

Cross-language identity resolution and approaches to its solution*

Z.V. Apanovich, D.N. Cherepanov, A.G. Marchuk

Abstract. This paper describes approaches to the vocabulary normalization and cross-language identity resolution problems that arise when the LOD datasets are used to populate the content of scholarly knowledge bases. We have proposed several new heuristics, using additional information extracted from the full text sources of data. The first heuristics uses the full record track of a person, the second uses self-citation networks and the third uses the textual analysis of documents. The dataset of the Open Archive of the Russian Academy of Sciences and several bibliographic datasets are used as test examples.

Keywords: Linked Open Data, SPARQL, ontology alignment, cross-language identity resolution, self-citation network, tf-idf, LDA.

1. Introduction

One of the projects carried out at the A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences (IIS SB RAS) is aimed at populating the Open Archive of the Siberian Branch of the Russian Academy of Sciences (SB RAS Open Archive, Open Archive)¹ [1, 2] with the data of the Open Linked Data cloud (LOD) [3]. A fragment of the Open Archive page devoted to Academician A.P. Ershov is shown in Figure 1. A four-step strategy for the integration of Linked Data into an application is proposed in [4]. The problems of access to linked data (1), vocabularies (schema, ontology) normalization (2), identity resolution (3), and data filtering (4) should be solved manually or semi-automatically in addition to specific problems of the application. Specialized tools for solving separate problems are becoming available [5–8]. However, according to [8], “the large scale processing, schema mapping and data fusion are still in their infancy”. In our experiments, we use a toolkit containing the previously developed ontology visualization program [9]. In our case, the problem of integration is complicated by the fact that we have to solve a *cross-language* identity resolution problem. Our experiments have shown that conventional methods of identity resolution usually fail in the cross-language environment. The idea of our approach is to involve additional information sources and apply additional methods of analysis using these sources.

*Supported by the RFBR under Grant 14-07-00386- and SBRAS under Grant 15/10.

¹<http://duh.iis.nsk.su/turgunda/Home>

<http://duh.iis.nsk.su/VirtuosoEndpoint/Home/Samples>.

The dataset of the Open Archive structured by the BONE Ontology and several bibliographic datasets structured by the AKT Reference ontology [10] are used as test examples. Several sources of full-text documents, such as the texts of papers available on-line, Academician A. Ershov's Archive², and the digital library SpringerLink³, are also used.

2. Experiments on ontologies alignment

The content of the SB RAS Open Archive provides various documents (photo documents mainly) reflecting information about people, research organizations and major events that have taken place in the SB RAS since 1957. The Open Archive contains information about the employments, research achievements, state awards, titles, and participation in conferences, academic and social events for each person mentioned in the Archive. The Open Archive has 20 505 photo documents and facts about 10 917 persons and 1519 organizations and events. The data sets of the Open Archive are available as an RDF triple store, as well as a Virtuoso endpoint⁴ for the SB RAS Archive. Its RDF triple store comprises about 600 000 RDF triples. The structure of the Open Archive knowledge base is organized with the so-called BONE Ontology described in OWL and comprising 44 classes.


	1976	1988	участник	заместитель заведующего кафедрой вычислительной математики ММФ	Новосибирский государственный университет
	1979	1988	первое лицо	председатель	Комиссия по системному математическому обеспечению Координационного комитета по Вычислительной технике СССР (КОСМО ККВТ АН СССР)
			участник	участник	Событие "Общим собранием Академии наук избраны п Сибирск..."
			организатор	организатор	Празднование 10-летия Отдела программирования
			участник		Третий всесоюзный симпозиум "Системное и теоретическое программирование"
	1964	1988	участник	заведующий отделом	Институт вычислительной математики и математической геофизики СО РАН
	1959	1964	участник	заведующий отделом программирования	Институт математики им. С.Л. Соболева СО РАН
отраж. в документе	Отражение				
					

Figure 1. A fragment of the Open Archive page devoted to Academician A.P. Ershov

We are working with different data sets of the RKBExplorer.com, which brings together data from many well-known scholarly datasets. For example,

²<http://ershov.iis.nsk.su/ershov/english/scient.html>

³<http://link.springer.com/>

⁴<http://duh.iis.nsk.su/VirtuosoEndpoint/Home/Samples>

the RKB Explorer DBLP dataset⁵ contains the data of the DBLP Computer Science Bibliography⁶; the RKB Explorer ACM⁷ dataset corresponds to the data extracted from the digital library of the Association for Computing Machinery (ACM)⁸; the RKB Explorer Citeseer⁹ dataset is taken from the digital library CiteSeerx¹⁰; the RKB Explorer IEEE¹¹ dataset contains information about IEEE publications¹²; and so on.

It is worth noting that the number of the RKBExplorer.com datasets is constantly increasing. It is also important to note that the data sets of RKB-Explorer.com are not the exact copies of the respective (reciprocal) libraries. For example, the RKB Explorer DBLP and DBLP Computer Science Bibliography have the same lists of publications and authors, but these data sets use different heuristics in the identification of synonyms and homonyms. However, all the errors of the identification of persons that occurred in the DBLP Computer Science Bibliography database are repeated in the dataset of RKBExplorer.com. We intend to use these datasets as a source of additional data for the Open Archive. For example, we would like to extend the Open Archive knowledge base with information about the publications by the people who previously worked at the A.P. Ershov Institute of Informatics Systems. Since these data sets are structured with the AKT reference ontology, we need to establish mappings between several classes and relations of the BONE ontology and the AKT Reference ontology. A specific feature of the BONE ontology is that it uses a Link Data modeling pattern called “qualified relation pattern”[11]. It means that the entities usually described by means of a relationship in other ontologies may be described as an instance of a class in the BONE ontology. This template compensates for the lack of attributes of the RDF predicates.

For example, using the “from-date” and “to-date” properties of the *bone:participation* class, we are able to specify the facts like “Academician A.P. Ershov was the head of a department at the Institute of Mathematics SB AS from 1959 to 1964 and the head of a department at the Computing Center SB USSR AS from 1964 to 1988”. For the same reasons, the classes like *bone:dating*, *bone:naming*, and *bone:authorship* are used in the BONE ontology instead of the predicates *akt:has-author*, *akt:has-date* or *akts:has-pretty-name* used in the AKT Reference ontology. This augmentation of expressive possibilities complicates the problem of data integration, since a

⁵dblp.rkbexplorer.com

⁶<http://www.informatik.uni-trier.de/~ley/db/index.html>

⁷ acm.rkbexplorer.com

⁸dl.acm.org

⁹citeseer.rkbexplorer.com

¹⁰<http://citeseer.ist.psu.edu>

¹¹ieee.rkbexplorer.com

¹²<http://ieeexplore.ieee.org/Xplore/guesthome.jsp>

need arises to systematically establish a correspondence between the groups of classes and relations of ontologies. More precisely, a correspondence between one or several groups of the form “Class1 - relation1 - Class2” of the AKT Reference ontology and one or several groups of the form “Class3 - relation2 - Class4 - relation3 - Class5” of the BONE ontology should be created. In particular, a new instance of the Class4 for every triple “Class1:instance1 - relation 1 - Class2:instance2” should be created. This kind of translation can be carried out by an appropriate SPARQL-query. A simplified template of a SPARQL query that generates instances of the Class4 is as follows:

```

PREFIX iis:<http://iis.nsk.su#>
PREFIX akt:<http://www.aktors.org/ontology/portal#>
PREFIX akts:<http://www.aktors.org/ontology/support#>
CONSTRUCT {
  _:p a iis:Class4.
  _:p iis:relation2 ?instance1.
  _:p iis:relation3 ?instance2.
}
WHERE {
  ?instance1 akt:relation1 ?instance2.
  ?instance1 a akt:Class1.
  ?instance2 a akt:Class2.
}

```

Since the needed SPARQL-queries are rather tedious, we have created a program that can generate this kind of queries using the visualization of two ontologies. An example of generating the instances of the *bone:participation* class with respect to the *akt:has-affiliation* relation is shown in Figure 2. The two ontologies are drawn side-by-side and several additional buttons are used to control the alignment process. When the “Fix matching” check box is activated, we can choose groups of classes and relations for alignment in both visualization panels. The group “<Person>has-affiliation <Organization>” of the AKT Reference ontology is selected in the left panel and the group “<sys-obj>participant <participation>in-org <org-sys>” of the BONE ontology is selected in the right panel. A SPARQL query that generates the instances of the *bone:participation* class is generated automatically. However, in order to use these queries effectively, we should select people described in the Open Archive from the RKBExplorer list. In other words, the identity resolution problem should be solved.

3. Cross-language identity resolution

The step of identity resolution is very important for populating the Open Archive content. We have to match up properly the persons described in the Open Archive with the data about these persons extracted from other

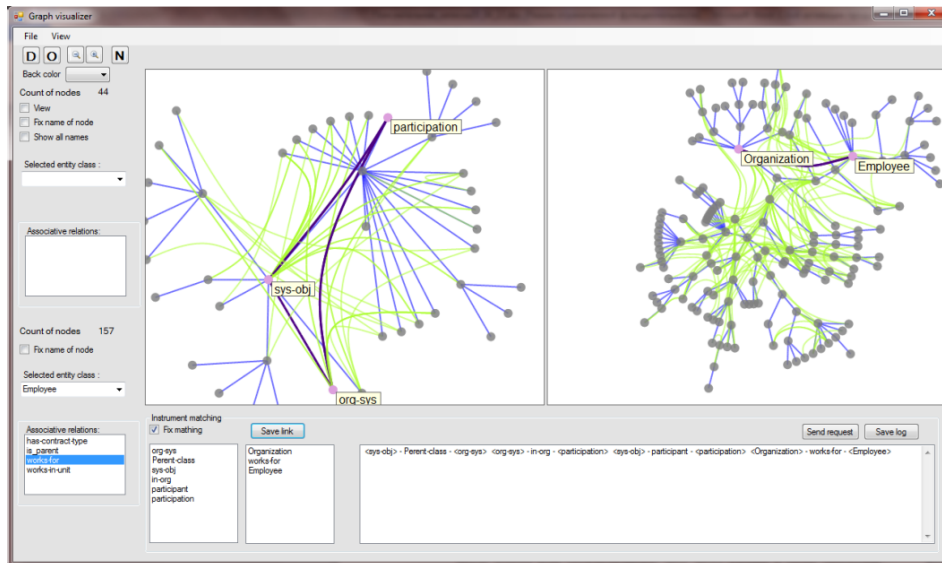


Figure 2. Interactive matching between two groups of classes and relations

data sets. The problem is complicated by the fact that the Open Archive uses names written in Cyrillic, and most of the RKBExplorer data sets use Latin names to identify the same persons. Thus, we have a *cross-language entity linking problem* to solve. The problem is known to have two other complicating features: entities can be referred to by using multiple name variants (aliases or misspellings) and several entities can share the same name (for example, there are many people called Petr Ivanov).

One might ask, of course, why not to use one of the Russian data sources, such as eLIBRARY.RU, to populate the database. Indeed, this digital library provides information about Russian researchers, but this is a relatively recent database, and the time period covered by eLIBRARY.RU is rather short. Therefore, it can be very useful for identifying the people and their publications pertaining to the last 10–15 years, but it becomes of little use when we are interested in somebody like Academician Andrei Petrovich Ershov, who died in 1988. Neither is there any information about B.A. Trakhtenbrot, V.E. Kotov or many other researchers who laid the foundations of the Soviet and Russian computer science. To compare, the DBLP contains information about publications dating from 1936. In the SB RAS Open Archive, all persons are specified by means of the *bone:name* attribute. The format of this attribute is <LastName, First Name Middle Name>. This attribute has two options: the Russian-language version and the English-language version. The English version is a transliteration of the Russian version. Every Russian name, however, can be transliterated in many ways. For example, the Russian family name Ершов can be spelt as Ershov, Yershov, Jerszow, and

the first name Андрей can be written as Andrei, Andrey, Andrew. Therefore, the first step of our identity resolution procedure is transliterating a name in the language of a dataset (i.e. English) and then performing a monolingual name matching in that language.

Prior to this, we tried to use VIAF [12], an important source for authority control at the international level which is becoming available. VIAF focuses mainly on national libraries and countrywide union catalogs. It is possible to find there many-language variants of the spelling of names of prominent people. VIAF mainly collects information about books, not about journals or conference papers. And the most frustrating fact about VIAF is that it also contains erroneous data. For example, several publications edited by or belonging to Academician A.P. Ershov are attributed in VIAF to a person identified as <http://viaf.org/viaf/196995053> and named Ershov, Aleksandr Petrovich. For these reasons, we have used the following procedure to identify the candidates for monolingual name matching with the RKBExplorer data sets:

- 1) Generate all possible transliteration variants for each normalized *bone:name*.
- 2) Generate all possible formats of *akt:full-name* for each *bone:name*.
- 3) Compare the generated full-name variants with the full-names of the RKBExplorer data sets.

Using string distance metrics for name-matching is not a new field of research and has an extensive bibliography [13–15]. Therefore, it suffices to say that a tokenized version of the Jaro-Winkler similarity metric was used at the third step of this procedure [13]. It means that the family names of the SB RAS Open Archive were compared with the *akt:full-name* of the RKBExplorer using a very high threshold value, and the variants of the first name and patronymic were compared with a lower threshold. At this stage, we are interested more in recall than in precision, and false positive results are quite common. Approximate name matching is a prerequisite for ensuring the completeness of the search, as unusual versions of the Latin spelling of the Russian names not covered by the rules of transliteration were regularly detected. For example, a person named A.P. Yersh'ov was found in the RKB Explorer DBLP.

The result of this procedure is a list of persons of the Open Archive who have matching persons in a dataset of the RKBExplorer. As a rule, a person of the Open Archive has several matches in the RKBExplorer with different spellings, and each person of the RKBExplorer has its own list of publications. For example, we have discovered 18 distinct persons whose *akt:full-name* is Andrei P. Ershov, two persons with *akt:full-name* Andrew Yershov, two more persons whose *akt:full-name* is A.P. Yersh'ov, one person named A. Ershov, and one person whose name is A. Yershov at the RKB Explorer DBLP. The person identified as <http://dblp.rkbexplorer.com/id/>

people1ac8593dbc7db6ec5766ea313914be4-1211d4d9974a0a977bd166da859d928f and named Andrei P. Ershov is the author of “Mixed computation in the class of recursive program schemata”. Another person identified as <http://dblp.rkbexplorer.com/id/people-e1ac8593dbc7db6ec5766ea313914be42fd1e3b39206345ab05fd9be97bc0d00> and named Andrei P. Ershov has a publication entitled “Time sharing: the need for re-orientation.” One more person identified as <http://dblp.rkbexplorer.com/id/people-8d3cb5ddb6e9bf9c359369b3cf3fb965955fb11eb7908a9dd2bb137161f8cb3d> and named Andrew Yershov is the author of the paper entitled “Unified Evaluation System for Audio Steganography Methods”.

By checking the DBLP Computer Science Bibliography, the counterpart of RKB Explorer DBLP, we can find the persons with the same names and the same publications. The only difference between RKB Explorer DBLP and the DBLP Computer Science Bibliography is that the DBLP Computer Science Bibliography brings together persons with equal full-names: there is one person per one variant of a name. Thus, there are five persons who might be a homonym of Andrei Petrovich Ershov in the SB RAS Open Archive: Andrei P. Ershov, Andrew Yershov, A.P. Yersh’ov, A. Ershov, A. Yershov. Each of these persons has its own list of publications. A person identified as A. Ershov has the publication “A. Ershov, A. Nariniy, I. Mel’chuk: RITA – An Experimental Man-Computer System On A Natural Language Basis. IJCAI 1975: 387-390”, and a person named A.P. Yersh’ov has the following two publications:

- 1) A. P. Yersh’ov: ALPHA – An Automatic Programming System of High Efficiency. J. ACM 13(1): 17-24 (1966)
- 2) A. P. Yersh’ov: One View of Man-Machine Interaction. J. ACM 12(3): 315-325 (1965)

Experiments with other persons of the Open Archive have shown that the publications of people with Russian names written in English are scattered between several “virtual” persons. On the other hand, the publications of several distinct persons are often attributed to one person. A similar situation is observed with the counterparts of the RKBExplorer data sets. It means that we have to answer the following questions:

1. Which of these identifiers correspond to the same physical object, and, therefore, can be connected by the relation *owl: sameAs* and which of them describe distinct physical objects?

2. Do all publications attributed to a person belong to this person? This is a well-known name ambiguity problem. Of course, name disambiguation has an extensive literature [16]. Personal names can be disambiguated by comparing the attributes and relations associated with each entity using the facts like e-mail, personal website, affiliation, etc. However, it is important to note that the datasets like the DBLP have their own authority control and

disambiguation procedures [17]. The existence of several “virtual persons” corresponding to one real-world person in this dataset indicates that the conventional methods of identity resolution have failed and that additional efforts in entities disambiguation need to be made.

First, we suggest using additional sources of information. The structured data stored by the RKBExplorer datasets are not sufficient for person identification because of the restrictive nature of the AKT Reference ontology. So we use the Open Archive itself as an additional source of information, as it provides the so-called “track records” – a list of affiliations with a related period for every person. Basically, these data could be used to disambiguate persons by comparing them with the data of RKBExplorer (if available). Regrettably, there is little or no information about affiliations, and job periods are not specified at all. This is the reason why we are trying to answer these questions by checking the full-text versions of publications.

Besides, authors usually cite their earlier publications. This might allow several people with distinct identifiers but related with the self-citation relation to be considered as a single person. The citation information is represented by the *akt:cites-publication-reference* relation on RKBExplorer. This information is also incomplete. The citation network generated on the basis of the *akt:cites-publication-reference* relationship between the publications of a single author is sparse, with many isolated nodes. However, this information can be also extracted from the text of a paper. Nowadays, plenty of full-text resources are becoming available, which enables an increase in knowledge discovery. The first group of our experiments was carried out using Academician A. Ershov’s Archive. A complementary data source was the digital library SpringerLink. Finally, many publications were available online. Currently we are checking the three heuristics:

1) Check the workplace and its dates. The publication date and authors’ affiliation are extracted from the textual version of the publication and compared with the person’s list of jobs of the OpenArchive.

2) Check the self-citation list. The name of the author of each publication is compared with the names of the authors of the cited publications. If a coincidence of the names is found, the current publication is added to the set of publications of the reference list. Then, the same procedure is applied to the added publications.

3) Apply textual analysis. The Tf-idf weighting with cosine similarity and Latent Dirichlet Allocation (LDA) [18] are used to estimate similarity between different texts in order to attribute these texts to the same person or to distinct persons. Here are several examples illustrating our approach.

Example 1. (A.P. Ershov and Andrew Yershov)

There are two persons named Andrew Yershov in rkbexplorer.com.

Their publications are dated by 2009-2011 and their affiliations (available as full-text papers) are Riga Technical University. These affiliations do not

intersect with the list of affiliations of Academician A.P. Ershov and, hence, these persons have nothing to do with the Open Archive.

Example 2. (A. Ershov vs. A. P. Yersh'ov)

The list of publications of the person named A. Ershov contains a single entry referred to as “A. Ershov, A. Nariniany, I. Mel'chuk RITA – An Experimental Man-Computer System on A Natural Language Basis”.

Does this publication belong to Academician A.P. Ershov from the SB RAS Open Archive and as such should it potentially fall into the list of all publications attributed to the RKB Explorer DBLP person with the *akt:full-name* attribute equal to Andrei P. Ershov? Since we have Academician A. Ershov's Archive at our disposition, it suffices to compare this title with the titles of the papers specified in this archive to answer positively. If such information was not available, we would have to apply a more complex procedure: to extract the author's name, title, affiliation, and the list of references from the text of the publication (available on-line). The author of the textual version is specified as “A. P. Ershov”; his affiliation is “Computing Center, 630090, Novosibirsk, USSR”, which matches the data of the Open Archive. In this way, we obtain another affirmative answer to our question. The text analysis of this publication has another benefit: it allows us to identify correctly another paper attributed to the person named A. P. Yersh'ov. This person has the publication entitled “One View of Man-Machine Interaction” mentioned in the list of references of the publication “RITA – An Experimental Man-Computer System on a Natural Language Basis”. Thanks to this information, persons with differing names can be merged and matched with a single person of the Open Archive. By checking the list of publications in Academician A. Ershov's Archive, we can validate this choice.

Example 3. (A paper not mentioned in Academician A. Ershov's Archive)

The paper “Axiomatics for memory allocation” is attributed to a person named Andrei P. Ershov by RKB Explorer DBLP. However, a paper with this title is not listed in Academician A. Ershov's Archive. Using the SpringerLink digital library as an additional data source, we can find out that the affiliation of the author of the paper “Axiomatics for memory allocation” is “Computing Center, 630090, Novosibirsk, USSR”. This confirms the fact that Academician A. Ershov can be its author. But the SpringerLink digital library provides another kind of useful information: the lists of references for every publication. We can find the paper “Axiomatics for memory allocation” in the reference list of another publication “The Transformational Machine: Theme and Variations” attributed to Academician A.P. Ershov by Academician A. Ershov's Archive and to Andrei P. Ershov by RKB Explorer DBLP. Moreover, the author of the paper “Axiomatics for memory allocation” is mentioned in the reference list as Ershov, A.P. Taking into account the additional information about affiliations and self-citations provided by

the SpringerLink digital library, it is possible to suggest that:

1. The author of the paper “Axiomatics for memory allocation” and the author of the paper “The Transformational Machine: Theme and Variations” named Andrei P. Ershov at the RKB Explorer DBLP are the same person (Figure 3, right column).
2. The paper “Axiomatics for memory allocation” belongs to Academician A.P. Ershov and needs to be added to the Open Archive (Figure 3, left column).

This situation is illustrated in Figure 3. The left column represents the data about these two publications in Academician A. Ershov’s Archive, the central column – in SpringerLink, and the right column – in RKB Explorer DBLP

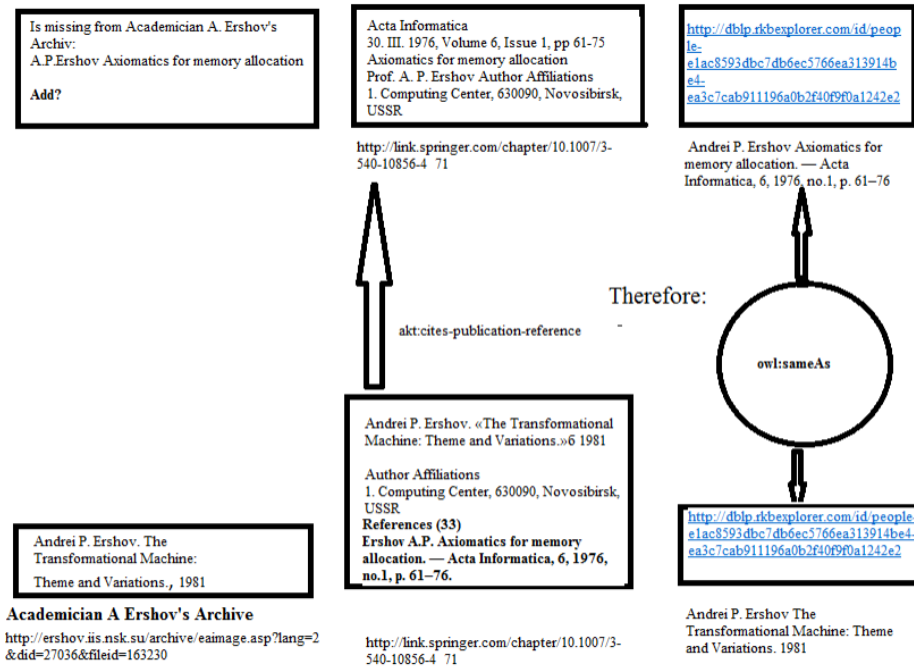


Figure 3. The relationship between the two publications in the datasets of Academician A. Ershov’s Archive, SpringerLink, RKB Explorer DBLP

4. Textual analysis for identity resolution

One of the most effective methods for the semantic analysis of textual data is Latent Dirichlet Allocation. A simpler and computationally cheaper alternative is to calculate document similarity using the tf-idf weighting with

cosine similarity. At the moment, we have implemented the first version of a program for identity resolution using additional data sources such as SpringerLink and additional methods for text analysis. Before computing the cosine similarity¹³, the text is cleaned by removing stop words and leaving only plain content-carrying words, then a stemming procedure¹⁴ is applied. After that, the collection of documents is treated as a graph. Each document is a node identified by its number in the list of documents and every pair of documents is connected by an edge whose weight (W) is given by the similarity between the two documents. A threshold is applied to the similarity matrix to ignore the links between documents with low similarity. The threshold depends on the number of nodes. For example, the threshold is equal to 0.05 for 30 nodes. The obtained graph is drawn by a usual force-directed placement algorithm so that similar documents are placed close to each other. In our case, the force of attraction and the repulsion force both depend on the weight of the edge between vertices.

Force of attraction = Temperature \times SpringForce(d) \times $W \times$ SpringForceK;
 Force of repulsion = Temperature \times ElectricForce(d) / $W \times$ ElectricForceK;
 SpringForce(d) = $2\log(d)$;
 ElectricForce(d) = $1 / d^2$;

where d is the distance, W is the similarity between two vertices, SpringForceK and ElectricForceK are parameters.

A placement of several articles attributed to the persons named A.P. Ershov in the SpringerLink digital libraries is shown in Figure 4.

Each line of the placement log shown on the right represents the information about the relation between two vertices (documents). For example, a line $U = \#27$ $V = \#28$ Distance = 1.078 $1/W = 4.033$. $K = W \times D = 0.267$ can be interpreted as follows: the vertex V represents document 27 and the vertex U represents document 28 in the list of documents. The distance between these vertices is equal to 1.078 when the placement step is finished. To estimate the placement quality, we calculate $K = W \times D$ for every edge and chose the placement with the smallest dispersion (K). For the current placement, the dispersion is 0.0052.

A user can get information about any paper by clicking on the respective graph node, and a click on a graph edge displays information about the words responsible for the papers similarity. Two sets of papers belonging to two different persons can be clearly identified. However, we are going to complement this program with a clustering procedure.

¹³<http://www.codeproject.com/Articles/12098/Term-frequency-Inverse-document-frequency-implemen>

¹⁴<http://snowball.tartarus.org/>

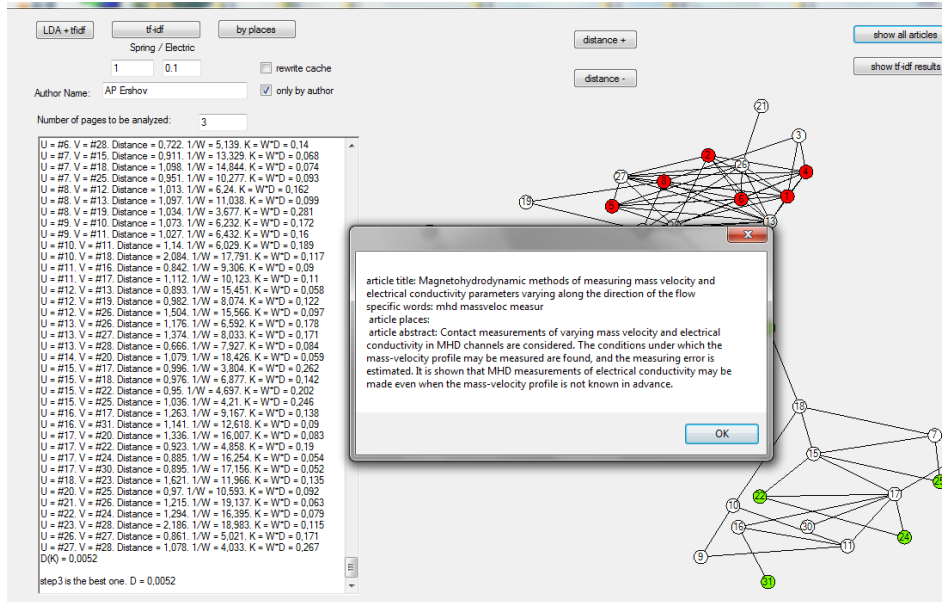


Figure 4. A placement of several articles attributed to persons named A.P. Ershov in the SpringerLink digital library

5. Conclusion

In this paper, we have considered several problems that arise when a Russian scholarly knowledge base is populated using the bibliographic data sets of LOD cloud and some approaches to their solution. It was demonstrated that the conventional tools used for name disambiguation as a part of identity resolution problem perform poorly in the cross-language context. We proposed several new heuristics using additional information extracted from the full text sources of data. The first heuristics uses the full record track of a person, the second uses self-citation networks and the third uses the textual analysis of documents. To verify this approach, we have compared the data extracted from the RKBExplorer datasets with the data of Academician A. Ershov's archive and eLIBRARY.RU. The experiments allowed us to identify many different persons of RKBExplorer.com as a single person. Besides, we have discovered several publications by Academician A.P. Ershov missing from Academician A. Ershov's archive and several papers that have never existed but were attributed to him. However, this approach needs further development and experiments. For example, we plan to extract full-sized citation networks for our experiments. Another problem encountered during our experiments is that the Russian authors usually cite their papers written in Russian translating their title into English. But the international catalogs

such as WorldCat¹⁵ use Latin transliteration of the publications written in Russian. Therefore, we have to compare the titles written in Russian with several variants of their translation into English. Besides, we are going to extend the set of data sources used in our experiments. But the most important observation is that the identity resolution problem is essentially the cross-language problem and, to solve this problem successfully, data sources written in native languages should be considered.

References

- [1] Marchuk A.G., Marchuk P.A. Specific features of digital libraries construction with linked content // Proc. of the RCDL'2010 Conf. – 2010. – P. 19–23 (In Russian).
- [2] Apanovich Z., Marchuk A. Experiments on using the LOD cloud datasets to enrich the content of a scientific knowledge base // Proc. of KESW 2013. – CCIS. – Springer Verlag, 2013. – Vol. 394. – P. 1–14.
- [3] Bizer C., Heath T. , Berners-Lee T. Linked Data – The Story So Far // Int. J. Semantic Web Inf. Syst. – 2009. – Vol. 5 (3). – P. 1–22.
- [4] Schultz A. et al. How to integrate LINKED DATA into your application // Semantic technology & Business Conference, San Francisco, June 5, 2012. – http://mes-semantic.com/wp-content/uploads/2012/09/Becker-et-al-LDIF_SemTechSanFrancisco.pdf.
- [5] Isele R., Jentzsch A., Bizer Ch. Silk Server – Adding missing Links while consuming Linked Data // 1st Internat. Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.
- [6] Ngomo A.-C. N., Auer S. LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data // IJCAI 2011: Proc. of the 22nd Internat. Joint Conf. on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. – P. 2312–2317.
- [7] Shvaiko P., Euzenat J. Ontology matching: state of the art and future challenges // IEEE Trans. on Knowledge and Data Engineering. – 2013. – Vol. 25(1). – P. 158–176.
- [8] Tramp S., Williams H., Eck K. Creating Knowledge out of Interlinked Data: The LOD2 Tool Stack. – <http://lod2.eu/Event/ESWC2012-Tutorial.html>
- [9] Apanovich Z. V., Vinokurov P. S. An extension of a visualization component of ontology based portals with visual analytics facilities // Bulletin NCC. Series: Computer Science. – 2010. – IIS Special Iss. 31. – P. 17–28.
- [10] Description of AKT Reference ontology. – <http://www.aktors.org/ontology>.

¹⁵<http://www.worldcat.org/>

- [11] Dodds L., Davis Ia. Linked Data Patterns. – <http://patterns.dataincubator.org/book/>
- [12] Bourdon, F., Boulet V. ‘VIAF: A hub for a multilingual access to varied collections’ // World Library and Information Congress: 78th IFLA General Conference and Assembly, August 2013, Singapore [online]. – <http://conference.ifia.org/past-wlic/2011/79-bourdon-en.pdf> (Accessed 13 July 2014).
- [13] Christen P. A Comparison of Personal Name Matching: Techniques and Practical // Issues. – TR-CS-06. – <https://digitalcollections.anu.edu.au/bitstream/1885/44521/3/TR-CS-06-02.pdf>.
- [14] Cohen W.W., Ravikumar P.D., Fienberg S.E. A Comparison of String Distance Metrics for Name-Matching Tasks // IIWeb. – 2003. – P. 73–78.
- [15] Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. Duplicate record detection // A Survey J. IEEE Transactions on Knowledge and Data Engineering. – 2007. – Vol. 19, Iss. 1. – P. 1–16.
- [16] Mann G., Yarowsky D. Unsupervised personal name disambiguation // Proc. of the 7th Conf. on Natural Language Learning at HLT-NAACL. – 2003. – Vol. 4. Association for Computational Linguistics. – P. 33–40.
- [17] Ley M. DBLP – Some lessons learned // PVLDB. – 2009. – Vol. 2(2). – P. 1493–1500.
- [18] Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // J. of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.