Experiments on Russian-English Identity Resolution

Zinaida Apanovich[1,2(✉)] and Alexander Marchuk[1,2]

[1] A.P. Ershov Institute of Informatics Systems, Siberian Branch
of the Russian Academy of Sciences, Novosibirsk, Russia
{apanovich,mag}@iis.nsk.su
[2] Novosibirsk State University, Novosibirsk, Russia

Abstract. The focus of this paper is on Russian-English identity resolution when English names of entities have been created by a transliteration or translation of Russian names. A new approach combining attribute-based identity resolution with the text analysis of publications attributed to these entities has been proposed. The dataset of the Open Archive of the Russian Academy of Sciences and digital library SpringerLink are used as test examples.

Keywords: Linked open data · Cross-language identity resolution · Authorship attribution · Self-citation network · Tf-idf · LDA · Jaro-Winkler

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] with the data of the Open Linked Data cloud (LOD) [2]. The problem of identity resolution is complicated by the fact that the Open Archive uses names written in Cyrillic, and other data sets use Latin names to identify the same persons. Our recent experiments [3] have shown that this problem has *language specific aspects* because sometimes a name in one language can be obtained by translation or transliteration of the name in another language. Several named entities with distinct English spellings and translations of their names may correspond in reality to the same Russian entity; on the other hand, several distinct entities may be homonyms and share the same name or some forms of this name.

Name ambiguity in the context of bibliographic citation records is a difficult problem that affects the quality of content in digital libraries. The library community has been working on it for a long time [4-6]. In the context of Linked Open Data and increasing data traffic on a global scale, the issues of data quality and confidence have become extremely important. In this environment, errors are promulgated, dispersed, and become difficult to discover and repair. As the number of homonyms and synonyms increases, it becomes crucial to have accurate data identifying various entities.

[1] http://duh.iis.nsk.su/turgunda/Home

R.B. Allen et al. (Eds.): ICADL 2015, LNCS 9469, pp. 12–21, 2015.
DOI: 10.1007/978-3-319-27974-9_2

An important aspect of this problem is multilingualism. Multilingual resources such as DBPedia, VIAF, WorldCat, etc., become increasingly common. Our experiments with several multilingual datasets have shown that Russian names admitting several transliterations are often treated as homonyms, and several persons with identical name variations are treated as synonyms.

Experiments with the RKBExplorer datasets[2] have shown that a person of the Open Archive has several matches in the RKBExplorer with different spellings and these matching persons have disjoint lists of their publications. For example, the publications authored or edited by Academician Andrei Petrovich Ershov have been attributed to 18 distinct persons whose names are Andrei P. Ershov, A.P. Yersh'ov, A. Ershov, and A. Yershov in DBLP RKBExplorer. By checking the DBLP Computer Science Bibliography, the counterpart of RKB Explorer DBLP, three distinct persons with publications belonging to Academician Andrei Petrovich Ershov have been identified. Their names are various forms of the Latin transliteration of "Андрей Петрович Ершов".

Experiments with WorldCat.org have shown that this resource, too, is not free from identification errors when Russian authors are considered. For example, the list of WorldCat Identities[3], containing descriptions of particularly prominent personalities, has a record dedicated to Academician Andrei Petrovich Ershov. It contains information about the books and papers authored or edited by Academician A.P. Ershov mixed with the publications authored by another A.P. Ershov (Alexander Petrovich) from Novosibirsk. Articles authored by A.P. Ershov and published between 1989 and 2012 have been described as "publications by Andrei Petrovich Ershov published posthumously" (Academician A.P. Ershov died in 1988). It is possible to find among these "posthumous publications" an article entitled "Capillary Flow of Cationic Polyelectrolyte Solutions". The text of this article indicates the affiliation of its author as the "Institute of Physical Chemistry, Russian Academy of Sciences". Academician A.P Ershov never worked for this organization, which means that this article is authored by another A.P. Ershov (Albert Petrovich).

Another example is VIAF, the Virtual International Authority File. Its web interface is available on http://viaf.org. The source files of VIAF include some of the most carefully curated files of names available. In addition, the bibliographic records using the files are professionally created, often reviewed and corrected by many libraries. In spite of the substantial work put into the creation and maintenance of the files, they still have inaccuracies. For example, VIAF has attributed several papers edited by or written by Academician A.P. Ershov to a person identified as http://viaf.org/viaf/196995053 and named Ershov, Aleksandr Petrovich. On the other hand, among the publications attributed to Academician Andrei Petrovich Ershov there are two books on economics (http://viaf.org/viaf/5347110), which can hardly belong to him.

This paper presents our approach to Russian-English identity resolution using text analysis methods in combination with attribute-based methods. The rest of this paper is organized as follows. Section 2 discusses specific features of our data sets that can

[2] www.rkbexplorer.com

[3] www.worldcat.org/wcidentities/lccn-n80162678

be used for identity resolution. Section 3 presents our algorithm along with some experimental results that show its effectiveness. Finally, Section 4 presents our conclusions and outlines future work.

2 Datasets and Evidence Used in the Disambiguation Task

The content of the SB RAS Open Archive provides various documents, mainly photos, 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, 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 1,519 organizations and events. The data sets of the Open Archive are available as an RDF triple store, as well as a Virtuoso endpoint for the SB RAS Archive. Its RDF triple store comprises about 600,000 RDF triples.

In the SB RAS Open Archive, all persons are specified by means of a normalized name. The format of a normalized name 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. However, several English name variations can correspond to a normalized Russian name. It can be < First Name Middle Name Last Name>, <First Name Last Name>, <First Name First letter of the Middle Name Last Name >, < First letter of the First Name First letter of the Middle Name Last Name >, etc. All this forms should be first generated in Russian and then transliterated in English. Again, every Russian name 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, in order to identify in an English knowledge base all the possible synonyms of a person from the Open Archive, we have to generate the most complete list of English spellings for each Russian name. This procedure is applied in the character by character manner, but some characters need special attention: the vowels such as “я”, “ю ”, and “е”, can be transliterated in many ways. For example, the character “я” can be spelt in English as “ia”, “ya”, and “ja”. The same is true of the consonants such as “й” and “щ”. Another source of multiple transliterations is such character pairs as “ья”, “ью”, etc.

All generated forms of names are used as key words to search for articles in a target database. The authors of the extracted articles can be both homonyms and synonyms. We have to process the list of articles and determine which of their authors are synonyms and which of them are homonyms. In other words, the list of articles should be clustered into subsets S_1, S_2,…, S_n such that each subset of articles is authored by a single person and all his or her name variations are synonyms. Subset S_1 should contain the articles authored by the person from the Open Archive.

A well-known system for attribute-based identity resolution in the context of the Open Linked Data is SILK [7]. The heuristics used by VIAF and DBLP for ambiguity detection are described in [8,9]. However, cross-language entity disambiguation requires taking into account differences in orthography between languages and differences in the way these words in different languages are used to express similar meanings.

Since we need to identify the persons of the Open Archive, we use the Open Archive itself as an additional source of information. It provides the so-called "track records" – a list of affiliations with a related period for every person. For example, the Open Archive contains such facts as "Academician A.P. Ershov was the head of a department at the Institute of Mathematics SB AS USSR from 1959 to 1964 and the head of a department at the Computing Center SB AS USSR from 1964 to 1988".

Another useful kind of information stored in the Open Archive is the list of names for every organization. For example, there is information that the Institute of Computational Mathematics and Mathematical Geophysics SB RAS was called Computing Center SB AS USSR from 1964 to 1991.

The author's affiliations, extracted from SpringerLink, are compared with the information about employment of persons of the Open Archive. Note that the English names of Russian organizations indicated by the authors of articles quite often do not correspond to their standard translations, and can contain unusual abbreviations and nonstandard word order. To tackle this problem, we have developed a cyclic modification of Jaro-Winkler algorithm [10].

Sometimes there is no information about author's affiliation or this information is rather general, such as "Russian Academy of Sciences". To classify this kind of article, we use semantic text analysis methods. Now a large number of publications are digitized, and the most important attribute characterizing each researcher is her or his publications. Nowadays, quite advanced methods of authorship attribution exist, including analysis of character, lexical, syntactic, semantic and application-dependent features [11-13]. The use of these methods is governed by the idea that authors unconsciously tend to use similar lexical, syntactic or semantic patterns. However, when comparing the English texts published by Russian authors, character, lexical, and syntactic methods do not seem to be the most appropriate because different texts of the same author can be translated by different translators who vary in their translation styles. On the other hand, semantic analysis of the texts can reveal, for example, their terminological similarity.

Thus, we suggest a combined use of articles metadata comparison and their text similarity estimation for the cross-language identity resolution.

As a source of detailed meta-data, the digital library SpringerLink has been chosen. SpringerLink is currently one of the largest digital libraries that holds more than 9,000,000 documents in various fields of research: computer science and mathematics, life sciences and materials, philosophy and psychology. Its wide range of fields corresponds well to the multidisciplinary orientation of the SB RAS Archive. SpringerLink contains the full texts of many articles. If the full text of a publication is not available, SpringerLink provides detailed meta-data about the publication, such as ISSN, abstract, the affiliations of its authors (if it is specified in the text of the publication), references etc. Finally, SpringerLink is one of the sources used by a part of the Open Linked Data cloud WorldCat.org (http://worldcat.org).

Basically, our program was used to identify the persons described in the Open Archive. Nevertheless, there is a mode which allows for use of the identity resolution program for persons missing from the Open Archive.

3 Algorithm for Identity Resolution

The general scheme of our algorithm is shown in Fig. 1.

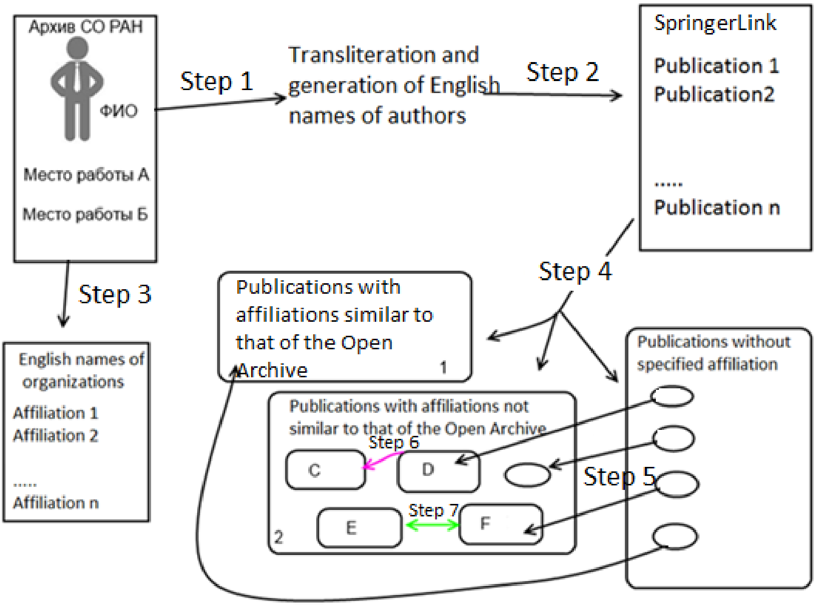

Архив СО РАН
ФИО
Место работы А
Место работы Б
Step 1
Transliteration and generation of English names of authors
Step 2
SpringerLink
Publication 1
Publication2
.....
Publication n
Step 3
English names of organizations
Affiliation 1
Affiliation 2
.....
Affiliation n
Step 4
Publications with affiliations similar to that of the Open Archive
1
Publications without specified affiliation
Publications with affiliations not similar to that of the Open Archive
Step 6
C
D
Step 5
E
Step 7
F
2

Fig. 1. The general scheme of our algorithm.

1. Our program takes as input a string *R_string,* corresponding to a normalized Russian name, and returns a set of all possible English transliteration and form variations *E_strings* as it is explained in Section 2. Initially, we used Google translate for the *E_strings* generation, but when we noticed some inaccuracies, we developed our own transliteration program. This step should allow the extraction of the most complete set of synonyms for the given person.
2. Each generated string s∈*E_strings* is used for key word search in SpringerLink. This search results in a list of documents where the key word can occur in the title of article, in the name of organization, in the reference list, etc. All the articles are filtered, and only publications having one of the key words as the author are retained. Each article is specified by a unique identifier. SpringerLink indexes several kinds of data formats (txt, PDF, PNG). For our experiment, however, we convert non-text formats into text and make use of plain text files. A set of meta-data such as citation_publisher, citation_title, citation_language, citation_pdf_url, citation_author, citation_author_affiliation, etc. are extracted and concatenated to create a text for analysis. The authors of the extracted articles can be both homonyms and synonyms. We have to process the list of articles and determine which of their authors are synonyms and which of them are homonyms. In other words, the list of articles should be clustered into subsets S_1, S_2,…, S_n such that each subset of articles is authored by a single

person and all his or her name variations are synonyms. Subset S_1 should contain the articles authored by the person from the Open Archive.

3. The publication date and authors' affiliations, provided by SpringerLink are compared with the person's list of affiliations specified by the Open Archive. Again, English names of organizations should be compared against their Russian counterpart. At this stage using transliteration is inappropriate, therefore Google translate and viaf.org web services are used for generating the English variants of the Russian names of organizations. For example, "Институт Систем Информатики" is identified in VIAF as VIAF ID: 159616561 (Corporate). Permalink: http://viaf.org/viaf/159616561. Its English name provided by VIAF is "A.P. Ershov Institute of Informatics Systems". However, only 10 percent of the names of organizations occurring in the Open Archive can be found at VIAF.org. If there is no English counterpart for the Russian name of an organization, its Google translation is used.
4. To distinguish persons whose affiliations specified in SpringerLink coincide with these indicated in the Open Archive, we generate a matrix that measures pair wise similarity between stemmed affiliations (names of organizations). Cyclic Jaro-Winkler distance [10] is used as the measurement. That is, a sequence of words that make up the name of the first organization, is cyclicly shifted to find the longest subsequence that matches the name of the second organization. Based on this comparison, the whole list of articles S is subdivided into three subsets P_1, P_2, and P_3, where P_1 is a set of articles whose affiliations are similar to one of the list of affiliations specified for the given person in the Open Archive, P_2 is a set of articles whose affiliations are not similar to any of the list of affiliations specified by the Open Archive for the given person, and P_3 is a set of articles that have no specified affiliation for the considered author. The publications of the subset P_2 are further subdivided into groups of articles $Group_2 \ldots Group_K$ such that each group of articles corresponds to exactly one place of work.
5. The program tries to distribute articles without specified affiliations among the groups of publications with specified affiliations. This procedure is based on the text similarity of articles. Two options are available at this stage. One of the most effective methods for the semantic analysis of textual data is Latent Dirichlet Allocation with Cullback-Leibler divergency [14,15]. A simpler and computationally cheaper alternative is to calculate document similarity using the tf-idf weighting with cosine similarity[4]. The results presented in this paper are mainly obtained using the tf-idf weighting with cosine similarity[5]. Before computing the text similarity, the text is cleaned by removing stop words and leaving only plain content-carrying words, then a stemming procedure[6] is applied. If the similarity value for an article A is below threshold for every group of articles $Group_1$, …, $Group_N$, the program creates a new group $Newgroup_{N+1}$, where N+1 is the serial number of the newly created group.

[4] http://www.codeproject.com/Articles/12098/Term-frequency-Inverse-document-frequency-implemen

[5] http://www.codeproject.com/Articles/12098/Term-frequency-Inverse-document-frequency-implemen

[6] http://snowball.tartarus.org/

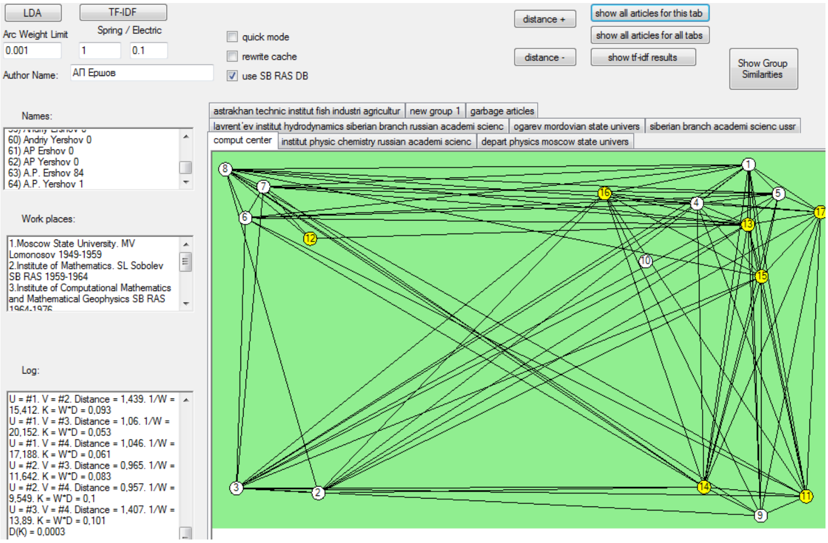
6. Some articles are specified by general affiliation such as “Russian Academy of Sciences”. The program tries to distribute articles with more general affiliations among the groups of articles with more specific affiliations. For example, “Siberian Division of the Russian Academy of Sciences” is considered to be more general with respect to “A.P. Ershov Institute of Informatics Systems of the Siberian Division of the Russian Academy of Sciences”. More general affiliation is a substring of a more specific one. If the author‘s affiliation specified in SpringerLink is considered to be a general name of an organization, the program tries to decide which of the more specific names of the organization can be used as author’s affiliation. Text similarity of the articles from groups with the exact names of organizations is used at this stage.
7. Text similarity measure is applied again to compare the generated groups of articles and if the similarity value for two groups of articles Group_i and Group_j exceeds the threshold value, the two groups are merged into one.
8. 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: $k\times N_{\text{nodes}}$, where the factor k was chosen experimentally as 0.0015. 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.

The result of the program is a set of SpringerLink articles subdivided into several groups. The first group of articles is attributed to the person described in the Open Archive.

4 Results

Experiments have shown that name variations generated by our program were more appropriate than that of Google translate. For example, 408 English variants have been generated by our program for the Russian name “Валерий Александрович Непомнящий,” among which only five variations have been discovered in Springer-link: V.A. Nepomnyashchii, Valery Nepomniaschy, V.A. Nepomniaschy, Valery A. Nepomniaschy, V.A. Nepomnyaschy. Google Translate created 160 variants of the same name, but some name variations existing in SpringerLink were absent from Google translate results. For example, the name variation “V.A. Nepomnyaschy” existing in SpringerLink was generated by our program but was not generated by Google Translate at the time of our experiments. A specific point of Google Translate was that along with the transliteration of names, it generated their translations. For example, Google translate generated variants like “Valery oblivious” for the Russian name “Валерий Непомнящий”, and “Vadim cats” for the Russian name “Вадим Котов”. For the Russian name «Андрей Петрович Ершов» 64 English variants have been generated.

An example of the program, searching for the articles authored by the Russian person "А.П. Ершов" in the digital library SpringerLink is shown in Fig. 2. English variations of a Russian personal name are displayed in the upper left tab. English versions of affiliations for the given person are displayed in the middle left tab. In the center, a graph representing the publications attributed by the algorithm to the person from the Open Archive is shown.


LDA
TF-IDF
distance +
distance -
show all articles for this tab
show all articles for all tabs
show tf-idf results
Show Group Similarities
astrakhan technic institut fish industri agricultur
new group 1
garbage articles
lavrent'ev institut hydrodynamics siberian branch russian academi scienc
ogarev mordovian state univers
siberian branch academi scienc ussr
comput center
institut physic chemistry russian academi scienc
depart physics moscow state univers
8
7
6
12
16
1
5
4
17
13
10
15
3
2
14
9
11

Arc Weight Limit: 0.001

Spring / Electric: 1 0.1

Author Name: АП Ершов

□ quick mode
□ rewrite cache
☑ use SB RAS DB

Names:

60) Andriy Yershov 0
61) AP Ershov 0
62) AP Yershov 0
63) A.P. Ershov 84
64) A.P. Yershov 1

Work places:

1.Moscow State University. MV Lomonosov 1949-1959
2.Institute of Mathematics. SL Sobolev SB RAS 1959-1964
3.Institute of Computational Mathematics and Mathematical Geophysics SB RAS 1964-1976

Log:

```
U = #1. V = #2. Distance = 1,439. 1/W =
15,412. K = W*D = 0,093
U = #1. V = #3. Distance = 1,06. 1/W =
20,152. K = W*D = 0,053
U = #1. V = #4. Distance = 1,046. 1/W =
17,188. K = W*D = 0,061
U = #2. V = #3. Distance = 0,965. 1/W =
11,642. K = W*D = 0,083
U = #2. V = #4. Distance = 0,957. 1/W =
9,549. K = W*D = 0,1
U = #3. V = #4. Distance = 1,407. 1/W =
13,89. K = W*D = 0,101
D(K) = 0,0003
```

Fig. 2. A placement of several articles attributed to the person named as Андрей Петрович Ершов in the Open Archive. Articles with specified affiliation are shown in a lighter color.

Each node of the graph corresponds to an article of the SpringerLink digital library. 10 white nodes correspond to the articles with a specified affiliation and 7 yellow nodes represent articles without a specified affiliation. This means that for this person 30 percent of publications is correctly attributed due to the text similarity measure. A total of 91 publications have been found for the "А.П. Ершов“ query. All these papers were written by several real persons. There were 84 papers belonging to persons named as A.P. Ershov, 5 papers belonging to a person named as Andrei P. Ershov, 1 paper belonging to a person named A.P. Yershov, and 1 paper belonging to Andrei Petrovich Ershov. Academician Andrei Petrovich Ershov (described in the Open Archive) has used all the above variants of his name, and the other persons (Albert Petrovich Ershov, Alexei Petrovich Ershov, Albert Petrovich Ershov...) have always named themselves A.P. Ershov. The program has correctly attributed 17 publications authored by Academician Andrei Petrovich Ershov, as well publications authored by another three A.P. Ershov who are not descibed in the Open Archive. Other groups of articles correspond to homonyms of the person described in the Open Archive**.**

5 Conclusion

The program has been tested on a test sample of 100 persons employed by the IIS SB RAS at various time periods (about 3,000 publications) and on the articles authored by Academician A.P. Ershov. The names of the IIS SB RAS employees were extracted automatically from the Open Archive, after that their publications were extracted from SpringerLink and the results of the identity resolution program were checked manually. To verify this approach, we have compared the data extracted from the SpringerLink digital library with the data of the Academician A. Ershov's archive and the digital library eLIBRARY.RU. Regarding the SpringerLink articles, a significant variation in the amount of available texts is detected (from a few lines to a few dozens of pages), which significantly affects the accuracy of identification. About eighty percent of the analyzed articles in SpringerLink had no information on the full names of their authors (only short forms were given) and approximately seventy percent of author's affiliations have been provided. Nevertheless, a joint comparison of attributes and text similarities have shown good accuracy, close to 93 percent.

The text similarity measure works quite well for homonyms operating in remote areas, but when their research fields are similar, errors are possible. This part of errors makes up the majority of all errors. We are going to improve the quality of this part of the program by further adjustment of text similarity comparison. Another problem for our algorithm arises when the Open Archive has no information concerning a change in affiliation, but the publications of the respective individual indicate a change in his/her place of work. Since in our program information on affiliations has higher priority than text similarity, our algorithm can produce several distinct persons for one real person. We have tried to use two additional well-known heuristics such as paper venue and information about co-authors for these cases [4]. However, quite often a change of affiliation correlates with a change or research partners and publication venue. In the near future we are going to develop a measure taking into account both text similarity and additional attributes such as paper references.

Acknowledgements. This work has been supported by the RFBR grant 14-07-00386.

References

1. Marchuk, A.G., Marchuk, P.A.: Specific features of digital libraries construction with linked content. In: Proc. of the RCDL 2010 Conf., pp. 19–23 (2010). (in Russian)
2. Schultz, A., et al.: How to integrate LINKED DATA into your application. In: Semantic Technology & Business Conference, San Francisco, June 5, 2012. http://mes-semantics.com/wp-content/uploads/2012/09/Becker-etal-LDIFSemTechSanFrancisco.pdf
3. Apanovich, Z., Marchuk, A.: Experiments on using LOD cloud datasets to enrich the content of a scientific knowledge base. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2013. CCIS, vol. 394, pp. 1–14. Springer, Heidelberg (2013)
4. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: Disambiguating author names. In: Large Bibliographic Repositories Conference: Internat. Conf. on Digital Libraries, New Delhi, India (2013)

5. Song, Y., Huang, J., Councill, I.G., Li, J., Giles, C.L.: Efficient topic-based unsupervised name disambiguation. In: Proc. of the 7th ACM/IEEE-CS Joint Conf. on Digital Libraries, pp. 342–351 (2007)
6. Godby, C.J., Denenberg, R.: Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC. http://www.oclc.org/research/publications/2015/oclcresearch-loc-linked-data-2015.html
7. Isele, R., Jentzsch, A., Bizer, C.: Silk server - adding missing links while consuming linked data. In: 1st Internat. Workshop on Consuming Linked Data (COLD 2010), Shanghai (2010)
8. Ley, M.: DBLP-Some Lessons Learned. PVLDB **2**(2), 1493–1500 (2009)
9. Hickey, T.B., Toves, J.A.: Managing Ambiguity In VIAF. D-Lib Magazine **20**, July/August, 2014. http://www.dlib.org/dlib/july14/hickey/07hickey.html
10. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: II Web, pp.73–78 (2003)
11. Stamatatos, E.: A survey of modern authorship attribution methods. J. of the American Society for Information Science and Technology **60**(3), 538–556 (2009)
12. Rogov, A.A., Sidorov, Y.: Vl. statistical and information-calculating support of the authorship attribution of the literary works. computer data analysis and modeling: robustness and computer intensive methods. In: Aivazian, S., Kharin, Y., Rieder, Y. (eds.) Proc. of the Sixth Internat. Conf. (September 10–14, 2001, Minsk), vol. 2, pp. 187–192. BSU, Minsk (2001)
13. Kukushkina, O., Polikarpov, A., Khmelev, D.: Using literal and grammatical statistics for authorship attribution. Probl. of Info. Trans. **37**(2), 172–184 (2001)
14. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
15. Steyvers, M., Griffiths, T.: Probabilistic Topic Models Handbook of Latent Semantic Analysis (2007)

http://www.springer.com/978-3-319-27973-2

Digital Libraries: Providing Quality Information
17th International Conference on Asia-Pacific Digital
Libraries, ICADL 2015, Seoul, Korea, December 9-12,
2015. Proceedings
Allen, R.B.; Hunter, J.; Zeng, M.L. (Eds.)
2015, XIV, 362 p. 88 illus. in color., Softcover
ISBN: 978-3-319-27973-2