.
- [31] The semantic web: An introduction. <http://infomesh.net/2001/swintro/>.
- [32] The semantic web (for web developers). <http://logicerror.com/semanticWeb-webdev>.
- [33] The semantic web in breadth. <http://logicerror.com/semanticWeb-long>.
- [34] Primer: Getting into rdf & semantic web using n3. <http://www.w3.org/2000/10/swap/Primer>.
- [35] The semantic web, taking form. <http://infomesh.net/2001/06/swform/>.
- [36] Student's t-tests. <http://www.physics.csbsju.edu/stats/t-test.html>. accessed on Sept 7, 2004.
- [37] Jamie Twycross and Steve Cayzer. An immune-based approach to document classification. In *Intelligent Information Processing and Web Mining, Proceedings of the International IIS:IIPWM03 Conference*, Advances in Soft Computing, Zakopane, Poland, 2003. Springer.
- [38] Patricia A. Vargas, Leandro N. de Castro, and Fernando J. Von Zuben. Artificial immune systems as complex adaptive systems. In *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, volume 1, pages 115–123, University of Kent at Canterbury, 2002. University of Kent at Canterbury Printing Unit.
- [39] World Wide Web Consortium. <http://www.w3.org/>.

Appendix A

Summer Project Declaration

A printed version of the summer project declaration is on the following page.

The University of Birmingham
School of Computer Science

MSc in Advanced Computer Science

Summer project

This form is to be used to declare your choice of summer project in course. Please complete this form, obtain the signature of your supervisor and post it in the appropriate assessed work pigeon hole.

Deadline: 16.00 hrs, 17th June 2004

Name: *RATNA KASHIF ALI*

Student number: *0578898*

Project title: *AIS AND SEMANTIC QUERY.*

Project supervisor: *PROF. XIN YAO*


The following questions should be answered in conjunction with a reading of the course handbook.

Aim of project	<i>TO EXTEND THE WORKING IN THE MINI-PROJECT AND TO CREATE A WEB-BASED APPLICATION THAT DEMONSTRATES AIS BASED QUERY EXPANSION</i>
----------------	--

Objectives to be achieved	<i>→ TO DEVELOP A WEB BASED APPLICATION IN JAVA THAT DEMONSTRATES QUERY EXPANSION ON THE IMMUNE SYSTEM PRINCIPLES. → TO PERFORM TESTING OF THE DEVELOPED APPLICATION. → TO PRESENT THE RESULTS IN THE FORM OF A REPORT</i>
---------------------------	--

Checked by: _____ /PTO
Date: _____

<p>Project management skills</p>	<p>→ MEETINGS WOULD BE HELD REGULARLY WITH THE SUPERVISOR</p>
<p>Briefly explain how you will devise a management plan to allow your supervisor to evaluate your progress</p>	<p>→ PROGRESS WOULD BE DOCUMENTED ON MY WEBSITE. → PROGRESS REPORT (WEEKLY OR BIWEEKLY) WOULD BE SUBMITTED TO THE SUPERVISOR.</p>

Signed (student) *Rana Koshif Ali* 

Date: *12th JUNE 2004*

Signed (supervisor): *Nis Goo*

Date: *12/6/2004*

<p>What is the project?</p>	<p><i>The project is to develop a web-based system for the University of Birmingham...</i></p>
<p>What are the objectives of the project?</p>	<p>→ To develop a web-based system for the University of Birmingham. → To ensure the system is user-friendly and easy to use. → To ensure the system is secure and reliable.</p>

Office use:	
Copy for	student / supervisor / file

Appendix B

Statement of Information Search Strategy

The inherent nature of the project required information searching in disparate areas, some relatively new, for example, semantic web and Artificial Immune Systems (AIS), while others a bit old, that is, Human Immune System (HIS). Since, the project attempts to combine different areas of research it was required, first, to establish basic concepts of the constituents of the project. A breadth-first approach was adopted initially to gain basic knowledge of all areas and then every area was explored in depth. Initially Science Citation Index (SCI) was used to find articles using keywords such as,

- semantic web AND query expansion
- semantic web AND AIS
- Artificial Immune System AND semantic web

None of the above returned any results on SCI. It was then decided to search Hap's website, since the project is basically an idea proposed by HP [14]. Various relevant resources were found at Dr. Cayzer's homepage [29]. He was then contacted and asked for supervision of this project, he agreed and guided alot in the information search process. The project is based on and extends the idea proposed by Lee et al in [17]. Dr. Cayzer was kind enough to mail me the article which was neither available in the library nor online, This article formed the starting point for the information search. Varying strategies were adopted for different areas of the project which are discussed below

B.1 Semantic Web

Before delving in to the details of the project, it was necessary to establish a basic foundation of the idea of semantic web. Initially two resources found on Dr.

Steve Cayzer's homepage were very helpful [31] and [30]. Further exploration of [30] revealed alot about semantic web and its core technologies for example RDF, N3, ontology, Jena etc. The sites I found useful for understanding semantic web are [39, 23, 22, 19, 12], Dr. Steve Cayzer emailed me his presentation that he delivered at Nottingham University which was also very helpful [27]. Beside all the above mentioned stuff the most popular and pioneering article on semantic web by Berners Lee that appeared in *Scientific American* was extremely helpful [2]. Later as I started exploring Jena and ontology following links were of immense value [16, 12, 35, 32, 33, 34].

B.2 HIS & AIS

Search for AIS was first performed on the SCI and some relevant articles were read. Most, in fact, all the papers that discuss AIS discuss HIS as well. Bibliographies of papers related to AIS revealed a good number of quality HIS resources, for example, [13]. Other resources of HIS were retrieved through Goggle using keywords like, *immunology*, *immune system*, *human immune system*, some of the useful resources found are [28, 21, 26, 20]

B.3 Query Expansion

In the area of query expansion, basic literature was required initially, searching on SCI returned results that dealt with advanced topics in this area. However after consulting the bibliography of some selected papers an excellent summary of research on query expansion was found [10]. Beside this only [3] was consulted, which proved sufficient for this work.

Appendix C

Data

Here we have added some of the data used in the experiments. Due to the huge size of the data, only small chunks of the files have been added. It is meant to give the reader a feel of how semantic data looks like in N3 format. The appendix is further broken down into four sections. The first section is an extract of data related to five organisations, followed by 5 items each from type and topic ontologies. Finally, there is a section on the input data sets, used to perform experiments, where only the names of the organisations are included. It is important to note that the input data sets are taken from the file *wwrite_tidy.n3*.

C.1 wwrite_tidy.n3

Semantic data for five organisations.

```
## N3 translation of wwrite tab-delimited data.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix vcard: <http://www.w3.org/2001/vcard-rdf/3.0#> .
@prefix swed_ot: <http://www.swed.org.uk/2004/02/swed/org_type#> .
@prefix swed_id: <urn:x-hp-swed:> .
@prefix swed_at: <http://www.swed.org.uk/2004/02/swed/activity_type#> .
@prefix swed_pt: <http://www.swed.org.uk/2004/02/swed/project_type#> .
@prefix swed_toi: <http://www.swed.org.uk/2004/06/swed_toi#> .
@prefix swed: <http://www.swed.org.uk/2004/02/swed#> .

swed_id:prorg0001 a swed:prorg;
  swed:prorg_number swed_id:prorg0001;
  swed:has_primary_prorg_name "RSA";
  vcard:TEL [rdf:type vcard:voice; rdf:value "0171-930 5115"];
  vcard:TEL [rdf:type vcard:fax; rdf:value "0171-839 5805"];
  swed:has_acronym_prorg_name "RSA";
  swed:has_year_formed "1754";
  swed:has_long_description "The RSA, also known as Royal Society for the Encouragement of
Arts, Manufacturing and Commerce) was founded over 200 years ago to anticipate change and
prepare for it by encouraging creativity, implanting new ideas and seeking to transform
outdated attitudes. The main areas of current activity comprise design, education, the
environment, manufacturers and commerce, and the useful arts. The RSA is independent of
```

```

special interests and, because of its breadth of interests and independent stance, it is
uniquely placed to bring together a wide range of other organisations to work in
partnership. Links with industry are strongly evident in the RSA's environment programme
organised by the RSA's Committee for the Environment.";
swed:has_organisation_type swed_ot:registered_charity;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#arts_and_crafts>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#business_and_commerce>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#design>;
swed:has_primary_url <http://www.rsa.org.uk/>;
swed:primary_contact [
    foaf:name "name withheld";
];
];
swed:has_postal_address [
    vcard:Street "8 John Adam Street";
    vcard:Locality "Adelphi";
    vcard:Region "London";
    vcard:Pcode "WC2N 6EZ";
    vcard:Country "England";
];
.

swed_id:prorg0002 a swed:prorg;
swed:prorg_number swed_id:prorg0002;
swed:has_primary_prorg_name "Vincent Wildlife Trust";
vcard:TEL [rdf:type vcard:voice; rdf:value "0171-283 2089"];
vcard:TEL [rdf:type vcard:fax; rdf:value "0171-929 0604"];
swed:has_acronym_prorg_name "VWT";
swed:has_year_formed "1977";
swed:has_long_description "The Vincent Wildlife Trust operates an otter rehabilitation centre
for orphaned or injured otters from throughout the UK with reintroductions occurring in
Northern Ireland and eastern England.";
swed:has_organisation_type swed_ot:registered_charity;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#animal_welfare>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#farming>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#farming_fish_and_other_aquaculture>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#resource_management>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#management_water>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#pollution_control_remediation>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#recreation>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#recreation_water-based>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#species>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#species_animals>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#species_mammals>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#wildlife_habitats>;
swed:has_primary_url <http://www.vwt.org.uk/>;
swed:primary_contact [
    foaf:name "name withheld";
    swed:contact_person_position "Secretary / Treasurer";
];
];
swed:has_postal_address [
    vcard:Street "10 Lovat Lane";
    vcard:Region "London";
    vcard:Pcode "EC3R 8DT";
    vcard:Country "England";
];
.

swed_id:prorg0003 a swed:prorg;
swed:prorg_number swed_id:prorg0003;
swed:has_primary_prorg_name "Impact Weather Services Ltd";
vcard:TEL [rdf:type vcard:voice; rdf:value "01556 611117"];
vcard:TEL [rdf:type vcard:fax; rdf:value "01556 611226"];
swed:has_previous_prorg_name "Weather Watchers Network Ltd";
swed:has_acronym_prorg_name "IWS";
swed:has_year_formed "1987";
swed:has_long_description "The aims of IWS are mainly educational. School parties regularly visit
the Weather Centre and many tourists come during the holiday season. The tour lasts one hour and
a fee is charged. IWS also helps pupils with their weather projects.";

```

```

swed:has_organisation_type swed_ot:registered_charity;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#education>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#information_services_enquiries>;
swed:has_primary_url <http://www.impactweather.co.uk/>;
swed:primary_contact [
    foaf:name "name withheld";
    swed:contact_person_position "M.D.";
];
];
swed:has_postal_address [
    vcard:Extadd "The Weather Centre";
    vcard:Locality "Millisle";
    vcard:Region "Dalbeattie";
    vcard:Pcode "DG5 4AX";
    vcard:Country "Scotland";
];
.

swed_id:prorg0004 a swed:prorg;
swed:prorg_number swed_id:prorg0004;
swed:has_primary_prorg_name "Free Form Arts Trust";
vcard:TEL [rdf:type vcard:voice; rdf:value "0171-249 3394"];
vcard:TEL [rdf:type vcard:fax; rdf:value "0171-249 8499"];
swed:has_year_formed "1969";
swed:has_long_description "Free Form Arts Trust offers skills in environmental design, community art and architecture, landscape architecture and planning to business, local authorities and communities. It organises and manages schemes which develop opportunities for the creative participation of client groups and individuals through consultation and work practise. It offers design and build schemes, community design and technical aid services, and education programmes. It promotes the use of arts, crafts and decoration in environmental improvement work. Northern Free Form carries out innovatory environmental arts projects in the North East, creating employment for local artists and craftspeople through partnerships in urban regeneration.";
swed:has_organisation_type swed_ot:registered_charity;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#arts_and_crafts>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#built_environment>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#design>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#education>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#education_training>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#landscape_and_landscape_design>;
swed:has_primary_url <http://www.freeform.org.uk/>;
swed:primary_contact [
    foaf:name "name withheld";
    swed:contact_person_position "Company Secretary";
];
];
swed:has_postal_address [
    vcard:Street "57 Dalston Lane";
    vcard:Region "London";
    vcard:Pcode "E8 2NG";
    vcard:Country "England";
];
.

swed_id:prorg0005 a swed:prorg;
swed:prorg_number swed_id:prorg0005;
swed:has_primary_prorg_name "London Green Belt Council";
vcard:TEL [rdf:type vcard:voice; rdf:value "0181-467 5346"];
swed:has_acronym_prorg_name "LGBC";
swed:has_year_formed "1954";
swed:has_long_description "The objective of the LGBC is to keep the Metropolitan Green Belt under review. Representations are made to the Department of the Environment; advice and assistance is given on planning matters and the preparation of appeals; representations are made to public enquiries; regular bulletins are issued to member organisations; and regular and ad hoc meetings are held in London for member organisations.";
swed:has_organisation_type swed_ot:voluntary_sector_organisation;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#landscape_and_landscape_design>;
swed:has_topic <http://www.swed.org.uk/2004/06/swed_toi#planning_system>;
swed:has_primary_url <http://www.londongreenbeltcouncil.org.uk >;
swed:primary_contact [
    foaf:name "name withheld";
];

```



```

    swed:contact_person_position "Honorary Secretary";
];
swed:has_care_of_prorg_address [
  swed:care_of_address_line "c/o 13 Oakleigh Park Avenue";
  vcard:Locality "Chislehurst";
  vcard:Region "Kent";
  vcard:Pcode "BR7 5PB";
  vcard:Country "England";
];
.

```

C.2 swed_org_type_skos.n3

Five itemd from the *type* ontology.

```

@prefix swed_ot: <http://jena.hpl.hp.com/2004/02/swed/org_type#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix : <http://owl.protege.stanford.edu#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3c.rl.ac.uk/2003/11/21-skos-core#> .

swed_ot:educational_organisation
  a skos:Concept ;
  rdfs:comment "Temporary documentation in the place of a full scope note for
  this term" ;
  rdfs:label "Educational Organisation"@en ;
  skos:broader swed_ot:organisation ;
  skos:prefLabel "Educational Organisation"@en .

swed_ot:charitable_organisation
  a skos:Concept ;
  rdfs:comment "In general a charitable organisation has purposes that are
  exclusively charitable [in law]. In England and Wales these include: the
  relief of financial hardship; the advancement of education; the
  advancement of religion and certain other purposes for the benefit of the
  community" ;
  rdfs:label "Charitable Organisation"@en ;
  skos:broader swed_ot:not_for_profit_organisation ;
  skos:prefLabel "Charitable Organisation"@en .

swed_ot:campaign_organisation
  a skos:Concept ;
  rdfs:comment "pressure groups, NGO" ;
  rdfs:label "Campaign Organisation"@en ;
  skos:broader swed_ot:organisation ;
  skos:prefLabel "Campaign Organisation"@en .

swed_ot:umbrella_organisation
  a skos:Concept ;
  rdfs:comment "Any kind of umbrella organisation including; associations,
  networks and federations." ;
  rdfs:label "Umbrella Organisation"@en ;
  skos:broader swed_ot:organisation ;
  skos:prefLabel "Umbrella Organisation"@en .

swed_ot:development_organisation
  a skos:Concept ;
  rdfs:comment "Organisations that work in the area of world development
  issues." ;
  rdfs:label "Development Organisation"@en ;

```

```
skos:broader swed_ot:organisation ;
skos:prefLabel "Development Organisation"@en .
```

C.3 wwrite_index_skos.n3

Five items from the *topic* ontology.

```
@prefix rss:      <http://purl.org/rss/1.0/> .
@prefix swed_wi:  <http://http://jena.hpl.hp.com/2004/02/swed/wwrite_index#> .
@prefix vcard:    <http://www.w3.org/2001/vcard-rdf/3.0#> .
@prefix dc:       <http://purl.org/dc/elements/1.1/> .
@prefix rdfs:     <http://www.w3.org/2000/01/rdf-schema#> .
@prefix jms:      <http://jena.hpl.hp.com/2003/08/jms#> .
@prefix daml:     <http://www.daml.org/2001/03/daml+oil#> .
@prefix rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix :         <http://jena.hpl.hp.com/2004/02/swed/wwrite_index#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix skos:     <http://www.w3c.rl.ac.uk/2003/11/21-skos-core#> .

:livestock
  a          skos:Concept ;
  rdfs:label "Livestock Farming"@en ;
  skos:broader :farming ;
  skos:externalID "OFALO" ;
  skos:prefLabel "Livestock Farming"@en .

:woodlands_and_forests
  a          skos:Concept ;
  rdfs:label "Woodland and Forest Habitats"@en ;
  skos:broader :wildlife_habitats ;
  skos:externalID "OWFO" ;
  skos:prefLabel "Woodland and Forest Habitats"@en .

:transport
  a          skos:Concept ;
  rdfs:label "Transport"@en ;
  skos:broader :wwrite_index_taxonomy ;
  skos:externalID "OTO" ;
  skos:prefLabel "Transport"@en .

:information_services
  a          skos:Concept ;
  rdfs:label "Information Services"@en ;
  skos:broader :wwrite_index_taxonomy ;
  skos:externalID "OIO" ;
  skos:prefLabel "Information Services"@en .

:captive_animals
  a          skos:Concept ;
  rdfs:comment "captive animals (zoos, circuses, etc)" ;
  rdfs:label "Captive Animals"@en ;
  skos:broader :animal_welfare ;
  skos:externalID "OAWCO" ;
  skos:prefLabel "Captive Animals"@en .
```

C.4 Input Data Sets

This section lists the names of the organisations in the 3 input data sets extracted from *wwite_tidy.n3*. The relation between an organisation and *topic* is one-to-many, hence there is some overlap between the three input data sets.

C.4.1 set1

Environment Focus (15 organisations in total)

Institute for Transport Studies
Tree Council, The
Media Natura
Butterfly Conservation
DOE (NI) - Environment and Heritage Service - Environmental Protection Directorate
Forum for the Future
John Muir Award
National Energy Action
BirdLife International
London and South East Regional Planning Conference
Scottish Solar Energy Group
Northern Ireland Environmental Standards Group
Convention of Scottish Local Authorities
Scottish Society for the Prevention of Cruelty to Animals
Nature in Art Trust

C.4.2 set2

Animal focus (15 organisations in total)

Vincent Wildlife Trust
National Trust for Scotland
Barn Owl Trust
North East Scotland Preservation Trust
Oxford Centre for the Environment, Ethics and Society
Zoological Society of London, The
Royal Highland and Agricultural Society of Scotland
British Lichen Society
Advocates for Animals
Council for Posterity
Countryside In and Around Towns Network
Scottish Solar Energy Group
New Economics Foundation
Cyclists' Touring Club
Royal Forestry Society of England, Wales and Northern Ireland, The

C.4.3 set3

Community focus (8 organisations in total)

Flood Hazard Research Centre
Regenerative Technology
Vision 21 - Action for Sustainable Communities
DOE (NI) - Environment and Heritage Service - Environmental Protection Directorate
John Muir Award
Linking Environment and Farming
London and South East Regional Planning Conference
Scottish Society for the Prevention of Cruelty to Animals

Appendix D

Source Code

Due to the big size of the source only selected source files are included in this section. Remaining code is accessible at the location specified in the chapter titled Online Resources.

D.1 RandomMutationOperator.java

```
/*
 * Created on 11-Jul-2004
 */
package edu.cs.bham.ac.uk.ais.core;

import java.util.Iterator;
import java.util.Random;

import edu.cs.bham.ac.uk.ais.common.Antibody;
import edu.cs.bham.ac.uk.ais.common.MutationOperator;
import edu.cs.bham.ac.uk.semanticsearch.core.Constants;
import edu.cs.bham.ac.uk.semanticsearch.core.DataModel;
import edu.cs.bham.ac.uk.semanticsearch.core.SemanticQuery;

/**
 * @author Rana Kashif Ali
 */
public class RandomMutationOperator implements MutationOperator {

    private Random random;

    public RandomMutationOperator() {
        random = new Random();
    }

    public Antibody mutate(Antibody antibody) {

        //isolate the genotype for mutation
        SemanticQuery query = (SemanticQuery)antibody.getGenotype();
        //a new genotype
        SemanticQuery mutantQuery = new SemanticQuery();

        //fetch all topics in the ontology file from the cache
        String[] allTypes = DataModel.getInstance().getTypes();
        double typeChangeProb =
            Double.parseDouble(Constants.getInstance().getPropertyValue(Constants.TYPE_CHG_PROB));
```

```

if(random.nextDouble() < typeChangeProb) {
mutantQuery.addType(allTypes[random.nextInt(allTypes.length)]);
}
else {
Iterator iter = query.getTypes().iterator();
while(iter.hasNext()) {
//retain the same type
mutantQuery.addType((String)iter.next());
}
}

//fetch all topics in the ontology file from the cache
String[] allTopics = DataModel.getInstance().getTopics();

// fetch the maximum no. of topics allowed in the query
int topicCount =
random.nextInt(Integer.parseInt(Constants.getInstance().getPropertyValue(Constants.MAX_TOPICS_IN_QUERY))) + 1;

//fetch the topics of the previous query
Object[] prevTopics = query.getTopics().toArray();

for(int i=0;i<topicCount;i++) {
double probability = Double.parseDouble(Constants.getInstance().getPropertyValue(Constants.INERTIA));
if(random.nextDouble() <= probability) {
//choose topic from previous topics
mutantQuery.addTopicToInclude( (String)prevTopics[ random.nextInt(prevTopics.length) ] );
}
else {
//choose topic from topic ontology
mutantQuery.addTopicToInclude( allTopics[ random.nextInt(allTopics.length) ] );
}
}

//return the mutant antibody
return new AntibodyImpl(mutantQuery);
}
}

```

D.2 ConstrainedMutationOperator.java

```

/*
 * Created on 31-Jul-2004
 */
package edu.cs.bham.ac.uk.ais.core;

import java.util.ArrayList;
import java.util.Collection;
import java.util.Iterator;
import java.util.List;
import java.util.Random;

import edu.cs.bham.ac.uk.ais.common.Antibody;
import edu.cs.bham.ac.uk.ais.common.MutationOperator;
import edu.cs.bham.ac.uk.semanticsearch.core.Constants;
import edu.cs.bham.ac.uk.semanticsearch.core.DataModel;
import edu.cs.bham.ac.uk.semanticsearch.core.SemanticQuery;

/**
 * @author Rana Kashif Ali
 */
public class ConstrainedMutationOperator implements MutationOperator {

private Random random;

public ConstrainedMutationOperator() {
random = new Random();
}
}

```

```

}

public Antibody mutate (Antibody antibody) {

//isolate the genotype for mutation
SemanticQuery query = (SemanticQuery)antibody.getGenotype();

//a new genotype
SemanticQuery mutantQuery = new SemanticQuery();

//fetch all topics in the ontology file from the cache
String[] allTypes = DataModel.getInstance().getTypes();
double typeChangeProb = Double.parseDouble(Constants.getInstance().getPropertyValue(Constants.TYPE_CHG_PROB));
if(random.nextDouble() < typeChangeProb) {
mutantQuery.addType(allTypes[random.nextInt(allTypes.length)]);
}
else {
Iterator iter = query.getTypes().iterator();
while(iter.hasNext()) {

//retain the same type
mutantQuery.addType((String)iter.next());
}
}

int prevTopicCount = 0;
Iterator iter = query.getTopics().iterator();

String[] prevTopics = new String[query.getTopics().size()];

while(iter.hasNext()) {
String topic = (String)iter.next();
mutantQuery.addTopicToInclude(topic);
prevTopics[prevTopicCount++] = topic;
}

String[] unusedTopics = getTopicsNotInQuery(DataModel.getInstance().getTopics(),query.getTopics());

//fetch the mutation rate, input parameter
double mutationRate = Double.parseDouble(Constants.getInstance().getPropertyValue(Constants.APPEND_PROB));

double appendProbability = random.nextDouble();
double deleteProbability = random.nextDouble();
double changeProbability = random.nextDouble();

List newTopics = new ArrayList();
List toDelete = new ArrayList();
for(int i=0;i<prevTopics.length;i++) {
//for each topic do some type of mutation
String topic = prevTopics[i];

if(appendProbability >= mutationRate) {
newTopics.add(unusedTopics[random.nextInt(unusedTopics.length)]);
}

if(deleteProbability >= mutationRate) {
toDelete.add(topic);
}

if(changeProbability >= mutationRate) {
toDelete.add(topic);
newTopics.add(unusedTopics[random.nextInt(unusedTopics.length)]);
}
}
}

```

```

newTopics.removeAll(toDelete);

//clear all the topics
mutantQuery.removeAllTopics();

//add the selected topic in the query
iter = newTopics.iterator();

while(iter.hasNext()) {
mutantQuery.addTopicToInclude((String)iter.next());
}

//return the mutant antibody
return new AntibodyImpl(mutantQuery);
}

//helper method to get unused topics in the query
private String[] getTopicsNotInQuery(String[] allTopics, Collection topicsInQuery) {
String[] unusedTopics = new String[allTopics.length - topicsInQuery.size()];
boolean used = false;
int index = 0;
for(int i=0;i<allTopics.length;i++) {
Iterator iter = topicsInQuery.iterator();
while(iter.hasNext()) {
if(allTopics[i] == (String)iter.next())
used = true;
}
if(used) {
//if used then dont add
used = false;
}
else {
//add unused topic
unusedTopics[index++] = allTopics[i];
}
}
return unusedTopics;
}
}

```