Methods of Machine Training on the Basis of Stochastic Automatic Devices in the Tasks of Consolidation of Data from Unsealed Sources

© Valeriy O. Kuzminykh © Alexander V. Koval © Mark V. Osypenko
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
Kyiv, Ukraine
vakuz0202@gmail.com avkovalgm@gmail.com mark.osypenko@gmail.com

Abstract

The article considers development of algorithms and methods that increase efficiency of search relevant information on request from open information sources. At the same time particular attention is paid to the questions of consolidation of data for their further use in information-analytical and information retrieval systems. In work considered process of consolidation in general and the most common types of information retrieval. Considered the main ways of consolidation of the information from open sources, in particular with use of machine training of based on stochastic model.

Such tasks of consolidation of the information from open sources require processing of big volumes of data, so they possess beside specific properties that requires development of specific methods of their realization. In the article opportunities of construction of algorithms of search of relevant information from diverse sources on the basis of analysis of probability information are considered which defines the evaluation of presence of relevant documents in these sources.

For search of relevant information on request the approach is used which is constructed on use of the evaluations of probabilities presence of relevant documents in the sources of the information and subsequent increase of the amount of chosen documents from the most promising sources for further analysis of their relevance to the request.

There is offered structure of programmed stochastic automaton for supplying of choice of the sources of the information and algorithm of information retrieval the most probable on parameters of relevance from the set of the sources of the information on the basis of stochastic automaton. There is presented the example of testing of algorithm on specialized test environment. Described algorithm with use of stochastic automaton for data consolidation allows to develop the complex of software, supplies enough complete decision of tasks of consolidation of data for various systems which implement information retrieval from open sources of data various on composition and the kind of representation.

1 Introduction

Data consolidation is considered as the complex of methods and procedures directed to extraction of data from diverse sources, transformation to uniform format, in which they can be loaded in the storage facility of given or analytical system. Usually consolidation described as process of search, selection, analysis, structuring, transformation, storage, cataloguing and granting to the consumer of the information on given topics is understood. The task of the information consolidation is one of urgent tasks of processing of big volumes of data [1].

In the base of consolidation procedure process of the collection and the organization of data storage as, convenient from the standpoint of their processing on particular platform lies. Important feature of data acquisition on the basis of unsealed sources is instability of information fullness of these sources, absence of reliable a priori information about their content and his relevance, low accuracy and efficiency of expert evaluations of their condition and conformity of these sources to the topics and enquiries parameters [2].

2 Problem statement

Therefore, for data processing from unsealed sources of the information it is necessary execution of effective consolidation of data with use of specialized software supplying:

1. adaptive choice of the sources of the information the most relevant for enquiry conformity;
2. accumulation of the information of condition of the sources during performance of the enquiry;
3. the account of opportunity of change of condition of the sources at repeat enquiry;
4. analysis of perspective from the standpoint of sources relevance;
5. construction of information evaluation of sources promising from the standpoint of relevance.

At development of the model and algorithm of consolidation of the information from diverse sources with plenty of documents it is set the task of choice of maximum quantity of documents relevant to the enquiry at minimum quantity of processed (checked to relevance to the enquiry) documents. It has been made possible through the expense of revealing of the sources of the information the most appropriate to the enquiry and their further use at further choice.

3 The task of consolidation

There are a range of information resources presenting information retrieval objects:

- mass media are various sort news sites (RSS channels) and semantic sites (or electronic mass media versions).
- electronic libraries are distributed information system enabling to preserve reliably and effectively to use electronic documents through global networks.
- databases are the set of files organized by special image, documents grouped on topics and with spreadsheets which are united to groups.
- sites are Internet-resources devoted some organization, company, enterprise, particular problem or person. They differ by completeness of the information, from pure fact-finding, surface to highly professional, which lights all parties of activity.
- information portals are the groups of sites to which it is possible to take advantage of various service services. They can contain a various scientific, political, economic and other information, as well as electronic letter boxes, catalogues, dictionaries, directories, weather forecast, TV-program, exchange etc. As a rule, their updating of which takes place in a real of the time.

There are numerous models of information retrieval, on the basis of which modern information-analytical and information retrieval systems are built.

Among them it is possible to allocate following the most common types of the models of information retrieval [3]:

1. Boolean model,
2. The model of indistinct plenty,
3. Vector model,
4. Latent-semantic model,
5. Probability model.

Boolean the model uses the apparatus of mathematical logic and the theory of plenty [4]. Advantages of the model are simplicity of the model and her realization. Model deficiencies is complexity of construction of enquiries without knowledge by Boolean algebra, is impossible to rank automatically documents and to scale search.

Traditional model of indistinct plenty is based on the theory of indistinct plenty and admits a partial element accessory to plenty [5]. In such model all document file is described as the set of indistinct plenty of terms. Advantage of the model - opportunity to rank results, simplicity of realization, is unnecessary big memory size. Model deficiencies are big computational expenditures, then in Boolean model, and smaller accuracy.

In vector model [6] documents and users' enquiries are presented as n-dimensional vectors in -dimensional vector space. At the same time space dimension corresponds total of various terms in all documents. Advantages of the model are simplicity of construction and opportunity of results ranking. Model deficiencies - it is required big volumes of data processing. It is necessary accurate of coincidence with enquiry parameters.

Latent-semantic model [7] is constructed on the basis of latent-semantic analysis (Latent Semantic Analysis - LSA). This method of extraction and representation of the values of words dependent on context with the aid of statistical processing of big volumes of textual documents. Advantage of the model - less space dimension, than vector model, is not provided with an accurate coincidence of enquiry parameters, is not required difficult set-up. To the model deficiencies it is possible to attribute - plenty of calculations and absence of the rules of choice of dimension, from which results efficiency depends.

Specific place is taken by probability models. These models of search are based on application of methods of the theory of probability and use statistics which define probability of the document relevance to search enquiry [8]. In the base of these models the principle of probability ranking on decrease of probability of their relevance to the user's enquiry lies. These models it differs considerably from each other by computational procedures. Advantages of the model are opportunity of adaptation of the model during use, absence of dependence of volume of calculations from the amount of enquiries parameters, high efficiency at work with the sources of the information which are constantly updated. Model deficiencies are necessity of constant training of the system during work of the model and regarding low efficiency on initial stages of work.

To date there are no models which would have obvious quantitative and qualitative advantages of some above others [3,9], however the largest interest is presented by the decision constructed on the basis of stochastic approaches which allow successfully to realize the principles of machine training, successfully to be adapted to varied conditions, is simple in construction and realization.

To meet the challenges of search and consolidation of the information from unsealed sources the most promising is use of methods of machine training based on operational analysis of condition of the sources of the information during consolidation problem solving [3].

Machine training is considered usually as separate section of the theory of artificial intellect studying methods of construction of algorithms, capable to be adapted and to learn. Machine training is on the joint of mathematical statistics, methods of optimization and classical mathematical disciplines, but has as well its own specific features connected with the problems of computational efficiency, adaptation and conversion training. Almost none research in machine training does not dispense with experiment on model or real data verifying practical serviceability of method which are developed.

Use of stochastic model allows effectively to realize a machine training in two ways [10]:

1. Training with reinforcement (reinforcement learning). At the same time the role of objects is played by the "situation and accepted decision". Results are value of functional quality characterizing correctness of accepted decisions (reaction of environment), that is, averaged evaluation of relevance. Here essential role is played by the time factor. The examples of applied tasks: choice of documents from bibliographic and abstract bases, information retrieval in open information resources, for example, scientific RSS-channels, self-learning of robotized systems, etc.

2. Dynamic training (online learning). Specific features that precedents arrive with flow. It is required to make decision immediately on each precedent and simultaneously re-training the model of dependence in view of new precedents. Here essential role is played by the time factor that assumes opportunity of constant adaptation of the model of choice to changes of environment during work of the system.

The structure of interaction during consolidation contains following basic elements:

1. The sources of the information.
2. Parameters of the content of the enquiry.
3. Documents packs.

The sources of the information - it information resources, important and are considered in construction of particular information system, is considered in determined case of search of the information.

Parameters of the content of the enquiry in most instances are difficult terms simple both, parameters determining various signs of the time, places (occurrences, edition, storage etc.), and many other specifications determining features of determined enquiry.

Documents packs are particular sets of information units (articles, records, tables, reports, newspapers, journals, books, the theses of reports, news, reviews and other electronic data).

We will consider basic specifications of description of consolidation of the information pursuant to stochastic model of choice of the sources of the information on the basis of the theory of stochastic automatic devices [11]. For the evaluation of relevance of the sources of the information to the enquiries model is used on the basis of the theory of indistinct plenty [5], which takes into account communications of parameters of the enquiry and the sources of the information.

Such model which admits partial conformity between enquiry and the source of the information. This conformity is evaluated by the value within the range of [0,1]. It is formed on the basis of definition of coincidence between enquiry parameters and specifications of information units entering particular source of the information.

After execution of sample and the evaluations of a few units of the information from determined source information average pursuant to the amount of chosen documents units.

The evaluation of conformity *i*-th source of the information *j*-th parameter of the enquiry defines his partial conformity, she corresponds range [0,1] and can be determined as:

$$k_{ij} = \frac{1}{V}\sum_{l=1}^{V} q_{ijl} \qquad (1)$$

where:

q_{ijl} – quantitative evaluation of conformity q_{ijl} *l*-th document (for*l= 1,..., V)* with*i*-th source of the information*j* -thparameter of the enquiry at one sample for the evaluation relevanceof the source *;*

$q_{ijl} = 0$ - when *j*-th parameter is not present in *l*-th document *i* -th source of the information;

$q_{ijl} = 1$ - when *j*-th parameter is present in *l*-th document *i* -th source of the information.

V – volume of one sample from one source of documents should correspond following limitations:

$$V \ll S_i \quad \text{для } i=1, \dots, n \qquad (2)$$

where:

S*i* – the amount of information units (documents) in each of *n* the sources of the information, are considered for given enquiry.

Then the evaluation of relevance *i* -th source of the information will be defined as:

$$R_i = \frac{1}{m}\sum_{j=1}^{m} k_{ij} v_j \qquad (3)$$

where: $0 < v_i < 1$

v_j–weightcoefficientj-th parameter of the topic of the enquiry, are defined on expert evaluations or on the basis of priorities, can define independently the customer of query.

At such consideration the evaluation of relevance *i*-thsource to determined enquiry will be defined in the interval [0,1].

4 The algorithm that is based on stochastic automaton machine

For the organization of procedures of effective choice of information sources of relevant pursuant to enquiries algorithms constructed on the basis of the theory of stochastic automatic devices can be used. Such by stochastic automatic device is type automatic device Mess [11]. It is automatic device, for which it satisfies following conditions for conventional density of probability

$$p(u',y/u,x)=p(u'/u,x)p(y/u) \qquad (7)$$

Considering type automatic device Mess, as automatic device with discrete time (discrete stochastic automatic device), where the moments of transition are defined as the number of iterations of process of information retrieval at ultimate the amount of iterations, can be recorded that

$$p(u(t+1),y(t)/u(t),x(t))=p(u(t+1)/u(t),x(t))p(y(t)/u(t)) \qquad (8)$$

For greater presentation stochastic type automatic device Mess can be described in canonical kind

$$u(t+1)=F(u(t),x(t+1)) \qquad (9)$$

$$y(t)=f(u(t)) \qquad (10)$$

where is t the variable which defines time, that is, the moments of operation of the automatic device. This time is defined as integer parameter, that is, *t = 1,..., N*, where *N* - given amount of iterations of information retrieval, on each of which is executed choice V documents from one of sources of the information elected on this iteration (*D1,..., Dn*).

From the standpoint of the system of information retrieval constructed on above described rules, these values can be defined so:

$u(t)$ – condition of the system on current iteration which defines probabilities of choice of the sources of the information (*D1,..., Dn*) on this iteration.

$u(t+1)$ – condition of the system on following iteration which defines probabilities of choice of the sources of the information (*D1,..., Dn*) on following iteration.

$x(t)$ – the input data of the system on current iteration determining results of the evaluation of sample of size V from elected on current iteration the sources of the information,

$y(t)$ – the output data of the system on current iteration determining elected source of the information (*D1,..., Dn*).

Realization of such automatic device can be constructed in such a manner that change of condition of automatic device *u(t + 1)* is defined as regular dependence, and his exit y (t) is defined in the kind of stochastic process.

Algorithm of the information consolidation constructed on the basis of use of such automatic device consists of following steps:

1. Initial state of stochastic automatic device u(t) for *t = 1* as probabilities vector is installed (with equal probabilities or on the grounds of the previous sessions of information retrieval)

$$P(t)=\{p_1(t),p_2(t),p_3(t),...,p_i(t),...,p_{n-1}(t),p_n(t)\} \qquad (11)$$

at the same time

$$\sum_{i=1}^{n} p_i(t)=1 \qquad (12)$$

2. Evenly distributed random variable *w* on the interval [0,1 is generated].

3. Depending on value random variable w is defined interval, the appropriate to one of n the sources of the information.

At the same time, if

$$\sum_{i=0}^{z-1} p_i(t)<w<\sum_{i=1}^{z} p(t) \qquad (13)$$

that realization of the automatic device will correspond the source of the information with number *z.*

4. Sample and processing *V* the units of documents from the source of the information on number *z* is executed.

5. The evaluation R_i (3) is led of relevance for V documents from the source of the information on number z .

6. On given algorithm recalculation of the values of probabilities is produced. Development of such algorithm is a separate investigation phase. At present a few algorithms with one, two and three level adaptations are developed. For simplification of account as algorithm recalculation table is presented.

Table 1. Table of recalculation of the vector of probabilities of stochastic automatic device.

R_i	<0.2	<0.4	<0.6	<0.8	<1.0
k_R	0.5	0.75	1	1.5	2

then

$$p_z(t+1) = p_z(t)k_R \qquad (14)$$

Calculation for vector P is led (t + 1)in such a manner that

$$D(t) = 1 - p_z(t) \qquad (15)$$

$$D(t+1) = 1 - p_z(t+1) \qquad (16)$$

$$p_i(t+1) = \frac{p_i D(t+1)}{D(t)} \qquad \text{для } i = 1,\ldots,n \text{ та } i \neq z \qquad (17)$$

7. We go to the step 2 for further sample of documents from the sources of the information or we finish process of choice in the case of performance of the conditions of his ending.

In the case of repeat information retrieval on the same enquiry, it is possible to use already present probability model of choice of the most relevant sources of the information that considerably reduces the time of search. At the same time and occurrence of other sources more appropriate to the enquiry is not excluded opportunity of change of the situation in due course.

5 Conclusions

Selection of parameters, methods, and optimization of their specifications with the use of real-life sources of the information practically is impossible, as access to real sources of the information containing necessary documents, for example, scientific articles, is limited to a considerable extent both on the amount of enquiries, and on their cost. Therefore, was developed test program environment on Python (Anaconda), which allows to conduct research of models for definition of parameters of adaptive stochastic models and comparisons of their efficiency with other models.

During performance of the program virtual sources of data are created in which preset per cent of valid models of documents are located. These values and distinction for all sources are specified before the beginning of generation. Valid model of the document is formed from beforehand present enquiry copy. Those parameters are used which can be specified by the user as a part of search enquiry, for example, the key words, the name of the author, the date of publication. Not valid documents are formed by the way of exception from the initial documents of part of parameters, the appropriate to the enquiry.

Use of test environment enables to obtain the appropriate data for definition of the most effective method of choice of the sources of the information, definition and set-up of his parameters for maximization of the amount of choice of relevant documents at limited amount of enquiries to the sources of the information.

Test environment can be used and for set-up of parameters of any other annexes requiring analysis of plenty of enquiries to diverse sources having probabilities of specification.

Table 2. Results of comparison of the models on test environment.

The amount of processed documents	Boolean model	Vector model	Stochastic model
200	7	10	14
400	15	16	33
600	19	22	51
800	26	31	72
1000	38	43	98

The example of testing of algorithm in test environment is presented consisting of ten sources of the information generated pursuant to enquiry each of which contains one thousand generated documents. In each of the sources they were located from 1% to 10% of relevant documents, enables of the evaluation of finding of determined amount of relevant documents with the use of various search algorithms. The example of results of testing they are listed in the table 2. Described stochastic algorithm at processing of plenty of documents (up to 1000) gave the twice best ratings on amount chosen relevant documents than algorithms of direct search constructed on the basis of Boolean and vector model.

Presented approach in development of algorithms with use of stochastic automatic device for data consolidation allows to create the complex of software for the decision of tasks of consolidation of data for various systems. One of promising directions of use of algorithm of consolidation based on use of stochastic automatic devices, is are used of their opportunities for construction of the systems of search of scientific and technical information from various scientific bibliographic and abstract bases and other unsealed sources. One of modifications of described algorithm was evaluated at development of pilot project of the system of geocoding on queries [12].

References

1. Cherniak L. Big Data - a new Theory and Practice—Open sustems.RDBMS—M.:Open systems, 2011— № 10—ISSN 1028-7493 (in Russian)
2. Shakhovs'ka N.B. Methods of processing of consolidated data using data space — Modelling problems—2011— № 4—p.72-84 (in Ukrainian).
3. Christopher D. Manning, Prabkhacar Ragkhavan, Hainrich Schutze. Introduction to Information Search (translate from English) — M.:JSC "I.D.Williams",2011—p.504 (in Russian).
4. Yaglom I.M. Boolean structure and its models —M: Soviet radio, 1980 —p.192 (in Russian).
5. Ukhobotov V.I. Selected chapters of the theory of fuzzy sets —Tutorial—Cheliabinsk: Publishing house of Cheliabinsk State University, 2011—p.245 (in Russian).
6. Dubinskiy A.G. Some questions of application of vector model of representation of documents in information search — Control systems and machines — 2001 — №4 — p.77–83 (in Russian).
7. Bondarchuk D.V. The use of latent-semantic analysis in the case of classification of texts by emotional coloring—Bulletin of research results—2012—2(3)—p.146—151 (in Russian).
8. Robertson S. E. The probabilistic ranking principle in IR— Journal of Documentation — 1977 — № 33 — p.294–304.
9. Lande D. V., Snarskiy A. A., Bezsudnov I. V. Internetika. Navigation in complex networks: models and algorithms—Moscow: Book house "Librocom"— 2009—p.264 (in Russian).
10. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005 ISBN 0-12-088407-0C. [11] O. V. Koval, V. O. Kuzminykh, D. V. Khaustov. Using stochastic automation for data consolidation - Research Bulletin of NTUU "KPI".Engineering. - 2017. - №2. - C. 29-36.
11. Rastrigin L.A., Ripa K.K. Automated random search theory—Riga: Zinatne,1973.— p.344 (in Russian).
12. Kuzminykh V.O., Boichenko O.S. The system of automatic geocoding of user requests— Environmental security clusters: energy, environment, information technology—Kyiv:"MP Lesia",2015—STUU "KPI"—p.217–222 (in Ukrainian).