Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 284–290

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Indexical structures to enable knowledge mining tasks

Sergey Kosikov[b], Larisa Ismailova[a], Viacheslav Wolfengagen[a,*], Alexander Marenkov[b], Mikhail Ermak[b], Igor Slieptsov[b], Vladislav Zaytsev[b], Andrey Shedko[b]

[a]*National Research Nuclear University "Moscow Engineering Physics Institute", Kashirskoye Shosse, 31, Moscow 115409, Russia*
[b]*NAO "JurInfoR", M. Pirogovskaya str., 5, Moscow 119435, Russia*

Abstract

In this paper, it is shown that semantic modeling technologies rely on data indexing systems. The semantic information theory known today is based on the method of using the principle of index expressions, using the division into real, possible and virtual individuals/data objects. Organization of a search on the Web for the concept of conceptual modeling of information processes based on variable domains that is being developed in the work is based on the natural requirements imposed on data organization. The model structure implements the separation of individuals, which manifests itself in the interaction of data organization, search and indexing. The identity of the data and the information processes designated by them is achieved in this model structure. A skeleton view of the organization of the search is being developed, leading to the maintenance of a network of variable domains and the possibility of deploying on them a conceptual modeling of interacting information processes.

Keywords: semantic information processing; computational model; variable domains

1. Introduction

Data indexing has become especially acute since the first steps of creating and using [11] database management systems, and now it has grown into one of the most significant tasks for information technology to identify valuable data hidden in the deep Web [2].

With the participation of the authors, a family of ideas about indexing and using index expressions for the needs of conceptual modeling of information processes has been developed. These include: an idea of displaced concepts [5]; maintaining the semantic stability of the information model in the development of information processes [6];

* Corresponding author. Tel.: +7-495-778-8726.
E-mail address: s.v.kosikov@gmail.com, lyui.ismailova@gmail.com, jir.vew@gmail.com, qwexly@gmail.com, demonnik2@yandex.ru, igor.slieptsov@mail.com, zvs2403@gmail.com, ayshedko@yandex.ru

1877-0509
10.1016/j.procs.2020.02.180

computational model for granularity of information processes [7] in order to ensure their semantic stability; providing representation of dynamics in recognizing properties in diagnostic studies [8]; presentation of dynamic conceptual dependencies [14]; model structure of representability for individual information processes [15]; advanced model structure for studying intertwined or entangled information processes [16].

In favor of indexing objects on the Web, more than enough weighty and convincing arguments have been made. The presentation of information images of objects on the Web, whatever their nature, is based on their appropriate indexing.

The deep web, invisible web, or hidden web, introduced into the practice of information retrieval, are parts of the World Wide Web that are not indexed by standard search engines. The opposite of the term "deep Web" is the surface network (surface Web), accessible to anyone who uses the Internet. Corresponding methods date back to Garcia-Molina [12], [10]. Existing search engines retrieve content only from a publicly indexed network, that is, a set of web pages that are accessible only via hypertext links, ignoring search forms and pages that require authorization or pre-registration. In particular, they ignore the huge amount of high-quality content "hidden" behind search forms in large searchable electronic databases.

Semantic information processing is based on a Web search with the deployment of the index infrastructure [9], [4].

In this paper, it is shown that the noted semantic modeling technologies need some kind of data indexing system. First of all, the semantic information theory known today is based on the method of using the principle indexical expressions of Carnap-Bar-Hillel [3], [1], which was given a modern look by D. Scott [13]. Organization of a search on the Web for the concept of conceptual modeling of information processes based on variable domains that is being developed in the work is based on the natural requirements imposed on data organization.

Section 2 discusses the requirements imposed by the model structure on the interaction of data organization, search, and indexing. Section 3 discusses the requirements for data identity and the information processes it identifies. Section 4 contains the requirement of a natural classification of data into actual, possible and virtual. Section 5 discusses the wireframe notion of organizing a search, which leads to maintaining a network of variable domains and the possibility of deploying conceptual modeling of interacting information processes on them.

2. Data, search engines and indexing

When discussing the possibility of extracting knowledge from data, it always turns out that we have to deal with *data*, from which knowledge is actually extracted, and the idea of data as a kind of *representation* has to be conducted and developed in a targeted way.

To begin with, the search is currently based on data indexing. Why are there no real alternatives to the few popular search index engine providers? Firstly, index providers face enormous technical difficulties due to the large number of documents received from the ever-changing nature of the Internet. Secondly, a significant problem is the cost of equipment, infrastructure, maintenance and personnel. Thirdly, the Web is huge, and the index search engine should match the task of covering as much of it as possible. While it is known that no search engine can reach the Web as a whole, modern search engines know the trillions of existing pages. Indexing these pages is just the beginning. The search engine must save its *current index*, which means that you need to update at least part of it every minute. This is an important requirement that is not satisfied due to any choice from current projects (such as Common Crawl) in order to index "snapshots" of some parts on the Internet [9], [4].

The diagram in Fig. 1 shows the Web infrastructure option for an index with searchability for the Open Web Index (OWI). Infrastructure part of the search engine (index) is separated from the service part, thereby allowing multiple services, regardless of whether they exist as search engines or not, work in a common infrastructure. The figure shows how the public infrastructure is responsible for scanning the network, indexing its contents, and providing an interface/API for index-based services.

The indexing phase is divided into basic indexing and advanced indexing. Basic indexing provides data in a form in which services built on top of the index can easily and quickly process this data. So thus, while services are allowed to perform further indexing to prepare documents, some advanced indexing is also provided by the open infrastructure. This provides additional information for indexed documents (e.g., semantic annotations). This requires an extensive infrastructure for data mining and processing. However, services must be able to decide for themselves to what extent

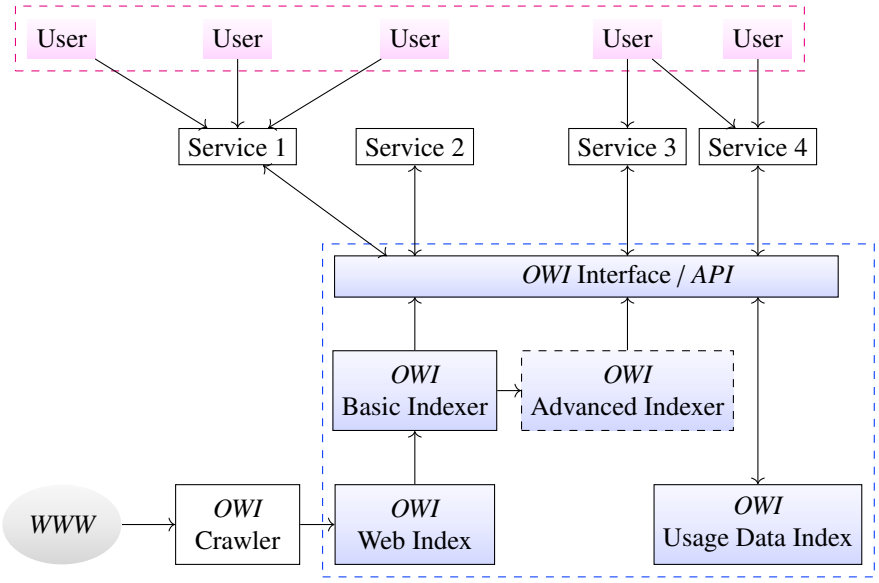

Fig. 1. Searchable Indexing.

they want to rely on the preprocessing infrastructure provided by the Open Web Index. The design solution should provide maximum flexibility of services.

3. Identity and variability of data

The problem seems to be that semantically immutable datasets can be changed while remaining unchanged.

3.1. Identity in the DBMS

In computations aimed only at adding, observations are recorded forever (or for a long time). The results are computed on demand (or periodically). This is an ordinary mode of operation of a DBMS (database management system), in which all changes are recorded in transaction logs, entered into the database. High-speed uploads is the only way to change such a log. From this point of view the contents of the database stores caching of the latest log entries. The database is cache subsets of the log. This cached subset is the last value of each record and the index value from the log.

As distributed systems scale in size and heterogeneity, identifiers increasingly link them. These can be identifiers, names, keys, numbers, URLs, file names, links, UPC (universal product codes) and many other terms. Often these terms refer to immutable things. In other cases, they relate to things that change over time. Identifiers are even used to represent the nature of computations running in untrusted systems.

The interesting thing about identifiers is that although they identify one and the same “thing” over time, this reference thing can be moving in meaning. Product descriptions, reviews, and inventory balances all change, but product identifiers do not. Reservations and bookings have identifiers that are not vary, while the material they identify may vary slightly over time. Identity and identifiers provide a consistent connection. Both sides of this relationship may change, but they provide the semantic consistency needed for business operations. No matter what you call it, identity is the glue that makes things connect and enables collaboration.

3.2. Web Identity

Identity is used for search, which is especially true for Web searches, as is done with Yahoo, Google, and Bing. At the same time, the search is performed by assigning unique identifiers to each of the documents on the Internet.

Document IDs, URLs, and search terms are intertwined. When these huge Web crawlers scan the URLs found to search for documents, they remember the URL of each document. These URLs form unique identifiers. Usually bind URLs to another unique identifier of the document, which is shorter.

When a document *is scanned* (viewed), sequences of words are retrieved for indexing. These word sequences (known as N-grams) match the search terms entered into the Web Search application. N-grams are divided into a large number of sections. When multiple search terms are included in a search, fragments are requested that may contain these terms. This returns sets of document identifiers from many *shards*. Sharding is a special technique for scaling data work, which consists in dividing (partitioning) the database into separate parts so that each of them can be transferred to a separate server. This process depends on the structure of your database and runs directly in the application, as opposed to replication.

Comparing the results of the search for document identifiers common to all search terms, you can get the final collection of document identifiers. Although it is very and extremely simplistic, the main thing is that search is all about identities. Search for an object-relational application slightly differs. Object-relational systems typically have application objects located on top of basic relational systems. Some object-relational systems offer search functions that find the identifiers of objects based on their contents and N-grams in them. This mechanism depends on the identifiers of the objects captured by the search engine and correlated with the objects. Although these identifiers may not be clearly understood by the underlying SQL database, they are understood by the object-relational system and the search engine located on top.

4. Data device

Such a flexible model structure is possible that the previous difficulties will be outside it, and instead of overcoming them, it will be possible to begin developing its important utilities and applications.

4.1. Individuals

It is sometimes assumed that “being an individual” is original undefined notion. As a working hypothesis, we assume that individuals are considered to be separable (allocated) in the (problem) domain. This is done through “individualizing functions”, so that the individual is put in correspondence with some set of individualizing functions. It is important to come to the assumption that individuals (or their representatives) can be assembled into one area, a set that we call D.

It is not necessary for all elements of D to require their constructive definition; only a certain property defining the set D can be considered given. If we assume that the set D is indefinite, then significant complications can arise. Since in this D can be expanded by adding new elements, then a lot of $D' \neq D$ arises, for which the previously obtained results may not remain unchanged, but need to be changed relative to the newly established area. The similar uncertainty of the set D when it changes occurs can take the form of a paradox, if you do not take into account that $D' \neq D$. One of them is in the form of *tangles*, when inside D, various traits of various key individuals merge, becoming indistinguishable.

The volume of the D area is large enough to cover what is meant by a Big Data Territory. At least this area is non-empty, which allows consider D as a region of *possible* individuals. Term “possible” is understood in the sense that an individual is possible with respect to some predetermined system of reasoning:

“possible individual” = “individual possible with respect to a priori notion.”

Such a notion may be conceivable, so that the idea of a possible individual is *relativized*.

Arising examples confirm this idea: “You can’t enter the same water twice”, “All fake news are being promoted to the Web in the same way”, “Everyone will like the current news page on the site no less than I do” .

The behavior of, say, people as individuals is interpreted when correlating them with the “world”, so that their existence can begin and end. This means that in the region of D individuals can appear, populating it, and also can disappear, stopping dwelling in it. However, the conceivable ability remains to compare any two individuals who lived in D at “distinct times” moreover, two specific individuals are not always compared, and this requires quantification over the entire set of possible D individuals.

This interpretation clarifies the difference between sentences, the second of which relates to the past tense, the third to the future, while the first is universal in time.

4.2. Virtual individuals

When working with information from the very first efforts, it becomes possible to introduce virtual objects, which is unlimited. The virtualization of individuals representing information processes remains largely still relatively unexplored. Virtual individuals should not be perceived as ghostly, since they are ideal objects that are introduced to increase the regularity of the language. When introducing into circulation and using the names of virtual objects, first of all, simplification of formulations is achieved, which can significantly reduce the addition of emerging special cases.

The simplest example of such objects is the "average salary of the supply department employees for the current month", obtained on the basis of the monthly salaries of employees of this department actually stored in the database. Of course, such an object can be eliminated from among the virtual ones by replacing it with the corresponding indicator stored in the database. But disposability should not be considered the main property of these objects.

The elimination of each individual occurrence of an expression corresponding to a virtual object shows a significant dependence on the context. Thus, a virtual object in various contexts behaves, generally speaking, in different images, since it does not consist in the same relations with all other objects. So there is a source that substantially generates inconsistent objects. This is often used in practice, for example, by forming *types* of a database that are focused on the particular needs of specific users. Thus, it is quite possible to consider expressions for virtual objects really naming abstract objects. The information processes of augmented reality are built on this.

Quantification is usually allowed with respect to variables running through the range of possible individuals. This is due to the fact that virtual objects regularize the structure of the base domain of possible individuals D without causing additional complications and problems. In practice, virtual objects can often be eliminated by replacing them with contextual expressions over individuals from D. In the case when virtual objects regularize the initial structure of the domain D, it may well be that they cannot be contextually eliminated. This means that virtual individuals have different rights than individuals from D.

For example, assume the view definition of PopularBooks.

```
CREATE VIEW PopularBooks AS
SELECT ISBN, Title, Author, PublishDate
FROM Books
WHERE IsPopular = 1
```

In the future, for the created view, its name can be used in statements containing SELECT. For example, you can list all the titles of popular books, sorting them by author:

```
SELECT Author, Title
FROM PopularBooks
ORDER BY Author
```

Another example is the use of aggregate functions SUM, MAX, MIN, AVG, etc. In addition, the functions of determining the cardinality of the sets included in the database are often used. Purely mathematical examples of virtual objects can be left aside for now, mentioning perhaps certain descriptions used in ontologies.

As a result, we arrive at a distinction between virtual, possible, and actual objects. Denoting by V the domain of virtual objects, by D the domain of possible objects, and by A the domain of real objects, we accept

$$A \subseteq D \subseteq V.$$

The difference between A and D usually does not appear in database models, but becomes a decisive factor when considering information processes on the Web. In the case of the Web, the set of actual individuals A is replaced by a family of domains $A_i \subseteq D$ for $i \in I$, since indexing mechanisms play a decisive role on the Web. In fact, for an

individual from the point of view of information processes, "becoming actual" actually means "being represented on the Web", which in turn is understood as "being indexed":

$$A = \bigcup_{i \in I} \{A_i \mid A_i \subseteq D\},$$

which provides accessibility for search engines. In fairness, we note that for published articles, their indexation is the main factor in the scientific activity of the researcher, but in this case we are talking about other indexing systems.

4.3. Indexing

Deep Web content is hidden behind HTTP forms and includes many common uses, such as webmail, online banking, and services that users have to pay for and that are protected by paid networks, such as video-on-demand and some online magazines and newspapers. Deep Internet content can be found and accessed by a direct URL or IP address, and a password or other secure access may be required outside of the public website page.

We relate the data of the deep, invisible or hidden Web to the dpmain D of possible individuals, while the data of the surface Web we relate to the domain of actual individuals A.

5. Web search engine

The search engine on the Web itself is that an indefinite reference to the index is implicitly used in the expression of the search language. For example, information on the P-property is searched, which corresponds to the expression $\forall x.P(x)$ when it is known in advance that $\forall x \in D$, i.e. the quantifier ranges through the set of possible individuals D. This leads to generation of the set $\{d \mid P(d)\} \subseteq D$ of possible individuals d such that the expression $P(d)$ is true, i.e. d has the P-property. This reasoning forces us to write an abstraction of the index i

$$\lambda i.\|\forall x.P(x)\|i = \{h \in D^I | P(h)\}.$$

Since we are dealing with indexed sets, the quantifier must range over the actual individuals $\lambda i.\forall_i$, which gives

$$\lambda i \in I.\|\forall_i x.P(x)\|i = \{h \in A^I \mid P(h)\} \subseteq A^I (\equiv H_A(I)),$$

but it still does not allow us to give a conclusion about the truth or falsehood of the expression $\lambda i \in I.\forall_i x.P(x)$ containing a reference to an indefinite index (only the fact of indexing by means of some set I is indicated). In the written expression, the quantifier $\forall_i x \in A_i$ ranges over the set of actual individuals A_i, $H_A(I) = \{h | h : I \to A\}$ is the construction of *variable domain*. In this case, the individual h is a *process* $h : I \to A$.

Taking a specific indexing of I and an index of $i \in I$, one can also instantiate information processes $h : I \to A$ by taking $h_i \in A_i$. Then

$$\|\forall_i x.P(x)\|i = \{h_i \in A_i \mid P(h_i)\} = C_i \subseteq A_i,$$

which leads to the formation of a subset of C_i, on the elements of which the predicate $P(h_i) = true$ gives a true expression.

6. Conclusion

The paper proposes an approach to the organization of knowledge extraction from data, the organization of which is superimposed on the model structure of index expressions. This is a natural indexing that takes into account actual, possible and virtual objects. An object is considered in a certain way selected set of defining properties. In other words, the object manifests itself in interaction and is an information process.

1) The organization of a search on the Web requires taking into account the capabilities of the applied indexing systems, and the search results are organized in the form of a specific model structure.

2) This model structure allows us to cope with the solution of the problem of data identity in the conditions of their variability.

3) A semantic framework of a sufficient general form has been established, which allows generating images of biased concepts that correspond to the concepts and dependencies between them discovered over the Web. This method of knowledge extraction can be based on a network of variable concepts, which entails the possibility of developing a family of parameterized computational models.

Acknowledgements

This research is supported in part by the Russian Foundation for Basic Research, RFBR grants 20-07-00149-a, 19-07-00326-a, 19-07-00420-a, 18-07-01082-a, 17-07-00893-a.

References

[1] Bar-Hillel, Y., 1954. Indexical expressions. Mind 63, 359–379. URL: http://www.jstor.org/stable/2251354.

[2] Bergman, M.K., 2015. The deep web: Surfacing hidden value. Journal of Electronic Publishing 7, 1–25. URL: http://www.press.umich.edu/jep/07-01/bergman.html, doi:http://dx.doi.org/10.3998/3336451.0007.104.

[3] Carnap, R., 1947. Meaning and Necessity. University of Chicago Press.

[4] Helland, P., 2019. Identity by any other name. Commun. ACM 62, 80–80. URL: http://doi.acm.org/10.1145/3303870, doi:10.1145/3303870.

[5] Ismailova, L., Kosikov, S., Zinchenko, K., Wolfengagen, V., 2020a. Environment of modeling methods for indicating objects based on displaced concepts, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 137–148.

[6] Ismailova, L., Wolfengagen, V., Kosikov, S., Parfenova, I., 2020b. Increasing of semantic sustainability in the interaction of information processes, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 149–156.

[7] Ismailova, L., Wolfengagen, V., Kosikov, S., Volkov, I., 2020c. Computational model for granulating of objects in the semantic network to enhance the sustainability of niche concepts, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 157–164.

[8] Kosikov, S., Ismailova, L., Wolfengagen, V., 2020. Dynamics of recognition of properties in diagnostics, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 246–259.

[9] Lewandowski, D., 2019. The web is missing an essential part of infrastructure: An open web index. Commun. ACM 62, 24–24. URL: http://doi.acm.org/10.1145/3312479, doi:10.1145/3312479.

[10] Mungamuru, B., Garcia-Molina, H., 2006. Beyond Just Data Privacy. Technical Report 2006-18. Stanford InfoLab. URL: http://ilpubs.stanford.edu:8090/783/.

[11] Palermo, F.P., 1974. A data base search problem, in: Tou, J.T. (Ed.), Information systems: COINS IV, Proceedingd of the 4th International Symposium on Computer and Information Sciences, Miami Beach, Florida, December 14-16, 1972. NY Plenum Press, New York, pp. 67–101.

[12] Raghavan, S., Garcia-Molina, H., 2000. Crawling the Hidden Web. Technical Report 2000-36. Stanford InfoLab. URL: http://ilpubs.stanford.edu:8090/456/.

[13] Scott, D., 1970. Advice on modal logic, in: Lambert, K. (Ed.), Philosophical Problems in Logic: Some Recent Developments. Springer Netherlands, Dordrecht, pp. 143–173. URL: https://doi.org/10.1007/978-94-010-3272-8_7, doi:10.1007/978-94-010-3272-8_7.

[14] Slieptsov, I.O., Ismailova, L.Y., Kosikov, S.V., 2020. Representation of conceptual dependencies in the domain description code, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 507–514.

[15] Wolfengagen, V., Ismailova, L., Kosikov, S., 2020a. Cognitive features for stability of semantic information processing, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 581–588.

[16] Wolfengagen, V., Ismailova, L., Kosikov, S., Maslov, M., 2020b. Mutable applicative model to prevent entanglement of information processes, in: Samsonovich, A.V. (Ed.), Biologically Inspired Cognitive Architectures 2019, Springer International Publishing, Cham. pp. 589–596.