UDC 615.47:616-085
DOI: 10.20535/1810-0546.2017.2.100011

O.V. Koval*, V.A. Kuzminykh, D.V. Khaustov
Igor Sikorsky KPI, Kyiv, Ukraine

USING STOCHASTIC AUTOMATON FOR DATA CONSOLIDATION

Background. Development of methods and algorithms for efficient search of relevant information on demand. The article deals with the consolidation of data for subsequent use in the information and analytical systems.
Objective. The aim of the paper is to identify capabilities and build relevant information search algorithms from disparate sources by analyzing the probability information identifying the possible presence of relevant documents in these sources.
Methods. To find the relevant information for search queries the approach based on the use of probability estimates of relevant documents available in the sources of further increasing the number of selected documents from these sources to analyze their relevance to the query is used.
Results. A stochastic programmable automaton structure to ensure selection of the most possible information sources by relevance parameters and information retrieval algorithm based on the use of stochastic automaton were developed.
Conclusions. The described algorithm using stochastic automaton for data consolidation allows developing a set of software tools, provides plenty full and holistic data consolidation problem-solving for diverse systems which search for information from information sources different in composition and presentation type.

Keywords: open data sources; data consolidation; information-analytical systems; information retrieval systems; probabilistic models; relevance; big data tasks.

Introduction

Directed information gathering based on open sources is considered as one of the standard methods of information gathering in different spheres of modern society. Traditionally, specialists collect and analyze information from the media, public statements, official data, press conference materials, public statements, professional and academic reports, conferences, reports, and articles. The transition to electronic media largely determines the directed search approaches and methods and the efficiency increase of both individual procedures and information search in general.

The amount of information produced by humanity, which is constantly increasing and greatly complicates the problem of storing large amounts of data and extracting from them relevant and important information. Every day the problem of processing such information assumes greater significance. A significant amount of electronic materials is going to make it difficult for users to find the necessary information [1]. The information consolidation can help solving these problems [2, 3]. In a broad sense the consolidation can be understood as the process of searching, selecting, analyzing, structuring, conversion, storage, cataloging, and providing consumer information on a given topic. The task of information consolidation is one of the most important problems of processing large amounts of data (big data) [4].

Consolidation is generally regarded as a set of techniques and procedures to extract data from various sources to ensure the necessary level of information content and quality conversion in a single format in which they can be loaded into the data warehouse or analytical system.

In some cases data consolidation is the initial stage of any analytical task or project [3]. The basis of consolidation is the process of collecting and storing data in a form optimal in terms of their processing on the specific analytic platform or specific analytical problem solving. The associated consolidation tasks are to assess the quality of data and their enrichment, to reduce the amount of information that has to be processed by data retrieval or information-analytical system.

The main results that data consolidation has to ensure for further processing depth are as follows:

- high access rate to large volumes of data;
- compactness of large volume data storage;
- support of a data structure integrity;
- monitoring of data consistency and relevance.

The feature of information gathering based on open sources is the instability of the informational contents of these sources, the lack of reliable prior information about their content and its relevance

*corresponding author: avkovalgm@gmail.com

and typically low accuracy and efficiency of expert assessments of state and compliance of the sources with the topics and query parameters.

Therefore, to process data from open information sources it is necessary to conduct effective consolidation of data using specialized software tools that will provide the following possibilities:

- automatic selection of the information sources, most relevant in compliance with the request;
- the possibility of accumulation of information on sources' state in the course of the request;
- consideration of the possibility of changing sources' state during the second request;
- analysis of the most promising sources from the point of view of the relevance as well as less relevant;
- creation of information assessment of the perspective sources in terms of relevance.

Problem statement

The purpose of the study is development of model and algorithm for consolidation of information from a certain quantity of heterogeneous sources with high content of documents. At the same time, the task to select the maximum number of documents relevant to the request is set in a case of the minimum number of the processed (checked for relevance to a request) documents by detection and use of information sources the most appropriate to the request.

Review of existing solutions

Today there are many information search models, that can form the basis for information consolidation system, that are based on various mathematical methods. Modern information-analytical and information retrieval systems based on these models and various modifications are built.

Among them are the following most common types of models [2]:

- Boolean model;
- fuzzy set model;
- vector model;
- latent-semantic model;
- probabilistic model.

Boolean model is based on the use of mathematical logic app and set theory [5]. The model based on matrix, which is the ratio between the document and the indexing terms [6].

The model advantages:

- construction simplicity;
- ease of program implementation.

The model disadvantages:

- difficulty of query building without knowledge of Boolean algebra;
- the search results must contain all terms of the user's request;
- virtually impossible to automatically rank the received documents;
- the search is badly scaled.

The fuzzy set model is based on fuzzy set theory, allowing the partial set element membership in contrast to the traditional set theory, which does not allow this [7, 8]. The whole array of documents is described as a set of fuzzy set terms in this model.

The model advantages:

- the ability to rank results;
- indexing and determining the relevance of the selected document by request requires less computing;
- easy to implement;
- no requirements for large volumes of memory.

The model disadvantages:

- computing costs and storage costs higher than Boolean model;
- lack of accurate search result distinction.

Vector model [8, 9] represents the documents and user queries as n-dimensional vectors in n-dimensional vector space. The dimension of the vector space n is the number of different terms in all documents [10].

The model advantages:

- construction simplicity;
- the ability to rank results;
- the model disadvantages;
- requires large amounts of data processing;
- requires exact word matching.

Latent semantic model is commonly referred to as latent semantic indexing in the information retrieval theory [11]. Latent semantic analysis (LSA) is a method of extracting and presenting context dependent word meanings by using statistical analysis of large text document sets. LSA consists of two stages – study and analysis of indexed data [12].

The model advantages:

- a space of much smaller dimension is used than the vector;
- no need to match the exact words;
- the model does not require a complicated setup.

The model disadvantages:

- complexity leads to a large number of calculations;

• there are no rules for choosing the dimension on which the effectiveness of obtaining results depended.

Probabilistic model. These search models are based on the use of probability theory methods. They used statistical indicators that determine the document relevance probability to the search query [13]. These models are based on the probabilistic ranking principle in descending probability of their relevance to the user request. Probabilistic models differ from each other in procedures for calculating estimates of probabilities [14].

The model advantages:

• the ability to adapt in the use;

• there is no volume calculations dependence on the number of terms, and other parameters;

• high efficiency when working with dynamically updated data sources.

The model disadvantages:

• the need for continuous learning of the system;

• low efficiency at the initial stages of work.

Today there are no models that would possess quantitative and qualitative advantages over the other [15, 16]. Therefore, the development of new algorithms and methods based on the combination of the advantages of the different approaches in building models is very relevant and requires new solutions. This article describes the approach in the implementation of a software system that is built on the use of elements of the stochastic automata theory in selecting information sources and fuzzy sets in assessing the relevant information sources.

The mathematical formulation of the problem

The structure of the of the main information element interaction, reflecting the consolidation of documents (information units) is shown in Fig. 1. It contains the following elements:

1. Information sources.
2. Query parameters.
3. Document packets.

The information sources are information resources that are important and considered in particular information system construction, or considered in a certain information search case.

Currently, there are following information resources:

1. The media which includes various kinds of news sites (RSS feeds) and semantic sites (or electronic versions of media).

2. Digital libraries – distributed information system that can reliably store and effectively use electronic documents through a global network.

3. Electronic database – a collection of files organized in a special way, documents, grouped by topic and spreadsheets, which are combined into groups.

4. Corporate and personal sites – online resources dedicated to any organization, company, enterprise, a particular issue or person. They are distinguished by the completeness of information that covers all aspects of the activity.

5. Information portals – a group of sites where you can use a variety of services. They can contain various scientific, political, economic and other information as well as electronic mailboxes, blogs, catalogs, dictionaries, directories, weather forecasting, television programs, exchange rates, etc. Generally, their renewal happens in real time.

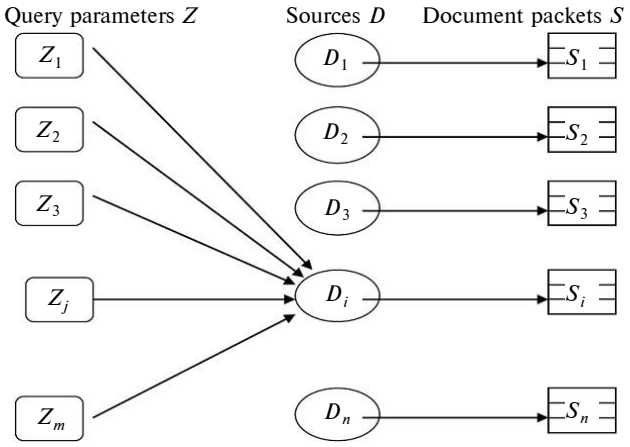

Fig. 1. The basic elements of the information consolidation

The request content settings – are mostly simple and complex terms (terms related by operators), the parameters that define the various features of the time, place (appearance, publication, storage, etc.), and many other characteristics that determine specific request characteristics.

The document packets – specific sets of information units (articles, records, tables, reports, newspapers, magazines, books, abstracts, news, reviews, etc.), with similar presentation format.

Let's consider the main information consolidation description characteristics in accordance with the stochastic model of information source selection based on the theory of stochastic automata [17].

To assess the relevance of the information sources to queries the model based on the theory

of fuzzy sets is used [7], which links the query parameters and information sources.

llows a partial matching between the request and the information source is used. This matching assesses values in the range [0, 1]. The value that estimates the correspondence between the request and the information source is based on the determination of coincidence between the query parameters and characteristics of information units included in a certain source of information in question.

Quantitative matching assessment of the l-th document (for $l = 1,...,V$) from the i-th information source to the j-th query parameter in single sampling for relevant source relevance evaluation is defined as

$q_{ijl} = 0$ – when j-th parameter is not present in the l-th document of i-th source, $q_{ijl} = 1$ – when j-th parameter is present in the l-th document of i-th source.

Then this value after sampling and evaluation of several pieces of information from certain information source is averaged according to the number of selected items (documents).

Quantitative assessment of compliance of the i-th source to the j-th query parameter determines the partial match, it corresponds to the range [0, 1] and can be defined as

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

where q_{ijl} is a quantitative matching assessment of the l-th document (for $l = 1,...,V$) from the i-th information source to the j-th query parameter in single sampling for source relevance evaluation; V – the volume of a single sample from one document source has to meet the following limitation:

$$V << S_i \text{ for } i = 1,...,n$$

where S_i is a number of information pieces (documents) in each n source considered for this request.

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

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

where $0 < v_i < 1$, v_j is a weight coefficient of the j-th topic query parameter determined by expert assessments or on the basis of priorities that can be independently determined by information request customer.

Thus evaluation of the i-th source relevance to specific request will be determined in the range [0, 1].

The consolidation algorithm based on the use of stochastic automaton

To organize selection procedures for effective information sources relevant for specific queries can be used algorithms that are based on stochastic automata. Such stochastic automaton is a Mealy type automaton [17]. This is the automaton for which the following conditions are satisfied for the conditional probability density

$$p(u', y/u, x) = p(u'/u, x)p(y/u, x)$$

where $x \in X$ is a stochastic automaton input values; $y \in Y$ is a stochastic automaton output values; $u, u' \in U$ is a stochastic automaton possible states.

For such stochastic automaton the next state, automatic automaton u' takes does not depend on automaton y output, and automaton y output does not depend on the state in which the automaton u' moves for any possible input value of stochastic automaton x and any possible condition of automaton u. So for stochastic Mealy type automaton the automaton y output and automaton u' state in which it moves are independent of one another. They depend only on the input value and the previous state.

If the stochastic Mealy type automaton runs the relation type

$$p(y/u, x, u') = p(y/u),$$

always when

$$p(u'/u, x) \neq 0,$$

then it is a Moore type automaton.

Stochastic Moore type automaton [17] is a special case of the stochastic Mealy type automaton. For such stochastic automaton the output depends on its condition and does not depend on the input, and the next state of the automaton is determined by its previous state and automaton input. So for stochastic Mealy type automaton can be determined that

$$p(u', y/u, x) = p(u'/u, x)p(y/u).$$

Considering this automaton as automaton with discrete time (discrete stochastic automaton) where the transition points are defined as the numbers of iterations of the information seeking process under the final number of iterations, we can write that

$$p(u(t+1), y(t)/u(t), x(t)) = p(u(t+1)/u(t))p(y(t)/u(t)).$$

With the perspective of a search system that is based on the principles described above, these values can be defined as:

$u(t)$ is a state of the system at the current time (the current iteration), which determines the probability of selecting information sources $(D_1,...,D_n)$ in the current iteration;

$u(t+1)$ is a state of the system at the next time period (the next iteration), which determines the probability of selecting information sources $(D_1,...,D_n)$ in the next iteration;

$x(t)$ is system input data on the current iteration, determining sampling estimation results with size V from the selected information source on the current iteration;

$y(t)$ is system output data on the current iteration, determining the selected information source $(D_1,...,D_n)$.

For greater clarity in some cases, stochastic Moore type automaton can be described in the following canonical form

$$u(t+1) = F(u(t), x(t+1)),$$

$$y(t) = f(u(t))$$

where t is the variable that determines the time, viz the automaton response times. This time is defined as the integer that is $t = 1,...,N$, where N is the given number of information search iterations, on each of which V document choice from one of selected at this information source iteration $(D_1,...,D_n)$ is performed.

Implementing the use of the automaton for selecting information sources for consistent consolidation can be constructed in such a way that the automaton u state change $(t+1)$ is defined as a regular constraint, and its output $y(t)$ is determined as a stochastic process.

The consolidation information algorithm based on the use of such automaton consists of the following steps:

1. The stochastic automaton initial state $u(t)$ for as a vector of probabilities is determined:

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

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

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

3. Depending on the value of the random value w the interval that corresponds one of the n information sources is determined as follows.

We accept that $p_0 = 0$, then if

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

the implementation of the automaton will correspond the information source by number z.

4. Sampling and processing of V information units (Documents) from the information source by number z is performed.

5. Assessment R_i (1) of the relevance for V information units (documents) from information sources by number z is performed.

6. Recalculated probability values are performed by a particular algorithm.

The example is given in Table.

Table. Stochastic automaton probability vector conversion coefficients

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

Indeed,

$$p_z(t+1) = p_z(t)k_R.$$

A calculation for the vector $P(t+1)$ is made in such a way that

$$D(t) = 1 - p_z(t),$$

$$D(t+1) = 1 - p_z(t+1),$$

$$p_i(t+1) = \frac{p_i D(t+1)}{D(t)}$$

for $i = 1,..., n$ and $i \neq z$.

Move to step 2 for further document (informative units) sampling of the information sources when choosing a source of accumulated information about their relevance of the request and its parameters that has been obtained in the preceding steps of selection.

As a result of described algorithm's action execution we get a sequence of probability vectors that reflects a consistent process of identifying the most relevant information sources for the request with the acceptability parameters. When you do another search for information on the same request we get the existing probabilistic model of the most relevant information source selection, which significantly reduces the search time, but does not preclude the situation changes over time among information sources and the possibility of other more relevant to the request sources, among those that considered.

The algorithm testing was conducted in a specialized test software environment that was developed for the analysis and testing of capabilities of relevant documents algorithm selection upon request. Test environment consists of ten formed in accordance with the request information sources, each of which contains one thousand generated documents. Each source got from 1 to 10 % of relevant documents, which allows evaluating the finding of a certain number of relevant documents by using different search algorithms.

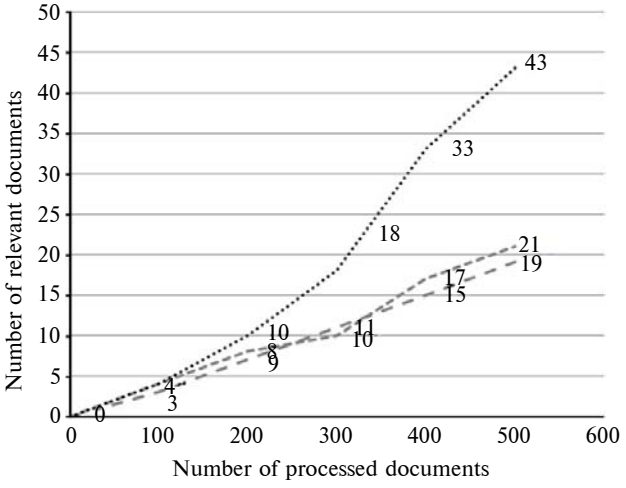

Fig. 2. An example of the results of testing an algorithm constructed on a stochastic (·····) automaton in comparison with Boolean (– – –) and vector (– – – –) models

Sample test results are shown in Fig. 2. The described stochastic algorithm when processing a large number of documents (more than 500) showed twice better performance on the number of selected relevant documents than direct search algorithms that are based on Boolean and vector models.

Conclusions

The use of stochastic automaton for data consolidation for further use in information-analytical systems as one of the most important problems of processing large amounts of data is presented. The use of an integrated approach based on stochastic information source selection model and fuzzy set model for assessing the relevance of certain documents is proposed. The paper describes the information consolidation algorithm implementing the possibility of using a stochastic automaton for selecting relevant documents. The structure of the interaction of open source information consolidation system basic elements is shown.

Described in the article algorithm of using the stochastic automaton for data consolidation was tested in the development of geocoding by information requests system pilot project [18, 19].

The developed algorithm of using the stochastic automaton for data consolidation allows developing a set of software tools that provides enough complete and holistic data consolidation problem solving for various systems searching for information from information sources different in composition and presentation type. One of the perspective directions of use of the consolidation algorithm, based on use of stochastic automata, is the utilization of their capabilities for building systems of search of scientific and technical information by one request from different scientific bibliographic and abstract bases and other open sources (SCOPUS, RSCI, UKRINTEI, scientific RSS feeds, etc.). It will give the chance to wide circles of scientists, graduate students and students considerably reduce the time and improve the quality of searching for materials, necessary for your profile, for scientific and educational activities.

Список літератури

1. *Хаустов Д.В., Кузьміних В.О., Шевченко О.М.* Стандартизація навчальних ресурсів на базі об'єктно-орієнтованого підходу // V(XXIX) Міжнар. міжвуз. школа-семінар "Методи і засоби діагностики в техніці та соціумі (МіЗД ТС-201)": Зб. матеріалів. – Івано-Франківськ: Факел, 2015. – С. 81–85.

2. *Певцов Г.В., Фастовский Э.Г., Олейник М.А.* Анализ методов консолидации информации и особенностей их применения // Вісник НТУ "ХПІ". Тем. вип. Інформатика і моделювання. – 2007. – № 39. – С. 45–153.
3. *Шаховська Н.Б.* Методи опрацювання консолідованих даних за допомогою просторів даних // Проблеми програмування. – 2011. – № 4. – С. 72–84.
4. *Черняк Л.* Большие Данные – новая теория и практика // Открытые системы. СУБД. – 2011. – № 10. – URL: https://www.osp.ru/os/2011/10/13010990
5. *Salton G., Fox E., Wu H.* Extended boolean information retrieval communications of the ACM // CACM. – 1983. – **26**, № 11. – P. 1022–1036.
6. *Яглом И.М.* Булева структура и ее модели. – М.: Сов. радио, 1980. –192 с.
7. *Ухоботов В.И.* Избранные главы теории нечетких множеств. – Челябинск: Изд-во Челяб. гос. ун-та, 2011. – 245 с.
8. *Baeza-Yates R., Rebeiro-Neto B.* Modern Information Retrieval. – Menlo Parl, California, New York: ACM Press, Addison-Wesley, 1999. – 501 p.
9. *Salton G., Wong A., Yang C.* A vector space model for automatic indexing // Communications of the ACM. – 1975. – **18**, № 11. – P. 613–620.
10. *Дубинский А.Г.* Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. – 2001. – № 4. – С. 77–83.
11. *Landauer T., Foltz P., Laham* D. An introduction to latent semantic analysis // Discourse Processes. – 1998. – **25**, № 2-3. – P. 259–284.
12. *Бондарчук Д.В.* Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. – 2012. – **2**, № 3. – С. 146–151.
13. *Robertson S.E.* The probabilistic ranking principle in IR // J. Documentation. – 1977. – **33**, № 4. – P. 294–304.
14. *Maron M.E., Kuhns J.L.* On relevance, probabilistic indexing, and information retrieval // JACM. – 1960. – **7**, № 3. – 216–244 p.
15. *Ландэ Д.В., Снарский А.А., Безсуднов И.В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: ЛИБРОКОМ, 2009. – 264 с.
16. *Маннинг К.Д., Рагхаван П., Шютце Х.* Введение в информационный поиск / Пер. с англ. – М.: ООО "И.Д. Вильямс", 2011. – 504 с.
17. *Растригин Л.А., Рипа К.К.* Автоматная теория случайного поиска. – Рига: Зинатне, 1973. – 344 с. – (ЛатАН).
18. *Кузьміних В.О., Бойченко О.С.* Система автоматичного геокодування інформаційних запитів // V(XXIX) Міжнар. міжвуз. школа-семінар "Методи і засоби діагностики в техніці та соціумі (МіЗД ТС-201)": Зб. матеріалів. – Івано-Франківськ: Факел, 2015. – С. 12–17.
19. *Кузьміних В.О., Бойченко О.С.* Система автоматизованого геокодування запитів користувачів // Екологічна безпека територіально-виробничих комплексів: енергетика, екологія, інформаційні технології. – К.: "МП Леся", НТУУ "КПІ", 2015. – С. 217–222.

References

[1] D.V. Khaustov *et al.*, "Standardization of educational resources based on object-oriented approach", in *Proc. V(XXIX) Int. Interuniversity School Seminar "Methods and Diagnostic Tools in Technology and Society (MiZD TS-201)"*, Ivano-Frankivs'k, 2015, pp. 81–85 (in Ukrainian).
[2] G.V. Pevtsov *et al.*, "Analysis of information consolidation methods and their application features", *Visnyk "KhPI". Special Edition: Information and Modelling*, no. 39, pp. 45–153, 2007 (in Ukrainian).
[3] N.B. Shakhovs'ka, "Processing methods of consolidated data using data space", *Problemy Prohramuvannya*, no. 4, pp. 72–84, 2011 (in Ukrainian).
[4] L. Cherniak. (2011). *Big Data – A New Theory and Practice* [Online]. Available: https://www.osp.ru/os/2011/10/13010990 (in Russian).
[5] G. Salton *et al.*, "Extended boolean information retrieval", *CACM*, vol. 26, no. 11, pp. 1022–1036, 1983. doi: 10.1145/182.358466
[6] I.M. Yaglom, *Boolean Structure and Its Models*. Moscow, SU: Sovetskoe Radio, 1980 (in Russian).
[7] V.I. Ukhobotov, *Selected Chapters of the Theory of Fuzzy Sets*. Cheliabinsk, Russia: Publ. House of Cheliabinsk State University, 2011 (in Russian).
[8] R. Baeza-Yates and B. Rebeiro-Neto. *Modern Information Retrieval*. Menlo Parl, California, New York: ACM Press, Addison-Wesley, 1999.
[9] G. Salton *et al.*, "A vector space model for automatic indexing", *CACM*, vol. 18, no. 11, pp. 613–620, 1975. doi: 10.1145/361219.361220

[10] A.G. Dubinskiy, “Some questions of application of vector model of document representation in information search”, *Upravljajushhie Sistemy i Mashiny*, no. 4, pp.77–83, 2001 (in Russian).

[11] T. Landauer *et al.*, “An introduction to latent semantic analysis”, *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998. doi: 10.1080/01638539809545028

[12] D.V. Bondarchuk, “The use of latent-semantic analysis in the case of text classification by emotional coloring”, *Bjulleten' Rezul'tatov Nauchnyh Issledovanij*, vol. 2, no. 3, pp. 146–151, 2012 (in Russian).

[13] S.E. Robertson, “The probabilistic ranking principle in IR”, *J. Documentation*, vol. 33, no. 4, pp. 294–304, 1977. doi: 10.1108/eb026647

[14] M.E. Maron and J.L. Kuhns, “On relevance, probabilistic indexing, and information retrieval”, *JACM*, vol. 7, no. 3, pp. 216–244, 1960. doi: 10.1145/321033.321035

[15] D.V. Lande *et al.*, *Navigation in Complex Networks: Models and Algorithms*. Moscow, Russia: LIBROCOM, 2009 (in Russian).

[16] C.D. Manning *et al.*, *Introduction to Information Search*. Moscow, Russia: I.D. Williams, 2011 (in Russian).

[17] L.A. Rastrigin and K.K. Ripa, *Automate Random Search Theory*. Riga, Latvia: Zinatne, 1973 (in Russian).

[18] V.O. Kuzminykh and O.S. Boichenko, “The information request automatic geocoding system”, in *Proc. V(XXIX) Int. Inter-university School Seminar “Methods and Diagnostic Tools in Technology and Society (MiZD TS-201)”*, Ivano-Frankivs’k, 2015, pp. 12–17 (in Ukrainian).

[19] V.O. Kuzminykh and O.S. Boichenko, “The user request automatic geocoding system”, in *Environmental Security Clusters: Energy, Environment, Information Technology*. Kyiv, Ukraine: “MP Lesia”, NTUU “KPI”, 2015, pp. 217–222 (in Ukrainian).

О.В. Коваль, В.О. Кузьміних, Д.В. Хаустов

ВИКОРИСТАННЯ СТОХАСТИЧНОГО АВТОМАТУ ДЛЯ КОНСОЛІДАЦІЇ ДАНИХ

Проблематика. Розробка методів і алгоритмів ефективного пошуку релевантної інформації за запитами. У статті розглядаються питання консолідації даних для подальшого їх використання в інформаційно-аналітичних системах.

Мета дослідження. Виявлення можливості та побудова алгоритмів пошуку релевантної інформації з різнорідних джерел на основі аналізу ймовірнісної інформації, що визначає можливість наявності релевантних документів у цих джерелах.

Методика реалізації. Для пошуку релевантної інформації за пошуковими запитами використовується підхід, побудований на використанні оцінок ймовірностей наявності релевантних документів у джерелах із подальшим збільшенням кількості вибраних із цих джерел документів для аналізу їх релевантності запиту.

Результати досліджень. Розроблено структуру програмованого стохастичного автомату для забезпечення вибору найбільш імовірних за параметрами релевантності джерел інформації та алгоритм пошуку інформації на основі використання стохастичного автомату.

Висновки. Наведений алгоритм використання стохастичного автомату для консолідації даних дає змогу розробити комплекс програмних засобів, що забезпечує достатньо повний і цілісний розв’язок задач консолідації даних для різноманітних систем, що здійснюють пошук інформації з різноманітних за складом і видом представлення джерел інформації.

Ключові слова: відкриті джерела даних; консолідація даних; інформаційно-аналітичні системи; інформаційно-пошукові системи; ймовірнісні моделі; релевантні документи; задачі обробки великих обсягів даних.

А.В. Коваль, В.А. Кузьминых, Д.В. Хаустов

ИСПОЛЬЗОВАНИЕ СТОХАСТИЧЕСКОГО АВТОМАТА ДЛЯ КОНСОЛИДАЦИИ ДАННЫХ

Проблематика. Разработка методов и алгоритмов эффективного поиска релевантной информации по запросам. В статье рассматриваются вопросы консолидации данных для дальнейшего их использования в информационно-аналитических системах.

Цель исследования. Определение возможности и построение алгоритмов поиска релевантной информации из разнородных источников на основе анализа вероятностной информации, которая определяет возможность наличия релевантных документов в этих источниках.

Методика реализации. Для поиска релевантной информации по поисковым запросам используется подход, который построен на использовании оценок вероятностей наличии релевантных документов в источниках с последующим увеличением числа выбираемых из этих источников документов для анализа их релевантности запросу.

Результаты исследований. Разработаны структура программируемого стохастического автомата для обеспечения выбора наиболее вероятных по параметрам релевантности источников информации и алгоритм поиска информации на основе использования стохастического автомата.

Выводы. Приведенный алгоритм с использованием стохастического автомата для консолидации данных позволяет разработать комплекс программных средств, обеспечивает достаточно полное и целостное решение задач консолидации данных для различных систем, которые осуществляют поиск информации из различных по составу и виду представления источников информации.

Ключевые слова: открытые источники данных; консолидация данных; информационно-аналитические системы; информационно-поисковые системы; вероятностные модели; релевантность; задачи обработки больших объемов данных.

Рекомендована Радою Навчально-наукового комплексу “Інститут прикладного системного аналізу” КПІ ім. Ігоря Сікорського

Надійшла до редакції 28 лютого 2017 року