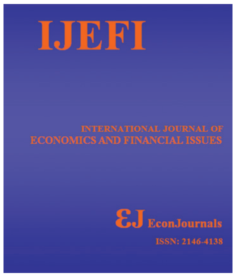


International Journal of Economics and Financial Issues

ISSN: 2146-4138

available at http: www.econjournals.com

International Journal of Economics and Financial Issues, 2016, 6(S8) 225-233.

Special Issue for "Fundamental and Applied Research in Economics and Management: New Perspectives"

Highly Pertinent Algorithm for the Market of Business Intelligence, Context and Native Advertising

Anna I. Guseva[1]*, Vasiliy S. Kireev[2], Stanislav A. Filippov[3]

[1]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation, [2]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation, [3]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation. *Email: aiguseva@mephi.ru

ABSTRACT

This article presents the study results of the business intelligence (BI) markets, the promote products on social media, and a new method for increasing the information pertinence in the scientific recommender systems (RSs), scientific information systems, analysis of the RSs that contain information about scientific publications, is represented. The prospects of using this method in the BI systems, content management systems for native advertising systems to find content on the Internet and assessed the current state of the market such systems.

Keywords: Context and Native Advertising Market, Business Intelligence Market, Highly Pertinent Algorithms, Recommender Systems
JEL Classifications: A11, M30, M37

1. INTRODUCTION

In modern conditions of the increased competition for the final consumer in all economy sectors, the special attention is paid to the personalization of the market offer. Who can predict more accurately the consumer intention, or form the needs as closely as possible to his common interests, the one gets a significant advantage. One of the measures to evaluate the quality of the generated proposals is to meet the information needs of the user unspoken-pertinence (Landia and Anand, 2009; Ivanova et al., 2011; Palchunov and Uljanova, 2010). The characteristics measurement accuracy is the limiting factor for the development of the real pertinence of proposals, and therefore the development of this topic is extremely important and perspective task.

The problems arising when forming the user inquiries can be connected with the ignorance of a set of the keywords that are unambiguously defining semantics of the required documents, lack of sufficient experience and the qualification of the search queries formation, or with absence accepted and the settled terminology in the interesting subject domain. Quite often, the person who is carrying out search has the approximate idea of the subject interesting him. All this causes relevance and the importance of the researches directed on the solution of one of the key problems of the information search-a problem of adequate display of information needs of users and, as a result, the search pertinent increase.

As the main potential areas for introduction of the proposed method of the pertinent and software-hardware increase the following solution is considered: The scientific recommender systems (RSs), the scientific information systems, the analytical RS, the business intelligence (BI) systems, the content control systems for the native advertising, the systems for the content finding on the Internet.

Context RSs are used in mass media to increase the depth of the users' viewings and their involvement. The scientific RSs-in the scientific organizations for creation of the intersubjective communications, in archives and libraries-for comparison of the saved-up information and navigation among large volumes of literature. Potential consumers of the RSs are the multinational corporations and large industrial enterprises, scientific organizations, mass media and news agencies.

The BI system is a set of methods and tools to translate raw data into meaningful, usable form. This data is used for business analysis. The BI technologies handle large volumes of the unstructured data, to find strategic business opportunities. The RSs integrated with the BI, from the analysts point of view issue the recommendations on the selection of the analysis tools in the formation of the data reporting spend supplement with a similar structure, carry out a comparison of the accumulated data.

The RSs development is very perspective as the BI products market in recent years is increasing. For example, the revenue growth of the largest Russian companies: I-Teco, RDTECH, the Forecast, CROCK, Helios IT on the Russian BI market in the 2013 relative to 2014 ranged from 8% to 28%, and in some cases (Force, Bars Group) reached 60%. According to International Data Corporation in April 2013, the world BI-services spending will increase on average by 14.3%, so that in 2016 they will amount to $70.8 billion. According to Gartner, by 2016, the market systems and analytical BI platforms will remain one of the fastest growing segments of the global software market. The average annual rate of the market growth of 7% in the period from 2011 to 2016. By 2016, the market size could reach $17.1 billion. At the same the BI market according to Gartner, April 2012, when considered in conjunction with the data warehousing and analytics into CRM, is growing even faster for 9%/year.

The RSs use for the native and context advertising is also very perspective direction of the pertinent increase method use. In this case, from the positioning point of view, the recommender system gives the improved masking of the recommended content among other materials, unites pertinently materials among themselves that lead to increase in time of viewing of the materials, which are removed in the right place and in due time.

The global market for native advertising is growing, and by HIS is estimated to reach 53 billion dollars by 2020. According to the US study IPG Media Lab and Sharethrough, in the case of the native advertising users are much more likely to associate themselves with the brand (42%) are more likely to notice (+52%) and share (+68%), communication over traditional banners (Outlook for national advertising in Russia, 2016). Native advertising is 1.5 times more stimulated to buy on the Internet. According to the eMarketer forecast, by 2017 more than half of the US brands and advertising agencies will invest in the native advertising on mobile devices by 25% more.

Separately it is worth mentioning so-called content discovery platforms (means to find certain content). One of the most popular representatives is the Outbrain and Taboola (Digital Advertisers Barometer, 2015). The Outbrain is used by large brands both for the production advertising, and for the recommendations publication from the sites. Among clients are such companies as TIME, CNN, Fast Company. According to the information from the company, this product is used more than on 35,000 sites and issues over 250 billion recommendations and 15 billion viewings of pages per month. These recommendations are seen by over 87% of the Internet users in the USA.

2. THE PROBLEM STATE

The work result assessment of the RSs can be carried out from the relevance and a pertinence point of view (Guseva et al., 2015). Relevance is defined as “the compliance to the received information to a query.” Pertinence is defined as compliance to the received information of the information requirement, i.e., pertinence is the compliance of the documents found the information retrieval system to information needs of the user irrespective of the fact how fully and precisely this requirement is expressed in the inquiry form. Pertinence is defined by the subjective perception of the person. Because of the pertinence subjectivity it is impossible to achieve exact coincidence: Any search engine is adjusted for information needs of the average, but not specific user. The searches insufficient pertinence can be caused by such reasons as the inquiries excessive decomposition: The user information need revealed in the form of very concrete inquiries series from 7 to 10 amount, or service by excessively broad inquiries: On one inquiry the subscriber receives from hundreds of thousands to hundreds of millions of documents, web pages though directly there corresponds to inquiry only the small part of information.

Pertinence of the information retrieval depends on two things: How precisely the formal request, drawn up by the user corresponds to its information needs, and how precisely the results of the issuance of the information system meet the formal request. Existing methods allow creating search patterns, with high relevance of answers, i.e., providing transmission of the information offers close within the meaning of the search query. Modern search engines have learned to reach sufficiently high values of this parameter. The numerical value of relevance depends on three parameters - the accuracy, completeness and ranking. Ranking is the order correctness, which presents a list of information retrieval. Accuracy is the proportion of the relevant resources among all the resources present in the issue, as fullness is the share of the relevant resources present in the issue among all relevant resources available on the Internet. However, from the user perspective the main criterion for the evaluation of information the search results are its pertinence.

The most important question of drawing up the pertinence information proposals is in the information systems of reference by virtue of the appointment. Most RSs operates explicit user profile (search queries and information about themselves), but about how to handle implicit profiles (collected on the basis of the analysis of implicit behavior-delta time of stay in a particular part of the text, the fact of copying, etc.) out of the question. While the commercial systems are actively implementing an implicit analysis of the system behavior, as in typical situations are expected (and it is already beginning to confirm the practice, for example, in the Amazon online hypermarket) that the recommender system can show information offers search results even before a person independently formulated the request, i.e. implement predictive search.

Since the pertinence is the user satisfaction with the results of information retrieval or selection of information units (IUs), which are given out by the recommender system, for the increase of this characteristic, when forming delivery it is necessary to predict interests of the user (information requirement). Thus, the current information needs of the user are understood as a set of such indicators, as to have idea of the events occurring at present; timely obtaining information on new achievements in the interesting subject domain; requirement of expeditious obtaining information necessary for decision-making. Retrospective information requirements are understood as information providing, for example, the beginning of performance of a new subject i.e., it is such information that most fully presents to the user the previous results of the researches on this subject.

The recommender system receives information about the user, first, at his registration in the scientific system. In that case, being guided by the specified interests, it is possible to pick up objects from appropriate sections. Thus, the text analysis can be necessary. The content filtration forms the recommendation based on the user behavior. For example, this approach can use retrospective information on viewings (what sources the user and characteristics of these sources reads). This content can be defined in the manual mode or is taken automatically based on other methods of similarity. However, this approach is rather unilateral, as the current preferences of the user can be not connected with previous. In addition, the problem of completeness and reliability of the specified information that leads to a low pertinence of the recommendation, which leans only on this way, can take place.

The next method is used in the trade systems and is based on the concept of "a market basket" in this case purchases in one check (trade operation) are analyzed, and there are most often found subject sets. Thus, if the user studies the IU at present, being on its page in system, he can recommend other units on the way "with these goods usually buy."

Unlike the method stated above, it is possible to consider behavior of other users in an explicit form. The main idea of this approach is in allocation of patterns of behavior definition to what of them the current user belongs, and then the recommendation formation based on the users average interests from the corresponding pattern. Patterns are defined on the basis of the analysis of the data containing in the system ravines. The collaborative filtration develops the recommendations based on the previous behavior model of the user. This model can be constructed only based on this user behavior or-that is more effective-taking into account behavior of other users with similar characteristics. When the collaborative filtration takes behavior of other users into account, it uses knowledge of group for the recommendations development based on the user's similarity. On the substance the recommendation are based on automatic cooperation of a great number of users and on allocation (a filtration method) of those users who show similar preferences or templates of behavior. As an example, we will assume that you create the website to offer it to visitors of the recommendation concerning blogs. Because of information from many users who subscribe for blogs and read them, you can group these users in their preferences. For example, you can unite in one group of users who read the same blogs. According to this information, you identify the most popular blogs among what are read by participants of this group. Then - to the specific user of this group-you recommend the most popular blog from on what it is not signed yet and which he does not read.

3. PRINCIPLES OF THE RSs CREATION

The RS represent software and methods which appointment is forecasting the user behavior concerning the object of the information search and formation of the recommendations for objects which it did not meet yet (Ricci et al., 2011; Guseva et al., 2016). The created recommendations help users with various decision-making processes, for example, what scientific article to choose, what discipline to study, etc. Such recommendations are under construction based on the characteristics of these objects and (or) the user profile. Formation of the recommendations possibly only because of the obtained data. The data used by the RS belong to three types of objects: Elements, users, transactions, i.e., relations between users and elements.

The scientific article, the book and the patent can be elements, or objects of the information search. The element can consist of the primary terms, i.e. IUs - the word, the author, the name of style, etc. Elements can be characterized by their complexity, value or usefulness. Value of the element can be positive if the element is useful to the user or negative if the element is not necessary, and the user made the negative decision choosing it.

The system user can have tastes and preferences. Information about the users can be collected in different ways. In the user model there is always the user's profile, obvious or implicit. The obvious profile is formed by means of filling of questionnaires, polls, etc. In this case, the user is personalized. The implicit profile of the user is formed at the expense of the accounting of its actions on the site.

Under the transaction is understood the recorded interaction between the user and the PC. For example, the transaction log may include the reference to the element selected by the user and context description (for example, a user request for information) to generate a recommendation. Such transaction may also reflect the presence of the explicit feedback that the user has provided, as an evaluation of the selected item.

Actually, estimates are the most popular form of the transaction data, which are collected by the RS. These estimates can be collected obviously or implicitly. In an obvious set of estimates of the user is asked to provide the opinion on an element on a rating scale. In the transactions collecting implicit estimates, the system seeks to output opinion of the user based on his actions. In dialogue systems, i.e., the systems supporting interactive process, more advanced models of transaction. I.e. the user can request the recommendation, and the system can make the list of the offer. Thus, the RS based on the additional user preferences reports to the user the best results.

In this article, it is offered to investigate the implicit profile of users that is formed in scientific and educational systems and to formulate rules of its processing for specification of pertinent answers in the field of scientific and educational information.

The RSs are classified as the content, collaborative and hybrid. At the content filtration, profiles of users and objects are created. Profiles of users can include demographic information or answers to a certain set of questions. Profiles of objects can include names of disciplines and training courses, names of teachers etc., - depending on the object type.

The content filtration is focused on the accurate classification as the users, and the objects appearing in the information offer. In the specified case, the direct compliance between the users and the objects based on their characteristics is established. In general, the strategy well works in areas with final and rather small amount of the criteria of the assessment following from the nature of things at big flows of the information and allows a large number of criteria at a small information stream. The main problem is the classification and creation of new information offers.

At the collaborative filtration information on the users' behavior in the past-for example, the information on orders or estimates is used. In this case does not matter, with what types of objects work is conducted, but thus implicit characteristics which would be difficult to be considered at creation of a profile can be considered. The main problem of this RSs type is the "cold start:" The data absence on the users who recently appeared in the system or objects.

The collaborative filtration is increasingly based the analysis of the user trajectory: What links moved, what and how appreciated what set a bookmark, where took the social buttons, what he was looking for, and implicit actions-which lingered longer, where less, analysis of cumulative actions and behavior, habits analysis, etc. This strategy is considered the most promising to date, including regularly held competitions among research groups to find the best algorithms.

4. THE RECOMMENDED APPROACH

In this work, the process of the obtaining recommendations by the end user breaks into several steps. As the main sources of data for the user, it is possible to consider his personal information, which he specifies in the questionnaire at registration on the site, and the ravine of activity, which automatically registers system, preceding from what actions the user executed on the site.

When forming the user profile, both obvious, and implicit, the used data sets contain in the structure not only quantitative, categorical and binary variables, but among them also can be presented other types-the text data (Kireev and Kuznetsov, 2016). Different references, questionnaires, recommendations and other unstructured text data can concern to them. Text data have the special nature of an origin and demand special approach to processing and a further clustering and classification. This stage is called as preprocessing data. The preprocessing stage includes removal of a punctuation and stop words, reduction of symbols to the lower register, reduction of words to a normal form and data binarization.

Recommendations are issued because of the user's belonging to some group of users with similar interests. Definition of such classes will help to create the need of the user more precisely, i.e., to raise a pertinence. This operation is called clustering. Reference of the new user to the most suitable class is carried out by the means of the classification operation. As a result, knowing what class the user belongs, and in what other participants with similar interests are interested, to the user the recommendation is issued. The general scheme is shown in Figure 1.

Despite all the uniqueness and singularity of each Internet user, there are a number of researches proving that all users can be divided into some classes (Kireev and Kuznetsov, 2016). For the subsequent classification, it is required to take behavioral patterns of users that can be received by application of the cluster analysis methods from the basic data. These methods have empirical character, and both work of many of them, and for the quality assessment of the received decisions requires promotion of the hypotheses both of the classes quantity, and about their preliminary structure (or physical interpretation).

4.1. Ensemble of the Clustering Algorithms

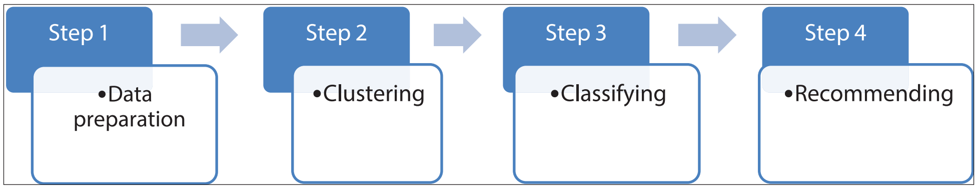
The data taken from the user existent profile of the RSs need to be structured, classified, and subjected to the careful analysis. In this case, the cluster analysis carries out segmentation of the certain associations of uniform elements given through allocation that are considered as the independent units possessing certain properties (Kireev and Bochkaryov, 2016).

Stability of decisions in the clustering problems can be increased thanks to the algorithms ensemble formation and construction with its help of the collective decision based on opinions of the ensemble participants, where under the algorithm opinion is meant its option of splitting data into clusters. These properties of the cluster analysis are especially actual during the work in areas with an intensive use of data when the subject domain is poorly formalized, for example, for the analysis of text documents, images, creation of implicit profiles of users of the RSs, etc. In case the considered area contains various types of data, it is necessary to apply not one certain algorithm, but a set of various algorithms to allocation of clusters. The ensemble (collective) approach allows to reduce dependence of the final decision on the chosen parameters of initial algorithms and to receive steadier decision even at large number of noise and emissions in data.

There are following main techniques of the algorithms ensemble acquisition: Finding of consensus splitting, i.e. the coordinated splitting at the available several decisions, optimum by some criterion and calculation of the coordinated matrix of the similarities/distinctions (co-occurrence matrix) (Kireev and Kuznetsov, 2016).

When forming the final decision the results received by various algorithms or by one algorithm with various values of parameters of different subsystems of variables, etc., are used. Applying the ensemble with the algorithms different set in accordance with their advantages and features, it is possible to create the most suitable scheme of clustering for a certain subject domain. Considering that fact that the choice of a concrete metrics of distances between objects is the important factor influencing

Figure 1: Sequence of actions for the recommendation issue



the clustering result it is possible to increase efficiency of the cluster analysis significantly.

The algorithms ensemble offered by the authors unites two considered above approaches and represents a combination of the consecutive algorithms of K-averages, each of which offers the splitting based on the changing metrics, and the hierarchical agglomerative algorithm uniting the received decisions by means of the special mechanism. The offered ensemble relies on the results of the preliminary research of the basic data that represent a small set of the objects marked by experts. Minimum necessary percent of the initial selection volume guaranteeing the set accuracy is the subject to further studying. For definiteness, in this work 0.5% are used that in case of increase in data volume, obviously, has to be the subject to revision.

On the first step each algorithm of K-averages, breaks data into clusters, using the distance metrics. Then, calculated the precision and the algorithm weight in the ensemble by the formula 1:

$$\omega_l = \frac{Acc_l}{\sum_{l=1}^{L} Acc_l} \qquad (1)$$

Where Acc_l - the l algorithm precision, i.e., the ratio of the correctly clustered objects number to the volume of the entire sample, and L - the algorithms amount in the ensemble.

For each received splitting the preliminary binary matrix of distinctions of the n × n size, where n is the amount of objects, necessary for definition, whether objects of splitting in one class are brought, is formed. Then calculated the coordinated matrix of distinctions, which each element is represented, weighed (with the use of weight from the formula 1) the sum of elements of preliminary matrixes. The received matrix is used as the entrance data for the algorithm of a hierarchical agglomerative clustering. Then by the means of standard practices, such as definition of jump of the agglomeration distance, it is possible to choose the most suitable cluster decision.

In this ensemble of the clustering algorithms five K-averages with the change of metrics (see formula 1), as one of the most demanded the clustering algorithms of huge data were used. For these algorithms were used such metrics, as:

- Euclidean distance $p\left(x,x^{'}\right)=\sqrt{\sum_{i}^{n}\left(x_i-x_i^{'}\right)^2}$ (2)
- Manhattan distance $p\left(x,x^{'}\right)=\sum_{i}^{n}\left|x_i-x_i^{'}\right|$ (3)
- Chebyshev's distance $p\left(x,x^{'}\right)=max\left(\left|x_i-x_i^{'}\right|\right)$ (4)
- Jacquard's coefficient $K(x,x^{'})=\dfrac{\sum_{i}^{n}x_i x_i^{'}}{\sum_{i}^{n}x_i^2+\sum_{i}^{n}x_i^{'2}-\sum_{i}^{n}x_i x_i^{'}}$ (5)
- Dynamic transformation timeline (DTW [dynamic time wrapping $\mathrm{DTW}(x,x')=\dfrac{\min\left\{\sum_{k=1}^{K}d(\omega_k)\right\}}{K}$ (6)

Where K - the length of the distance between x and x', which is calculated on the special matrix of transformations.

For the best partition into clusters, it is necessary, as mentioned above, to make a binary similarity\differences matrix L for each partition in the ensemble:

$$H_i = \{h_i(i,j)\} \qquad (7)$$

Where $h_i(i, j)$ is equal to zero, if the element i and the element j are in one cluster, and 1 if not.

The next step in the preparation of the ensemble of the clustering algorithms is to formulate a coherent matrix of binary partitions.

$$H^* = \{h^*(i,j)\} \qquad (8)$$

$$h^*(i,j)=\sum_{i=1}^{L} w_l h_l(i,j) \qquad (9)$$

Where w_l - the algorithm weight.

To form the best partition at the coordinated matrix the nearest neighbor algorithm was selected. The principal components method (principal component analysis) was chosen to reduce the dimension of the source data. As the number of component selection criteria, the Kaiser criterion was selected (eigenvalue >1 component).

The next step was revealed the accuracy of each algorithm by comparing the resulting partition into two clusters each cluster algorithm, tagged expert way.

After obtaining the values of accuracy of each algorithm, the algorithm weight was calculated by the formula 1.

4.2. Ensemble of the Classification Algorithms

To increase the accuracy of the classification algorithm work in (Kireev and Kuznetsov, 2016) it was offered the content classification by the authors.

The underlying hypothesis for this study was the assumption of the results types in the scientific and research activities. Among the scientific papers there are several groups that most full describing the results presented in these papers. The immediate assumption of the existence of such a structure can be recovered from the research results separation to the theoretical and practical component. In addition to this, you can select a particular type of work, which as a result is an overview of the results of other people's research activities. This approach is consistent with the results of the latest research in the field of scientific classification system users.

This hypothesis was formalized and adapted for the classification problems solution in the field of the RSs for the scientific social networks. On its basis, the set of the users' types who can be interested in receiving content of one of the results groups of the scientific activity was made. In addition, for each type phrases and keywords, which most precisely describe each of them, were allocated:

1. The observers-accumulates the result received earlier and knowledge of the subject domain, there is no own development-compares approaches and methods, gives pluses and minuses, generalizes the facts, draws conclusions.
2. The analysts-the purpose of this type is carrying out the comparative analysis, receiving a new method based on the earlier known approaches. Are limited only to theoretical aspects, form theoretical base for future experts-researchers.
3. The practical -this type tries to find practical application for the existing methods and algorithms. Receive the existing decision ready for the commercial operation.
4. These groups may be found in not only pure form, but also their symbiosis with other groups is possible.

To solve the classification problem the authors propose an approach based on the use of several different algorithms, voting independently. This approach can improve the final accuracy of the classifier and reduce the computational complexity of the algorithms. Such set of algorithm, as a rule, consists of a simple machine learning algorithms and conditionally called ensemble-voting algorithms. Gain simple qualifiers-the approach to solve the classification problem (recognition), by combining primitive weak classifiers in one strong. Under the classifier force in this case is meant the efficiency (quality) of the classification problem solutions.

At creation of the ensemble various combinations based on the indication of the various scales of algorithms, application of the identical algorithms with different parameters, segmentation of data under different algorithms, etc., are used. For the ensemble creation of the voting algorithms the following simple algorithms will be used:

- Naive Bayes qualifier;
- The basic vectors with a linear kernel method;
- The method of the basic vectors with a kernel of radial basic function of Gauss;
- Random forest (RF).

The "naïve" Bayes qualifier uses Bayes's formula for the probability calculation. The algorithm idea consists in calculation of the conditional probability of the object belonging to a class at equality of its independent variables to certain values. The purpose of classification consists in understanding the document therefore we need not probability, but the most probable class belongs to what class. The naive Bayes qualifier unites model with the rule of the decision. One general rule has to choose the most probable hypothesis; it is known as the a posteriori rule of the decision-making (MAP).

The support vector machine (SVM) method is a set of similar learning algorithms with the teacher, used for the classification problems and regression analysis (Krivenko and Vasilyev, 2013; Philippov et al., 2015). The method basic idea is translating the original vectors in high-dimensional space and search the separating hyperplane with the maximum clearance in the space. Two parallel hyperplanes are constructed on both sides of the hyperplane, dividing our classes. The divide is the hyperplane that maximizes the distance to the two parallel hyperplanes. The algorithm works on the assumption that the greater the difference or distance between the two parallel hyperplanes, the smaller the average error of the classifier. To solve nonlinear problems, a method of creating a non-linear classifier, which is based on the transition from scalar products to arbitrary cores, allowing to build nonlinear separators. The most common kernels are polynomial (uniform); radial basis function; Gauss radial basis function; sigmoid.

The RF qualifier uses the decisive trees ensemble. The decisive tree in itself does not provide sufficient accuracy for this task, but differs in speed of construction. The RF algorithm trains k of the decisive trees on the parameters, which are incidentally chosen for each tree then on each of tests vote among the trained ensemble, is taken. The idea that if to aggregate data from a large number of various weak algorithms, having reduced them in the uniform answer is the cornerstone of this algorithm creation, the result, most likely, will be better, than at one strong algorithm. The objects classification is carried out by the vote: Each tree of the committee refers the classified object to one of the classes, and wins against a class for which the greatest number of trees voted. The optimum number of trees is selected so that to minimize a qualifier error on tests selection. In case of its absence, out-of-bag error assessment is minimized: A share of the examples of the training selection that is incorrectly classified by committee if not to consider a voice of trees on the examples entering their own training subsample.

The offered model of the voting algorithms ensemble is based on the following approach.

1. For each algorithm, it is necessary to define a measure of the classification uniformity by each class on the training selection. At calculation of the uniformity measure for the classification algorithms will be used entropy (see Formula 1) for two independent casual events "x" with "n" possible states (from 1 to "n"), where "p" - the probability function:

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \tag*{10)}$$

2. For each article the probability of reference to a class calculates, and these values are serially multiplied by a uniformity measure for each class. The maximum value gets out of the received works.
3. In this case, the uniformity measure acts as the correction coefficient. The classification result is calculated by formula 2, where "*h*" - the algorithm uniformity measure for a class, and "*p*" - the probability of the object reference to a class:

$$C = argmax\ (h_i * p_i),\ i \in (\text{from 1 to 4}) \qquad (11)$$

4.3. The Automatic Offer Algorithm of the Information Proposal

The automatic formation algorithm of the information offer uses the behavioral data for satisfaction of the current implicit information need of the user.

Besides the user information need identification, it can be created, and in this case it is possible to use the long-term and short-term trends concept (further "top" and "trend"). It is possible to carry to the "top" IUs, what are looked for or looked through most often by other users during the long-term period (quarters and half-year), short-term - what are popular within several days and weeks. The viewport can be adjusted more precisely, depending on the sphere of the human activity to which the IUs belongs.

In this paper, to improve the pertinence is proposed to use a combined approach in which recommendations are formed taking into account the above. The length of the tuple should be recommended to take into account features of the human nature, which determine the number of objects that can operate at the same time people in 7 ± 2. The train will include the following parts:

1. The IU received on the users similar to the current one;
2. The IU received from the most often found sets;
3. The IU relating to long-term trends ("tops");
4. The IU relating to the short-term trends, actual ("trends").

Points 1 and 2 belong to the restored information requirement and are assumed making the greatest contribution to the recommended set (about 60-70%). Points 3 and 4 belong to the formed information requirement and have to occupy 30-40% of the train volume. The formation scheme of the final train is given below in Figure 2.

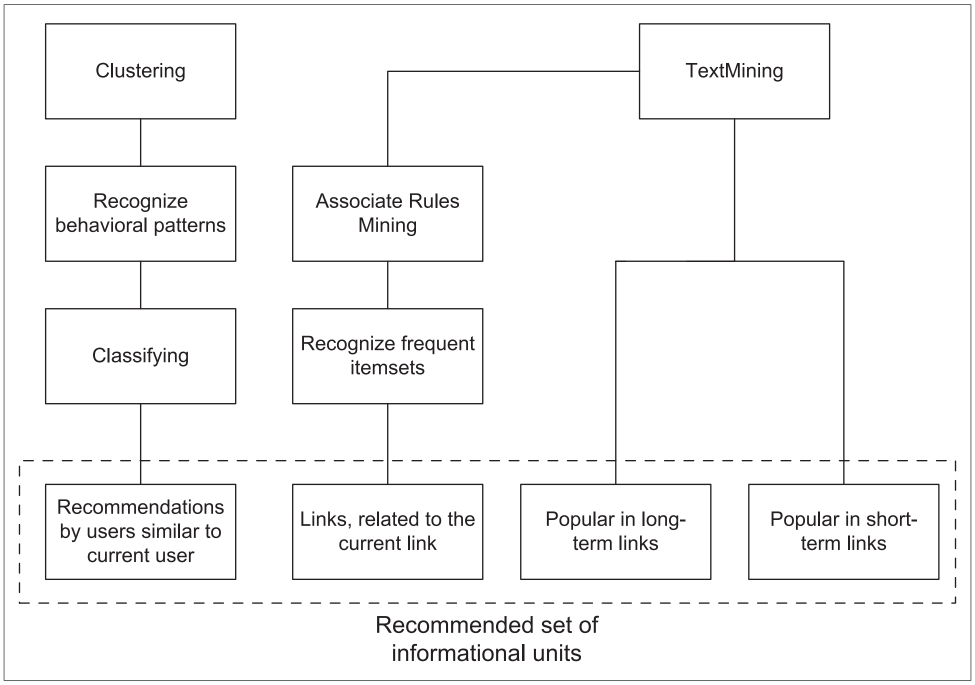
In the construction of the scientific or educational system, it is necessary to take into account that they have their own specifics. As can be seen from the description of the scientific networks of the above, the user of these systems interacts with the IUs - scientific results.

5. THE RESULTS AND DISCUSSION

For the testing and evaluation of the clustering algorithm ensemble was used the RapidMiner software, with which you can decide how the research (modeling) and the application (real) problems of data mining, including text analysis, media analysis, data flow analysis, which is suitable for testing ensemble clustering algorithms. As the data used for clustering with open data from the UCI web-site (UCI Machine Learning Repository, 2016).

Applying the offered the clustering algorithms ensemble, it is possible to increase the reliability of splitting data into groups. It is essential that this method can be applied in various areas. The clustering algorithms ensemble offered in these project levels shortcomings of metrics of distances for algorithms of K-averages, thereby increasing reliability of splitting. By means of the offered ensemble, it was succeeded to provide accuracy not <90%, and in each parameter not <80%.

Figure 2: Algorithm of the pertinent set formation


Clustering
TextMining
Recognize behavioral patterns
Associate Rules Mining
Classifying
Recognize frequent itemsets
Recommendations by users similar to current user
Links, related to the current link
Popular in long-term links
Popular in short-term links
Recommended set of informational units

For carrying out the experiment on the classification algorithm approbation of the basic data about publications on two Russian scientific public electronic resources - elibrary.ru and cyberleninka.ru were collected. The specialized program- crawler that collected the main data on articles-such as the name, author, keywords, summary, etc. was created for this purpose. The obtained data (data about more than 500 thousand articles) in a special way were pretreated for allocation of the main words and terms relating to the developed ontological model. After normalization of words, the part of articles relating to various subjects, such as physics, mathematics, informatics, economy and management was marked by means of the semi-automatic approach including the clustering method of k-means and an expert assessment. From the processed data for experiment, selection of scientific publications which volume made 5000 objects, in equal quantity for each direction was taken. For formation of the training and test selection, initial selection broke in a percentage ratio 70:30.

The offered algorithmic ensemble efficiency is rather high and reaches in the best case 88% of accuracy. Nevertheless, in experiment were considered only the part of possible values of the algorithms parameters and not all possible subjects of articles, therefore the additional researches in this area are required. The offered approach showed high efficiency in sense of the classification accuracy of the scientific content.

To test the automatic generation algorithm of the information offers was developed a prototype software package increasing pertinence of information as follows:

- The subsystem of collecting and storage of the behavioral data creating the user implicit profile based on these data.
- The subsystem of the implicit profile analysis allowing to cluster and classify the user profiles and to establish correlations between them, as well as to update this information, depending on the changes implicit user profile over time.
- The subsystem of the IUs' analysis using the subsystems of collecting and storage of the behavioral data.
- The subsystem of the information offer formation based on the analysis of the behavioral data that is carried out by the other subsystems.

The software package was implemented using the following programming languages and technologies. The server part: Linux OS, Ruby programming language, RubyonRails platform of the application creation, web server Nginx and Passenger, non-relational database (MongoDB), relational database of the intermediate data (MySQL) actual for 2014 versions.

The client part is realized with the use of HTML5, CSS3, the Javascript language with jQuery library, the Ajax technologies.

During the experiment, the software package provided the processing 100 million pieces of information, the size of 1024 bytes each. The volume data warehouse to collect behavioral data provided storage more than 10 billion implicit user actions for not <10 million users at the rate of 1000 activities on average per user. Indicators of performance enhancing method pertinence showed stability under load of up to 100,000 users/day with a peak before 1000 user actions per second. System response time at this load was not more than 0.9 s.

6. CONCLUSION

The study showed that increasing pertinence of the information in various information systems is extremely relevant and useful task. The proposed method of increasing pertinence and developed on this basis prototype software system proved its efficiency and prospects of use. Its successful implementation will significantly improve the quality of user requests and the documents quality. Application of such methods is extremely broad, this study is located the distinguished scientific and analytical RSs; content management system for native advertising system content.

7. ACKNOWLEDGMENT

This study was financially supported according to Federal Target Program "Research and development on priority directions of scientific-technological complex of Russia for 2014-2020" (grant No. RFMEFI57614X0068).

REFERENCES

Digital Advertisers Barometer. (2015), Sostav.ru. Available from: http://www.sostav.ru/publication/digital-advertisers-17101.html. [Last retrieved on 2016 Jul 01].

Guseva, A., Kireev, V., Bochkarev, P. (2015), Scientific and educational recommender systems. In: "Proceedings of the International Scientific-Practical Conference "Information Technologies in Education of the XXI Century, (ITE-XXI). Moscow: Knowledge.

Guseva, A., Kireev, V., Bochkarev, P., Smirnov, D., Filippov, S. (2016), The formation of user model in scientific recommender systems. International Review of Management and Marketing, 6(S6), 214-220.

Ivanova, O., Gromov, Y., Didrih, V., Polyakov, D. (2011), Fuzzy approach to determining the pertinence of search results and the choice of optimal query. Bulletin of Voronezh Institute of the Federal Penitentiary Service of Russia, 2, 49-54.

Kireev, V., Bochkaryov, P. (2016), Development of the clustering algorithms ensemble based on varying distances metrics. In: Data Analytics and Management in Data Intensive Domains, XVIII International Conference DAMDID/RCDL'2016. Moscow: Federal Research Centre "Information and Management" of the Russian Academy of Sciences.

Kireev, V., Kuznetsov, I. (2016), Development of algorithms ensemble in case of the solution of the task of statistical classification in recommender systems. International Journal of Applied Engineering Research, 11(9), 6613-6618.

Krivenko, M., Vasilyev, V. (2013), Metodyi Klassifikatsii Dannyih Bolshoy Razmernosti. Moscow: IPIRAN. p204.

Landia, N., Anand, S. (2009), Personalised tag recommendation. In Proceedings of the 2009 ACM Conference on Recommender Systems. Moscow: Publisher Moscow University.

Palchunov, D., Uljanova, E. (2010), Methods for automatic generation of search heuristics. Novosibirsk State University Journal of Information Technologies, 8(3), 6-12.

Perspektivy Nativnoj Reklamy v Rossii Outlook for National Advertising in Russia. (2016), Sostav.ru. Available from: http://www.sostav.ru/publication/perspektivy-razvitiya-native-programmatic-v-rossii-19422.html. [Last retrieved on 2016 Jul 01].

Philippov, S., Zakharov, V., Stupnikov, S., Kovalev, D. (2015) Organization of big data in the global e-commerce platforms. Ceur Workshop Proceedings. In: XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015). Vol. 1536. Obninsk: Federal Research Centre "Information and Management" of the Russian Academy of Sciences.

Ricci, F., Rokach, L., Shapira, D., Kantor, P. (2011), Recommender Systems Handbook. Berlin: Springer Science.

UCI Machine Learning Repository. (2016). Available from: https://www.archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. [Last retrieved on 2016 Aug 04].