Extraction of Design Variables using Collaborative Filtering for interactive Genetic Algorithms

Tomoyuki HIROYASU
and Hisatake YOKOUCHI
Department of Life and
Medical Sciences,
Doshisha University

Misato TANAKA
Graduate School of Engineering,
Doshisha University

Mitsunori MIKI
Department of Science and
Engineering,
Doshisha University

***Abstract*—Interactive Genetic Algorithm (iGA) is one of evolutionary computations in which the design candidates are evaluated by human. Using iGA, the sensibility and subjective feelings of humans can be optimized by learning the user's evaluation of presented individuals. In this research, iGA was applied to product recommendation on shopping sites. One of the most difficult points to be addressed in construction of a product recommendation system is to taking a long time to extract and assign values to design variables from all of the actual products on the site. It is also difficult to define product design variables appropriately. To address these problems, we propose a method to generate design variables automatically based on a lot of users' preference data on the Web. We constructed the design variables using the relevance of products obtained by Collaborative Filtering and discussed them. Through the simulation experiments, the effectiveness of the proposed method is discussed.**

I. Introduction

Recently, online shopping sites have been attracting more users than conventional physical stores. Online shopping sites can provide vendors with increased sales opportunities and present tremendous choices for consumers. Many of these shopping sites, such as Amazon[1], adopt product recommendation schemes, such as Collaborative Filtering [1], [2] and support vector machine (SVM) [3], to lead users to products purchased by other users with similar preferences and encourage their purchase. Product recommendation is a technique that can present products that may be bought by users. In this technique, information based on a user's preference is utilized.

Here, we consider the application of interactive Genetic Algorithms (iGAs) [4] for product recommendation. The iGA is an optimization method in which users evaluate the solutions instead of a fitness function in a Genetic Algorithms (GA) [5]. It has been confirmed that iGA can present products that suit each user's preferences.

Here, we discuss a basic requirement for application of iGA to product recommendation. Many problems must be addressed for practical use of iGAs. One of these is how to design products. In iGA, the colors, patterns, shapes, materials, etc. of products are defined as the design variables that constitute the chromosome. However, as these components of products vary considerably, it is difficult to determine the structure of the chromosome common to all products. Moreover, the measurement and input of characteristics of products such as color, pattern, etc. are associated with significant costs to the vendor.

Based on this background, we proposed a technique for constructing the design variable space of products automatically based on users' preference data on the Web. On the web, a lot of users' personal information and history log are accumulated, and provided as collective knowledge [6]. We think that the average taste information can be extracted from these collective knowledge and others. We define this information as collective preference. In this paper, we treated the recommendation relation on the Web as collective preference.

II. Interactive Genetic Algorithm

A. Applying interactive GA to product recommendation

The iGA is an algorithm derived from GA. The evaluation operation in GA is replaced with the user's subjectivity. The iGA searches for an optimal solution using the user's subjective evaluation. Therefore, it can analyze a complex structure of human sensibility, and this approach is often applied to problems that are difficult to evaluate quantitatively, such as hearing aid fitting and fashion design. It has been confirmed that iGA can recommend products that suit a user's preference [7], and we applied this approach to product recommendation on online shopping sites.

Fig. 1 shows the flow of product recommendation by iGA. The user evaluates the products shown on the Web interface. Based on the user's evaluation, the iGA system performs the genetic operations (i.e., evaluation, selection, crossover, and mutation), and presents the user the evolved products again. The presentation of products can be optimized by repeating these operations.

B. Design Variables

To design products, the design variables should be defined in the iGA. For example, this section focuses on T-shirt designs. In this system, a T-shirt has various components, such as color, pattern, shape, material, etc. One T-shirt design is determined by the combination of these components. The iGA can determine the optimal combination of these modules using optimization techniques. Each candidate of the module

[1]http://amazon.com/

978-1-4244-3597-5/09/$25.00 ©2009 IEEE

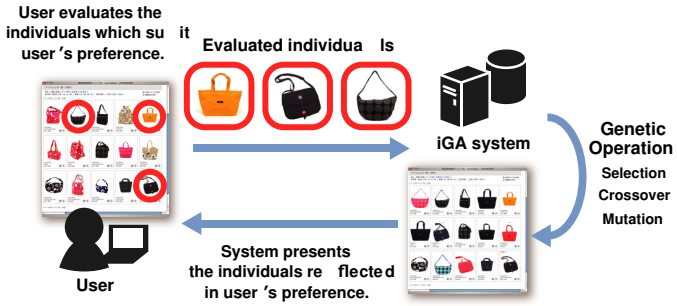

Fig. 1. Flow of product recommendation by iGA

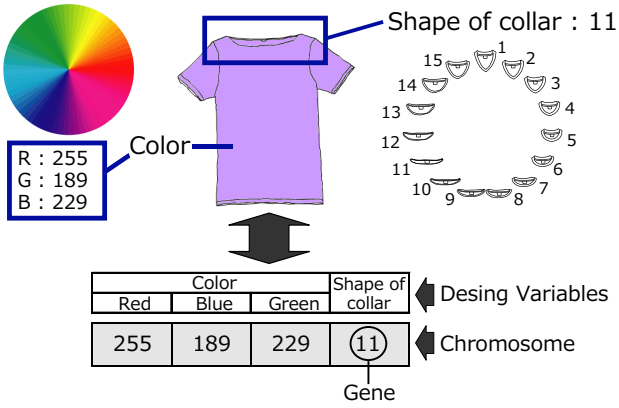

Fig. 2. Design variables in T-shirt iGA system

is converted to a number, which is called the chromosome in iGA.

III. Extraction of Design Variables using Collaborative Filtering for Interactive Genetic Algorithms

A. Defects of iGA

In obtaining the chromosomes of actual products on shopping sites, the following problems are considered.

1) The vendors must measure the values of the actual products for each defined design variable and input them as genes. Although there is already technology which acquires the dominant color of products, it is difficult to analyze products which have some patterns and decorations. Therefore, the extraction of the proper design variable is a human-intensive process.

2) In addition to components that can be evaluated easily, such as color and size, there may also be those that are difficult to quantify and whose neighborhoods are difficult to determine. The neighborhood of the design variables in iGA must reflect human sensibility. For example, when people look at Pattern A and Pattern B and think that both are alike, the distance between both in the design variable space is close. On the other hand, the distance is greater when Pattern C and Pattern D are evaluated as different.

3) The developers must take the properties of design variables into consideration when developing an iGA system. The various differences among design variables influence implementation in addition to constraint conditions. For example, scale, continuity, and whether values are residue class. Therefore, if the definition of design variables is changed by exchange of products, complicated work becomes necessary to change the system.

Of the above problems, the first may be the most serious in cases where there are large numbers of products. To overcome these problems, a technique is needed to define the design variables automatically and compute the actual values of each product with as little human intervention as possible. iGA is the contents filtering based on the attribute of contents. On the other hand, Collaborative Filtering does not reflect any attribute of contents because of using only user's history logs.

In this research, to generate the design variables automatically for iGA, we used the relevance of the products based on collective preference accumulated on the Web. In this paper, the recommendation relations by Collaborative Filtering are treated as collective knowledge, because they are easy to extract the degree of association between goods. Specifically, we propose a method for obtaining the adjacency matrix from the recommendation relations among the products by Collaborative Filtering, computing the genes by Principal Components Analysis (PCA) and optimizing the product recommendation by iGA.

B. Extraction of Design Variables using Collaborative Filtering for interactive Genetic Algorithms

Collaborative Filtering is a technique for analyzing a user's past action record and predicting the taste of others who take similar actions. Many shopping sites, such as Amazon, use product recommendation algorithms based on this approach [8].

The proposed method generates the design variable space and chromosomes of individuals by computing the adjacency matrix that shows the relevance of products based on the recommendation relations of the products obtained from such sites, and by applying principal component analysis to this matrix. We regard a feature of which of other products recommend a product as a chromosome of products. However, when using this feature as a chromosome directly, the gene length is dependent on the number of products. Therefore, we use PCA to adjust genes to a suitable length.

C. Algorithm

The algorithm for automatic generation of the design variable space by Collaborative Filtering is as follows:

1) Obtain the recommendation relations of products on Collaborative Filtering from sites on the Web.

2) Generate the following symmetric matrix based on the recommendation relations of products (Individual: Ind). When Ind_1 recommends Ind_2, the values of $Ind_{1,2}$ and $Ind_{2,1}$ are 1.

$$\begin{array}{c} \\ Ind_1 \\ Ind_2 \\ Ind_3 \\ \vdots \\ Ind_N \end{array} \begin{array}{c} \begin{array}{ccccc} Ind_1 & Ind_2 & Ind_3 & \dots & Ind_N \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & \dots & 0 \\ 1 & 1 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 1 \end{pmatrix} \end{array}$$

On the other hand, these are set to 0 if there are no recommendation relations. Each row of a matrix shows the raw genotype of each product individual and each column indicates the design variables of products.

3) Reduce the number of dimensions of the design variable of product individuals by PCA.

   a) Calculate the eigenvalue and eigenvector from the matrix obtained in 2.

   b) Define the number of dimensions of the design variables, extract the eigenvectors of the defined number in descending order of the absolute values of the eigenvalues, and generate a rotation matrix.

   c) Obtain the principal component scores as genotypes of each product by multiplying the original matrix by the rotation matrix.

When adding a new product, the principal component score of the new product is calculated by the rotation matrix obtained in 3b.

i) Generate a vector that shows the relevance of products by obtaining a recommendation relations between a new product and the others used when computing the rotation matrix in the algorithm used for automatic generation of the design variable space.

$$\begin{array}{c} \\ Ind_{N+1} \end{array} \begin{array}{c} \begin{array}{ccccc} Ind_1 & Ind_2 & Ind_3 & \dots & Ind_N \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & \dots & 1 \end{pmatrix} \end{array}$$

ii) The principal component score of the new product is computed by multiplying the vector obtained in III-C by the rotation matrix, and it is considered the gene of the new product.

IV. Experiments

In this section, we investigated the space generated by the proposed method from Collaborative Filtering and discussed the results.

A. Overview of the Experimental System

A system that implements the proposed method was developed for the preliminary experiment. The requirements for the system ware as follows:

1) Ability to acquire the recommendation relations based on Collaborative Filtering
2) Ability to determine appropriate gene length

The recommendation relations of products can be acquired in 1 from Amazon Web Service (AWS) [2], which provides access to Amazon's product database and technical platform. For example, it is possible to obtain the details of a certain book, i.e., title, authors, publisher, price, user reviews, etc. By implementing this service, this system obtains the list of other books recommended by a given book, and generates an adjacency matrix based on the recommendation relations based on Collaborative Filtering [8].

The details of the preliminary experiment are described in IV-B. In IV-C, we showed the results that we investigated the feature of the design variable space obtained by the proposed method.

B. Preliminary Experiment for Dimension Number

[The Goal of This Experiment]
In this preliminary experiment, the appropriate gene length to reduce the dimensions of the design variable space is derived.

[The Experiment Procedures]
The suitable gene length is defined as the number of eigenvalues when the contribution ratio calculated by principal component analysis is added to the cumulative contribution in descending order and exceeds 0.6. The definition is based on standard principal component analysis determining the number of dimensions by the cumulative contribution. Although various values are regarded as suitable cumulative contribution, we selected a value of 0.6.

Based on the above definition, the experiment was performed using the following procedures:

1) Search books from a category and a keyword, and sets the top result to the target product.
2) Obtain 10 products recommended by the target product.
3) Add the obtained products to the queue of target products.
4) The product at the head of the queue is set to the next target product, and the operations from 2 to 4 are repeated until the number of the products getting the recommendation relations exceeds the defined number.
5) Generate an adjacency matrix from the recommendation relations, and compute eigenvalues and eigenvectors by using principal component analysis.

[2]http://developer.amazonwebservices.com/

6) Calculate the cumulative contribution by adding contribution ratio in descending order.

The preliminary experiment was performed under the conditions shown in TABLE I.

[The Experimental Result]
TABLE II shows the list of the contribution ratios and cumulative contributions in descending order. This experimental result indicated that the appropriate number of dimensions was seven. Therefore, this number was used in all the experiments conducted on the same scale.

C. The Preliminary Experiments for Design Variable Space

[The Goal of This Experiment]
The experiment objective was to investigate the feature of the obtained design variable space based on Collaborative Filtering and probe about the requirements on implementing this space in an iGA system.

[The Experiment Procedures]
The preliminary experiment in IV-B showed that the suitable number of dimensions was seven. However, to visualize the design variable space, we contracted the design variable space to two dimensions by the proposed method.

[The Experimet Result]
Fig. 3 shows the two-dimensional design variable space. This experiment was performed under the conditions shown in TABLE I. A book was expressed as a point based on the values of the gene.

The distribution of books and authors shows relevance in Fig. 3. Some authors were distributed over a specific part of the graph.

[The Discussion]
In Fig. 3 shows the design variable space that was generated. In this space, the books of the same author are gathered in the same area. It may happen that many users tend to buy the same author's books. Thus, Collaborative Filtering recommends the products to other users who bought the same products. From this way, when there are many users who buy the same author's books, these books are tend to be recommended.

In this distribution, while two authors, Kyogoku and Hatanaka, are near to Miyabe, they are distant mutually. Moreover, although Kogetsu overlaps with Kyogoku, he is far from Miyabe. Therefore, it is thought that the neighborhood of the author in the design variable space can be obtained from Fig. 3.

TABLE I
PARAMETERS FOR THE PRELIMINARY EXPERIMENT

Search Keyword	Search Category	The number of the products
Miyuki Miyabe	Books	100

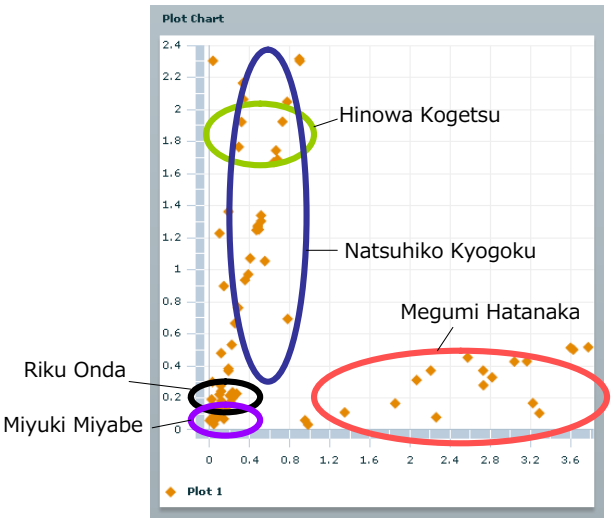

Fig. 3. Results with reduction of the dimension of the design variable space to two

The element which a user thought as important when the user purchased products appears strongly in the recommendation based on Collaborative Filtering. We could obtain the neighborhood of elements such as author. It is necessary to confirm the validity of neighborhood and the tendency in other genres.

In this distribution, Kyogoku and Hatanaka were distinct each other. Therefore, if conventional iGAs search, iGA cannot recommend suitable books to the users who like both of authors. In this case, we think users have multimodal preference. To search the users' multimodal preference, the preference area needs to be divided as shown in Fig.5. In each are, the search should be performed.

For this search, we have proposed the method in which new individuals are generated by clustering [9]. Firstly, products are clustered into sub groups. Secondly, In each sub group, the minimum and maximum values of the included individual of each design variable are obtained. Finally, new individuals are

TABLE II
THE CONTRIBUTION RATIOS AND CUMULATIVE CONTRIBUTIONS IN THE PRELIMINARY EXPERIMENT

dimension	contribution ratio	cumulative contribution
1	1.41962827	0.18965670
2	0.81778887	0.29891004
3	0.68218989	0.39004790
4	0.51423215	0.45874727
5	0.49074888	0.52430937
6	0.43704136	0.58269636
7	0.34048466	0.62818376
8	0.31997330	0.67093092
9	0.23922675	0.70289067
10	0.17271718	0.72596499

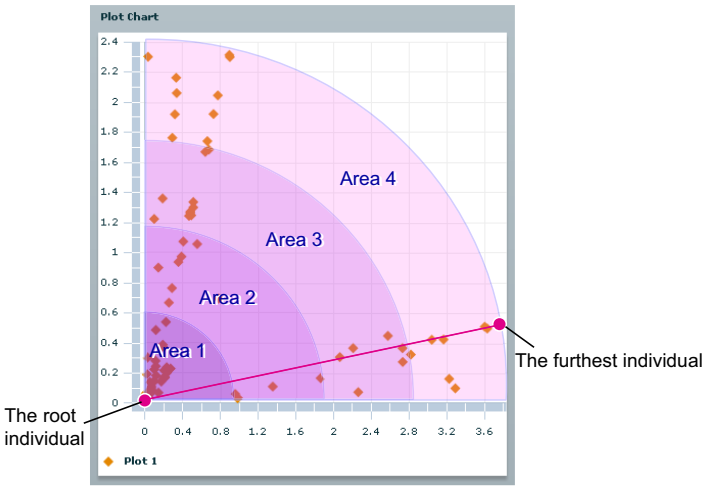

Fig. 4. The design variable space divided into four areas

generated within the super-cube based on the derived values.

D. The Subjective Experiment

[The Goal of This Experiment]

In this experiment, we checked the variable space generated by the proposed method was similar to the user's mental design variables space.

We compared two individuals which are close in the derived space and checked that users feel these two individuals are really close in their subject.

[The Experiment Procedures]

We generated the design variable spaces according to the procedure of IV-B. Three keywords were prepared as shown in TABLE III.

The top product of search result in 1 of IV-B is called a route individual. The recommendation relations of the products which generate design variable space are based on this route individual. The Euclidean distances of the route individual and other individuals were calculated, and they were arranged in ascending order. The distance of the route individual and the furthest individual was divided into plurality, and the individual group with each distance was determined as the individual group of Area1, Area2 and so on from the direction near a route individual. The example which divided two-dimensional design variable space into four areas was shown in Fig. 4.

The subjects judged which two individuals chosen from different areas at random was more similar with the route individual. The followings are the question given to the subject and five choices of a reply. A and B means the product individuals chosen at random from different area.

Question: Which products of A and B do you think that the route individual is more similar with?

Answer1: "A" is more similar with the route individual than "B".

Answer2: It seems that "A" is rather alike the route individual than "B".

Answer3: Either.

Answer4: It seems that "B" is rather alike the route individual than "A".

Answer5: "B" is more similar with the route individual than "A".

The number of subjects was eight. There were seven men and there was a woman. An order of the combination of the area and the order of individuals presented at random maintains counterbalance.

[The Experimental Result]

TABLE IV shows the result of the experiment. In TABLE IV, the individual with the shorter Euclidean distance from the route individual is set to the A side. The subjects answered the question shown in IV-D about all the combination of areas, and the averages of all the subjects were taken. If the average value is closer to one, it means that more subjects had judged that A is more similar with the route individual than B.

[Discussion]

TABLE III
PARAMETERS FOR THE EXPERIMENT

Dataset	Search Keyword	Search Category	The number of products
1	Harry Potter	Books	100
2	Economy	Books	100
3	Programming	Books	100

TABLE IV
AVERAGE OF THE SUBJECTS' ANSWER

Dataset	A	B	average
1	1	2	1.125
	1	3	2.750
	1	4	2.000
	2	3	3.875
	2	4	4.375
	3	4	2.375
2	1	2	2.125
	1	3	1.000
	1	4	1.625
	2	3	2.000
	2	4	2.625
	3	4	4.500
3	1	2	2.375
	1	3	1.500
	1	4	1.750
	2	3	2.375
	2	4	2.750
	3	4	3.500

The average of all the data (144 set) is 2.479166667. We tested the hypothesis that the average is 3 by the one-sided testing. The significance level was 0.01. In this result, the rejection region was given as $\mu < 2.53972996$, so this hypothesis is rejected significantly. Therefore, the result was obtained significantly that the design variable space generated by the proposed method is similar with the users' mental design variable space. However, the average of Area2 of Dataset1 and Area4 of Dataset2 is not so good. The cause is considered as following:

- Product individuals other than books, such as DVD, exist near the route individual DVD was selected as an individual of the Area2 in Dataset1.
  DVD, CD record, etc. are contained in the recommendation products of books in Amazon. The interview showed that DVD tends to give the users the impression that DVD is not similar to a route individual because most of individuals presented to the users are books.
- The distance from the route individual is near.
  Here, the data set is created by linking the recommendation relations from the route individual until the number of products exceeds 100. By obtaining the recommendation products of ten from one product, the number of goods increases exponentially in proportion to the recommendation distance from a route individual. Therefore, in these data sets, the products which had only four recommendation links from a route individual could be obtained. Some user also commented that it was difficult to feel the difference among the presented products.
- The number of dimensions
  Here, we asked for the number of dimension in the Preliminary Experiment. However, it is necessary to examine a number of data to reference and to examine much data.

We considered that the negative result about IV-D was caused by the expression of the question: "whether it is similar or not". When the users actually use iGA system, they evaluate products on the basis of their taste: "whether you like or not"or "whether you want to buy or not", so we thought this problem was not important.About IV-D, we need to treat not only 100 but 1000 or more products in generating the design variable space in order toget the product individuals which have longer distances of the recommendation relations from a route individual. Moreover, the same experiment as this needed to be performed in the recommendation distance from a route individual. In a subjective experiment, it is necessary to confirm that products are similar in a user's sensitivity if recommendation distances from the route individual are short. The suitable number of dimensions should be examined based on the experimental result.

V. Conclusions and Future Work

In this research, iGA was applied to product recommendation on shopping sites. One of the most difficult points in construction of a product recommendation system using iGA is to define the design variables of products appropriately. In this paper, we proposed a method of generating the design variable space automatically using the information on shopping sites. The proposed method obtains an adjacency matrix from the recommendation relation of products based on Collaborative Filtering, and generates genes of each product by principal component analysis. When treating many products, this method is thought to be able to reduce the cost of defining the design variable space and measuring the values of the genes of products.

Here, we used the recommendation relations of the products based on Collaborative Filtering. The data of Collaborative Filtering can be transformed easily to a matrix. We consider it possible to analyze text data, such as product descriptions, to obtain the relevance rate of products and implement it in generation of the design variable space.

In order to examine whether the design variable space generated by the proposed method was similar to the user's mental design variables space, we performed the experiment with the subjects. The experiment results showed that the both of design variable spaces were similar significantly.

In the future work, it is necessary to perform a subjective experiment in order to confirm the similarity of the recommendation relation of the individual before generating design variable space and a user's mental design variable space. Moreover, the user's preferences are considered by product recommendation using iGA, although Collaborative Filtering recommends products with high relevance among items or users. Therefore, further development of product recommendation systems that use iGA or Collaborative Filtering is required to determine which algorithm shows better performance in recommending a user's favored products.

References

[1] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: applying collaborative filtering to usenet news," *Communications of the ACM*, vol. 40, no. 3, pp. 77–87, March 1997.

[2] B. Sarwar, G. Karypis, J. A. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM New York, USA, 2001, pp. 285–295.

[3] T. Joachimes, "Text categorization with suport vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*. Springer-Verlag London, UK, 1998, pp. 137–142.

[4] H. Takagi, "Interactive evolutionary computation: fusion of the capabilities of ec optimization and human evaluation," in *Proceedings of the IEEE*, vol. 89, no. 9, 2001, pp. 1275–1296.

[5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

[6] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.

[7] H. Takagi, T. Unemi, and T. Terano, "Interactive evolutionary computation," in *Genetic Algorithm*. Asakura Publishing, 2000, vol. 4, ch. 11, pp. 325–365.

[8] G. Linden, B. Smith, and J. York, *Amazon.com Recommendations Item-to-Item Collaborative Filtering*. IEEE Computer Society, January 2003.

[9] F. Ito, T. Hiroyasu, M. Miki, and H. Yokouchi, "Discussion of offspring generation method for interactive genetic algorithms with consideration of multimodal preference," in *Simulated Evolution and Learning*, ser. Lecture Notes in Computer Science, vol. 5361. Springer, 2008, pp. 349–359.