AUTOMATIC SEARCH OF INDICATORS OF TEXT AUTHORSHIP

Shevelev Gennady Efimovich, Shevelev Oleg Gennadyevich
Tomsk Polytechnic University
Lenina St. 30, Tomsk, 634050, Russia
Tel: +8(3822)558155
E-mail: osh@inet.tsu.ru

Abstract:
This paper is the work in the field of text authorship analysis – stylometry. The main assumption underlying stylometric studies is that all ripe authors of texts have an unconscious aspect to their style. Stylometrics use enormous variety of text features for determining authorship of texts. But there are no definite rules from which indicators to choose and finding ones for a specific set of texts seems to be a black art. In our work we have tried to realize the automatic feature extraction algorithm based on the genetic search. The special code allowing carrying out searching of indicators of text authorship and the function for determining quality of obtained variants have been developed. The details, results and conclusions of the experiment are presented.

Keywords: Indicators of authorship, Authorship attribution, Genetic algorithms, Stylometry, Statistical analysis of author style.

1. Introduction

With increase of textual information and development of electronic libraries objective necessity of automatic classification and analyzing of the text information has appeared. One of possible classifications is attribution of author's style, when on the basis of available patterns or with the help automatic clustering to each text (from a set of suggested) the author is put in conformity. Available methods of an attribution of authorship demand knowledge of author's invariant – a special set of quantitative characteristics of the text equals for one author and essentially differing for texts of different authors. The statement about existence author's invariant is based on the assumption that each text possesses a set of the latent properties, unique for each mature author. These properties are brought unconsciously, therefore practically do not give in to imitation. The problem consists that each researcher tries to offer his own author's invariant and usually bears this offer unsubstantially [1, 2].

In our work we offer a method which is attempt of an automatic presence of an author's variant. Its idea is consists that gets out enough the big set of texts with known authors and with the help of genetic search is carried out searching every possible characteristics. For realization of searching we have developed the special code, allowing varying rules of calculation of characteristics and covering as many as possible of them. Searching with the said code we have realized in Visual C ++ language and tested it on a various data set. Though results while cannot be named sensational, and the received characteristics carry more local (adhered to the concrete data), rather than universal character, the method has shown the serviceability and great potential for development.

2. Genetic algorithms

Completely clearly, that search of characteristics of authorship (author's invariant) should have global character. Moreover, the space of various sets of characteristics has vaguely high complexity, therefore can have infinite quantity of local extrema. Methods which would work with such huge amount of the information and similar restrictions, it is not a lot of. The most obvious is full searching, beginning with the minimal set of characteristics and consistently changing it aside infinity. The similar approach in view of enormous volume of the analyzed information is not necessary: it can reach the purpose never. Methods of consecutive approach the purpose (even if the step of approach will vary, we are expected with set of problems in the correct organization of search) for the same reason do not approach. Idle time searching of randomized decisions does not possess convergence and can not find also anything valuable. In our case attempt of using methods of an artificial intellect, and in particular to genetic algorithms has been made.

Genetic algorithms are adaptive methods of search, last time frequently used for the decision of tasks of functional optimization [3, 4]. They are based on idea of genetic transformations of biological organisms. The populations, submitting to laws of natural selection (« survives the most adapted »), develop during several generations. Imitating this process genetic algorithms are capable "to develop" decisions of real tasks if those are in appropriate way coded.

3. Coding of characteristics of authorship and the organization of searching

The choice of the coding in genetic algorithms plays a main role. In a case of successfully chosen coding better decisions are for much smaller time interval. It is important, that all possible variants of a code line completely covered space of decisions, and better even coincided with it. If coding assumes generation of beforehand incorrect variants which do not satisfy to restrictions of a problem, that is probability that genetic algorithms will carry out search not the best decisions, and in general possible (for example, among ten received decisions only two satisfy to conditions of a task). And in case of an incomplete covering of space of decisions some decisions will not be considered by code space at all.

Thus, a choice of a code is an uneasy task. It should be capacious enough to reflect utmost feature space, compact that it was not necessary to analyze each time a bulky code, universal to work with any quantity a symbol (and any language potentially), and also full (that change or transformation of any part of a code formed also a correct code).

The code developed by us always has even length. Numbers on odd positions are numbers of atoms, whereas following for each of them number on an even position their attribute. Before the program can process such code, it is necessary for it to load atoms and consequently, each code is adhered to the set of atoms. The set of atoms is a range of definition of genomes on odd positions. There is a base set of atoms which is realized as the list and includes all possible atoms incorporated in the program. In our case it is all ASCII the table of symbols, some compound punctuation marks (a dots, a question, an exclamation mark with dots, etc.), and service atoms. If necessary the program easily extends for work with symbols Unicode.

The working list of atoms is the list containing those atoms which participate in formation of a concrete genetic line, or the atoms adhered to already generated genetic line. Besides ordinary atoms, there are service atoms. One of the most important of them is a divider of characteristics. If it meets in line it signals the beginning of the new characteristic. All genetic line has been named us a variant because it is only a variant of a set of characteristics. Thus, the variant can include some characteristics, divided among themselves a code of a divider of characteristics.

The genetic string (variant) sets itself a pattern of gathering of statistics. Each characteristic in it as a result of processing one text gives a vector - column of values (on one on each sample of the text which participates in the analysis). From here follows, that all variant sets a pattern of gathering of statistics which result is the file of vectors-columns with number of columns equal to number of characteristics in a variant, and with number of lines equal to quantity samples which participate in the analysis. The pattern of gathering of statistics - only a trope, in practice each characteristic is Boolean expression which is consistently applied to each word of the text. If the condition is carried out (we shall consider while only one sample) the counter of the characteristic increases for unit. When the characteristic contains some words to the current word of the text the first word of the characteristic is applied, to the following after current - the second, and so on. The counter of the characteristic increases just in the case that all Boolean expressions for all words of the characteristic were true. After gathering statistics the file of vectors-columns with the saved up values of characteristics which each element is divided on length of sample turns out.

Besides dividers there is one more service atom is a length of a word. Its attribute is a size of length, attributes of other atoms (letters) specify their site in a word.

Atoms participate in genetic search from the working list only, therefore it is possible to set an any kind of a line. The format of a code has turned out universal as allows to set any character set for searching so both any language, and capacious, time with its help it is possible to set as much as complex descriptions, included patterns not only the whole offers, but also pieces of the text.

4. Realization of criterion function

Essential part of genetic algorithm is the choice of criterion function (fitness - function). Fitness-function is responsible for an estimation of decisions. Its task to determine quality of the decision proceeding from a kind of a concrete chromosome. The first part of calculation of fitness-function of any variant in our concrete case of search of a set of author's characteristics consists in formation of a file of vectors-columns of statistics on the basis of that pattern that gives this variant. Analysis of a genetic line and the description of provisional algorithm of gathering of statistics it have been given in the previous section. It is necessary to open, how the file of vectors-columns of the saved up statistics addresses in a singular - value of fitness-function of a chromosome.

We had been found the simple and effective decision of this problem. This one of few probable decisions, also is possible, that it not the most optimum. Fitness-function in this case is calculated as the relation of a dispersion of classes of authors to an average dispersion inside classes of authors:

$$\frac{1}{n-1}\sum_{i=1}^{n} M(\overline{A},\overline{A}_i)^2 \Bigg/ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k-1}\sum_{t=1}^{k_i} M(\overline{A}_i, A_{it})^2\right), \qquad (1)$$

Where n - number of various authors;

k_i - Number samples in texts of i-th author;

A_{it} - A vector of coordinates of t-th text of i-th author in stylometric space for the found feature set (has dimension equal to quantity of characteristics);

$\overline{A_i}$ - Selective average of coordinates samples under texts of i-th author in stylometric space for the found feature set (has dimension equal to quantity of characteristics);

$\overline{A}$ - Selective average of selective average coordinates samples under texts of all writers in stylometric space for the found feature set (has dimension equal to quantity of characteristics);

$M()$ - Function of a finding of distance between vectors.

Result of gathering of statistics is the file of vectors-columns, on column on each characteristic, and on a line on each sample. That is, actually is under construction multivariate feature space (do not confuse to space of a set of characteristics!) . Before start of algorithm texts on which training is made get out, and the author of everyone is beforehand known. Thus, any lines of space can be related to one author (they correspond to samples of products of this author), any to another. Each line corresponding to concrete sample is a point in multivariate feature space. The initial variant (genotype) is generating for this space. Value of fitness-function of the variant generated the given space, the above, than are better grouped points-samples of one writer, and than further groups different writers are located from each other. Therefore value of fitness-function is calculated as the relation of a dispersion clusters of writers (disorder clusters in feature space the more, the better) to an average dispersion inside classes of authors (the disorder of points-samples of one writer - the is less, the better).

5. Realization of genetic search on the basis of library GALib References

Programming of a nucleus of genetic algorithm is not a simple and as a matter of fact unnecessary task. Instead of developing from zero all libraries of classes, we had been carried out search of suitable library in the Internet, and what completely meets our requirements has been found.

Library GALib is a full-function library of objects and methods for development of genetic algorithms [5]. It includes types of the data, various classes of genotypes, genetic algorithms, populations, and circuits of selections, classes of gathering and accumulation of statistics on work of algorithm and classes of generation of random numbers. All rights on distribution belong Massachusetts Technological Institute and the author of the program – Matthew Wall.

In an ideal, to the programmer using library GALib, it is enough to choose a class of genetic algorithms and genotypes which are necessary for the decision of his problem, to establish the necessary parameters, and to write fitness-function. In our case all has turned out a little bit more difficultly. For correct work of genetic algorithm it was necessary to copy procedure of initialization of genotypes, and also two main genetic operators - crossover and mutation. In library opportunities of redefinition of base functions without necessity of a spelling of a hereditary class are incorporated.

As type of a chromosome (genotype) class GA1DArrayGenome with integer elements (integer) has been chosen. It is a scaled file, all functions for work with which are incorporated in library GALib. As to a class of the most genetic algorithm we had been chose class GASteadyStateGA. It is the algorithm using overlapped populations with the set user in factor of overlapping. It means that the next population contains the set percent of individuals from the previous population, and other members are collected from among the genetic operators formed as a result of work.

The first what it was necessary to face during work with GALib, besides a choice of classes and adjustment of parameters, was realization of fitness-function. Not including fine tunings to given to genetic library and debugging, with it have not arisen problems - all evaluation stages of quality of a genetic line have been thought over and realized by us beforehand. In the beginning there was an estimation of a correctness of a variant (in case of mistakes in line zero value of fitness - function came back), formation of a file of vectors-columns further followed on the basis of gathering statistics, and at last stage for an estimation of quality of the found author's invariant the arrangement clusters of writers in multivariate space of attributes was considered. Besides realization of fitness-function by us own operators of a mutation, crossover and the initialization adapted to our specific target have been written.

6. First results

During testing a method by us various groups and volumes of texts have been considered. The universal author's characteristic, allowing dividing any authors, by the moment of a spelling of work have not been found yet as process of search has global character and works for a long time enough, but intermediate results are available already now. For example, the style space for seven texts of three writers (A. Beljaev, I. Ilf and E. Petrov, A. Solzhenitsyn) and the found set of characteristics looks as you can see at the Fig. 1.

Clusters are allocated still insufficiently precisely, but the tendency to their formation already is now appreciable. We carried out also more scale experiments and experiments on the other set of texts, all of them have shown the results confirming that the method is effective, but it is necessary to make its further improvement.

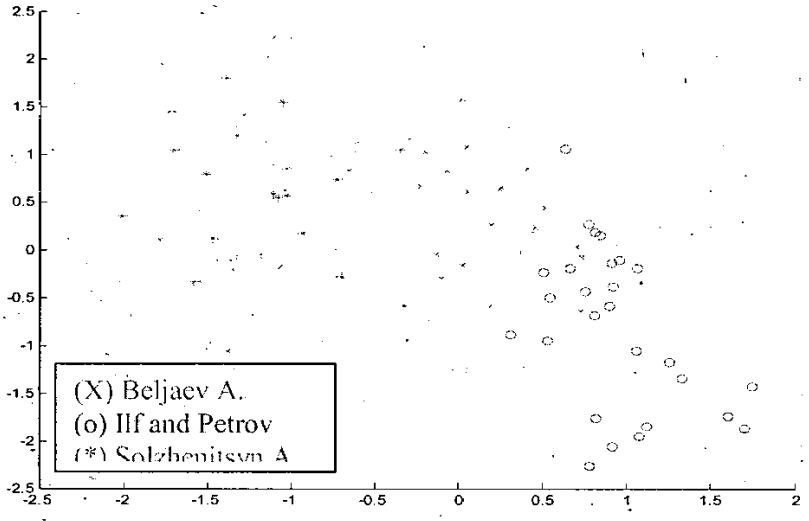

Fig. 1 Style space for the found set of characteristics

7. Conclusions. Merits and demerits of the approach

The suggested method of search of author's invariant is unique (searches of similar methods in the Internet have not crowned success) and possesses a number of advantages in comparison with standard intuitive selection of characteristics. First of all, due to atomic structure of a genetic line, it is universal. The described search of a set of characteristics allows working with any character set, so with any sign language. Theoretically the method can process any information (not only texts) - for example, it is enough to copy the block of the lexical analyzer and possible to establish authorship of computer programs. That manual search of author's characteristics is very labor-consuming and unreliable procedure.

In case of a successful finding of a universal set of characteristics it is possible to process any number of authors and texts. It is enough to renew search with an additional set of products and to estimate, as they cooperate with earlier present.

With the help of a universal set of characteristics it is possible to process the big files of the information quickly enough. The given opportunity can be used in search on the Internet, or the organization of electronic libraries. The method can make automatic clusterization of unknown texts by its authors.

The found sets of characteristics can give interpretation in simple human language, or as patterns. Thus, the program is capable to receive the descriptions of author's style competing to descriptions of stylists. Only as against rules of stylists they will be really unique for the author (are statistically proved).

One of serious lacks of a method is the impossibility of forecasting of successful result. Genetic search on the set of texts can find never a good variant for division of characteristics. There is no criterion of, whether in a correct direction search goes, whether truly it does gallop, whether necessary saves the information on researched space.

Other problem of a method is its labor input. The number of the loaded texts which directly influences quality of search, demands the big resources from the computing system (great volume of memory and the powerful processor). For a finding of rather universal characteristics it is necessary to process not one ten mbyte of texts that it was possible to declare their universality with confidence. As to memory at present we specially did not examine a question of its optimum distribution, but it is potentially possible to develop a variant with smaller loading.

As a whole, the results received during research give the basis to believe, that unconscious characteristics of author's style exist, and genetic algorithms possess necessary potential for their finding.

References

[1] Farringdon. Jill M. A brief introduction to cusum analysis // How to be a literature detective: Authorship attribution;

[2] Tweedie F.J.; Singh, S.; Holmes, D.I. Neural Network Applications in Stylometry: The Federalist Paper // Computers and the Humanities 30, 1996, p. 1-10;

[3] Goldberg D.E. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading, 1989;

[4] Isaev S. Genetic algorithms – evolution methods of search;

[5] Matthew W., GALib: A C++ Library of Genetic Algorithm Components, version 2.4, Documentation Revision B, – Massachusetts Institute of Technology, August 1996. – 104p.