Int. J. Advance Soft Compu. Appl, Vol. 9, No. 1, March 2017
ISSN 2074-8523

Adaptive User Profiling: Experimenting on Multiple Interests and Changing Interests

Nurulhuda Firdaus Mohd Azmi[1,2], Suriani Md. Sam[1], Saiful Adli[1], Nilam Nur Amir Sjarif[1]

[1] Advanced Informatics School (UTM AIS)
Universiti Teknologi Malaysia, Jln Sultan Yahya Petra, 54100 Kuala Lumpur
[2] UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research
Universiti Teknologi Malaysia, 81310 Johor Malaysia
e-mail: huda@utm.my,suriani.kl@utm.my,
saifuladli@utm.my, nilamnur@utm.my

Abstract

Self-adaptation is an important capability of applications deployed in dynamically changing environments, such as adaptive information filtering (AIF). Immunological inspiration is typically applied to the computational perspective to promote software with characteristics such as self-organization, adaptation and learning. In this paper, we experiment the features of dynamic clonal selection (DCS) focusing on the gene libraries to produce reasonable coverage of detectors to detect changes of multi interest of topics. We experiment how the capability of DCS through diversity of detector in gene libraries classify multi topics of user's interests. Moreover, we further investigate the ability of the DCS to forget a previous topic when it starts to process a new topic. The results shows that the antibody-antigen interaction of B cells and the introduction of the diverse coverage of detectors helps further improve the capability for identifying new and potentially uninteresting topics thus have a positive effect on the adaptability of the profile to changes in user interests.

Keywords: *Adaptive Information Filtering, Dynamic Clonal Selection, Gene Libraries, Profile Adaptation, Dynamic User Profile*

1 Introduction

Conventional information filtering systems can be considered *static*, in that the components of the system work in deterministic ways. However, these conventional static systems are unable to maintain the quality of output as the inputs change character; for these applications, an *adaptive* approach is needed.

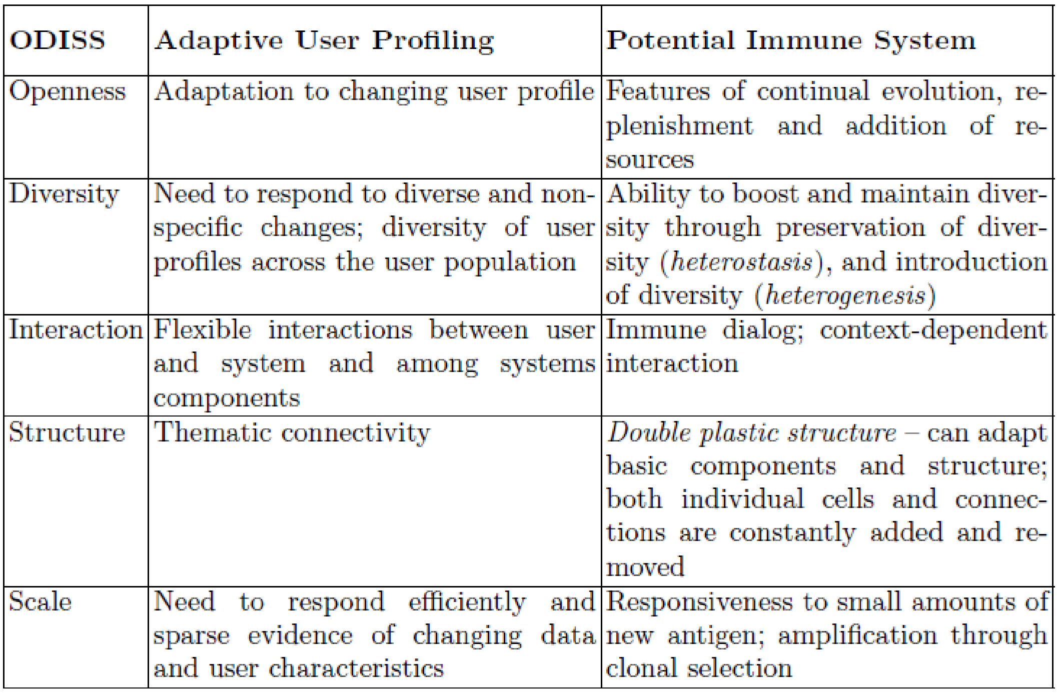
Tauritz [11] defined an Adaptive Information Filtering (AIF) system as ``a system that is capable of adapting to changes in both the data stream and the information needs". Adaptation in information filtering involves the process of filtering incoming data streams in such a way that only relevant data (information) is preserved. The relevance of the data is dependent on the changing (adaptive) needs of a particular person or group of persons with a shared interest. In addition, the users' interest cannot be assumed to be constant. Therefore, a filtering system must be responsive to dynamic user interests, to users with multiple topic interests and to users with changing interests. Changes in the user interest may be caused by changes in the user's environment and knowledge. The combination of these parameters causes a variety of changes (dynamic) and renders the profile adaptation a challenging research area.

This paper is concerned with a study into experimenting Artificial Immune Systems (AIS) towards profile adaptation in AIF. We describe Dynamic Clonal Selection (DCS) algorithm that is used to maintain the profiles and the used of gene libraries in maintaining sufficient diversity for the set of terms that can be added to the profile during mutation. The goal is to adapt our multi-topic profile both to short-term variations in the user's need and to progressive, but potentially radical changes in long-term interests.

2 Principled Abstraction of Immune Inspired to Adaptive User Profiling

What is missing from existing research on bio-inspired adaptive systems is a principled and systematic way to identify appropriate biological inspiration. Stepney et al.[10] propose a conceptual framework explicitly as a guideline for designing bio-inspired algorithms. Andrews [1] attempts to apply the principled approach to AIS algorithm design, from a study of part of the mammalian immune system. He concludes, not unexpectedly, that the principled approach has potential, but it needs to be clear what sort of application is being designed. In our work, we are seeking an immune inspiration for adaptive user profiling, an algorithm that able to adapts and continuously to changing of user interests. As a first step towards identifying appropriate immune-inspiration, we follow Stepney et al.[10], that the underlying properties of classes of models can be abstracted by meta-probes known as *ODISS meta-questions*: Openness, Diversity, Interaction, Structure and Scale. The ODISS approach leads to a comprehensive set of requirements which allow the identification of an appropriate property of the immune system and the problem domain studied as described in [2,3]. From the mapping in Table 1, this work suggested that profile adaptation can be developed by incorporating ideas from aspects of DCS with the use of gene libraries to maintain sufficient diversity. Furthermore, through DCS, it can inherently maintain and boost diversity and can dynamically control the size of the immune repertoire by means of selection, cloning, and mutation procedures. Further implementation of the approach in the design of the algorithm and the experiment is presented in the next section.

Table 1: ODISS Characteristics of Immune Inspired Towards Adaptive User Profiling [3]



ODISS	Adaptive User Profiling	Potential Immune System
Openness	Adaptation to changing user profile	Features of continual evolution, replenishment and addition of resources
Diversity	Need to respond to diverse and non-specific changes; diversity of user profiles across the user population	Ability to boost and maintain diversity through preservation of diversity (*heterostasis*), and introduction of diversity (*heterogenesis*)
Interaction	Flexible interactions between user and system and among systems components	Immune dialog; context-dependent interaction
Structure	Thematic connectivity	*Double plastic structure* – can adapt basic components and structure; both individual cells and connections are constantly added and removed
Scale	Need to respond efficiently and sparse evidence of changing data and user characteristics	Responsiveness to small amounts of new antigen; amplification through clonal selection

3 Profile Adaptation through Dynamic Clonal Selection (ProAdDCS)

Profile adaptation is a challenging problem with distinct characteristics and requirements. The user profile must be capable of continuous learning and forgetting. A profile that only learns and does not forget will eventually become saturated with irrelevant features. Forgetting is necessary for maintaining an up to date representation of the user's interest. The dynamic nature of profile adaptation invites the application of biologically inspired approach. Of interest is the AIS which can inherently maintain and boost their diversity and can dynamically control the size of the immune repertoire. The clonal selection theory in the immune system has received the attention of researchers and given them inspiration to create algorithms that evolve candidate solutions by means of selection, cloning, and mutation procedures.

3.1 The Algorithm

A diagrammatic of ProAdDCS flow is shown in Fig.1. The gene libraries features of ProAdDCS is defined in Algorithm 1. While, the initialisation, the affinity function and the clone and mutation function is described in Algorithm 2, Algorithm 3 and Algorithm 4 respectively.

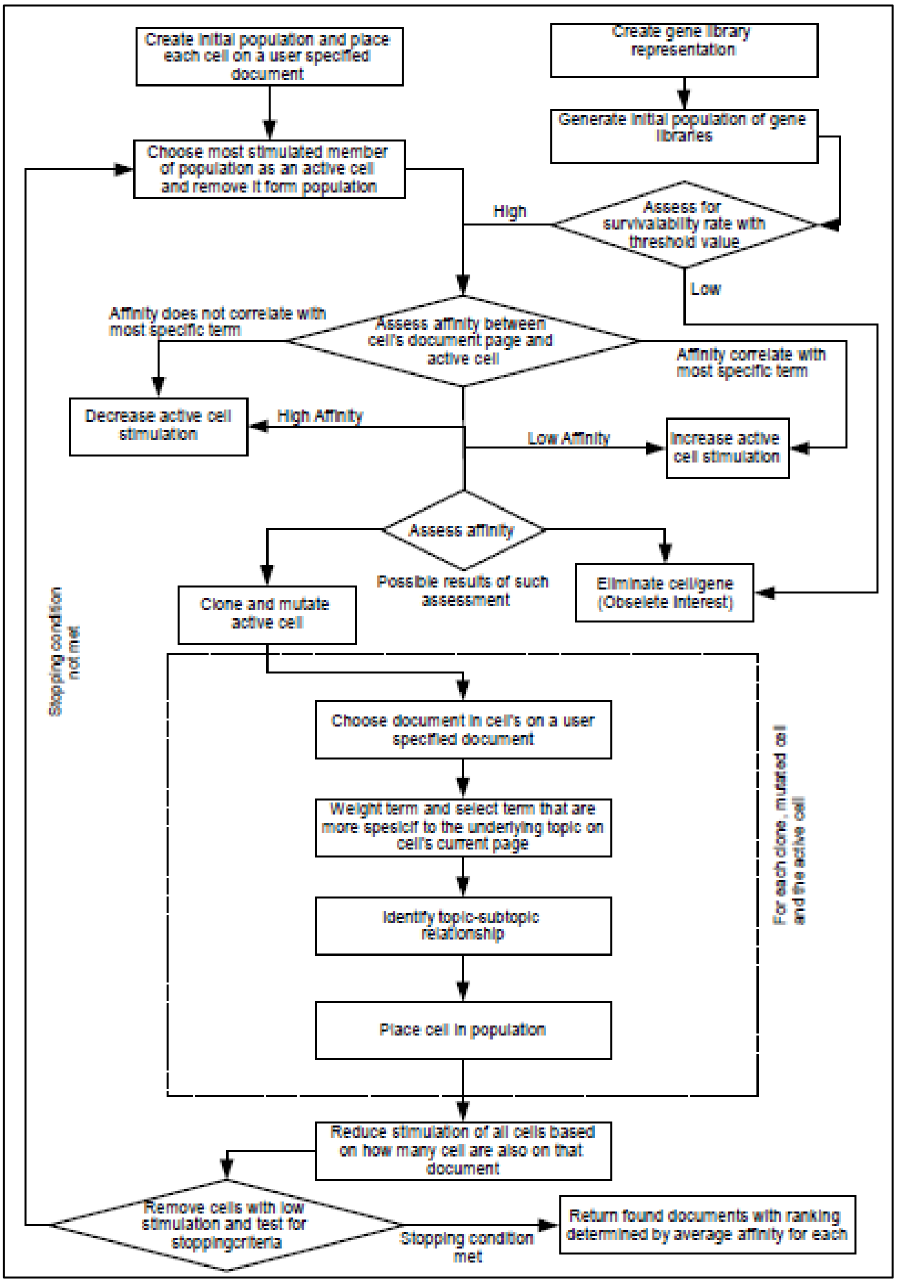
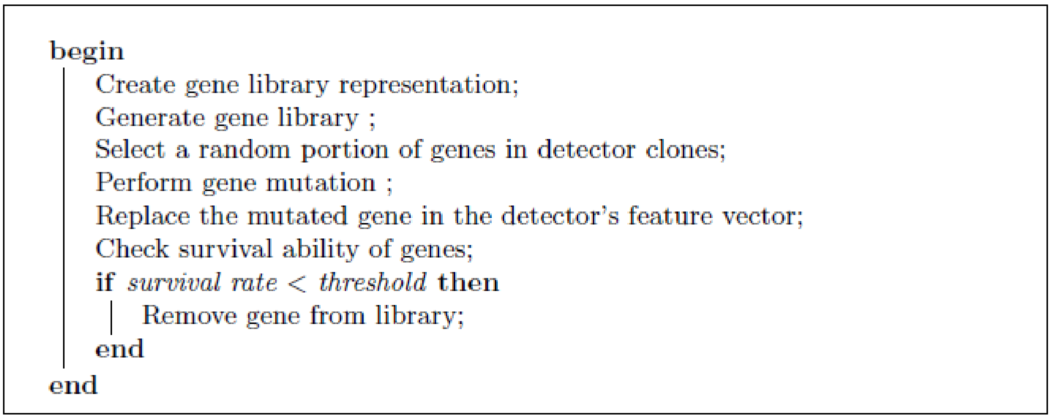
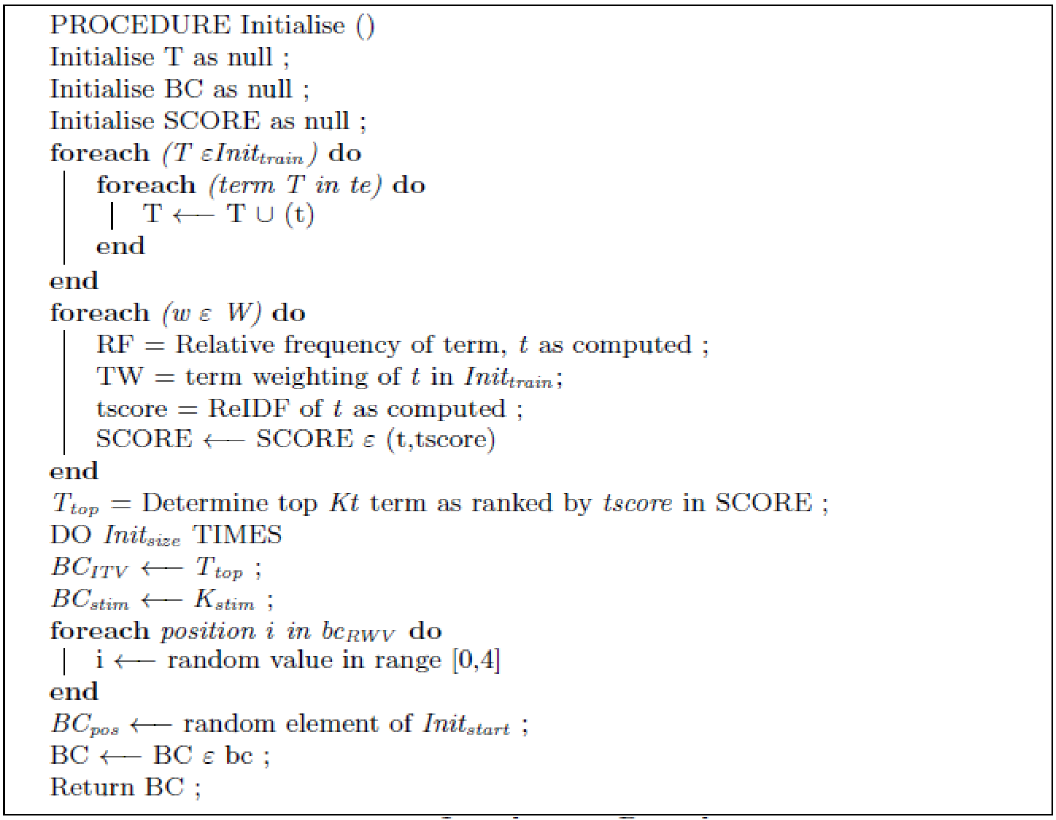

Create initial population and place each cell on a user specified document
Create gene library representation
Generate initial population of gene libraries
Choose most stimulated member of population as an active cell and remove it form population
High
Assess for survivability rate with threshold value
Low
Affinity does not correlate with most specific term
Assess affinity between cell's document page and active cell
Affinity correlate with most specific term
Decrease active cell stimulation
High Affinity
Low Affinity
Increase active cell stimulation
Assess affinity
Possible results of such assessment
Clone and mutate active cell
Eliminate cell/gene (Obselete interest)
Stopping condition not met
Choose document in cell's on a user specified document
Weight term and select term that are more spesicif to the underlying topic on cell's current page
Identify topic-subtopic relationship
Place cell in population
For each clone, mutated cell and the active cell
Reduce stimulation of all cells based on how many cell are also on that document
Remove cells with low stimulation and test for stoppingcriteria
Stopping condition met
Return found documents with ranking determined by average affinity for each

Fig. 1: The ProAdDCS Flow Chart



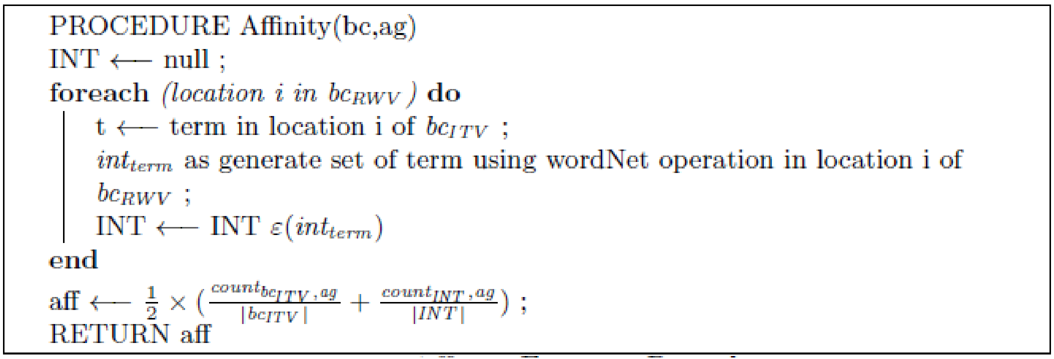
```
begin
    Create gene library representation;
    Generate gene library ;
    Select a random portion of genes in detector clones;
    Perform gene mutation ;
    Replace the mutated gene in the detector's feature vector;
    Check survival ability of genes;
    if $survival\ rate < threshold$ then
        Remove gene from library;
    end
end
```

Algorithm1: Gene Library for Maintaining Diversity in ProAdDCS



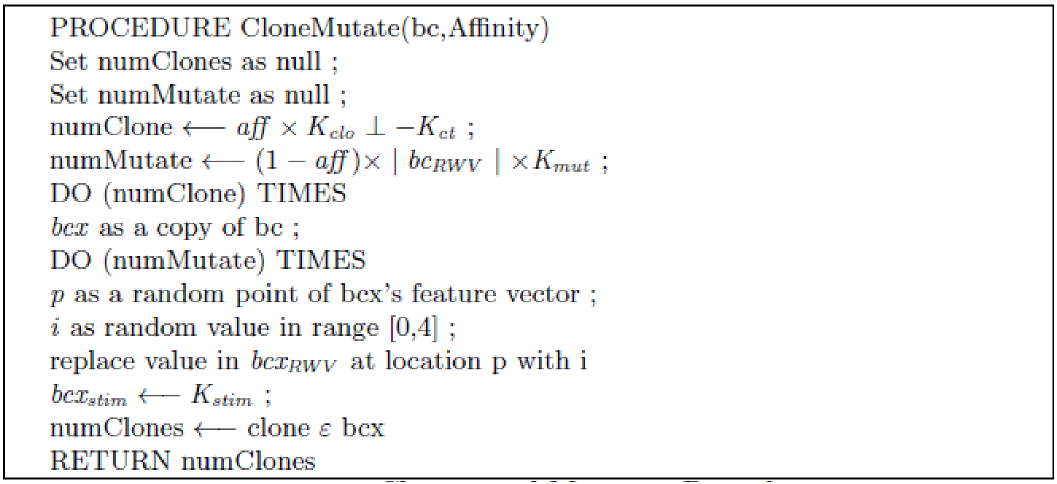
```
PROCEDURE Initialise ()
Initialise T as null ;
Initialise BC as null ;
Initialise SCORE as null ;
foreach $(T\ \varepsilon Init_{train})$ do
    foreach $(term\ T\ in\ te)$ do
        T $\longleftarrow$ T $\cup$ (t)
    end
end
foreach $(w\ \varepsilon\ W)$ do
    RF = Relative frequency of term, $t$ as computed ;
    TW = term weighting of $t$ in $Init_{train}$;
    tscore = ReIDF of $t$ as computed ;
    SCORE $\longleftarrow$ SCORE $\varepsilon$ (t,tscore)
end
$T_{top}$ = Determine top $Kt$ term as ranked by $tscore$ in SCORE ;
DO $Init_{size}$ TIMES
$BC_{ITV} \longleftarrow T_{top}$ ;
$BC_{stim} \longleftarrow K_{stim}$ ;
foreach $position\ i\ in\ bc_{RWV}$ do
    i $\longleftarrow$ random value in range [0,4]
end
$BC_{pos} \longleftarrow$ random element of $Init_{start}$ ;
BC $\longleftarrow$ BC $\varepsilon$ bc ;
Return BC ;
```

Algorithm 2: Initialization Procedure



```
PROCEDURE Affinity(bc,ag)
INT ⟵ null ;
foreach (location $i$ in $bc_{RWV}$) do
    t ⟵ term in location i of $bc_{ITV}$ ;
    $int_{term}$ as generate set of term using wordNet operation in location i of
    $bc_{RWV}$ ;
    INT ⟵ INT $\varepsilon(int_{term})$
end
aff ⟵ $\frac{1}{2} \times (\frac{count_{bc_{ITV},ag}}{|bc_{ITV}|} + \frac{count_{INT,ag}}{|INT|})$ ;
RETURN aff
```

Algorithm 3: Affinity Function Procedure



```
PROCEDURE CloneMutate(bc,Affinity)
Set numClones as null ;
Set numMutate as null ;
numClone ⟵ $aff \times K_{clo} \perp -K_{ct}$ ;
numMutate ⟵ $(1 - aff)\times \mid bc_{RWV} \mid \times K_{mut}$ ;
DO (numClone) TIMES
$bcx$ as a copy of bc ;
DO (numMutate) TIMES
$p$ as a random point of bcx's feature vector ;
$i$ as random value in range [0,4] ;
replace value in $bcx_{RWV}$ at location p with i
$bcx_{stim} \longleftarrow K_{stim}$ ;
numClones ⟵ clone $\varepsilon$ bcx
RETURN numClones
```

Algorithm 4: Cloning and Mutation Procedure

4 Experimenting ProAdDCS in Dynamic User Interests

In this section, ProAdDCS is tested in order to evaluate its performance via real corpus of web document. The profile is tested for their ability to adapt over time in the content of documents. Evaluation is carried based on simulation procedure whereby it involves the use of *virtual* or *synthetic user* which is used to simulate such radical changes. Given a pre-classified collection of documents, a virtual user's current interests are defined by a subset of the classification topics. Training documents that relate to the topics in the subset comprise the positive feedback. Interest changes can then be simulated by modifying this subset. To simulate the loss of interest in a topic, it is removed from the subset. Similarly, the emergence of a new topic of interest can be simulated by adding a new topic to the subset. System can therefore be tested against radical drifts in the topic of interest. The experiment is investigated based on Reuters-21578 Document Collections. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. and include 21578 news stories. The documents have been manually classified according to 135 topic categories, but for these experiments we concentrate only on the 23 topics with at least 100 relevant documents. Usually, in text classification experiments, for each topic category a classifier is first trained using relevant documents from the training set and is subsequently evaluated against the test set. However, for the purpose of this experiment we use the collections in a different way. The documents in Reuters 21578 are ordered according to publication date and their topicality changes accordingly. We exploit this ordering to test the ability of the algorithm to continuously learn, but also to forget. The 23 topic involved in the experiments are ordered according to decreasing size, i.e., the number of documents in the collection that are relevant to a topic.

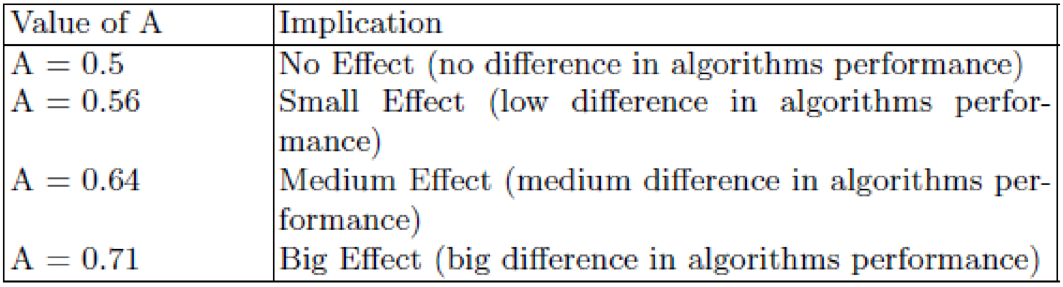
4.1 Experiment Methodology

We followed a methodology proposed in [8]. This methodology was proposed for evaluating the ability of a user profile for continuous adaptation in a dynamically changing environment. We adopt this methodology because we believed that this methodology can be used for testing the ability of adaptive systems, such as AIF, for online learning (and forgetting) in a complex, multi-dimensional and dynamic environment. Furthermore, iterating over the same set of documents and suddenly switching between topics of interest, causes discontinuities. Nevertheless, how the system reacts to these discontinuities is an interesting test for the system's dynamics.

In this experiment, each experimental run starts with an initially empty profile. The profile then evaluates (i.e., assigns a score to) the 21578 documents in order. The empty profile assigns a zero score to documents until it encounters a document relevant to the first of the 23 topics (i.e., earn) and then it is initialized. The initialized profile assigns a score to the remaining documents in the collection and every time it encounters a document that belongs to topic *earn*, it evaluates the document and then uses it as positive feedback and adapts based on it. When all 21578 documents have been evaluated, they are ranked according to decreasing relevance score and the Average Un-interpolated Precision (AUP) for each of the 23 topics is calculated on the ordered list of documents. After the first evaluation period, the process is reinitiated and the profile, which now represents the first of the 23 topics, starts anew to evaluate the document collection. This time, however it uses as positive feedback documents that belong to the second of the 23 topics (i.e., acq). In other words, the profile has to forget the no longer interesting first topic (*earn*) and learn the new topic of interest (*acq*). Once all documents have been evaluated, a new set of AUP values is calculated and the process is repeated for the third topic (*moneyfx*). The experiment finishes once all 23 topics have been used as positive feedback.

For multi-topic experiments, the same process is applied. Multi-topic experiments include the two-topic. In the two-topic experiment, the profile initially learns the first two topics in parallel. The system then forgets the first and learns the second and third topics, and so on. To evaluate the performance of the algorithm, we carried out a statistical analysis based on the non-parametric Mann-Whitney-Wilcoxon or rank-sum test [14] to test whether two algorithms had different distributions (each having a different median), and the Vargha-Delaney A statistics [13] to measure the effect of size between these algorithms. Table 2 presents a list of the A values and the implication on the tested sample size. We used this guideline to assess the effect size.

Table 2: The Vargha-Delaney A Statistics Value and Its Implication on Sample Size [13]



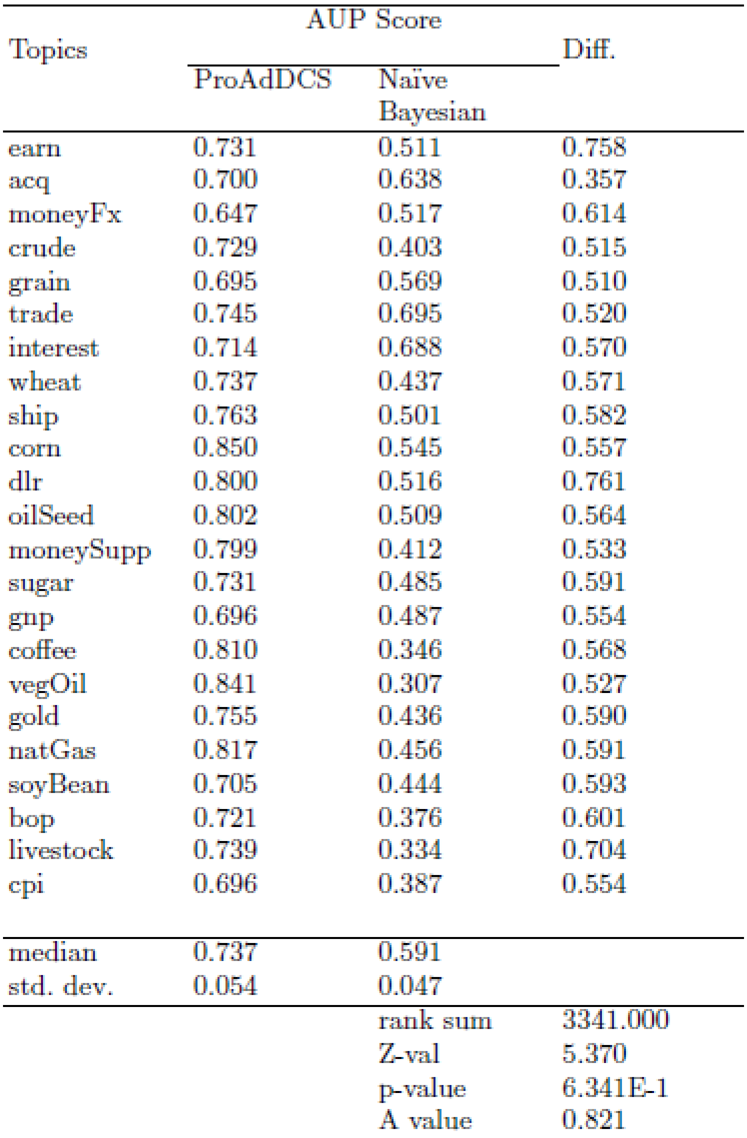
Value of A	Implication
A = 0.5	No Effect (no difference in algorithms performance)
A = 0.56	Small Effect (low difference in algorithms performance)
A = 0.64	Medium Effect (medium difference in algorithms performance)
A = 0.71	Big Effect (big difference in algorithms performance)

4.2 Experiment Result and Discussion

In this section, the results of the experiment are discussed. Table 3 and Table 4 present the complete comparative experiment results for the single-topic and two-topic experiments. The fourth columns present the differences between the ProAdDCS with Naive Bayesian. The combined values of AUP scores for the multiple-topic are calculated based on the aggregate set of relevant documents for all constituent topics. The statistical analysis for the tested algorithms is presented in the last row of the list of the topics. The analysis includes the median, standard deviation and the non-parametric statistical analysis based on the Rank-Sum and the Vargha-Delaney A statistics.

In the single-topic experiment (Table 3), the ProAdDCS profile produces better AUP score than Naive Bayesian approach for the overall topics tested. However, as already discussed, representing a single topic of interest is a relatively simple problem and does not accurately reflect a real situation. In reality, a user is typically interested in more than one topic in parallel. In contrast, in terms of statistical non-parametric analysis, the p-value indicates with 95% confidence that there was a statistically and scientifically significant difference between the ProAdDCS's profile with Naive Bayesian approach. In summary, these samples had a different distribution which indicates that their medians were different. The A value from the Vargha-Delaney A statistics shows that the ProAdDCS's profile

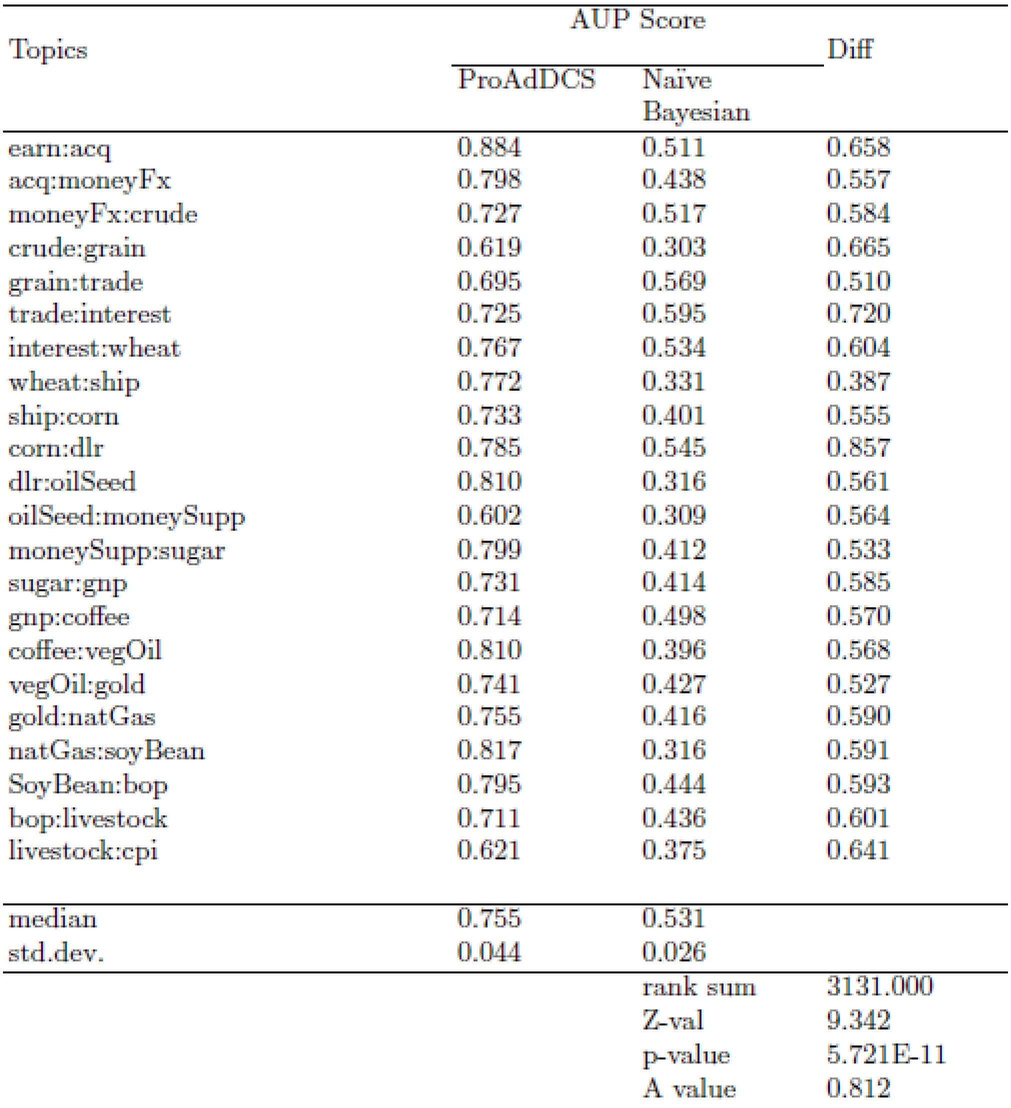
Table 3: Results for Single-Topic Experiments



Topics	AUP Score		Diff.
	ProAdDCS	Naïve Bayesian	
earn	0.731	0.511	0.758
acq	0.700	0.638	0.357
moneyFx	0.647	0.517	0.614
crude	0.729	0.403	0.515
grain	0.695	0.569	0.510
trade	0.745	0.695	0.520
interest	0.714	0.688	0.570
wheat	0.737	0.437	0.571
ship	0.763	0.501	0.582
corn	0.850	0.545	0.557
dlr	0.800	0.516	0.761
oilSeed	0.802	0.509	0.564
moneySupp	0.799	0.412	0.533
sugar	0.731	0.485	0.591
gnp	0.696	0.487	0.554
coffee	0.810	0.346	0.568
vegOil	0.841	0.307	0.527
gold	0.755	0.436	0.590
natGas	0.817	0.456	0.591
soyBean	0.705	0.444	0.593
bop	0.721	0.376	0.601
livestock	0.739	0.334	0.704
cpi	0.696	0.387	0.554
median	0.737	0.591	
std. dev.	0.054	0.047	
		rank sum	3341.000
		Z-val	5.370
		p-value	6.341E-1
		A value	0.821

with the comparative algorithm showed a large effect with the A value above 0.71. This indicates that the performance between the ProAdDCS's profiles with the baseline approach shows a large effect when tested on the single-topic of Reuters 21578 document collections. In addition, results for the two-topic (Table 4) clearly shows that as the complexity of the user's interest increases, the preservation of diversity become increasingly important. The ProAdDCS's profile performs better compared with Naive Bayesian for overall topic combined. Moreover, in all cases, the Mann-Whitney-Wilcoxon or rank-sum test showed that the differences between these algorithms were statistically and scientifically significant. This can be summarized as showing that the tested algorithms had different distributions and showed a large performance effect with the A value above 0.71. This indicate that ProAdDCS's adaptation, which involves the introduction and preservation of diversity in the gene libraries are able to respond to short-term variations and occasional radical changes in the composition of a stream of feedback documents.

Table 4: Results for Two-Topic Experiments



Topics	AUP Score		Diff
	ProAdDCS	Naïve Bayesian	
earn:acq	0.884	0.511	0.658
acq:moneyFx	0.798	0.438	0.557
moneyFx:crude	0.727	0.517	0.584
crude:grain	0.619	0.303	0.665
grain:trade	0.695	0.569	0.510
trade:interest	0.725	0.595	0.720
interest:wheat	0.767	0.534	0.604
wheat:ship	0.772	0.331	0.387
ship:corn	0.733	0.401	0.555
corn:dlr	0.785	0.545	0.857
dlr:oilSeed	0.810	0.316	0.561
oilSeed:moneySupp	0.602	0.309	0.564
moneySupp:sugar	0.799	0.412	0.533
sugar:gnp	0.731	0.414	0.585
gnp:coffee	0.714	0.498	0.570
coffee:vegOil	0.810	0.396	0.568
vegOil:gold	0.741	0.427	0.527
gold:natGas	0.755	0.416	0.590
natGas:soyBean	0.817	0.316	0.591
SoyBean:bop	0.795	0.444	0.593
bop:livestock	0.711	0.436	0.601
livestock:cpi	0.621	0.375	0.641
median	0.755	0.531	
std.dev.	0.044	0.026	
		rank sum	3131.000
		Z-val	9.342
		p-value	5.721E-11
		A value	0.812

6 Conclusion

This paper is concerned with a study into experimenting AIS towards profile adaptation to changes on user interests in AIF. We have argued that the user interests are by nature dynamic where a combination of parameter causes a variety of changes. Frequent changes in the user’s short-term needs contribute to progressive changes in the user’s long term interests and vice versa. The user's interest may shift frequently between different topics or related subtopics. New topics and subtopics of interest emerge and the interest in a certain topic might be lost. To achieve adaptation on single and multi-topic profile subject to variety of changes in the user interests, we have been inspired by biological theories of DCS. DCS can inherently maintain and boost diversity and can dynamically control the size of the immune repertoire by means of selection, cloning, and mutation procedures. Moreover, diversity in the population is enabled by means of the receptor editing process. Here, we suggested that profile adaptation can be developed by incorporating ideas from aspects of DCS with gene libraries to maintain sufficient diversity through transformation of synset relationship based on WordNet. To test our approach to profile adaptation, we have synthesized virtual users based on web document corpus namely the Reuters-21578 document collection. We made the assumption that a user's interest and changes in them are reflected by the feedback that the user provides. On these ground, we have carried an experiment to test the ability of the profile in learning and forgetting topics based on identified corpus. We may argue that the experiment results have been positive. Furthermore, the result on ProAdDCS's profile has also been positive for task comprising the unrelated topics learned in parallel as well as not learned in parallel. Our next concern is to further experiment the algorithm with other algorithms such as the Rocchio's learning algorithm and the self-organize Nootropia's profile model. Comparative performance is important in order to assess whether the works are incremental improvements on the state of the art or evolutions of existing work.

ACKNOWLEDGEMENTS.
The authors would like to thank Advanced Informatics School, Universiti Teknologi Malaysia for the support of the resources. This work is currently funded by a FRGS Grant from Ministry of Higher Education Malaysia (Vot Number:4F266).

References

[1] P. S. Andrews. An Investigation of a Methodology for the Development of Artificial Immune Systems: A Case-Study in Immune Receptor Degeneracy. PhD thesis, Department of Computer Science, University of York, 2009.

[2] N. F. M. Azmi, J. Timmis, and F. Polack. Profile Adaptation in Adaptive Information Filtering: An Immune Inspired Approach. In IEEE Int. Conf. on Soft Computing and Pattern Recognition (SoCPaR), 414 - 419, 2009.

[3] N. F. M. Azmi, J. Timmis, and F. Polack. Towards A Principled Design Of Bioinspired Solutions To Adaptive Information Filtering. In 15th IEEE Int. Conf. on Engineering of Complex Computer Systems (ICECCS), 315 - 316, 2010.

[4] S. Cayzer and J. Smith. Gene Libraries: Coverage, Efficiency and Diversity. In Int. Conf. on Artificial Immune Systems, 136-149, 2006.

[5] S. Cayzer, J. Smith, J. A. R. Marshall, and T. Kovacs. What Have Gene Libraries Done for AIS? In Int. Conf. on Artificial Immune Systems, 86-99, 2005.

[6] Y. Dinga, H. Suna, and K. Hao. A Bio-Inspired Emergent System for Intelligent Web Service Composition and Management. Knowledge-Based Systems, 20(5):457-465, 2006.

[7] K. De Jong. Adaptive system design: A genetic approach. IEEE Transactions on Systems, Man and Cybernetics, 10(9):566-574, 1980.

[8] N. Nanas, M. Vavalis, and L. Kellis. Immune Learning in a Dynamic Information Environment. In 8th Int. Conf. on Arti_cial Immune Systems, 192-205, 2009.

[9] R. Pfeifer, M. Lungarella, and F. Iida. Self-Organization, Embodiment, And Biologically Inspired Robotics. Science, 318(5853):1088-1093, 2007.

[10] S. Stepney, R. Smith, J. Timmis, A. Tyrrell, M. Neal, and A. Hone. Conceptual Frameworks for Artificial Immune Systems. Unconventional Computing, 1(3):315-338, 2006.

[11] D. R. Tauritz and I. G. Sprinkhuizen-Kuyper. Adaptive Information Filtering Algorithms. In Advances in Intelligent Data Analysis, 513-524, 1999.

[12] C. Teuscher, D. Mange, A. Staufier, and G. Tempesti. Bio-Inspired Computing Tissues: Towards Machines That Evolve, Grow, And Learn. Biosystem, 68(2-3):235-244, 2003.

[13] A. Vargha and H. D. Delaney. A Critique and Improvement Of The Common Language Effect Size Statistics Of McGraw And Wong. Journal on Educational and Behavioral Statistics, 25(2):101-132, 2000.

[14] F. Wilcoxon. Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6):80-83, 1945.