AIS and Semantic Query

Rana Kashif Ali and Steve Cayzer

[1] The University of Birmingham, Birmingham B15 2TT, UK
[2] HP Laboratories, Bristol, UK

Abstract. 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.

1 Introduction

The presented work combines disparate areas of research namely, semantic web, Artificial Immune Systems (AIS), Query Expansion (QE) and Information Retrieval (IR). In this section, we introduce these concepts before outlining the structure of the remaining paper.

The Semantic Web is an extension of the current World Wide Web (WWW) in which resources are connected semantically rather than through hyperlinks. This semantic connectivity is achieved by making metadata about resources available for machine processing. Metadata is typically written in Resource Description Format (RDF: [4]) to conform to a model, or ontology. Both the metadata and ontologies are available to all. Different frameworks exist to manipulate the RDF metadata, for example Jena [3], which supports the Resource Description Query Language (RDQL; [5]) for querying metadata.

Computational models and problem solving approaches inspired from the Biological Immune System (BIS) are called AIS [7]. The presented work focuses on the *clone-and-refine* paradigm of the BIS, according to which, a body exposed to antigen produces various antibodies, some of which (those showing a higher affinity

A. Gómez-Pérez and J. Euzenat (Eds.): ESWC 2005, LNCS 3532, pp. 333 – 345, 2005.

to the antigen) are more suitable to overcome the infection. These antibodies undergo affinity-related cloning and mutation to produce novel, but similar, antibodies, some of which might be an improvement over the original antibody and can better tackle the infection. However, some antibodies may be self-reactive and hence must be destroyed or they will cause an autoimmune reaction. We bring this idea into the realm of query expansion by establishing an analogy between antibodies and queries, refining the search process with clonal expansion, mutation and screening of self-reactive queries.

The next section gives an overview of the related work, followed by the details of the AIS and query expansion. In section 4 we describe the experimental plan and the obtained results. In the final section we discuss future directions and present our conclusions.

2 Related Work

AIS is a relatively new area of research with a diversity of applications such as data mining, computer security and robotics. A full survey can be found in Jon Timmis' and Leandro de Castro's book [7], but here we describe work relevant to our application.

The notion of AIS for semantic queries was first proposed by Lee et al [1]. They show, using the Gene Ontology (GO) as an example, how data can be retrieved based on the principles of immunity by expanding queries. Their work does not involve a concrete implementation but does provide a useful conceptual framing for our work. We have applied the idea to a new domain, filled in some details and provided a real application that we evaluate.

Efthimiadis's work [2] provides a sound foundation of traditional query expansion, drawing a distinction between manual, automated and interactive approaches. However the methods he describe are predominantly keyword based - that is, not semantic. Thus, we aim to demonstrate the feasibility of using semantic queries within a principled query expansion framework.

Our work is grounded using real semantic web data. The Semantic Web Environmental Directory, SWED [6], provides a decentralised, RDF-backed portal for storing the details of environmental organisations in the UK. SWED provides a novel 'facet browse' mechanism that enables users to navigate to the organisations of interest using conjunctive combinations of metadata attributes (for example, "Not for Profit organisations based in Bristol that are concerned with animal welfare"). Our semantic query mechanism facilitates a different approach, akin to a semantic "More Like This" utility.

3 AIS for Semantic Query Expansion

A web based semantic search utility was developed with the AIS infrastructure embedded in it. A high level view of the utility is shown in Figure 1. It is also important at this stage to establish the mapping between the AIS and BIS. In our AIS

we regard the irrelevant results as self and relevant results as non-self. Antibodies are semantic queries and antigens are a collection of the non-self (relevant results). Finally, mutation is equivalent to query expansion. Thus, mutation of a query may result in queries that are better suited to answer a particular search criterion. On the other hand, mutation may result in queries that return irrelevant results; these are deemed self-reactive and hence are destroyed.

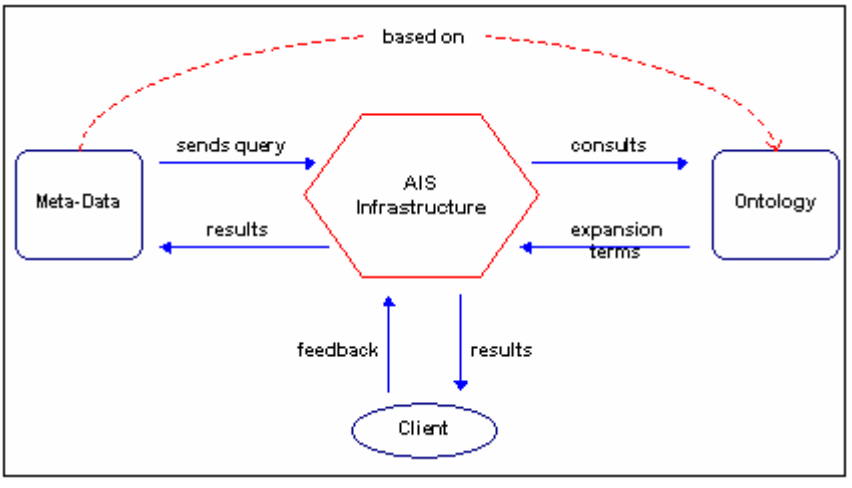

Fig. 1. AIS infrastructure and flow of information

3.1 User Interface

The interface to the search process is designed so as to let the user know how the query expansion is being done. Initially, the user chooses a particular organisation of interest, the details of which appear as shown in figure 2. This is a single record drawn from the SWED dataset that will seed the semantic queries. The only feedback at this stage involves the user clicking on the 'FIND SIMILAR' link. Upon the feedback the AIS initialises, fetches relevant organisations and presents them grouped by queries (figure 3). This new interface lets the user build sets of self and non-self by specifying results as either relevant (non-self) or irrelevant (self). These sets act as an evaluation mechanism for query refinement, guiding the semantic query population towards novel relevant results. The interface also bears links at the top that allow users to view/edit sets of self and non-self. There is also an option to view the state of the query pool that gives a good insight into the expansion process. This feature may however be removed from a commercial application to avoid complexity.

3.2 AIS Algorithm

The algorithm for the AIS is given below followed by explanation of its constituents.

```
begin
    take initial user feedback
    initialise Q, query population based on feedback
    while (halting criteria not met)
        display the results of queries in Q and take user feedback
        add relevant results to non-self and irrelevant to self
        evaluate fitness of queries in Q
        select queries with highest fitness (Q_s) using fitness
proportionate selection
        perform clonal expansion on the selected queries to form Q_c
        apply mutation operator to transform Q_c to Q_m
        replace the previously selected queries Q_s with Q_m
    end while
end
```

Organisation Details

Vincent Wildlife Trust

Organisation number	prorg0002
Acronym	VWT
Year formed	1977
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.
Type	registered_charity
Topics	animal_welfare | farming | farming_fish_and_other_aquaculture | resource_management | management_water | pollution_control_remediation | recreation | recreation_water-based | species | species_animals | species_mammals | wildlife_habitats
Telephone	voice: 0171-283 2089 fax: 0171-929 0604
Email	contact@vwt.org.uk
URL	http://www.vwt.org.uk/
Primary Contact	Secretary / Treasurer
Postal address	10 Lovat Lane, London , EC3R 8DT, England

FIND SIMILAR

Fig. 2. User interface for semantic query invocation

Initial User Feedback

This involves a user specifying one organisation of interest and saying that s/he wants to find similar organisations.

Initialisation of AIS

When the user click the 'FIND SIMILAR' link the AIS is initialised with the query population equal to the input parameter `INIT_POP_SIZE` (5 in our case). This initial query population is generated randomly using two ontologies used in the SWED data, namely *organisation_type* and *topic*. The pseudo code for the initialisation of the AIS is given below

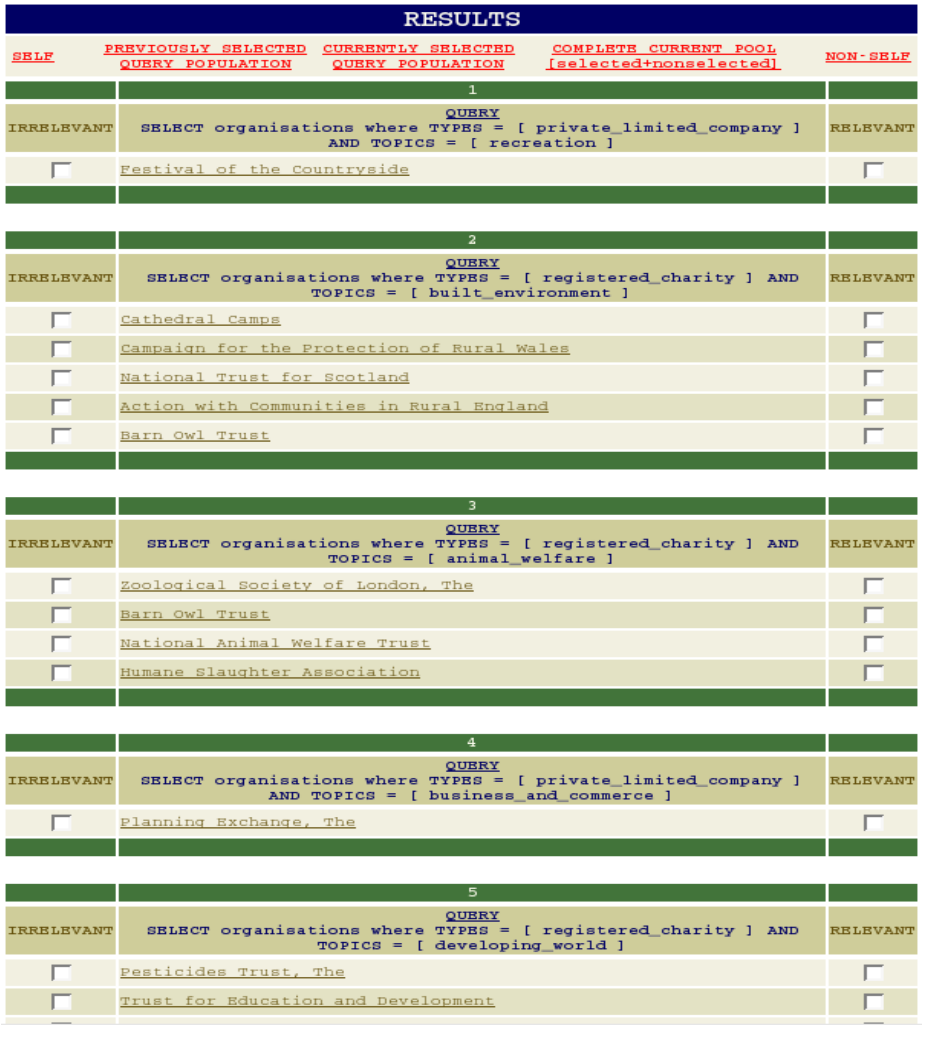

Fig. 3. User interface for semantic query expansion

```
set generatedQueries = 0
while(INIT_POP_SIZE > generatedQueries) {
    randomly select a organisation_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 organisation_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. As mentioned, the user may give feedback by specifying whether a particular result is irrelevant or relevant.

Fitness Evaluation
Once the user has given feedback, the antibodies/queries need to be evaluated. The fitness of an antibody in our case is the measure of how well it binds to the non-self while avoiding self. This is equivalent to a search for queries returning many relevant and few irrelevant results. The following formula was used to evaluate the antibodies

$$affinity = \frac{NonSelf \times w_{pos} + Self \times w_{neg} + New \times w_{neutral}}{total\_number\_of\_results} \quad (1)$$

Queries can return results that are relevant (`NonSelf`), irrelevant (`Self`) or have unknown relevance (`New`). The numbers of each result set are weighted and combined into a fitness function, whose weights are: $w_{pos} = 1$, $w_{neg} = 0$ and $w_{neutral} = 0.4$. The choice of the values for different weights was empirical.

Selection
In a pure AIS individuals are selected so as to maximise the collective affinity against the antigen called *affinity maturation*. Affinity maturation is fitness proportionate and thus can be modelled as roulette wheel selection.

Clonal Expansion
This is a two step process the first step involves generation of the clones based on fitness and the second step involves mutation of the clones using mutation operators. Any cloned antibody/query should fulfil the following constraint.

$$\{all\_results\} - \{Self \cup NonSelf\} \neq \varnothing \quad (2)$$

In other words a query should return some previously unseen results.

Mutation Operators
The two query mutation operators that we used for evaluation of the AIS are as follows:

ConstrainedMutationOperator
This operator appends, deletes or changes various characteristics of the individual, retaining all others 'as is'.

```
    generate a random number between 0 and 1, random
    if(random < TYP_CHG_PROB) {
      replace existing organisation_type with a one randomly chosen
from the ontology
```

```
  }
  else {
    retain the old organisation_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
    }
  }
```

RandomMutationOperator

This is a more exploratory operator, which allows a considerable degree of novelty in the generated query.

```
  generate a random number between 0 and 1, random
  if(random < TYP_CHG_PROB) {
    replace existing organisation_type with a one randomly chosen
from the ontology
  }
  else {
    retain the old organisation_type
  }
  generate a random number between 1 and MAX_TOPICS_IN_QUERY, count
  while(count != 0) {
    generate a random number between 0 and 1, 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
    }
    decrement count
  }
```

Replacement Strategy and Halting Criteria
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. There are two possible halts to the search process. Firstly, when all the desired results have been found. Secondly, if further query expansion is not possible.

4 Experimental Setup and Results

We compared the proposed mutation operators in terms of precision, recall and convergence. In the first set of experiments the performance of the two operators was

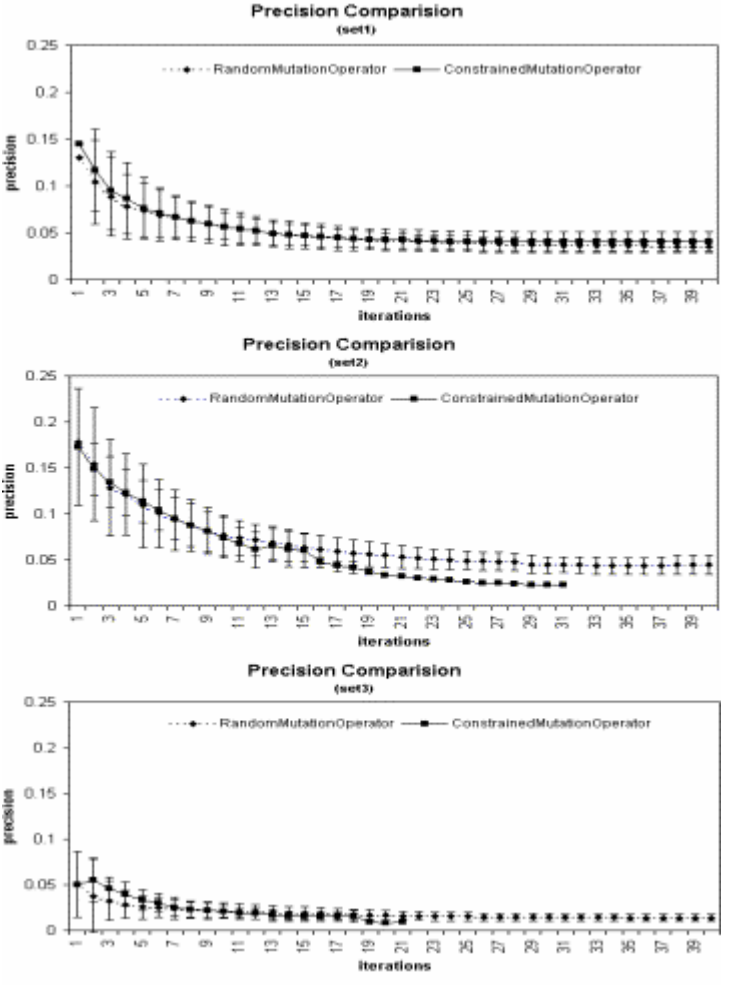

Fig. 4. Precision comparison on different input data sets averaged over 30 runs. Bars show standard deviation

observed on three different input data sets. The second set of experiments was aimed towards finding the change in performance with changing mutation rate on only one input data set. We implemented an automated test script to simulate a user interacting with the system. The script was controlled by various parameters for example the maximum number of iterations and results to be marked as relevant or irrelevant in every iteration. For the first set of experiments, three input data sets were selected by a real user of the system, each containing around 10 relevant and 90 irrelevant items. The task of the AIS was to find organisations in a particular input data set in minimum number of iterations. The starting point for the AIS was one randomly chosen organisation from the set. Figure 4 shows the precision for both operators on three different input data sets. The precision was measured cumulatively:

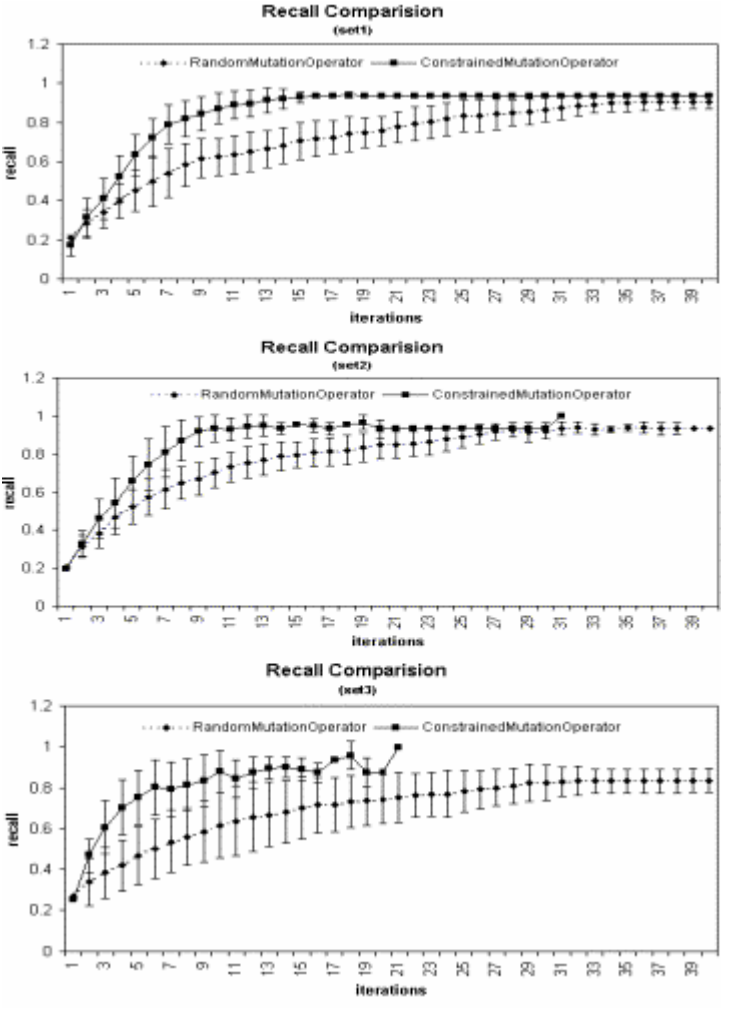

Fig. 5. Recall comparison on different input data sets, averaged over 30 runs. Bars show standard deviation

$$precision = \frac{relevant\_results\_so\_far}{all\_results\_so\_far} \tag{3}$$

This cumulative precision decayed asymptotically, as it becomes progressively harder to find the remaining relevant items. In the early stages, the constrained operator was significantly superior (e.g. iteration 4, dataset 1: p-value < 0.05, Student's t-test).

Figure 5 shows the recall comparison. Again, recall was calculated cumulatively, so it increases asymptotically to a theoretical maximum of 1.0.

$$recall = \frac{relevant\_results\_so\_far}{all\_relevant\_results} \tag{4}$$

The *ConstrainedMutationOperator* clearly performs better since it reaches a higher value of recall more quickly (iteration 10, all datasets: p-value < 0.0001).

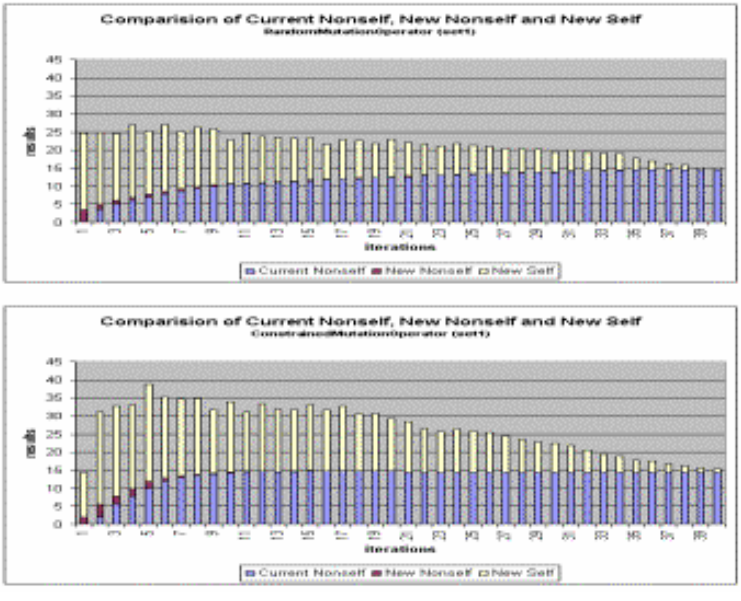

Fig. 6. Convergence comparison on input data set 1, averaged over 30 runs

Figures 6, 7 and 8 compare the convergence between the two operators. *RandomMutationOperator* exhibits delayed convergence and finds fewer relevant and irrelevant results. The *ConstrainedMutationOperator* on the other hand is aggressive in nature and converges quickly.

For the second set of experiments we selected the input data set 2 and changed the mutation rate from 0 to 1. We found no significant different between the operators in terms of precision which remained under 0.2. However, in case of recall we found that the two operators behave in an opposite way (figure 9). The exploratory

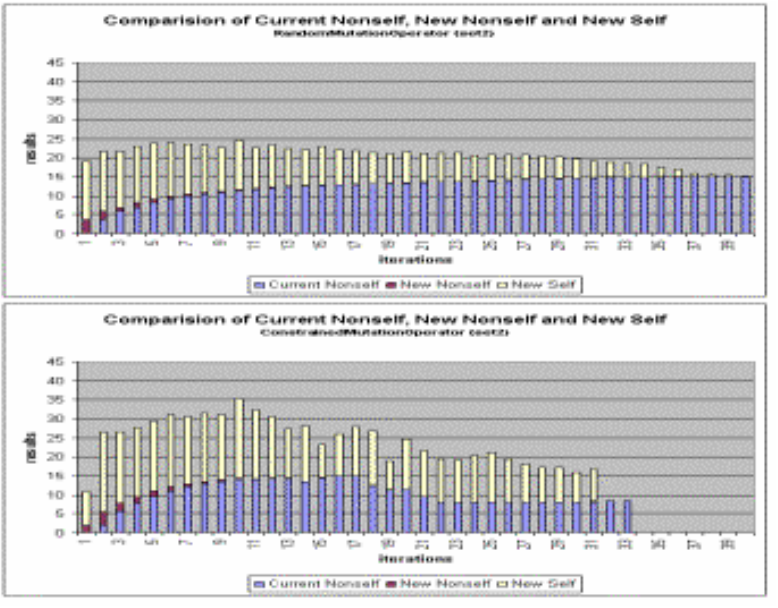

Fig. 7. Convergence comparison on input data set 2, averaged over 30 runs}

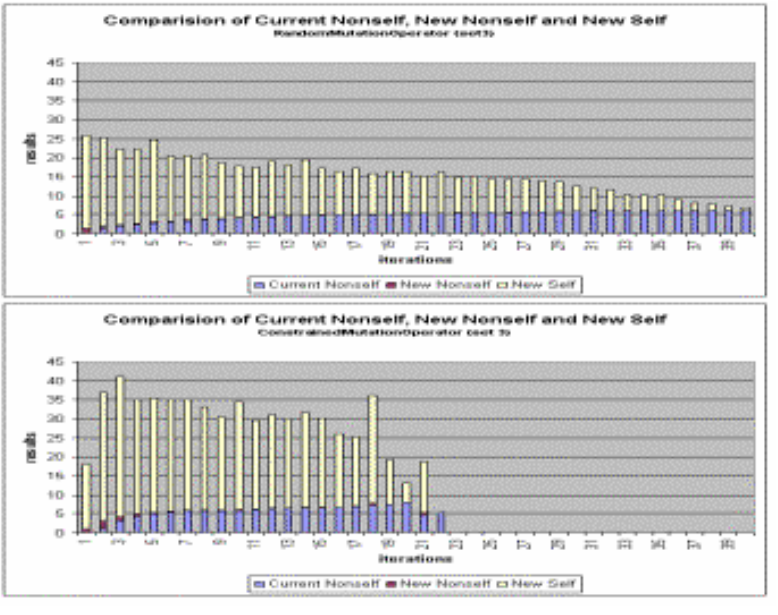

Fig. 8. Convergence comparison on input data set 3, averaged over 30 runs}

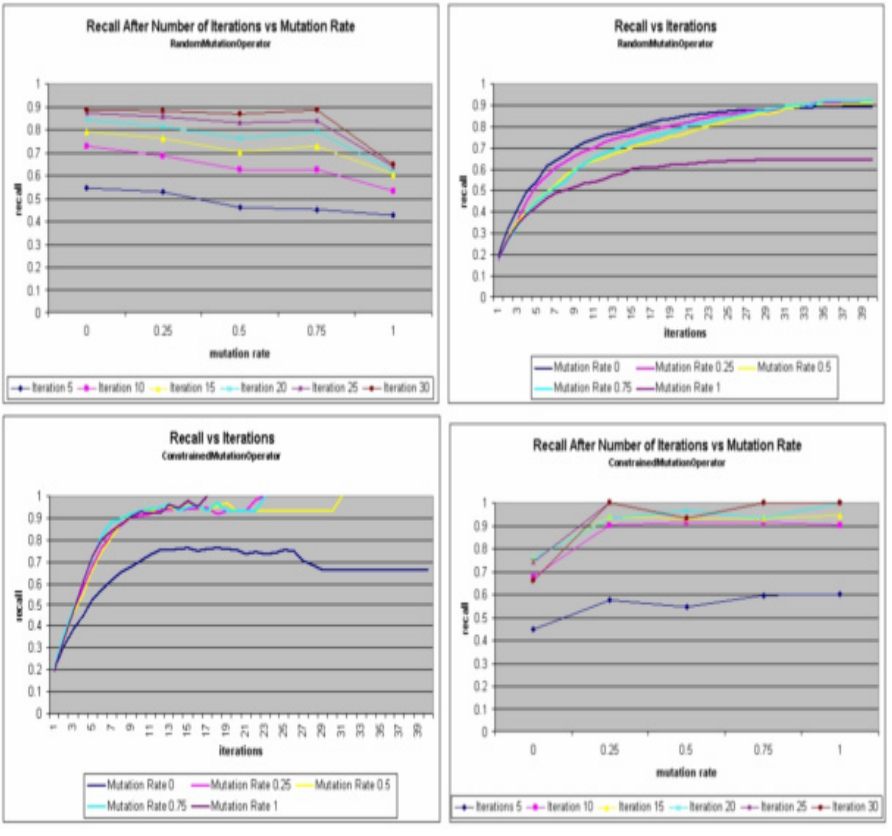

Fig. 9. Recall comparison with changing mutation rate

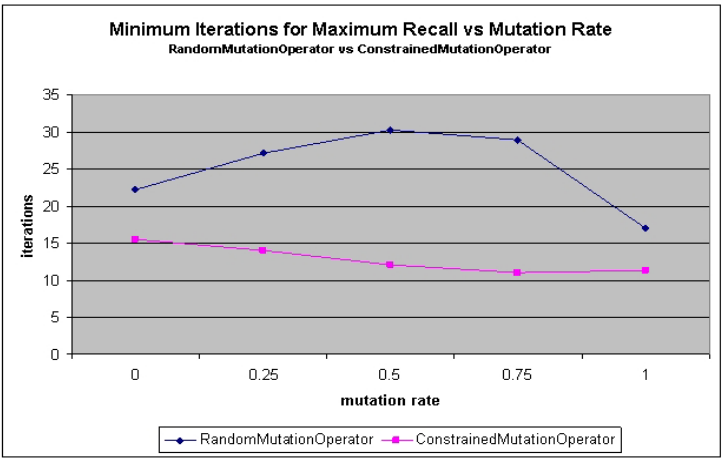

Fig. 10. Minimum iterations for maximum recall vs. mutation rate, averaged over 30 runs

RandomMutationOperator benefits from a low mutation rate, whereas the more aggressive *ConstrainedMutationOperator* requires some mutation in order to avoid premature convergence. These results are underlined by our experiments investigating the effect of mutation rate on speed of convergence (figure 10).

5 Conclusion and Future Directions

We have shown in this paper that AIS are a useful metaphor for query expansion on the semantic web. Our initial mutation operators demonstrate ways of exploring and exploiting the query space. An obvious next step would be to try the operators on a larger dataset (more than 100 organisations) with more sophisticated semantic markup (more than two ontologies). Another fruitful direction would be a study to explore suitable metaphors for the user interface. Finally it would be possible to integrate this work into the SWED portal and to provide value to a real semantic web community.

References

1. Lee, D., Kim, J., Jeong, M., Won, Y., Park, H., Lee, K.: Immune-Based Framework for Exploratory Bio -Information Retrieval from the Semantic Web. Artificial Immune Systems: Second International Conference, ICARIS 2003, Edinburgh, UK, September 1-3, 2003, Proceedings **2787** (2003) 128--135 Lecture Notes In Computer Science, Springer.
2. Efthimiadis N.E.: Annual Review of Information Systems and Technology (ARIST) Query Expansion **31** 1996 121--187 Information Today Inc Medford, NJ
3. Jena Semantic Web Framework http://jena.sourceforge.net/
4. Resource Description Framework (RDF) http://www.w3.org/RDF/
5. RDQL - A Query Language for RDF W3C Member Submission 9 January 2004
6. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/
7. SWED - The Semantic Web Environmental Directory
8. http://www.swed.org.uk/swed/index.html
9. de Castro, L.N., Timmis, J.: Artificial Immune Systems: A New Computational Approach Sept 2002 Springer-Verlag London. UK.