Modification of user profile using the genetic algorithm

Hidekazu Yanagimoto
Department of Computer and Systems Sciences
Osaka Prefecture University
1-1, Gakuencho, Sakai
Osaka 599-8531, Japan

Sigeru Omatu
Department of Computer and Systems Sciences
Osaka Prefecture University
1-1, Gakuencho, Sakai
Osaka, 599-8531, Japan

({hidekazu, omatu}@cs.osakafu-u.ac.jp)

Abstract

We propose a method which modifies a user profile (user's interests) using a genetic algorithm (GA). In information filtering, a method which constructs linear function using cosine distance, for example relevance feedback is often used. This discriminant linear function represents a hyperplain in vector space model (VSM). We focus on the classification hyperplane in the VSM and discuss a selection method of the most optimal hyperplane. Our aim is to make a selection criterion based on the above discussion. To decide the hyperplain which satisfies the criterion, we explore search space using a GA. Finally we describe experiments to evaluate a proposed method and discuss effectiveness of it.

Information filtering, Genetic Algorithm

1 Introduction

Users can have access to vast store of information through the Internet recently, as the Internet becomes widespread. When they search information they desire with search engines, they get a lot of search results and hardly find ones they desire. To cope with this situation, many researchers have paid attention to developing information filtering systems. The information filtering systems automatically select information depending on user's interests (user profile). To select information correctly, the systems must keep appropriate user's interests.

In this paper, we propose a method which modifies a user profile using a genetic algorithm (GA). To make the user profile is to decide a hyperplain in vector space model (SVM). There are many hyperplains which can separate training data according to their labels. Hence, we introduce a new criterion which can select many documents correctly and maximize a distance between the hyperplain and each training data. We use the GA to make the user profile which satisfies the criterion.

To evaluate our proposed method, we carry out an expriment and discuss effectiveness of the criterion in the view of selection ability.

2 Previous work

The proposing system consists of an information filtering and a GA to create user profile. We describe related works in detail.

An information filtering system is a system which selects information depending on user's interests. A content-based filtering system[1] chooses information analyzing a content of information. The content-based filtering system generally selects texts. A social filtering system[2] is a system which selects information using the relationship between a user and an information creator. Since social information is used without analyzing a content of information, the systems can be expanded to a systems which deal with non-text data. A collaborative filtering system[3] is a system which selects information depending on ratings voted by many other users. The system calculates a similarity among users and recommends information which other users who have the same interests are interested in. The system is applied to texts, graphical data, music and so on.

A GA is an algorithm which simulates biological evolution and is proposed by Holland[4]. The GA is often applied to combinational optimization problems. NewT[5] is an information filtering system using a GA.

3 Methodology

In this section, we explain key concepts of a proposed method. They are a problem setting, VSM and a GA.

3.1 Problem setting

We consider a linear discriminant function without a threshold and training data without noise in a classification problem. Then the discriminant function represents a hyperplain and weights of the discriminant function represents a normal vector. The normal vectors form a hypersphere in VSM. The normal vectors which can select training data are correctly included in a part of the hypersphere. Hence, to decide the optimal discriminant function, we must select one of the normal vectors included in the part of hypersphere using a criterion.

We introduce a criterion to select one of discriminant functions which select training data correctly. Using the criterion, we focus on the nearest trainig data from the hyperplain and select the hyperplain maximizing a distance between the hyperplain and the training data. To make discriminant function (user profile), this criterion is applied to a fitness function of a GA.

3.2 Vector Space Model

In VSM, documents written with natural language is represented as vectors to deal with them on computers. The vector is called a document vector below. An element of the document vector denotes a word included in a document and is calculated by TF-IDF.

$$w_j^i = tf_j^i \ log\frac{N}{df_j} \tag{1}$$

where w_j^i is a weight of the jth term T_j in ith document Doc_i, tf_j^i is a term frequency of T_j in Doc_i, df_j is a document frequency of T_j and N is the total number of documents.

In the VSM, a similarity between two documents can be defined by a distance between them. The distance is generally calculated by an inner product.

$$sim_i = profile \cdot doc_i \tag{2}$$

where sim_i is a similarity between a user profile and a document, doc_i is a document vector of the ith document Doc_i and $\cdot$ denotes an inner product.

In this paper, a user profile, which represents user's interest, is the same vector as a document vector. As the user profile represents user's interest, it is close to document vectors of interesting documents and far from ones of an uninteresting documents. Recommending documents according to user's interest, our proposed method selects a document depending on the similarity between the user profile and the document.

Given initial genes at random. For $t = 0, \cdots, T$:
- ·Select two genes as parents.
- ·For $n = 0, \cdots, N$
  - ·Produce two child genes using UNDX.
    - $\mathbf{C}_1 = \mathbf{m} + z_1\mathbf{e}_1 + \sum_{k=2}^{n} z_k\mathbf{e}_k$
    - $\mathbf{C}_2 = \mathbf{m} - z_1\mathbf{e}_1 - \sum_{k=2}^{n} z_k\mathbf{e}_k$
    - $\mathbf{m} = \frac{(\mathbf{P}_1+\mathbf{P}_2)}{2}$
    - $z_1 \sim N(0, \sigma_1^2)$, $z_k \sim N(0, \sigma_2^2)$
    - $\sigma_1 = \alpha d_1$, $\sigma_2 = \frac{\beta d_s}{\sqrt{n}}$
    - $\mathbf{e}_1 = \frac{(\mathbf{P}_1-\mathbf{P}_2)}{|\mathbf{P}_1-\mathbf{P}_2|}$, $\mathbf{e}_i \perp e_j (i \neq j)$
- ·Select two genes as survivors from all child genes.

Figure 1: The GA algorithm.

3.3 Genetic Algorithm

We make a user profile using a GA since to make the user profile is to find a vector which maximizes a similarity between a gene and an interesting document vector and minimizes a similarity between a gene and an uninteresting document vector. In the GA, a gene represents the user profile and is a real valued vector.

To carry out a GA, a fitness function must be defined. The fitness function is

$$F_k = |D_c| + \min_{D_i \in I} sim_i - \min_{D_i \in U} sim_i \tag{3}$$

where $|\cdot|$ is the number of elements in a set, D_c is a set of documents selected correctly, I is a set of interesting documents and U is a set of uninteresting documents. The first term in Equation 3 denotes ability to select documents correctly, the other terms denote distances between a hyperplain and a document. This fitness function denotes that the fitness value becomes a maximum value when a gene selects documents according to their ratings.

Our crossover is unimodal normal distribution crossover (UNDX) proposed by Ono et al [6]. The UNDX is one of crossover techniques implemented for a real-coded GA and explores a search space depending on a distribution of genes. Pseudocode for the GA is shown in Figure 1.

Table 1: The tasks used in the experiment.

Task No.	all	relevant	relevant/trial
108	371	64	16.1
109	453	21	3.8
110	624	34	6.3
111	540	40	7.5
114	616	20	3.1
115	931	94	9.3
119	387	20	4.9
121	501	63	11.9
126	279	22	7.7
132	210	22	10.5
138	253	24	9.0
139	386	142	36.7
140	309	23	6.6
147	591	80	14.6

4 Experiment

4.1 Data

To evaluate a proposed method, we use NTCIR2 test collection which consists of many evaluated documents. In the NTCIR2, many tasks, for example queries on information retrieval, homesickness syndroand and singular point, are prepared. In each task, a document has a label which denotes whether it is relevant or not.

In this experiment, we regard relevant documents as interesting documents and make a user profile. We pick up 14 tasks which have more than 20 relevant documents and use them as experiment tasks. Since a GA is probabilistic optimizing method, we carry out 10 trials per one task. Traning documents consist of 100 documents drawn at random from all documents and test data consist of the other documents. Each trial uses different training data and test data. In Table 1, we show the number of all documents, all relevant documents and average relevant documents in one trial.

4.2 Simulation and Result

To discuss our proposed method, we compare it with relevance feedback. The relevance feedback is used in the VSM and uses an innar product as a similarity between a user profile and a document. The relevance feedback is defined below.

$$profile = a \sum_{D_i \in I} doc_i - b \sum_{D_j \in U} doc_j \qquad (4)$$

Table 2: The Parameters of GA.

population	Generations	Crossover	α	β
5000	50000	20	0.5	0.35

Table 3: The 11-point average precision.

Task No.	GA($\hat{\mu}_{GA}$)	cosine($\hat{\mu}_{COS}$)
108	0.283	0.287
109	0.235	0.126
110	0.121	0.065
111	0.166	0.188
114	0.157	0.120
115	0.225	0.174
119	0.254	0.084
121	0.300	0.313
126	0.257	0.290
132	0.240	0.176
138	0.737	0.592
139	0.618	0.760
140	0.309	0.246
147	0.336	0.339

where both a and b are arbitrary positive numbers. We decide both a and b using leave-one-out.

Since our proposed method uses a GA, the parameters of theGA for two methods are shown in Table 2. The GA needs more population and generations than usual to explore wide search space.

In Table 3, 11-point average precision is shown. The 11-point average precision is calculated with 11 precision on 11 recall levels.

4.3 Discussion

In Table 3, we realize that our proposed method is superior to relevance feedback in 8 tasks. However, in the other 6 tasks, relevance feedback is superior to the proposed method. We discuss this 6 tasks in detail especially.

Since a genetic algorithm is probabilistic method, the proposed method depends on various conditions (initial genes, generation of child genes and selection). Hence, we discuss a difference of 11-point average precision using t-test. The t-test is a statistical method to consider whether this difference is significant or not. In this case a null hypothesis H_0 is $\mu_{GA} = \mu_{COS}$ where μ_{GA} is an expectation of 11-point average precision for the proposing method and μ_{COS} is an expectation of 11-point average precision for the relevance feedback. First, the difference of the 11-point average precisions

Table 4: The result of t-test.

Task No.	t
108	0.208
109	2.627
110	2.545
111	1.008
114	1.088
115	2.443
119	3.176
121	0.684
126	0.883
132	1.724
138	2.703
139	3.463
140	1.418
147	0.096

is transformed with next formula in each task.

$$t = \frac{\hat{\mu}_{GA} - \hat{\mu}_{COS}}{s\sqrt{\frac{1}{n_{GA}} + \frac{1}{n_{COS}}}} \tag{5}$$

where s is standard deviation of all 11-point average precisions for all trials, n_{GA} is the number of 11-point average precision for the proposing method ($n_{GA} = 10$) and n_{COS} is the number of 11-point average precision for the relevance feedback ($n_{COS} = 10$). Since we adopt the level of significance of 5% in this t-test, t becomes 2.101. When t is more than 2.101, the null hypothesis H_0 is rejected. In table 4 we show the result of t-test. When a value of relevance feedback is bigger, the t is underlined.

In Table 4, we find that in 5 tasks the proposed method is superior to the relevance feedback in the view of t-test and in one task inferior.

5 Conclusion

In this paper, we proposed information filtering using a GA. We introduced a new criterion which regards a distance between a discriminant hyperplain and training data. We confirmed that the proposed method was superior to relevance feedback in some tasks with significance test.

We will apply this proposed method to other situations and evaluate the proposed method in detail.

Acknowledgements

This study is supported by Grants-in-Aid for Scientific Research. We would like to thank MEXT for the supports.

References

[1] T. W. Malone, K. R. Grant and F. A. Turback(1987), The Information Lens: As Intelligent System for Information Sharing in Organizations, Proceedings of ACM CHI'86, pp.1–8.

[2] P. Maes(1994), Agents that Reduces Work and Information Overload, Communication of ACM, Vol. 37, No. 7, pp.30–40

[3] J. A. Konstan, B. N. Miller, D.Maltz, J. L. Herlocker, L. R. Gordon and J. Riedl(1997), GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM, Vol. 40, No. 3, pp.77–87.

[4] J. H. Holland(1962), Outline for a Logical Theory of Adaptive Systems , Journal of the Association for Computing Machinery, vol.3, pp.297–314.

[5] B. Sheth and P. Maes(1993), Evolving Agents for Personalized Information Filtering, Proceedings of the Conference on Artificial Intelligence Application, Vol. 9, pp.345-352.

[6] I. Ono and S. Kobayashi(1997), A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, Proceedings of 7th International Conference on Genetic Algorithms, pp.246–253