Extracting and Classification the Semi-Structured Data of Web-Systems

Irina Pelekh

Lviv Polytechnic National University, Lviv, Ukraine

presty@i.ua

Abstract. The extracting and classification of semi-structured data of web-systems is described. The definition of semi-structured data is given and the main characteristics are defined. The variety of tasks text information processing is grouped into the eleven large classes related to the analysis of text data. The traditional models of knowledge representation are considered. An algorithm for the web-sources, from which data will to be obtained, ontological model integrating creating is proposed. The process of data extracting using the query language to the markup language elements is characterized.

Keywords: semi-structured data, web-system, extracting, classification, ontology.

1 Problem formulation

In our time of global computerization, the information search using the Internet has long been extremely popular and most used. In the era of the Internet the systems of an entirely new type - services began to appear. These are standalone programs that implement certain functionality that developers can use in their applications. The service itself became a separate unit, separated from the user programs internal structure. The service can be written using another programming language, can be operated under the control of another OS and the server on which it operates, can be physically located anywhere in the world. A service whose requests are generated and transmitted over a local area network or the Internet is called a web service. Especially popular, today, are web-services that integrate the data from several information resources.

In WWW, information is provided to the user in the form of web pages that do not have a clearly defined structure. There is an urgent problem of obtaining data from such sources for further work with them. The application of web-scraping method generates the problem of analysis and identification of semi-structured data. Semi-structured data is characterized by the lack of strict table structures and relationships in relational database models, however, this data form contains tags and other markers

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

for the separating of semantic elements, as well as to provide a hierarchical structure of records and fields in data sets [6].

2 State of arts

The Internet is the largest source of data, most of which are presented as web-pages that do not have a strictly formalized structure. To make a quality search you need to use the complex mathematical models, semantic analysis and other methods of information analysis. Therefore, the data, that is receiving should be structured in a certain way. However, most web-sources databases are presented in the form of unique structures, which makes it difficult to obtain structured data from similarly semi-structured web-pages.

Extracting and classification of structured data from web-pages is reduced to the following tasks [8]:

- searching and receiving of available information resources for data extracting (navigation problem);
- recognition of areas containing the required data (data recognition problem);
- search of the found data structure (the problem of finding a common data structure);
- ensuring of the extracted data homogeneity (the problem of matching the attributes of the extracted data);
- data integration from different sources (problem of data integration).

Traditionally, in the systems of text analyses for knowledge representing are used four types models: productive, formal-logical, framing and semantic-network model. Based on these models the solutions are described and the main prospects for their using [1].

It should be noted the prospects of using functional control systems to solve the problems of text analysis. In this case, the core of such systems may be became the intelligent information systems that include elements of artificial intelligence, based on the methods and means of the intelligence theory [2, 9-12]. In [3] authors propose a system for obtaining data from the science-metric databases that uses the language of queries for markup language elements.

The following basic methods are used to solve the problem of text information processing [2, 7, 8, 13-18]: data mining; associative rules; production model; formal-logical model; framed model; semantic-network model; decision tree; clusters; mathematical functions; etc. [19-49]

The analysis of this methods shows that each of them has a well-developed formal system that allows you to make a sufficiently full description of the entities and processes of various subject areas. However, they all have a significant drawback in their usage. They do not allow you to describe the illogical, incomplete or contradictory of text, which is a reflection of the natural language. There is a paradox associated with the limited capabilities of logical formal systems and the needing to describe logical features and not logical knowledge and data contained in the texts.

3 Algorithm of extracting and classification the semi-structured sources data

The quality of extracting and classification the necessary information from semi-structured sources depends on the methods and technologies, which are working with such data. So the most effective to solve this problem, in our time, are considered semantically oriented technologies [4-6].

To classify semi-structured data from web-sources, we offer an algorithm for creating an integration ontological model of all web sources from which data will to be extracted.

To classify semi-structured data from web-sources, we offer an algorithm for creating an integration ontological model of all web sources from which data will to be extracted.

Consequently, in the first stage, the writing procedure of the semi-structured html format information into a pre-created database is performed.

The process of extracting data from the HTML page is as follows. Web page returned by the server is formatted using the markup language (mostly HTML), for further displaying in one form or another using a special program (web browser). In fig. 1 is shown an example of a web browser's visualization of a particular data area (social network community) and the source code of this data.

Here, for example, the data description on the community wall is as follows:

```
…
<div class="post_content">
<div class="post_info">
<div class="wall_text"> …
```

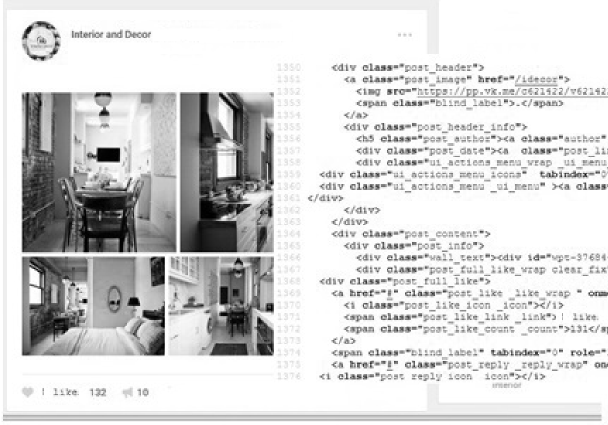
A certain number of different classes have been allocated that responsible for their data area. So, the class "wall_text" describes the record recording on the wall. The text importance in the record is determined by the presence of such characteristics: bold type, italic, underline, hashtag, etc.

To get such data, is made the searching for the classes names that are responsible for describing the necessary information and obtaining their content. Thus, one of the database table results fields is filled in. To automate this process, the data extraction program uses the query language for the markup language elements (Xpath).

The next stage is the direct construction an integration ontological model of all web sources from which data will to be extracted, on the basis of the received and created databases. When constructing any algorithm, the primary task is to determine the input and output data. The input data for the algorithm for constructing an integration ontological model of all web sources from which data will to be extracted are:

- structural schemes of the web-sources databases;

- domain ontology.



```
    <div class="post_header">
      <a class="post_image" href="/idecor">
        <img src="https://pp.vk.me/c621422/v621422532/a5bc/znRa PcyHs4.jpg" data-post-id="-3768443
        <span class="blind_label">.</span>
      </a>
      <div class="post_header_info">
        <h5 class="post_author"><a class="author" href="/idecor" data-from-id="-37684434" data-pos
        <div class="post_date"><a  class="post_link"  href="/wall-37684434 339924"  onclick="retur
        <div class="ui_actions_menu_wrap _ui_menu_wrap " onmouseover="uiActionsMenu.show(this);" c
  <div class="ui_actions_menu_icons"  tabindex="0" aria-label="Дії" role="button" onclick="uiActic
  <div class="ui_actions_menu _ui_menu" ><a class="ui_actions_menu_item" onclick="wall.markAsSpam(
</div>
      </div>
    </div>
    <div class="post_content">
      <div class="post_info">
        <div class="wall_text"><div id="wpt-37684434_339924" class="_wall_post_cont"><div class="w
        <div class="post_full_like_wrap clear_fix">
  <div class="post_full_like">
    <a href="#" class="post_like _like_wrap " onmouseover="Wall.likesShow(this, '-37684434_339924'
      <i class="post_like_icon _icon"></i>
      <span class="post_like_link _link">I like  </span>
      <span class="post_like_count _count">131</span>
    </a>
    <span class="blind_label" tabindex="0" role="link" onclick="Wall.likesShowList(this, '-3768443
    <a href="#" class="post_reply _reply_wrap" onclick="return Wall.showEditReply('-37684434_33992
  <i class="post reply icon  icon"></i>
```

Fig.1. The data of records description on "Interior and Decor" community wall of soc. network

The domain ontology is usually developed pre-emptively, with the participation of a domain expert and a knowledge expert specialist in ontological format. The process of creating such a model takes a long time, but it only needs to be done at the initial stage of integration. With the further addition of new systems are working in this field, the ontology itself does not require any additional changes.

Output data is a general ontological model that describes the structure of all web-sources within their domain and the relationship between elements of different systems. Such model will be modeled using RDF language, its RDFs extension and OWL language.

The algorithm for constructing an integration ontological model of all web sources from which data will to be extracted contains 6 main steps, namely:

- Representation of the database structure in the RDF form (sequential displaying of the scheme S in the RDF format.
- Adding semantic properties and creating the ontology. This step is realized by using the procedure of determining the database elements common features and adding links between them.
- Adding ontologies of the upper levels and domain ontology. We implement this step by using the OWL language by using the owl: import command. By usage the transitivity rule in RDF, additional ontologies are extending the subject areas and add new concepts and properties.
- Checking of the created ontology. This step is implemented by checking and analyzing the elongated ontology for "connectivity", that is, we check whether there is a lack of semantic links anywhere. If so, then go to step five, if not - go to step six.
- Editing an extended ontology by usage the ontology editor and adding links between concepts. Next, go back to step 4.

- Storing the received total ontology of the system structure in the RDF metadata repository.

4 Conclusion

The need to develop methods and tools for extracting and processing the semi-structured data of web-systems has become relevant in the context of the information searching using Internet. In this paper the extracting of semi-structured data of web-systems has been described. The problems of the obtaining and classification of structured data from web-pages have been highlighted. Methods for solving the problems of the obtaining and classification of structured data from web-pages have been considered. The traditional models of knowledge representation are showed. The process of data extracting using the query language to the markup language elements is characterized. An algorithm for the web-sources, from which data will to be obtained, ontological model integrating creating is proposed.

References

1. Bondarenko M.F., Shabanov-Kushnarenko Yu.P.: Theory of intelligence. Textbook, X.: Izdvo SMIT, 576 p. (2007).
2. Buileaar P., Eigner T.: Topic extraction from scientific literature for competency management. In The 7th International Semantic Web Conference PICKME 2008, Karlsruhe, Germany, 55-67. (2008)
3. Kolada A.S., Gogunsky V.D.: Automation of information extraction from the science-computer databases, Management of the development of complex systems, No. 16 (2013)
4. Kushniretska I., Kushniretska O., Berko A.: Designing of Structural Ontological Data Systems Model for Mash-UP Integration Process, Applied Computer Science, 11(1) (2015)
5. Kushniretska I., Kushniretska O., Berko A.: The ontological model of knowledge of scientific and technical information system, Computer Science and Information Technologies (CSIT'2014): proc. of the IX-th Intern. Scientific and Techn. Conf., Lviv, Ukraine / Min. of Education and Science of Ukraine (2014)
6. Kushniretska I.: Semi-structured data dynamic integration Mashup system, Computer Science and Information Technologies (CSIT'2016): proc. of the XI-th Intern. Scientific and Techn. Conf., Lviv, Ukraine, Min. of Education and Science of Ukraine, 220-221 (2016)
7. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval, Cambridge University Press, ISBN 0-521-86571-9, (2008).
8. Lytvyn, V., Pukach, P., Bobyk, I., Vysotska, V.: The method of formation of the status of personality understanding based on the content analysis. In: Eastern-European Journal of Enterprise Technologies, 5/2(83), 4-12 (2016)
9. Kravets, P.: The game method for orthonormal systems construction. In: The Experience of Designing and Application of CAD Systems in Microelectronics (2007).
10. Lytvyn, V., Vysotska, V, Veres, O., Rishnyak, I., Rishnyak, H.: Content linguistic analysis methods for textual documents classification. In: Computer Science and Information Technologies, Proc. of the XI-th Int. Conf. CSIT'2016, 190-192 (2016)

11. Zhao Li, Wee Keong Ng, Aixin Sun: Web data extraction based on structural similarity, Journal Knowledge and Information Systems archive, Vol. 8, Issue 4, 438-461 (2005)
12. Zhou L.: Ontology Learning: State of the Art, Information Technology and Management, 8 (3), 241-252 (2007)
13. Chen, J., Dosyn, D., Lytvyn, V., Sachenko, A.: Smart Data Integration by Goal Driven Ontology Learning. In: Advances in Big Data. Advances in Intelligent Systems and Computing. – Springer International Publishing AG 2017. P. 283-292 (2017).
14. Su, J., Vysotska, V., Sachenko, A., Lytvyn, V., Burov, Y.: Information resources processing using linguistic analysis of textual content. In: Intelligent Data Acquisition and Advanced Computing Systems Technology and Applications, Romania, 573-578, (2017)
15. Vysotska, V., Chyrun, L., Chyrun, L.: Information Technology of Processing Information Resources in Electronic Content Commerce Systems, CSIT, 212–222 (2016)
16. Vysotska, V., Hasko, R., Kuchkovskiy, V.: Process analysis in electronic content commerce system. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT 2015, 120-123 (2015)
17. Vysotska, V.: Linguistic Analysis of Textual Commercial Content for Information Resources Processing. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET'2016, 709–713 (2016)
18. Basyuk, T.: The Popularization Problem of Websites and Analysis of Competitors. Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, Cham pp. 54-65 (2017)
19. Vysotska, V., Chyrun, L., Lytvyn, V.: Methods based on ontologies for information resources processing. Germany: LAP LAMBERT Academic Publishing (2016).
20. Vysotska, V.: Tekhnolohiyi elektronnoyi komertsiyi ta Internet-marketynhu. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
21. Vysotska, V., Lytvyn, V.: Web resources processing based on ontologies. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
22. Vysotska, V., Shakhovska, N.: Information technologies of gamification for training and recruitment. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
23. Vysotska, V.: Internet systems design and development based on Web Mining and NLP. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
24. Vysotska, V.: Computer linguistics for online marketing in information technology : Monograph. Saarbrücken, Germany: LAP LAMBERT Academic Publishing (2018)
25. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum O.: Intelligent System Structure for Web Resources Processing and Analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56-74 (2017)
26. Lytvyn, V., Vysotska, V., Wojcik, W., Dosyn, D.: A Method of Construction of Automated Basic Ontology. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 75-83 (2017)
27. Lytvynenko, V., Lurie, I., Radetska, S., Voronenko, M., Kornilovska, N., Partenjucha, D.: Content analysis of some social media of the occupied territories of Ukraine. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, 84–94 (2017)
28. Shepelev, G., Khairova, N.: Methods of comparing interval objects in intelligent computer systems. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, (2017)
29. Orobinska, O., Chauchat, J.-H., Sharonova, N.: Methods and models of automatic ontology construction for specialized domains (case of the Radiation Security). In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 95–99 (2017)

30. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30 (2017)
31. Grabar, N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19 (2017)
32. Hamon, T.: Biomedical text mining. In: Computational Linguistics and Intelligent Systems, colins.in.ua/wp-content/uploads/2017/04/2017COLINS-THAMON-keynote.pdf
33. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: Computational Linguistics and Intelligent Systems, COLINS, (2017)
34. Lande, D.: Creation of subject domain models on the basis of monitoring of network information resources. In: 1st International Conference Computational Linguistics and Intelligent Systems, http://colins.in.ua/wp-content/uploads/2017/04/Lande.pdf (2017)
35. Protsenko, Y.: Intuition on modern deep learning approaches in computer vision. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/protsenko.pdf (2017)
36. Kolbasin, V.: AI trends, or brief highlights of NIPS 2016. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
37. Kersten, W.: The Digital Transformation of the Industry – the Logistics Example. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/CoLlnS_TuS.pdf (2017)
38. Shalimov, V.: Big Data – Revolution in Data Storage and Processing. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/BigData_eng.pdf (2017)
39. Hnot, T.: Qualitative content analysis: expertise and case study. In: 1st Inter. Conference Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Qualitative-content-analysis_expertise-and-case-study.pdf (2017)
40. Romanyshyn, M.: Grammatical Error Correction: why commas matter. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-why-commas-matter.pdf. (2017)
41. Yukhno, K., Chubar, E.: Gamification: today and tomorrow. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 139–140 (2017)
42. Pidpruzhnikov, V., Ilchenko, M.: Search optimization and localization of the website of Department of Applied Linguistics. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 137–138 (2017)
43. Olifenko, I., Borysova, N.: Analysis of existing German Corpora. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 135–136 (2017)
44. Kolesnik, A., Khairova, N.: Use of linguistic criteria for estimating of wikipedia articles quality. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, (2017)
45. Kirkin, S., Melnyk, K.: Intelligent data processing in creating targeted advertising. In: 1st Inter. Conf. Computational Linguistics and Intelligent Systems, COLINS, 131–132 (2017)
46. Hordienko, H., Ilchenko, M.: Development and computerization of an English term system in the fields of drilling and drilling rigs. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 129–130 (2017)