Study of folklore and mythological traditions using intellectual data mining

Article *in* Pattern Recognition and Image Analysis · December 2009

DOI: 10.1134/S1054661809040099

4 authors, including:

Svetlana Borinskaya
Vavilov Institute of General Genetics
58 PUBLICATIONS **603** CITATIONS
SEE PROFILE

Anna Victorovna Kuznetsova
Russian Academy of Sciences
41 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Oleg V. Sen'ko
80 PUBLICATIONS **732** CITATIONS
SEE PROFILE

Study of Folklore and Mythological Traditions Using Intellectual Data Mining

Yu. E. Berezkin[a], S. A. Borinskaya[b], A. V. Kuznetsova[c], and O. V. Sen'ko[d]

[a] *Peter the Great Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy of Sciences, Universitetskaya Embankment 3, St. Petersburg, 199034 Russia*

[b] *Vavilov Institute of General Genetics, Russian Academy of Sciences, ul. Gubkina 3, Moscow, 119334 Russia*

[c] *Emmanuel Institute of Biochemical Physics, ul Kosygina, Moscow, 119334 Russia*

[d] *Dorodnicyn Computing Centre, Russian Academy of Sciences, ul Vavilova 40, Moscow, 119333 Russia*

e-mail: senkoov@mail.ru

Abstract—This work is devoted to the development and substantiation of intellectual data mining as applied to studying folklore and mythological traditions. The approach is based on use of the functions of distance between traditions. The examples of application of the methods developed to investigate the interrelation between folklore traditions of the American continent are considered.

Key words: Intellectual data mining, anthropology, cluster analysis.

DOI: 10.1134/S1054661809040099

INTRODUCTION

The goal of this work is to develop and substantiate intellectual data mining for effective use in investigation of folklore and mythological traditions by using the set of the motifs presented. The motifs are repeating images, episodes, or their combinations of maximal length encountered in narratives of two and more traditions.

The representative database containing information on the motif occurrence was developed in the Peter the Great Institute of Anthropology and Ethnography (Kunstkamera), Russian Academy of Sciences, in 2007 and included data on the occurrence of 1355 mythological motifs in 337 traditions (in February 2009 it had 1463 motifs in 453 traditions). Moreover, the presence or the absence of each motif in the literature and archive sources is recorded in binary form.

The following goals of mining were set:

—estimation of the mutual closeness of traditions by the whole set of motifs;

—revelation of the tradition groups that are homogeneous by the motif occurrence;

—estimation of statistical reliability of differences by occurrence between the prespecified tradition groups by the whole set of the motifs presented in them;

—revelation of some motifs or their combinations showing reliable differences by occurrence in the prespecified tradition groups. The results of mining should help answer questions about the ways of peopling of various territories, directions of ancient migrations, as well as genesis of cultures and civilizations.

METHODS FOR ANALYSIS

Pairwise Closeness between Traditions

Two traditions T_j and $T_{j'}$ can be related by the closeness function $S(T_j, T_{j'})$ showing the closeness of their motifs. Mathematically, the tradition T_j can be interpreted as a random function t_j specified on the set of motifs and taking values 0 and 1 depending on whether the presence of the motif is recorded or not. One should stress that the availability of 0 in a certain position of the tradition does not indicate the obligatory real absence of the motif because of insufficient understanding of some traditions. The latter fact does not allow for using standard Euclidian and Hemming metrics as the closeness functions since they assume summing up by all components, which would lead to high closeness between weakly studied traditions. In this connection, alternative closeness functions were suggested. Let us assume that the total number of motifs in the question base is N, then

(1) Function $S_k(T_j, T_{j'}) = 1 - 0.5[K(T_j, T_{j'}) + 1]$,

where $K(T_j, T_{j'}) = \dfrac{\sum_{i=1}^{N}(t_j(i) - \hat{t}_j)(t_{j'}(i) - \hat{t}_{j'})}{\sqrt{\sum_{i=1}^{N}(t_j(i) - \hat{t}_j)^2}\sqrt{\sum_{i=1}^{N}(t_{j'}(i) - \hat{t}_{j'})^2}}$

Received October 16, 2008

is the common (Peirce) coefficient of double correlation between binary functions t_j and $t_{j'}$. Here $\hat{t}_j$ is the average value of t_j found on the whole set of motifs;

(2) Function $S_C(T_j, T_{j'}) = 1 - C(T_j, T_{j'})\frac{1}{N}$, where $C(T_j, T_{j'}) = \frac{m_j^0(\hat{t}_{j'}^0 - \hat{t}_{j'})^2 + m_j^1(\hat{t}_{j'}^1 + \hat{t}_{j'})^2}{(1-\hat{t}_{j'})\hat{t}_{j'}}$ is the value of the chi-square criterion in testing the distribution equality of function $t_{j'}$ values in the groups formed by the indicator function t_j. Here, the m_j^0 is the number of motifs with $t_j = 0$;

m_j^1 is the number of motifs with $t_j = 1$;

$\hat{t}_j$ is the average value of t_j by the whole set of motifs;

$\hat{t}_{j'}^0$ is the average value of $t_{j'}$ by all motifs with $t_j = 0$;

$\hat{t}_{j'}^1$ is the average value of $t_{j'}$ by all motifs with $t_j = 1$.

Both criteria are measures of the similarity of the t_j and $t_{j'}$ distributions widely used in statistics. Along with this, the value of the above closeness functions can be affected by the number of presented motifs in the traditions to be compared, often associated just with the degree of their comprehension. To illustrate the aforesaid, we note that by randomly and independently excluding motifs in two very close traditions we may get a situation in which the subsets have a small area of intersection or do not have it at all.

In order to estimate the effect of the degree of filling (fullness below) of traditions with motifs on the values of the closeness functions between them, we have undertaken the following study. We randomly chose 10% from all 56 616 possible pairs for further analysis. For these chosen pairs, using correlation analysis, we studied $K(T_j, T_{j'})$ versus m_j, $m_{j'}$, and $\sqrt{m_j m_{j'}}$, where m_j, $m_{j'}$ are the numbers of registered motifs for the tradition T_j, $T_{j'}$.

The correlation coefficients characterizing the linear dependence of $1 - S_k(T_j, T_{j'}) = 0.5[K(T_j, T_{j'}) + 1]$ on the parameters are the following: m_{left} is fullness of the tradition in the left-hand position of the chosen pair; m_{right} is fullness of the tradition in the right-hand position of the chosen pair; and $\sqrt{m_j m_{j'}}$ and m_{rand} randomly equiprobably chosen from the pair $(m_j, m_{j'})$ are shown in Table 1. It is seen that a relatively weak but statistically significant linear relation between the distance between traditions and the degree of their comprehension does exist.

Moreover, the explicitness of the dependence of $K(T_j, T_{j'})$ on the degree of comprehension turned out to decrease with increasing the latter. Table 2 shows the correlation coefficients calculated by the tradition pairs presented by the number of motifs in each not less than 50, 70, and 100.

Table 1. Correlation of the distance between traditions and fullness

m_{left}	m_{right}	m_{rand}	$\sqrt{m_{\text{left}} m_{\text{right}}}$
0.17	0.15	0.16	0.26

Table 2. Correlation of the distances between two traditions and their fullness for different fullness intervals

	m_{left}	m_{right}	m_{rand}	$\sqrt{m_{\text{left}} m_{\text{right}}}$
≥50	0.11	0.09	0.11	0.15
≥70	0.10	0.08	0.10	0.13
≥100	0.06	0.06	0.05	0.08

The dependence of $1 - S_k(T_j, T_{j'})$ on fullness m_{rand} is shown in Fig. 1.

The tangent of the slope angle of the straight line corresponding to the linear dependence of $K(T_j, T_{j'})$ on m_r is 7.2×10^{-5} per motif (less that 0.01 per 100 motifs). In other words, when comparing the closeness of traditions T_1 and T_2 to the tradition T, the consideration of the fullness effect allows for explaining $K(T_1, T) - K(T_2, T) \le 0.01$ at $m_1 - m_2 = 100$. The standard deviation is about 0.0495. From here we may conclude that the variation of the coefficients $K(T_j, T_{j'})$ can be an effective tool to reveal the differences between traditions by motif composition.

Revealing Homogeneous Tradition Groups

In order to reveal the tradition groups with a similar mode of occurrence of mythological motifs, the widely known method of hierarchical grouping was

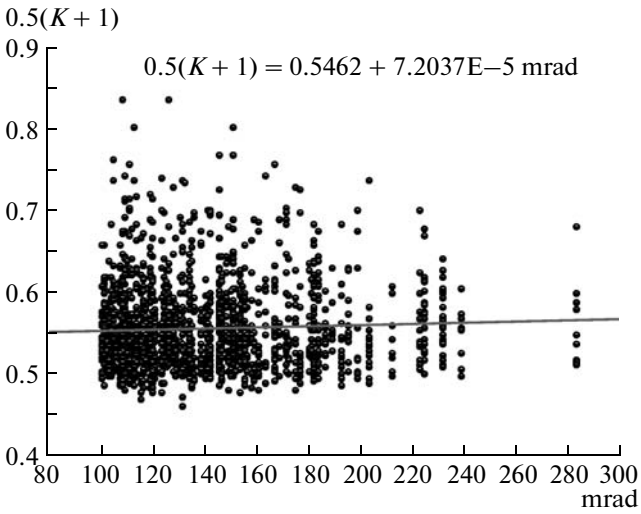

Fig. 1.

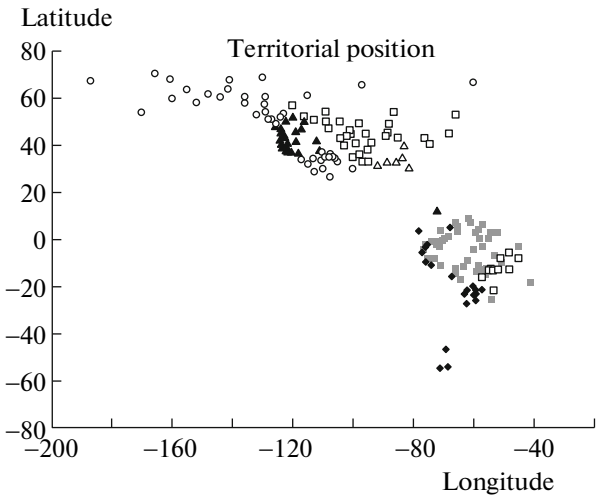

Fig. 2. Territorial distribution of clusters of close traditions for the American continent.

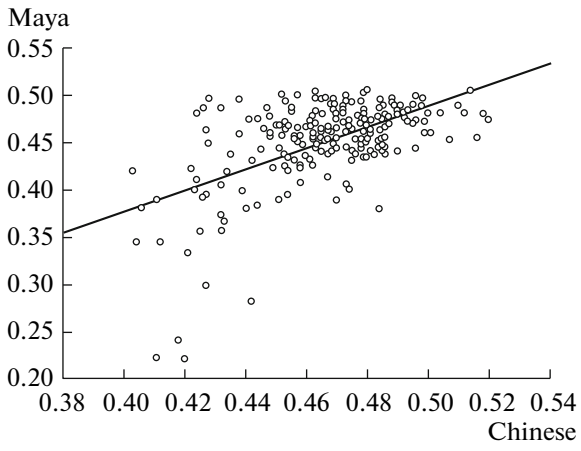

Fig. 3. The dependence of the distance to the set of traditions of the Mayan language group versus the distance to the folklore tradition of China.

used. At each step, the clusters with the maximal value of the averaged (by all object pairs from different clusters) closeness function were united. Studies have shown the traditions in clusters obtained according to the similarity of mythological motifs appear, as a rule, also to be close geographically. Territorial distribution of clusters for the American continent is shown in Fig. 2.

It is seen that, with few exceptions, the clusters are geographically compact groups.

Statistical Reliability of Differences between Tradition Groups by the Set of Motifs

In order to estimate the statistical reliability of the difference between the groups, the permutation test was used. Let m be the total number of traditions in the groups under comparison. m_1 is the number of traditions in the group G_1, m_2 is the number of traditions in the group G_2. The value of the quality functional $F_M(G_1, G_2) = \sum_{j_1 \in G_1 j_2 \in G_2} \sum S(T_{j_1}, T_{j_2}) - \frac{2m_1m_2}{m(m-1)} \sum_{j_1=1 j_2=j_1+1}^{m} \sum^{m} S(T_{j_1}, T_{j_2})$ using real data is compared to the value of the same functional on the sampling obtained from the initial real sampling by random permutations of the indicator function of classes. The share of permutations for which F_M exceeds F_M using real data is taken as the statistical measure of reliability of differences (p-value). One should note the high degree of substantiation of usage of the permutation test according to which the mode of probabilistic distribution is not taken into account and the size of sampling is unlimited [2]. The results of the usage of the method under consideration in this division to compare language groups of native Americans are shown in Table 3. In the table cells, D_1 is the average distance between traditions inside group 1; D_2 is the average distance between traditions inside group 2; D_{12} is the average distance between traditions of groups 1 and 2; p are values calculated by the permutation test. It is seen from the table that the permutation test has not revealed any meaningful differences between geographically close Algonquian and Sioux and Southern Utah Aztec and Mayan language groups.

Table 3. Difference between two linguistic groups of traditions by the set of motifs

Group 1	Group 2	D_1	D_2	D_{12}	p
Algonquian	Sioux	0.37	0.35	0.37	0.14
Eskimo-Aleuts	Caribbean	0.34	0.39	0.47	0
Southern Utah Aleuts	Maya	0.38	0.32	0.34	0.51
Chichewa	Macro-Tupi	0.39	0.39	0.43	0.002
Maya	Chichewa	0.32	0.39	0.41	0.025

Table 4. Relation between Mayan language group and some non-American traditions

Tradition	$K(R_1, R_2)$
Handza, Sandawe (East Africa)	0.31
Chukchi	–0.17
Baikal–Amur Evenki	0.01
Ainu of Sakhalin and Hokkaido	–0.09
Papuan in New Guinea	0.31
China	0.57

Difference between Groups by Individual Motifs

In order to quantitatively estimate the degree of differences in the motif presentation in groups G_1 and G_2, the functional $F_Q^U(G_1, G_2) = m_l(\nu_{1i}^l - \nu_1)^2 + m_r(\nu_{1i}^r - \nu_1)^2$ is used, where ν_1 is the share of objects from the group G_1 in the general set of traditions $G_1 \cup G_2$, ν_{1i}^l is the share of objects from the group G_1 inside the set of traditions for which the i-th motif was not recorded, ν_{1i}^r i is the share of objects from the group G_1 inside the set of traditions for which the i-th motif was recorded, and m_r, m_l are the numbers of traditions with the recorded and unrecorded i-th motif. By analogy with the estimation of statistical reliability of differences between tradition groups, the permutation test was used to estimate the statistical reliability in this case.

Estimation of Closeness of Two Groups of Folklore Traditions by Degree of Correlation of Distances between Them

Calculation of the linear correlation coefficient $K(R_1, R_2)$ between two respective functions $R_1(T)$ and $R_2(T)$ specified on a set of traditions T can be an additional way to assess the closeness of two groups of folklore traditions G_1 and G_2. In this case $R_i(T) = \frac{1}{m_i}\sum_{j_i \in G_i} S(T_{j_1}, T)$, where m_i is the number of objects in the group G_i. One should note that the usage of the coefficients $K(R_1, R_2)$ to assess the degree of similarity of traditions allows us to a great extent to avoid the distorting effect of the tradition fullness as well as to show more effectively the statistical reliability of the closeness and explicitly present the study results. Figure 3 shows the distance to the set of traditions of the Mayan language group as a function of the distance to the folklore tradition of China obtained on the set of all American traditions. The correlation coefficient $K(R_1, R_2) = 0.57$. Table 4 shows the values of the coefficients characterizing the relation between Mayan and some non-American traditions calculated in a similar way.

CONCLUSION

This work presents a number of methods for statistical analysis of information on the occurrence of different mythological motifs in folklore and mythological traditions developed in terms of the general approach based on the function of paired looseness of tradition. The performed studies have shown the significance of the territorial factor for closeness of mythologies. They also prove the high stability of the analysis results in database modification.

REFERENCES

1. *Yu. E. Berezkin, Myths Colonize America. Areal Distribution of Folkloric Motifs and Early Migrations into the New World* (Izdatel'stvo OGI, Moscow, 2007) [in Russian].
2. O. V. Sen'ko, "Permutation Test in the Method of Optimal Partitions," Zh. Vych. Matem. i Matem. Fiz., No. 9, 1438–1447 (2003).
3. Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko, *Recognition. Mathematical Methods. Software System. Applications* (Fazis, Moscow, 2006) [in Russian].

Yurii E. Berezkin, doctor of history, head of the Department of American Nations, Peter the Great Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy of Sciences. Scientific interests: America peopling, setting of early civilizations of the Old and New World, developing of catalogues of folklore and mythological motifs.

Svetlana A. Borinskaya was born in 1957. In 1980 she graduated from the Faculty of Biology of Moscow State University. She received her candidate's degree in biology in 1999. She is a leading researcher of Vavilov Institute of General Genetics. Scientific interests: human genetics and adaptation, biological, social and cultural evolution, cross-culture researches.

Anna V. Kuznetsova was born in 1961. In 1986 she graduated from the 2nd Moscow Medical Institute. From 1991 to 1994 she attended post-graduate courses and received her candidate's degree in biology in 1990. She is a senior researcher of Emmanuel Institute of Biochemical Physics. Scientific interests: pattern recognition, intellectual methods of data mining, developing and forecasting algorithms in medicine, and biological and medical applications in medicine and other fields.

Oleg V. Sen'ko was born in 1957. In 1981, he graduated from the Moscow Institute of Physical Technology, in the years 1985–1989 he attended post-graduate courses and received his candidate's degree in 1990. At present he is a leading researcher at the Dorodnicyn Computing Centre, Russian Academy of Sciences. In 2007 he became a full doctor of physics and mathematics. Scientific interests: data mining, mathematical models of pattern recognition, classification, forecasting, and practical applications in medicine and other fields.