Journal of Universal Computer Science, vol. 19, no. 13 (2013), 2013-2033
submitted: 30/10/12, accepted: 28/6/13, appeared: 1/7/13 © J.UCS

An Item based Geo-Recommender System Inspired by Artificial Immune Algorithms

Antonio Cabanas-Abascal
(Computer Science Department, University Carlos III Madrid
Av. Universidad, 30, Leganés, 28911, Madrid, Spain
antonio.cabanas@uc3m.es)

Eduardo García-Machicado
(Computer Science Department, University Carlos III Madrid
Av. Universidad, 30, Leganés, 28911, Madrid, Spain
eduardo.garcia.machicado@uc3m.es)

Lisardo Prieto-González
(Computer Science Department, University Carlos III Madrid
Av. Universidad, 30, Leganés, 28911, Madrid, Spain
lisardo.prieto@uc3m.es)

Antonio de Amescua Seco
(Computer Science Department, University Carlos III Madrid
Av. Universidad, 30, Leganés, 28911, Madrid, Spain
antonio.amescua@uc3m.es)

Abstract: Nowadays, one of the most relevant features provided by in almost every web site is a recommender system. However, they are usually focused on the common characteristics of several items which are shared among the users without taking into account that there are other very important features, such as geo-position. To face this lack of such relevant factors, authors propose the usage of a useful system that will aid in tasks related to pattern detection and fast adaptability to changes: Artificial Immune System. A combination of both systems and the addition of a geographic component will provide a new solution to this problem, which will solve as well these issues as other ones like comparison tasks in big data.

Keywords: Recommender Systems, Artificial Immune Systems, Geo-localization, Item, Big data
Categories: H.1.1, H.1.2, H.3.1, H.3.2, H.3.3, H.3.5

1 Introduction

Along this document authors are going to deal with two main types of technologies which are the base of the proposed study. These technologies are recommender systems and bio-inspired technics. Specifically, Artificial Immune Systems (from now on AIS) used in pattern detection problems. Thus, the aim of this study is to provide a new recommender system based on item and inspired by AIS, which will take advantage of the current technologies applied to field of tourism [Colomo-Palacios et al. 2012].

Nowadays every single recommender system based on item with a relevant number of instances has to face two main problems that can make it not be as accurate as it should be. On one hand there is the challenge of including new important parameters, which provide additional information such as geographic position. It is obvious that as well as people characteristics are strongly influenced by the place they belong to, certain trends or likes may be grouped according to different location of the recommended items. Nevertheless, this geographic component has to be taken into account, as it is always relevant in recommender systems for tourism [Tran and Cohen 2000], and it will only be available in domains where such items can be placed in a geographic position, i.e. point of interest recommendation.

On the other hand there is the challenge of being able to deal with a big amount of data in systems based on collaborative filtering. To execute one of these algorithms it should be necessary to process every structure associated to user's likes, what is very tedious when the system has previously stored thousands of them. Today, there are many solutions which face this problem by using, for instance, clustering algorithms so they group user's likes structures in order to only have to use the representative clusters instead of the whole data.

Nevertheless, it is easy to appreciate that there is no based reason for that grouping beyond the common characteristics they have, so such grouping may be wrong. In this paper a similar clustering is carried out. This clustering will save a lot of time in processing tasks but with a clear base that will also be useful in the extraction of additional information, in terms of trends which are taking place in the different places where the system is deployed.

Thus, this work shows the development of a reliable system capable of handling the recommender system scalability in terms of number of users, as well as providing new techniques for solving recommendation issues which will be inspired by AIS beyond the current work related to Semantic Technologies [Casado-Lumbreras et al., 2012], [García-Crespo et al. 2012], [García-Crespo et al. 2011], [González-Carrasco et al., 2012]. The remaining of the paper is structured as follows: section 2 contains the State of the Art (SoA) about artificial immune systems and recommender systems. Following, section 3 details the proposed solution. Section 4 contains the validation of the proposal. In section 5 the most relevant conclusions relating to design and development of the solution are presented, and finally section 6 details the future lines of work derived from this study.

2 State of the art

In this section SOA of recommender systems and AIS will be summarized, since they are the two main technologies used for the implementation of the final system. In first place the most relevant characteristic of AIS as well as the typical modules implemented in its construction will be explained. In second place, a review of the different recommender systems implementations and their main characteristics will be offered.

2.1 Artificial Immune Systems

"Artificial immune systems can be defined as computational systems inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving" [Nunes and Timmis 2002]. AIS can be included in the set of bio-inspired algorithms o in bio-informatics (neural networks, ant colonies, genetic algorithms…) which were introduced in 80's by Bersin, Farmer, Packard, Perelson and Varela.

Natural Immune Systems have structures divided by layers where there are established different complementary mechanisms one another. In the first level physical barriers such as skin and mucous membranes are found. In the second level there are the biochemical ones: tears, saliva, sweat… and, finally, two more complex systems: innate immunity and acquired immunity. AIS base their implementation in those two structures so they are described below:

Innate immunity is the capability of elimination a limited set of existent antigens, is distributed all over the body and formed by several types of cellules: macrophages, monocytes, polymorphonuclear neutrophils. In addition, innate system collaborates with adaptive system since it is responsible of its activation and control.

Acquired or adaptive immunity allows recognizing and eliminating antigens which had not been faced previously. It also allows remembering this type of antigen in order to easily eliminate it in the future. This is possible thanks to lymphocytes that can be T type or B type. T Cellules can be activated by any other antigen presenter molecule, as a macrophage. When a T cellule is activated it proliferates in order to harden the system. Moreover, B cellules, which are also antigen presenter through MHC complex, collaborate with T cellules which were previously activated so B cellules generate antibodies that will remember the immune response if it happens in the future.

Finally, T lymphocytes recognize and destroy own cellules which are infected by virus. There are different characteristic of the immune systems that make them especially interesting for its use in computation:

- **Learning**: It can be defined as the acquisition of a behavior through the experience. The AIS are able to adapt to different antigens and to eliminate then easier and easier.
- **Memory**: "the ability to remember information, experiences, and people". Antibodies which recognized antigens in the past, last in time.
- **Decentralization**: It takes place thanks to antibodies' autonomy which can carry out their main function without needing any kind of central control.
- **Pattern recognition**: It is one of the AIS' main characteristics and it allows recognizing the similarity grade between an antigen and an antibody.
- **Parallelism**: Due to the antibodies' autonomy, pattern recognition, cloning and the rest of actions can be carried out within the system in a parallel way.
- **Diversity**: Mutation allows increasing the amount of different individuals in the system, as well as in other type of genetic algorithms, so it also increases the number of patterns which are recognized.
- **Self-regulation**: Cellules' time life, as well as their collaboration, makes the system regulate on its own.

Every AIS has a similar structure defined in [Nunes and Timmis 2002] and which is shown in the next image. This universal structure is composed by "a representation of the system components, a set of mechanisms to evaluate every interaction between individuals or between them and the environment and, finally, an adaptation process that guide the system operation. This is, the algorithm itself.

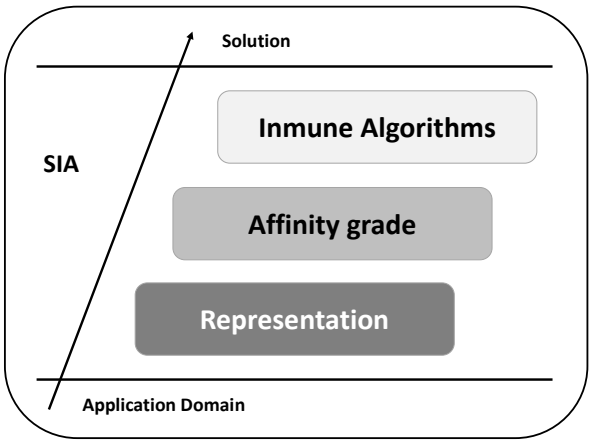

Figure 1: AIS main structure

Representation is the assimilation of domain components which are going to be faced with the main components that a typical SI has. What is an antigen? What is an antibody? What is the knowledge region? Does it exist? How are they stored?

On the other hand, the affinity grade must be analyzed, designed and, finally, it is necessary to develop one or many functions that allow knowing the similarity among two individuals. Affinity function design is very important since it will be responsible for the system proliferation due to the fact that cloning will be directly related to affinity grade. There are many ways to study the affinity, i.e. Euclidean distance or Manhattan distance.

Moreover it is necessary to take into account that, in order to favor affinity and diversity in AIS as well as in other bio-inspired algorithms, mutation operators are used. In nature, such mutation is only observed in B type cells.

Finally, a selection function must be defined which, as it was mentioned before, has to be proportional to affinity. Some of the most common selection functions that can be used for this purpose are:

- **Elitist Selection**: It consists of keeping the best individual or a bigger amount of the best tones in the population.
- **Ranking based selection**: It consists in the assignation of a reproduction or cloning probability depending on the affinity grade of a given individual. This is, its quality.
- **Bi-class selection**: A percent of best and worst individual are kept in the population whereas the others are chosen randomly.

- **Tournaments based selection**: An n individual size set is selected in a random way, and its individuals compete among them to check out which ones have more affinity since they will be kept for the next generation. This process is repeated a certain number of iterations.

In the last module there is the structure which define the global immune algorithm. Three main models can be cited:

- **Bone Marrow models**: used for population AIS. They and are mechanisms that generate the initial data set of immune cellules and/or system cellular receptors. The simplest ones are based in random generation taking into account their own nature.
- **Thymus models**: As well as in real immune systems thymus generates a set of cellules and molecules that allow differentiating between own and foreign cellules. To generate this set of cellules it is essential to carry out two steps, a positive selection which generates a set of T cellules and which are compared to those that belong to the own S set. Then, if any of those cellules overcomes a defined similarity threshold, it is added to available set A. On the other hand in order to carry out the negative selection and to eliminate auto-reactive clones it is also generates a T cellules set and then the affinity with the own S set is compared. If the affinity between T and an own peptide overcomes a defined threshold it is eliminated. If it is smaller than that defined threshold, T cellule is added to available set A.
- **Clonal Selection Models**: In this kind of model, the algorithms used for selecting elements to be cloned, when carrying out the mutation and also in other policies that enrich the algorithm performance is very important. There exist several algorithms, where two of the most relevant ones are *ColnalG* and *Immunos*.

AIS application is wider and wider, what makes it to be found in many environments. One of them is IT Security, which takes advantage of pattern recognition characteristic of this kind of system for virus and anomalies detection [Greensmith et al. 2004]. It can even be applied in Ad-hoc mobile networks protection [Mazhar and Farooq 2007]. AIS have also been used in optimization algorithms as in [Coello and Cruz 2006], where clonal selection is used for solving multi-objective optimization problems in conjunction with Pareto dominance.

On the other hand, there are several classification tasks which use AIS like e-mail classification [Secker et al. 2003], or a generic framework for multi-class problems [Goodman et al. 2002].

AIS have been used in much more fields but, in this document, it is really important those related to recommender systems where these kinds of techniques have been most relevant. Worth noting a web page recommender which does not use mutation [Morrison and Aickelin 2002] and a film recommender [Cayzer and Aickelin 2005] where their evaluation results are really interesting for this work, so they will be presented in following sections.

2.2 Recommender systems

There are many technologies and systems which have appeared since the emergence of the Internet due to the fact that computer networks allows to information regarding groups with the same likes to get over geographic barriers. In addition, the emergence

of Web 2.0 has contributed to collaboration among users, leading to huge amount of data generation in the net. This is probed by a study carried out by IDC1 (International Data Corporation) which estimates that in 2015 the whole digital store will be of 8 Zettabytes what is, approximately, 8·1012GB. This reveals the great Internet expansion as well as the information overload problem.

So, due to this humongous amount of information, it is necessary to rely on technologies which are capable of filtering available data and allow the search for valuable information [Morrison and Aickelin 2002]. Thus, online users have the need of having tools which help them to face the huge amount of data available in the World-Wide-Web. Recommender systems have demonstrated that they are a good solution for overload problems by providing more dynamics and personalized searching services [O'Donovan and Smyth 2005].

One of the fields where recommender systems (RSS from now on) have been widely applied is in e-commerce area. It is so that there are authors that RSS as a part of this new way of trading on the Internet: "Recommendation systems are intelligent e-commerce applications that help users in their information searching tasks by offering them personalized recommendations in their interactions with the system" [Adomavicius and Tuzhilin 2005]. RSS inclusion in this environment is logical since products or services recommendation for users is intrinsically related with e-commerce. However, it is possible to find this type of system in other nonprofit fields such as news recommendation based on user profile [Phelan et al. 2009] or social network contacts recommendation [Hannon et al. 2010].

First recommender systems work come from 90's. These systems were them called filtering system and their purpose was selecting those pieces of news which were potentially interesting for a user [Kim et al. 2010]. Once studied the main fundaments of this technology, there are many works which classify RSS as social filtering or collaboration systems, recommended systems based on content and, finally hybrid systems [Pazzani and Billsus 2007] which are a combination of the previous ones. [Barragáns et al 2010]

2.3 Content based recommender systems

Recommender systems based on content analyze those characteristics from the elements to identify which ones can have a special interest for the user [Pazzani and Billsus 2007]. In this kind of systems, the utility function $u(c, s)$ for element s and user c estimates the utility of the assignation $u(c, s_i)$ for user c of the elements $s_i \in S$ that are similar to s [Adomavicius and Tuzhilin 2005]. Utility is the adequacy grade of the element to the user based on the similarity with other items.

In RS based on content, there are three common tasks in every implementation. First of all, it is necessary to create a formal representation of the elements which are going to be recommended: films, books, news, o users to follow. Once the formalization has been done in order to store it in an information system, it is necessary to create a user profile (UP).

This profile is a fact base generated in an explicit or implicit way where data which represent users' likes is stored. The creation a pattern which is capable of extracting this kind of information will be the last objective within a RS. Such patter is known as user model (UM). It is a part of UP and can be created by using machine learning techniques.

In systems based on content, every item is represented by a characteristics vector. The characteristic are constituted by a numeric or nominal value which represents a field of the element such as color, price, etc. [Debnath et al. 2008].

User profile generation is another important part within a RS based on content. UP must store relevant information which allows carrying out an abstraction from the likes, needs and desires from the client to build the user model. User profile can be divided in two types of information: preferences and iterations [Pazzani and Billsus 2007].

- Preferences describe the type of the element the user is interested in, and can be automatically generated.
- On the other hand, an historic dataset of the interactions between the user and the system is also stored.

The last step in the UP generation is the creation of a user preferences model, which can find the probability of user c to be interested in item s. These models can be automatically generated by using learning machine techniques with a historic dataset as a training dataset. Pazzani and Daniel highlight the following models as the most important ones: decision trees [Bala and Agrawal 2009] and rule induction [Banfield et al. 2007], nearest neighbor, feedback relevance and Rocchio algorithm, linear classifiers and probabilistic methods and Bayesian classifiers.

Finally, it is important to highlight the two main problems of this kind of RS:

- Item's digital representation: Usually, users choose a product taking into account a lot of information which is not completely stored in the model since it would suppose a huge amount of data. This problem can be solved by implementing algorithms which automatically structure the dataset although its complexity will always depend on the domain [Hannon et al. 2010].
- The second one is called over-specialization, very common in machine learning techniques, and it happens when the algorithm is too much adapted to the training dataset so a new data which is not included in such dataset will not be processed correctly [Lathia 2009]. Many solutions inspired by biologic computation have been proposed to solve this problem.

2.4 Collaborative filtering based recommender systems

Recommender system based on collaborative filtering try to group users who have a common profile, taking into account the valuations and opinions from those users about the different elements and obviating their content. Thus, the concept "neighborhood" arises. It can be defined in a formal way as follows: Utility function $u(c, s)$ of an element s for a user c is estimated by basing it on the utility $u(c_j, s)$ for the element s of the users $c_j \in C$ who are similar to user c [Adomavicius and Tuzhilin 2005].

Nowadays, collaborative filtering RS are the most used recommender systems and probably they are also the most analyzed by the scientific community. Along all those approaches, the best classification may be the one shown in Figure 2. This is, collaborative filtering RS are divided in model based and memory based systems based on (which are also subdivided in user based and item based systems) [Lathia 2009].

Candiller (2009) also offers a formal representation for these systems including a set of qualifications R unlike Adomaviius y Tuzhilin (2005). The complete formalization is: "U is a set of N users, I is a set of M items and R a set of qualifications r_{ui} for users $u \in U$ of an item $i \in I$. $S_u \subseteq I$ for the set of items that user u has valuated. Thus, the main objective for all these different approaches is the estimation of the user a valuation p_{ai} for an item i."

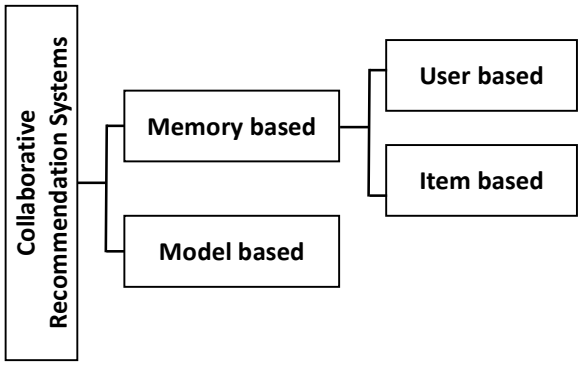

Figure 2: Collaborative filtering recommender systems classification

2.5 Memory based recommender systems

Memory based filtering systems are one of the most important methods in the recommendation generation [Drachsler et al. 2007]. They are known as memory based since they suppose the users who have likes in common in the past to keep those similarities. Within this kind of systems is possible to find those focused on user and those which give more importance to recommended elements.

On one hand, systems based on user try to calculate the valuation that a user would give to an element taking into account the valuations that other similar users gave to such element. These types of system define similarity measures to determine the valuation given to an item by a user being more relevant those users who are more similar. The estimation of that measure can be carried out by calculating the mean of every valuation. Thus, it should be considered that entire neighborhood is at the same distance each other. It is defined as [Adomavicius and Tuzhilin 2005]:

$$p_{ai} = k \sum_{a' \in A} sim(a, a') \cdot p_{a'i}$$

So p_{ai} is estimated by calculating the mean of every known valuation multiplied by the similarity between the new user and the users who have previously evaluate the item. Where "k" is a normalization factor.

$$k = {}^{1}/_{\sum_{a' \in A} sim(a, a')}$$

Similarity measure between two users is a heuristic component in these systems. A good approach of this measure is crucial to achieve an appropriate performance. Depending on the problem, it can be defined in different ways, i.e. one of them is based on the rest of the valuated items, as several author purposes.

Thus, these item subsystems based systems need to define a similarity measure between elements in order to be able to create items neighborhoods, so the estimation will be only done by taking into account the likes belonging to the nearest neighbors.

2.6 Model based recommender systems

Finally, model based systems are able to get a faster prediction performance than the previous ones by generating probabilistic off-line data models to forecast as soon as possible on-line valuations [Candiller et al. 2009].

One of the first approach is based on creating an off-line way a model which represents the users grouping in clusters by using different techniques (k-mean, Bayesians models, SVM…), avoiding the similarity estimation with every user, and doing it by only using his neighbors.

There exist two main problems in collaborative filtering systems when making the recommendation. The first one is related to the system when it has no enough data from the user to compare it to other users. The second one is related to new items which has no previously valuations by any other user so the system cannot establish similarities between users likes. The solution is the usage of hybrid systems which are explained in the following section.

2.7 Hybrid recommender systems

Like in other technologies where there are many ways to achieve the same objective, hybrid systems try to solve problems by a new model which integrates the best characteristics of current ones. In recommender systems, collaborative ones and based on content ones are going to be combined in order to eliminate the previously named weaknesses [Hannon et al. 2010].

There are several approaches which have been included in the literature. First studies simply carried out both recommendation types in a simultaneous way and then they offered both results [Hannon et al. 2010]. However, in other studies, when knowledge base is small, content based systems are used in first place and latter collaborative ones. Other hybrid systems get to combine both types of recommendations through an intelligent interface [Tran and Cohen 2000]. Other models aim at using item formalization and so know characteristics of them to enlarge chosen item in the collaborative filtering.

Basing on the same idea, other authors like Wang et al., (2006) try to compare users not only by common valued items but also by those items which are similar. All these works and other can be classified by the proposed specification in [Burke 2002]. In first place we find weighted systems where valuation or vote predictions techniques are combined to produce one qualification, as in. Commutation systems change the recommendation technique depending on the situation [Tran and Cohen 2000]. Mix systems use different and finally they offer the obtained results [Smyth and Cotter 2000]. Other types of system are them which combine different characteristics from different recommendation algorithms into a single algorithm.

Cascade systems are also very relevant since they use more than one algorithm in a sequential way where the output of one of them is the input of another, so they refine the results more and more. Finally there exist meta-level systems which use the learned model instead of reusing the characteristics [Schwab et al. 2001].

3 Solution

3.1 Antigen and antibody geo-representation with PostGIS

As it was explained in the SoA section, for the development of AIS it is necessary to define a representation for antigens as well as antibodies. In this point it is going to be shown how it has been carried out and why it has been done. The most interesting characteristic in the representation is the usage of a spatial database, *PostGIS*, as in other current research [Yan and Wang 2012]. This decision arises from the need of adding a geographic position to antigens and antibodies, so it is needed a system which ease spatial queries that require the use of areas, distances, etc. I.e. get those antibodies that are closer than a "x" distance from a given antigen.

Before the physic data model description, it is important to highlight the logical representation of an antibody and an antigen (fundamental parts in our algorithm) as well as the explanation of decisions taken.

Logic representation depends on the type of recommender system to develop so the different individuals to store could be oriented to a collaborative filtering, profile based or content based. Due to the found dataset, as it will be described in the following section, a collaborative filtering was forced to be carried out, leaving the content based representation for future works. It is necessary to highlight that content based representation could enrich the algorithm thanks to generalization over the individuals, what will make them specific localization independent.

An antibody, as well as an antigen, is represented by a chain composed by touristic sites identifiers and their valuations as follows:

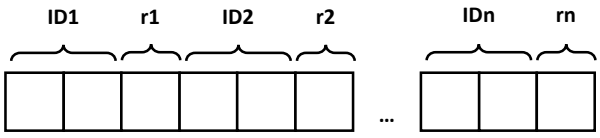

Figure 3: Antigen and antibody representation

Where rn represents the previously placed POI valuation and it will allow the comparison between an antigen and an antibody, taking into account which valuations do they have in common.

Moreover, physic data model is used for storing this information in a permanent way and it allows to generate spatial queries. This model is shown in the diagram below.

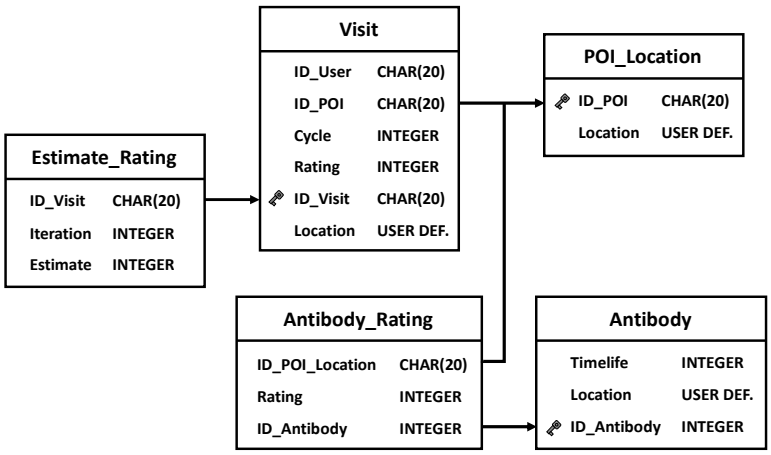

Figure 4: Data model in spatial database

3.2 Building the recommender system

In this section the recommender system description is presented from the point of view of the common architecture of an Artificial Immune System. This architecture, as it was discussed in the SoA, is composed by three general modules, representation, affinity grade and immune algorithms. However, the system has only been described in terms of implementation without taking into account its algorithmic foundation. That is why the following paragraphs will explain in a detailed way how every part of the system has been carried out.

3.3 Affinity estimation

One of the most relevant part what was explained in the state of art was the affinity grade, since it is going to determine how fast the antibody proliferation will be in the system. Although there exist several classic estimation techniques, such Euclidean distance or Manhattan distance, in this problem will be needed a little bit more complex affinity function due to the antigen and antibody representation.

In common Artificial Immune Systems, which are closer to the classic design, the antibody and antigen representation can be considered as a list composed by 1's and 0's. This is, a binary number list without any additional element. Nevertheless, in this domain, can be very difficult to define a binary representation of every visited place as well as its valuation without losing a lot of relevant information or getting it in a compact way.

This is why the codification has been defined as it was explained in the previous section, looking for other methods to calculate the similarity taking into account that, In future works, there will be possible to use other different algorithms.

To carry out such similarity estimation, Pearson's correlation coefficient applied in the comparison between users' valuations to different element (films in this case) used by [Cayzer and Aickelin 2005] was used. This coefficient measures the linear

correlation between two random variables independently from their measurement scale. Such measure is calculated dividing both variables' covariance by the product of their standard deviation. The formula what describes it is:

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

Where:

- **"ui":** Represents "u" user's valuation for "i" hotel
- **"ū" :** Represents mean valuation for "u" user
- **"vi":** Represents "v" antibody's valuation for "i" hotel
- **"v" :** Represents mean valuation for "v" antibody
- **"n" :** Represents the number of hotels that "u" user and "v" antibody have in common

So "r" value will be between -1 and 1, and its meaning is:

- **"r = -1":** Correlation between both variables is inverse.
- **"-1 < r < 0":** There is some type of inverse correlation between both variables.
- **"r = 0":** There is no lineal correlation between both variables.
- **"0 < r <1":** There exists some type of correlation between both variables.
- **"r = 1":** There is a perfect linear correlation between both variables.

In the context of the problem, it is essential to understand this correlation factor since it will determine what antibodies will be used for the recommendation estimation. Finally, we have to pay attention to several special cases which will modify in a very meaningful way the value of Pearson' correlation. So it is necessary to be careful when defining them because if not, the algorithm convergence could be affected. These cases are:

- **No Overlap:** It takes place when in the comparison between an antibody and an antigen, they have no valuations in common. Thus, the algorithm returned value is 0, meaning that there is no correlation between both elements.
- **Zero Variance:** It appears when the denominator is 0, so any of two standard deviations are 0. The established value for the Pearson measure is also 0 since it was the value which offered best results in the paper this correlation measure was based on.
- **Overlap penalty:** It is a penalization value which affects the final Pearson measure value in terms of how many valuations both particles have in common. Its main function is avoiding the proliferation of those antibodies which have a few number of valuations in common since it is easier to be more similar to an antibody with a few number of valuations than to another antibody with a big number of valuations. It is necessary to say that in this work this value has not been taken into account since it delayed the algorithm convergence in the beginning of it execution.

3.4 Recommendation estimation

The objective of this section is to explain the algorithm inner working as well as the first decisions in terms of parameters values. Despite the objective of providing a description as close as possible to the original AIS one, those behaviors that just belong to this problem will be considered.

Thus, given an initial antibodies population randomly generated and a set of 239 antigens which have carried out at least three visits, the algorithm will get the antibodies to be adapted to the antigens likes for every algorithm cycle.

It is worth mentioning that as well as in other kind of algorithm, the convergence is reached after a number of iteration of a single cycle. In this problem, such cycle is composed by the following points:

- The following antigen is selected from the list of antigens which this system has to be adapted to.
- For every valuation of such antigen:
    - It is deleted and stored in an auxiliary variable.
    - The set of nearest antibodies to the antigen is obtained from the spatial database.
    - For every antibody belonging the set composed by the nearest ones, the Pearson measure is calculated between such antibody and the current antigen.
    - The most similar antibodies (from the nearest ones) are used for calculating the recommendation value.
    - The difference between the estimated and the real value is stored.
    - The life of those antibodies which were present in the antigen action range but were not used for calculating the recommendation is decreased.
    - If the recommendation error is lower than a fixed value (which means that the recommendation is accurate), the antibodies used for calculating the recommendation are replicated. Although they are replicated, the location of the new antibodies is slightly changed as well as the valuation of a hotel.
    - The deleted valuation is restored.
    - Before executing the basic cycle, the population size is checked in order to ensure that its size is not lower than a fixed value.
- The root mean square is calculated in order to have control of the algorithm convergence.

In order to describe every function which was used for the algorithm implementation in a more detailed way, the table below has been include. It merges all those details.

As it can be observed in the table, along the algorithm execution, no antibody is deleted from the database although its life is 0. However, the database, in “nearest ones” query will only return those antibodies whose life is higher than 0. This decision was taken to have a control about the generated antibodies that will let us to carry out techniques such as recovering antibodies that in the past were successful and could be beneficial for the antibodies population.

FUNCTION	DESCRIPTION	RELEVANT PAREAMETERS
RETRIEVE NEAREST ONES	Function which given and antigen with a location returns every antibody which belongs a defined influence range.	InfluenceRange = 100
CALCULATE MOST SIMILAR ONES	Obtain the set of "n" most similar antibodies to an antigen.	AntibodySet = nearest Antigen = current n = 5
CALCULATE RECOMMENDATIO N	(*) It is explained in the following sections	
DRECREASE LIFE	Antibody's life is decreased in an established amount	Decreasing value = 1
REPLICATE ANTIBODY	Antibody is replicated and mutated so a new one is created with the same characteristics but with a location and valuations slightly changed.	Location change = 5 Valuation change = 2 Error to be replicated = 0.5
COMPLETE POPULATION	If the antibody population size is smaller than a given value, it is regenerated by creating new antibodies.	Population size = 1000

Table 1: Relevant functions

3.5 Algorithm

Despite AIS main objective is not recommendation, in this problem it is, so this section is aimed to explain every mechanism used for calculating the estimated valuation for a given antigen.

Once again, the work done by [Cayzer and Aickelin 2005] must be referenced, since the estimation method for the valuation of the point of interest will be based on their way to estimate the valuation for their set of movies. The used formula is:

$$p_i = \bar{u} + \frac{\sum_{v \in \mathbb{N}} r_{uv}(v_i - \bar{v})}{\sum_{v \in \mathbb{N}} r_{uv}}$$

Where:

- **"pi":** Estimated valuation for hotel i
- **"ū":** Mean of antigen's "u" valuations
- **"ruv":** Pearson's correlation measure between the antibody and the antigen.
- **"vi":** Valuation given by antibody "v" to hotel "i"
- **"v":** Mean of antibody's "v" valuations
- **"N":** Set of antibodies that take part in the recommendation.

As in the previous section, there are a set of special cases that make the Pearson measure fail. In the valuation estimation the problem arises when the summation in the denominator is 0. In this case, it has been decided to set the estimation of valuation as the mean of all antigens' valuations since it is considered to be the most accurate measure for it.

However it is also true that such estimation will not be as accurate as it should be, so in the future cycles of the algorithm, other antibodies will be able to provide better measures to make the current ones disappear. Thus, an acceptable estimation for the first cycles will be gotten, although it will disappear when more accurate ones will be obtained.

4 Validation

4.1 Dataset description

Touristic places domain is pretty common in big systems (Google Places, FourSquare, TripAdvisor…), however, datasets managed by these platforms are not completely available. In this paper, the selected dataset merges information about the visits carried out by tourists to several Irish hotels with an amount of 27575 visits.

The distribution of tourists as well as their visits is shown in the graphic below:

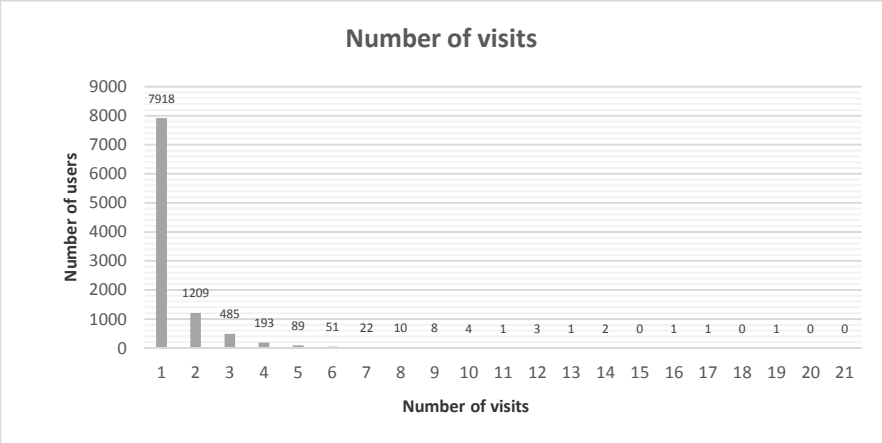

Figure 4: Number of visits

This dataset is the root of the antigen generation for the previously explained physic model. However, locations are not available in the original dataset so they must be generated as well as the initial set of antibodies. The table below shows a summary of tasks for this dataset preparation.

Name	**Input**	**Output**
Description		
Clean DB	*boolean experiments*	*boolean*
It deletes all the DB content in order to reboot the algorithm. There is an option to keep all the information from the previous experiments		
Load hotels file	*Point top_rigth*	*boolean*
It loads the data file into antigens with a random geographic component which will be placed in a location within the limit zone.		
Generate antibodies	*int life* *int number* *int visits per antibody* *Point top_rigth*	*boolean*
It generates the given quantity of randomly located antibodies with a certain number of visits per each one, since visits are also random. Initial life is defined by the parameter "life" and the location restriction is fixed by the furthest point where a particle can be placed within the geographic region.		

Table 2: Database filler

4.2 Chosen parameters

Once the recommender system has been implemented, a test has been carried out with the following selected parameters:

PARAMETER	**VALUE**
Distance to nearest ones	100
Number of similar elements to estimate recommendation	5
Minimum population size	1000
Antibody's life decreasing value	1
Distance in mutation mobility	5
Change grade in valuation mutation	2
Antibody's life	40
Valuations per antibody	20
Default value for nooverlapdefault in Pearson measure	0
Default value for nooverlapdefault in ZeroVarianceDefault	0
Default value for overlap penalization	There is no penalty
Recommendation value when it can be generated	Antigen's valuations average
Number of antigens	239
Number of cycles	100

Table 3: Parameter configuration

Although some parameters are randomly fixed, those values are established to ensure a constant population size. Thus, the global amount of generated antibodies after 100 cycles is 100.000 with a partial size between 1100 and 1200 antibodies alive in every cycle what means that their creation and destruction is balanced.

4.3 Results

Obtained results show the root mean square error of every prediction for every antigen's valuation (y axis) in every cycle (x axis).

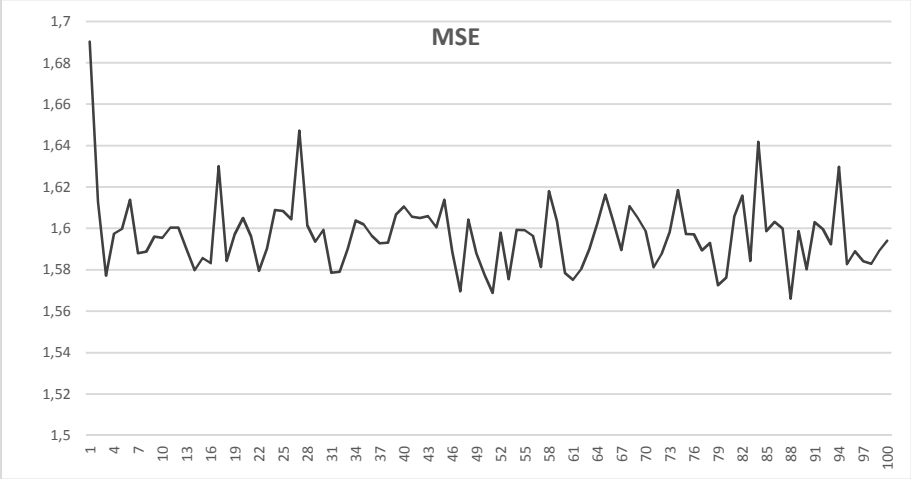

Figure 5: Prediction error

As it can be observed, the algorithm presents convergence since it gets a reduction of the root mean square error. However, the evolution is not as fast as it was supposed to be, so maybe several improvements to the algorithm or making a different parameter configuration could result in better solutions.

4.4 Precision, recall and F1 measures

In order to evaluate the goodness of the algorithm, precision (positive predictive value), recall (sensitivity) and F1 measures (harmonic mean of evenly weighted precision and recall values) have been calculated using results from several executions.

Precision represents the fraction of retrieved elements that are relevant. In this case, there will be considered relevant elements those predictions which has an error rate below 0.5, 1, 1.5 and 2 units. Applying the precision formula:

$$precision = \frac{|\{relevant\ elems.\} \cap \{retrieved\ elems.\}|}{|\{retrieved\ elems.\}|}$$

In simulations of 48 cycles, with 1496 predictions results are:

Error	0,5	1	1,5	2
Avg. precision	0,272	0,512	0,747	0,864

Table 4: Average precision for different relevant instances

Recall represents the fraction of relevant retrieved elements. In this case, the relevant elements are determined by different chosen error thresholds, so relevant elements will be always in the retrieved elements, giving recall values of 1 according to the recall formula:

$$recall = \frac{|\{relevant\ elems.\} \cap \{retrieved\ elems.\}|}{|\{relevant\ elems.\}|}$$

Error	0,5	1	1,5	2
Avg. recall	1	1	1	1

Table 5: Average recall for different relevant instances

Finally, F1 score for the obtained values can be calculated by this formula:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Error	0,5	1	1,5	2
Avg. recall	0,429	0,678	0,855	0,927

Table 6: Average recall for different relevant instances

5 Conclusions

In this section a set of conclusions extracted from the implementation of this recommender system is presented. However it must be taken into account that this paper shows the initial version of the system implementation, which will need several adjusts and improvements for its implementation in a production system.

- Using AIS in combination with collaborative filtering to solve problems is strongly affected by pattern detection, so recommendation based on similarities between users is a good choice.
- By combining both solutions, a recommender system which takes advantage of their strengths has been developed and implemented.
- Due to the fact that this problem has a big amount of antibodies, it is necessary to add any mechanism which allows the selection of those which are optimal for the recommendation estimation, avoiding an excessive computational load that could make the algorithm non-viable.
- The inclusion of a geographic component is crucial to enhance system's performance. On one hand it is going to allow a more efficient data

processing and, on the other hand, it is going to reach AIS much closer to a real one where the particle's location is also very important.

- Since training set is not dynamic, if there are some antibodies which will be never under the influence range of an antigen, they are not going to be eliminated. For system training this is harmful, however, in a real system they should not be deleted as there could be having zones with no antibodies what would mean that any new antigen in such zone could not be recommended.
- Mutation is crucial since it is going to allow the generation of new antibodies providing more diversity.

Finally, it is worth mentioning that although the system performance is not as good as it should be, the main objective has been achieved: Demonstrate that a recommender system based on artificial immune system, which take advantage of the main characteristics from collaborative filtering with a geographic component is viable and it works.

6 Future work

Along the algorithm description some alternatives to the implementation have been commented. Future work will consist of adjusting the algorithm in order to improve its convergence and its efficiency. Additionally, this work will be focused on making the system independent from the dataset. These will be the main actions to take:

- System flexibility improvement: Redesign the system focusing on modularity to be able to adapt it to any dataset with a geographic component.
- Optimum parameters estimation: Calculate the most suitable parameters for the algorithm convergence through any existing technique like Taguchi tables or genetic algorithms.
- Antibody recovery: Mechanism to recover those antibodies which in the past were successful but died.
- Mutation addition: Addition of new mutation operators which modify not only their valuation and their location but also their ID of valuated places.
- New affinity grades: Addition of new formulas to estimate the affinity grade between antibodies and antigens. Thus it will be possible to compare them with Pearson measure.
- New recommendation estimation: Addition of new operator for the recommendation estimation which is based on Pearson measure.
- Constant decrease of antibody's life: Decrease antibody's life, although it does not take place in the recommendation or in the influence area (nearest antibodies), to ensure the removal of those antibodies which are not used in the system owing to antigens are never close enough to them.
- New dataset: Used dataset has no real locations but it was generated randomly. Thus, although it was possible to prove that algorithm can adapt its antibodies set, it will be more relevant if such dataset has real locations.
- New recommendation type: It should be desirable providing, as well as the best places to visit, an itinerary for a given zone as in [Gavalas et al. 2012].

- Generation of additional information: Study the composition of antibody set and extract additional information such as the economic status [Quan et al. 2011] or likes belonging to the different tourists.

References

[Adomavicius and Tuzhilin 2005]. Adomavicius and Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions"; IEEE Transactions on Knowledge and Data Engineering. 17(6), 734 – 749.

[Banfield et al. 2007] Banfield, Hall, Bowyer, and Kegelmeyer. "A Comparison of Decision Tree Ensemble Creation Techniques"; Pattern Analysis and Machine Intelligence, 29(1), 173 - 180.

[Barragáns et al 2010] A.B. Barragáns-Martínez, E. Costa-Montenegro, J.C. Burguillo, M. Rey-López, E.A. Mikic-Fonte, A. Peleteiro. ,"A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition"; Information Sciences, 180 (22) , pp. 4290 – 4311

[Burke 2002] Burke, R. "Integrating Knowledge-Based and Collaborative-Filtering Recommender Systems"; Artificial Intelligence for Electronic Commerce: Papers from the AAAI Workshop, (pp. 69 - 72)

[Casado-Lumbreras et al., 2012] Casado-Lumbreras, C., Rodríguez-González, A., Álvarez-Rodríguez, J.M., & Colomo-Palacios, R. (2012). PsyDis: Towards a diagnosis support system for psychological disorders. Expert systems with applications, 39(13), 11391-11403.

[Cayzer and Aickelin 2005] Cayzer, S., and Aickelin, U. "A Recommender System based on Idiotypic Artificial Immune Networks"; Journal of Mathematical Modelling and Algorithms, 4(2), 181 - 198.

[Coello and Cruz 2006] Coello, C. and Cruz, N. "Solving Multiobjective Optimization Problems Using an Artificial Immune System"; Genetic Programming and Evolvable Machines, 6(2), 163 - 190.

[Colomo-Palacios et al. 2012] Colomo-Palacios, R., Rodríguez-González, A., Cabanas-Abascal, A. and Fernández-González, J. "Post-via: After Visit Tourist Services Enabled by Semantics"; In On the Move to Meaningful Internet Systems: OTM 2012 Workshops (pp. 183 - 193). Springer Berlin Heidelberg.

[Drachsler et al. 2007] Drachsler, H., Hummel, H., & Koper, R.. "Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning"; Proceedings of Workshop on Social Information Retrieval for Technology-Enhanced Learning (SIRTEL'07) at the EC-TEL conference. September 17 - 20, Crete, Greece.

[García-Crespo et al. 2012] García-Crespo, A., López-Cuadrado, J.L., González-Carrasco, I., Colomo-Palacios, R., & Ruiz-Mezcua, B. (2012). SINVLIO: Using Semantics and Fuzzy Logic to provide individual investment portfolio recommendations. Knowledge-Based Systems, 27(1), 103-118.

[García-Crespo et al. 2011] García-Crespo, Á., López-Cuadrado, J., Colomo-Palacios, R., González-Carrasco, I. and Ruiz-Mezcua, B. "Sem-Fit: A semantic based expert system to provide recommendations in the tourism domain"; Expert Systems with Applications, 38(10), 13310 - 13319.

[Gavalas et al 2012] Gavalas, D., Kenteris, M., Konstantopoulos, C. and Pantziou, G. "Web application for recommending personalised mobile tourist routes"; Software, IET, 6(4), 313 - 322.

[González-Carrasco et al 2012] González-Carrasco, I., Colomo-Palacios, R., López-Cuadrado, J.L., García-Crespo, A., & Ruiz-Mezcua, B. (2012). PB-Advisor: a private banking multi-investment portfolio advisor. Information Sciences, 206, 63-82.

[Goodman et al 2002] Goodman, D., Boggess, L. and Watkins, A. "Artificial immune system classification of multiple-class problems"; Proceedings of the artificial neural networks in engineering ANNIE, (pp. 179 - 183)

[Greensmith et al. 2004] J. Greensmith, U. Aickelin, J. Twycross. "Detecting danger: Applying a novel immunological concept to intrusion detection systems"; 6th International Conference in Adaptive Computing in Design and Manufacture (ACDM'04), Bristol, UK.

[Kim et al. 2010] H.N. Kim, A.T. Ji, I. Ha, J.S. Jo. "Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation"; Electronic Commerce Research and Applications, 9 , pp. 73 – 83.

[Mazhar and Farooq 2007] Mazhar, N. and Farooq, M. "BeeAIS: Artificial Immune System Security for Nature Inspired, MANET Routing Protocol, BeeAdHoc"; Lecture Notes in Computer Science, 4628, 370 - 381.

[Morrison and Aickelin 2002] Morrison, T. and Aickelin, U. "An Artificial Immune System as a Recommender for Web Sites"; 1st International Conference on Artificial Immune Systems, pp 161 - 169, Canterbury, UK.

[Nunes and Timmis 2002] Nunes, L. and Timmis, J. "An Introduction to Artificial Immune Systems: A New Computational Intelligence Paradigm"; Congress on Evolutionary Computation, 131 – 138.

[Quan et al. 2011] Quan, J., Dattero, R., D. Galup, S. and Dhariwal, K. "The Determinants of Information Technology Wages"; International Journal of Human Capital and Information Technology Professionals, 2(1)

[Schwab et al. 2001] Schwab, I., Kobsa, A. and Koychev. "Learning User Interests through Positive Examples Using Content Analysis and Collaborative Filtering"; Internal Memo. Augustin.

[Secker et al. 2003] Secker, A., Freitas, A. and Timmis, J. "AISEC: an artificial immune system for e-mail classification"; (pp. 131 - 138)

[Smyth and Cotter 2000] Smyth, B. and Cotter, P. "A Personalized TV Listings Service for the Digital TV Age"; (pp. 53 - 59)

[Sudurama et al. 2013] Sudurama, A., Piarsa, N. and Buana, W. "Design and Implementation of Geographic Information System"; International Journal of Computer Science Issues, 10(2), 478 - 483.

[Tran and Cohen 2000] Tran, T. and Cohen, R. "Hybrid Recommender Systems for Electronic Commerce"; Knowledge-Based Electronic Markets, Papers from the AAAI Workshop.

[Yan and Wang 2012] Yan, X. and Wang, Y. "Development of Zaozhuang Tourism Information System Based on WebGIS"; International Journal of Computer Science Issues, 9(3), 249 - 252.