Use of Semantic Similarity Estimates for Unstructured Data Analysis

© Julia Rogushina

Institute of Software Systems of National Academy of Sciences of Ukraine, Kyiv, Ukraine
ladamandraka2010@gmail.com

Abstract. The paper discusses problems related to unstructured data analysis in order to acquire implicit knowledge from them. Semantic similarity estimations are used as one of instruments for such analysis. We use the portal version of the Great Ukrainian Encyclopedia (e-VUE) to demonstrate some examples where ontologies and semantic Wiki markup are used for generation of semantically similar concepts (SSCc). Semantic similarity in these examples is defined in domain-specific way. Grouping of concepts into SSCs is based on high-level ontological classes, and semantic properties and their relations are used for construction of attribute space. Various sets of SSCc are applied for navigation and search to increase functionality. Every e-VUE article is represented by Wiki page with unstructured and semi-structured natural language (NL) and multimedia content pertinent to some concept. Ontological model of e-VUE is considered as a domain knowledge base that simplifies processing of e-VUE article content and defines semantic relations between concepts.

Keywords: Ontology, Unstructured data, Wiki technology, semantic similarity.

1 Unstructured data analysis

Now the largest part of stored information (more than 80% of all stored data, and their number is increasing by an order of magnitude faster than structured data) is represented by unstructured data (USD) [1], so methods and means of their analysis are evolving rapidly. These situation causes transformation of disengaged USD analysis implementations in integral scientific research area. Great information new knowledge source provided by USD are potentially of but their use involves problems deal with storage and analysis. Automation of such processing needs in USD transformation into structured information that can be processed automatically in various ways.

USD are usually considered as information collected without any predefined data model or data organization. Major portion of USD is represented by textual information – arbitrary-length sets of natural language (NL) words combined by weakly formalized linguistic rules. Such USD may also contain dates and numbers. Examples of textual USD are NL documents in various formats, information from social networking, data from mobile devices and content of the Web sites.

USD is not well defined term on account of complications in distinction be-

tween structured and unstructured data without formally defined structure or with structure that cannot be used for automated processing [2]. One of the criteria for structured data determining is a possibility to create data element parser. Thus, we consider data as USD if accessible information about their structure cannot make data analysis more efficient.

Unstructured information can be stored in the form of objects (files or documents) that have their own structure. For example, the body of email and email attachment are USD, but location of attachment into the mail is determined by structure. The combination of structured and unstructured data are considered also as USD.

1.1 Properties of unstructured data

Main properties of USD are:

- *Heterogeneity*. USD can be generated by different ways, in various formats, from various information sources, and these data cannot be structured and placed in any DBMS by various reasons;
- *Ambiguity*. Equal phrases of different persons can have different meanings that depend on their individual experience, views, etc. (for example, an expert phrase "I don't understand this article" means the poor quality of the article, and the same statement of student means the inadequate education), and the same idea may be expressed by different words .
- *Contextual dependency*. Interpretation of word or name differs in different contexts (for example, meanings of "model" term differs in technology and mathematics).
- *Dynamics of value.* Words can rapidly change their meaning, for example, previously unknown "Wuhan" is associated now with coronavirus and gains additional meaning.

Often USD are created directly by humans (in contrast to structured data), and therefore systems oriented on USD analysis have to take into account the "human factor". Technologies such as Data Mining, Natural Language Processing, and Text Mining provide different methods for finding structure in USD. Common text structuring methods usually include manual metadata tagging for further structuring. The Unstructured Information Management Architecture (UIMA) standard provides a general framework for processing this information to make sense and create structured data.

1.2 Text Mining as a basis for NL USD analyzing

From the late 1990s Text Mining becomes separate scientific area [3]. Early approaches regarded text as a "bag of words" such as abbreviations, plurals and word combination, as well as multiple word terms known as n-grams. Basic lexical analysis takes into account the frequency of words and terms to perform such tasks as document classification by topic. But this approach doesn't consider documents semantics. Now Text Mining is looking for hidden relations and other complex structures into the text data sets.

Text Mining technology is based on linguistics and Data Mining. Initially it was oriented on recognition of personal and geographical names, dates, phone numbers

and email addresses in the text. Now more sophisticated methods provide retrieval of concepts and relations between them and even emotions.

Big Data spread increase the urgency of USD structuring [4]. In the most general form, the solution of this the complex scientific problem consists in construction of USD content marked graphs and matching of such graphs. Another aspect of this problem is related to retrieval of relevant knowledge for USD markup.

Text Mining should provide a transition from USD to structured information. Most often, this process ignores a lot of the specific features of the NL that are used only at the previous stage of text parsing, and the following phases of analysis use the "bag of words" model where the order of words is not important [5].

1.3 Elements of text document structuring

From some points of view (distinct from automated analysis) NL text document can be considered as a structured object. For example, from a linguistic point of view each document contains a large amount of semantic and syntactic structure that is hidden in the text. In addition, markup elements (punctuation signs, capital letters, numbers and special characters, etc.) and formatting elements (tables, columns, paragraphs, etc.) can be considered as "soft markup" language to help identify important document subcomponents, such as title, names of authors, subdivisions, etc. Word sequence can also be a structurally significant characteristic to a document. In addition, some text documents may contain embedded metadata in the form of formatted markup tags that are automatically generated by text editors.

Documents that have relatively few such structuring elements (for example, scientific publications and business reports) are called *free-format* or *weakly structured*. Documents with relatively more structuring elements (such as e-mail, HTML web pages) are called *semistructured*.

Pre-processing Text Mining operations allow to take into account various NL document elements to convert it from an USD with implicit structuring into explicitly structured data. However, the potentially great number of words, phrases, sentences and formatting elements contained even in a small document (not even considering the potentially large number of different meanings of these elements in different contexts) causes the need in identification of a simplified subset of document properties (features). Such set of features is called a *representative model* of a document: individual documents are characterized by the sets of features contained in their representative models. But each individual document in the collection has an extremely large number of properties even in the most effective representative models. Therefore, problems associated with high dimensionality of characteristics (ie, the size and scale of possible combinations of feature values for data) are usually much more significant in Text Mining systems than in classic Data Mining systems.

Structured representations of NL documents have a much larger number of potentially representative features – and therefore a greater number of possible combinations of their meanings – than in relational or hierarchical databases. For example, relatively small collection of 10-15,000 documents contains more than 25,000 non-trivial words. The number of attributes in a relational databases analyzed in Data Mining tasks are usually much smaller. The high dimensionality of potentially representative

properties leads to pre-processing of the text aimed at creating of simplified models of representation .

Another feature of NL documents is *feature sparsity* – only a small subset of the properties available for document collection as a whole appear in each individual document, and thus, when a document is presented as a binary feature vector, almost all vector values are zero.

1.4 Properties of individual NL document

Symbols, words, terms and concepts define properties of individual NL document. Text Mining algorithms process document representation through the set of properties, not directly the documents themselves, and therefor we need in compromise between two important goals.

The first goal is to achieve a correct classification of the volume and semantic level of the properties to accurately represent the document during the pre-processing operation. The second goal is to select the property definition that is most computationally efficient and practical for pattern detection. This choice can be supported by validation, normalization, or use properties from controlled vocabularies or external sources of knowledge, such as dictionaries, thesauruses, ontologies, or knowledge bases, to create smaller sets of properties with greater semantic significance.

Although many potential properties can be used to present NL documents, the following four types are most commonly used:

- *Symbols*. Letters, numbers, special characters and spaces are the building blocks of higher-level semantic features such as words, terms and concepts. Symbol-level views may include a complete set of all characters for document or some filtered subset. Character-based representations without position information ("bag-of-characters" approaches) are usually have very limited utility for Text Mining. Views that include some level of positional information (such as bigrams or trigrams) are more useful.

- *Words*. Specific words selected directly from the NL document are the baseline for semantics. Every word-level property has at most one linguistic marker.

- *Phrases* and multiword expressions do not constitute separate properties at the word level. Word-level document representation includes features for each word in that document, that is, the text of the document is represented by complete set of word-level properties. Therefor some representations of word-level document collections contain the great number of unique words in their feature space. However, most document submissions at this level have at least some minimal optimization and are therefore composed of subsets of representative properties that are filtered from such elements as stop words, symbols and meaningless numbers.

- *Terms*. Terms are individual words and multiword phrases selected directly from the source NL document body using the term extraction methodology. Term-based document submission consists of a subset of terms in that document. Various methodologies for extracting terms that convert the raw text of a document into a sequence of normalized terms (tokenized and lemmatized word forms) tagged with the relevant parts of the language can be used. Sometimes, an external vocabulary is also used to normalize terms to provide a controlled vocabulary. Term extraction techniques use different approaches to generate and filter a list of the most relevant document terms

from this set of normalized terms.

- *Concept*. Concepts are properties created for NL document using various categorization techniques. Concept-level properties can be created manually, but now more commonly they are retrieved from documents through complex pre-processing procedures that identify individual words, multiword expressions, entire sentences or even larger syntax units, which then relate to specific concept identifiers.

Terms and concepts levels represent properties more significant for semantics. Term-level representations are easier to generate automatically from text than concept-level ones. However, concept-level representation is much more useful for processing synonymy and polysemy.

Many categorization methodologies include the referencing to external knowledge source. For example, some statistical methods can use as externals source an annotated collection of documents. For manual and rule-based categorization, cross-referencing and validation of perspective properties at the concept level typically involve interaction with external databases, such as domain ontology, vocabulary or formal hierarchy. In contrast to word-level and term-level properties, concept-level document properties may consist of words not contained in this document. Concept-based representations allow using very complex concept hierarchies and diverse domain knowledge provided by ontologies and knowledge bases. But concept-level representations have several potential drawbacks: 1) the relative complexity of using heuristics in pre-processing operations, 2) the dependence of concepts on the domain specifics.

1.5 Using background knowledge in Text Mining

Background knowledge can be used for pre-processing to improve the acquisition of domain concepts. Domain in Text Mining is a specialized area of interest represented by ontologies, lexicons, dictionaries, thesauri, taxonomies etc. Text Mining systems can use information from formalized external knowledge sources for these domain to improve document pre-processing and knowledge discovery. Concepts used in Text Mining systems are connected not only to the descriptive attributes of a particular document, but also to domains.

Access to background knowledge – while not absolutely necessary for creating concept hierarchies in the context of a single document or document collection – can play an important role in developing more meaningful, consistent and normalized concept hierarchies.

Text Mining uses background knowledge more than Data Mining: properties of USD are not just elements in a flat set, as is often the case with structured data, because they are linked through lexicons and ontologies to support advanced queries.

Although Text Mining pre-processing operations play an important role in transforming unstructured content of raw document collection into more handy concept-level data representation, the core functionality of such systems is oriented on analysis of concept co-occurrence models in documents collection. Text Mining uses algorithmic and heuristics to consider distributions, frequent sets and various associations of concepts on inter-documentary level to identify the nature and relations of concepts represented by this collection.

For example, if news collection contains many articles about both event X and

company Y at the same time, as well as articles that deal with company Y and product Z at the same time, then Text Mining analysis indicates the relation between X and Z, notwithstanding this relation is not present in any document.

In classic Data Mining, background knowledge from external sources is used to limit search. Text Mining systems can use information from external sources of knowledge in pre-processing and concept testing operations. In addition, access to background knowledge can play an important role in developing meaningful, consistent, and normalized concept hierarchies.

Domain knowledge can also be used by other components of the text extraction system. For example, an important application of background knowledge is the construction of significant constraints on knowledge discovery operations. Similarly, background knowledge can also be used to formulate constraints that allow users to increase the flexibility of viewing large result sets or formatting data for presentation.

Text Mining systems can utilize background knowledge that is represented as domain ontology describing the set of all important facts, classes and relations between these classes. Domain ontology can be used as a vocabulary designed to be both human-readable and machine-readable.

Well-known ontology used in Text Mining is WordNet developed by Princeton University for NL modeling.

2 Problem formulation

Traditional Text Mining approach is not effective enough to process Big Data, and it causes the need in intelligent methods of USD analysis that use background domain knowledge and specialized ontologies for semantic markup of NL texts. We propose to use Wiki technologies and their semantic extension as a source of domain knowledge. This knowledge can be used in estimations of domain concepts semantic similarity for NL elements structuring of Big Data metadata.

3 Estimations of semantic similarity

It is advisable to apply the domain ontological knowledge for estimation of the semantic similarity of domain concepts. The sets of semantically similar concepts can be used as a base for structuring of USD by linking of data fragments with ontological elements. Such knowledge makes it possible to quantify the substantive similarity of both the domain concepts and the NL words and phrases corresponding to these concepts.

The values of similarity estimations depend either on estimation methods or on the choice of the domain ontology and on ontology pertinence to user conception about domain. Various ontologies that represent different perspectives on the same domain can be used for this purpose. Such ontologies formalize the contexts of the user task but their use necessitates the integration and reconciliation of these ontologies. It should be noted that the integration of independently created ontologies is a non-trivial problem that cannot be fully automated and requires the participation of domain experts to establish correct relations between ontological concepts.

The definition of semantic closeness between domain concepts is quite closely

related to the problem of displaying independently constructed ontologies of this domain.

The task of domain ontologies mapping consist of two separate sub-tasks: *local* representation of concepts that require matching between classes and instances of these ontologies; *global* concept mapping – an analysis of the entire set of local ontology element mappings. Global mapping provides additional information about pairs of different ontology concepts from information about their relation with other elements of ontologies.

Similarity analysis of hierarchical (taxonomic) relations is probabilistic. The assessment of the similarity of concepts from different ontologies may be based on the positions of these concepts in the hierarchy of classes for which similarity has already been determined: if the superclasses and subclasses of these concepts are similar, then the same concepts may also be similar.

The similarity of two entities depends on similarity estimation of: direct superclasses of these concepts; all superclasses of these concepts; subclasses of these concepts; and instances of these concepts.

One of the tasks of mathematical semantics is the measurement of semantic distances between NL words. Estimation of the semantic distance allows us to assess the density of semantic and associative-semantic relations between words and concepts of the dictionary, between units of text and, in the framework of more complex tasks, between fragments of text.

The value of semantic distance plays an important role in determining the meaning with taking into account the context of several sentences. The semantic meanings of words in a sentence should create semantic unity, therefore, the meanings of the concepts (and sems) of words that stand side by side in a sentence should be in the optimal range of semantic proximity.

Determining the coefficients of semantic connectivity of relations between language units makes it possible to assess the correspondence of NL fragment and its phrases to the points of a multidimensional space of a potentially generated ordered set – classification of NL phrases. Estimation of the distance between language units is also applicable in other tasks: constructing computer thesauruses where computer automatically constructs a coherent and meaningful text, the work of expert systems, determination of the text subject etc. Semantic similarity of NL fragments A and B can be calculated taking into account the frequency of words typical for A and B. Various types of linguistic relations between words in a language, such as homonyms, synonyms, hyperonyms, antonyms, equonyms, have to receive an accurate numerical estimation based on a uniform scalar value.

A special case of ontologies is taxonomies. They are a fairly common and convenient source of knowledge for analyzing the semantic closeness of NL concepts and words.

3.1 Use of taxonomies for semantic similarity evaluation

Evaluations of semantic similarity based on domain knowledge network representations has long history that was started with the spread of the activation approach [6, 7]. Some researchers consider similarity evaluation in semantic networks using one taxonomic relations “is-a” and exclude other types of relations [8]; other analyze also

relations “part-of-part” [9]. A common and long-known way of semantic similarity evaluation in taxonomy lies in measuring the distance between net nodes that correspond to the elements being compared – the shorter path from one node to another means their higher similar. If elements are connected by multiple paths between then the shortest path length is used [10, 11].

Although researchers identify many similarity criteria but many of them are rarely accompanied by an independent characteristic of the phenomenon that they measure, in particular for those that are is used in software (for example, similarity of documents in information retrieval, similarity of cases in case-based considerations). On the contrary, the value of such measures depend on their usefulness for particular tasks.

However, this approach is compounded by the notion that all connections in the taxonomy represent homogeneous distances. Unfortunately, it is difficult to define taxonomy distance uniformity because real taxonomies have great variability of distances covered by a single taxonomic relation, especially if some taxonomy subsets are much denser than others. For example, WordNet [12] contains a lot of direct links between either fairly similar concepts or relatively distant ones. An alternative way of evaluating semantic similarity in a taxonomy is based on the concept of informational content and is not sensitive to the distances sizes between relations [13].

3.2 Similarity and informative content

One of the key factors in the taxonomy concepts similarity is the degree of their information sharing that defines by the number of highly specific terms that applies to both of these concepts. The edge-counting method takes it into account indirectly, because if the long minimum connection path by “is-a” relations between two nodes means that it is necessary to ascend more in taxonomy to general abstract concepts to find the smallest upper bound – a concept to which both concepts under review relate.

Following the standard argumentation of information theory the probability of the concept increases then it’s information content decreases. This quantitative characterization of information provides a new way of semantic similarity measuring. The more information is shared by two concepts, the more similar they are, and the information shared by two concepts is determined by the informational content of the concepts in taxonomy.

In practice, we often need to measure the similarity of words, not concepts. Such measure can be based on word representation through the set of taxonomy concepts that represent meanings (contents) of these words.

Although there is no standard way of evaluating computational measures of semantic similarity, it is appropriate to use estimates that are consistent with human similarity estimates for this purpose. We can use computational similarity measure of word pairs and compare them with human-constructed similarity ratings of the same pairs. In [14] similarity measures are based on maximum taxonomy depth the and the shortest path length in the taxonomy between the concepts. Another points of view for comparing the concept similarity is based on the use of the concept probability rather than information content. Probability-based similarity estimations consider the word occurrence frequency more important than information content.

4 Wiki technology as a means of information structuring

Wiki-technology is the Web-based technology for building of distributed information resource (IR) that allows users to submit and edit materials without additional software and specialized skills. specify explicitly links between individual pages through hyperlinks and define their categories [15].All content changes become accessible immediately, but users can turn up to the later versions [15].

The Wiki page format uses simplified markup language to distinguish various structural and visual elements. Now a large number of Wiki engines and information resources on their base are realized. The largest and most well-known of them is Wikipedia. The main elements of Wiki markup are hyperlinks and categories. Their use makes it easy to convert USD into partially structured data. In addition, the analysis of the Wiki resource structure at the level of words and concepts allows acquiring knowledge for structuring other USD.

Wiki resources can be used as an external source of features for text categorization and determining the semantic relatedness between NL texts [16].

4.1 Semanticization of Wiki resources

Semantic MediaWiki (SMW) is an add-on to the MediaWiki [17]. SMW advantages are semantic processing of information, the availability of group knowledge management tools, relatively high expressive power and reliable implementation. SMW allows to integrate information from different Wiki pages by the knowledge level retrieval and to generate ontological structures on Wiki Pages that can be used by other intelligent software.

In addition to categories, SMW structures information by semantic properties. They allow to link Wiki pages semantically with each other and with other data. Each semantic property has a type, a name and a value, as well as own Wiki page in a special namespace that allows it to determine its place in the property hierarchy and to document how that property should be used.

In terms of ontological analysis, each Wiki page is an ontological element, that is, an element of one of the RDF [18] classes – Thing, Class, ObjectProperty, DatatypeProperty, AnnotationProperty. In addition, each page has its own URI. Usually, Wiki pages are instances of OWL [19] ontology classes, Wiki categories are classes, and Wiki semantic properties are object properties and data properties of ontology.

Therefore, an appropriate OWL/RDF file can be generated on request for any SMW page or the set of pages. Semantic Wiki resources can be used as a basis for automated generation of distributed knowledge bases in RDF format. Exporting to OWL / RDF is a means of ensuring external reuse of data from Wiki, but only practical application of this feature can show the quality of the RDF generated. To this end, system developers have used a number of Semantic Web tools to issue RDFs.

SMW is compatible with the OWL DL knowledge model, therefor external ontologies can be used in Wiki resources. There are two ways to do this: ontology importing allows to create and modify Wiki pages to represent the relations specified in an existing OWL DL document; and reusing the dictionary allows users to match Wiki pages with elements of existing ontologies.

4.2 Use of semantic similarity estimations in online version of the Great Ukrainian Encyclopedia

Theoretical principles for development of semantic search and navigation means are implemented into e-VUE – the portal version of the Great Ukrainian Encyclopedia (vue.gov.ua). This resource is based on ontological representation of knowledge base [20]. To use a semantic Wiki resource as a distributed knowledge base we develop knowledge model of this resource represented by Wiki ontology [21]. Using this model for semantic markup provides the formation and software implementation of an appropriate set of hierarchically related categories, templates for typical IOs, their semantic properties and the queries that use them [22].

Application of semantic similarity estimation for this ontology provides the functional extension of Encyclopedia by new ways of content access and analysis on the semantic level.

An ontological model of the structure of the e-VUE is used to support semantic navigation on the portal. One of the significant advantages of e-VUE as a semantic portal is the ability to find *semantically similar concepts* (SSC). This search is based on the following assumptions: 1. concepts that correspond to Wiki pages that belong to the same set of categories are semantically closer to each other than other concepts on the portal; 2. concepts that correspond to Wiki pages that have the same or similar meanings of semantic properties, are semantically closer to each other than concepts that correspond to Wiki pages with different values of semantic properties or those ones with not defined values of semantic properties; 3| concepts defined as semantically similar by the both preceding criteria are more semantically similar than concepts similar by one of criteria.

For e-VUE, user need to locate the SSP if he (she) is unable to select correctly the knowledge field of concept or enters it with errors. In such cases, user can find similar concepts and then go to desired concept. For example, user wants to find information about a writer or artist whose last name he does not remember accurately, and is not able to accurately determine the style of his (her) works of art, but may indicate the name of more famous person who worked in the same sphere. In some cases, the problem of SSP search is solved by search of the semantically similar words into the NL definition of concept.

In order to extend e-VUE functionality related to search and navigation we propose means of retrieval of semantically similar close IOs – both *globally similar* (by the full set of features – either categories and values of semantic properties) and *locally similar* (only by some subset of these features). Concepts of e-VUE are matched with the current Wiki page.

To demonstrate the capabilities of the described above approach we propose the following examples of local SSPs retrived by: 1. the fixed subset of categories of current page; 2. the values of the fixed subset of semantic properties of current page; 3. the combination of categories and values of semantic properties of current page.

The semantic closeness of the search terms is determined relative to the characteristics of current Wiki page that the user is viewing, that is, the categories and properties of this page are analyzed as parameters of such calculation.

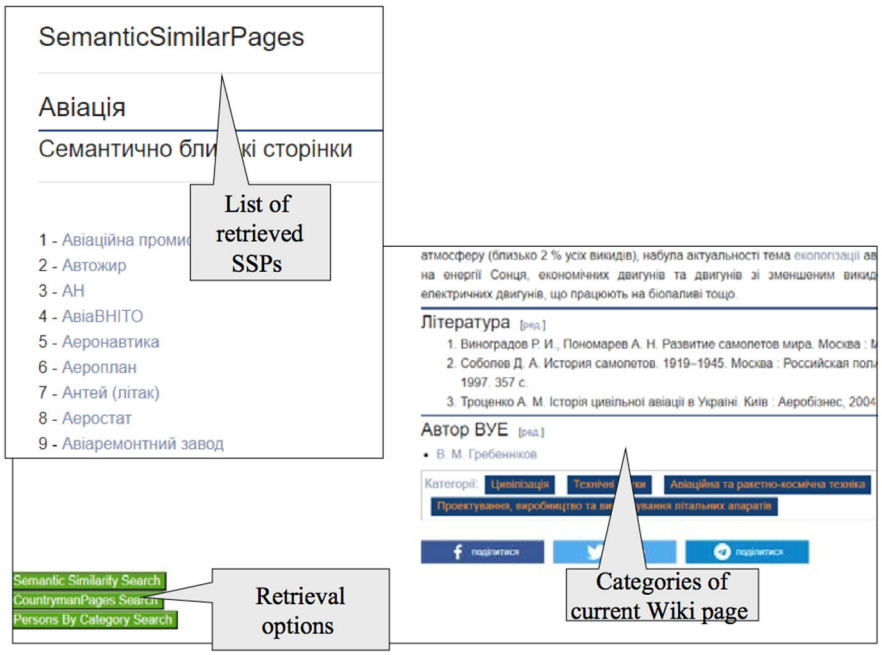

Fig. 1. Search for SSPs for e-VUE page "Aviation".

In the first case, for the current Wiki page you need to find the concepts o e-VUE that are assigned to the categories of the current page. Now this retrieval considers categories and sub-categories of knowledge fields and typical IOs categories (for example, “Countries”, “Rivers”) IS, and don’t take into account service categories related to the form of publication (e.g. “VUE, Volume 1”) (see Fig. 1).

It should be noted that all these cases of SSP search (locally and globally) cannot be performed automatically by Semantic MediaWiki's built-in tools. User can achieve the same result with Semantic MediaWiki tool of semantic retrieval by manual entering of all matching parameters – categories and semantic properties copies from selected Wiki page. Therefore these searches are provides by special software code for analyzing the page content.

The second case deals with the search of e-VUE pages corresponding to fixed typical IOs – personalities, cities, countries, etc. Such searches take into account the values of some semantic properties specific to this IO, and match them with values of these properties for the current page. For example, for a typical IO “Person” it is possible to search for persons born in the same place, work in the same field, contemporaries, etc. The third case search option allows to search SSP of the selected category (or the set of categories) with a set of semantic properties defined for selected page. For example, you can find persons (pages from category "Person") who specialize in the fields relates to current page (categories of this page) (see Fig. 2).

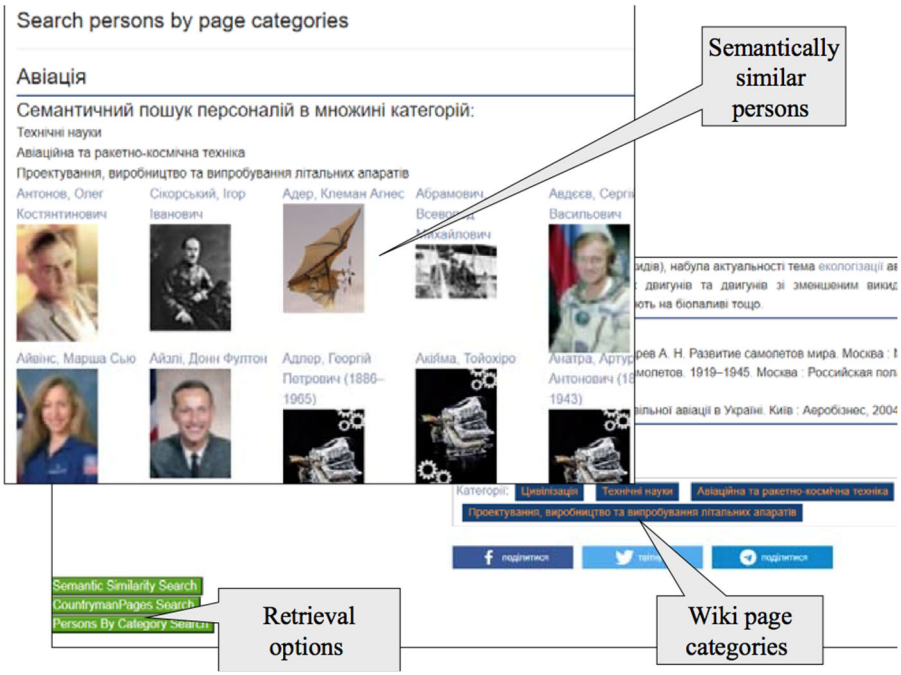

Fig. 2. Search for specialists (by set of current page category) for e-VUE pages.

SSP search generate the Wiki pages with the sets of locally similar concepts. These SSPs can be used in analysis of textual part of USD.

5 Conclusion

We propose to use semantic Wiki markup of IR as a source for generation of domain ontologies and groups of SSCc from these domains. The evaluation of semantic similarity can use such characteristics of pages as categories (and their taxonomies), semantic properties and their values, NL content and links with other pages. Then these SSC sets can be used as a base for analysis of NL unstructured data. The implementation of proposed approach needs in creation of relevant Wiki resources with appropriate domain knowledge. Use of semantic Wiki-technologies for distributed information resources development simplifies the process of NL text structuring and also generates background knowledge source for the analysis of arbitrary NL texts from corresponding domains. The models and methods proposed in the work allow to improve this process.

References

1. Grimes S.: Unstructured Data and the 80 Percent Rule, Clarabridge, Bridgepoints, (2008), http://breakthroughanalysis.com/2008/08/01 /unstructured-data-and-the-80-percent-rule/.
2. Unstructured_data, https://en.wikipedia.org/ wiki/Unstructured_data.

3. Grimes S.: A Brief History of Text Analytics, http://www.b-eye-network.com/view/6311.
4. Buneman P., Davidson S., Fernandez M., Suciu D.: Adding structure to unstructured data. In: International Conference on Database Theory, pp. 336-350. (1997).
5. Feldman R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data (2007), https://wtlab.um.ac.ir/images/e-library/text_mining/The%20Text%20Mining%20HandBook.pdf.
6. Quillian M. R.: Semantic memory. In: Minsky, M. (Ed.), Semantic Information Processing. MIT Press, Cambridge, MA, (1968)
7. Collins, A., Loftus, E.: A spreading activation theory of semantic processing. In: Psychological Review, 82, pp.407-428, (1975) .
8. Rada R., Mili H., Bicknel E., Blettner M. (1989) Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, 19(1), P.17-30.
9. Richardson R., Smeaton A. F., Murphy J.: Using WordNet as a knowledge base for measuring semantic similarity between words. In: Working paper CA-1294, Dublin City University, School of Computer AppUcations, Dublin, (1994).
10. Lee J. H., Kim M. H., Lee Y. J.: Information retrieval based on conceptual distance in IS-A hierarchies. In: Journal of Documentation, 49(2), pp.188-207, (1993).
11. Rada R., Bicknell E.: Ranking documents with a thesaurus. In: JASIS, V.10(5), pp.304-310, (1989).
12. Fellbaum C.: WordNet. In: Theory and applications of ontology: computer applications, pp. 231-243). Springer, Dordrecht, (2010).
13. Resnik P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language // Journal of Artificial Intelligence Research 11, P.95-130. (1999).
14. Miller G. A., Charles, W. G.: Contextual correlates of semantic similarity. In: Language and cognitive processes, 6(1), pp.1-28, (1991).
15. Wagner C.: Wiki: A technology for conversational knowledge management and group collaboration. In: The Communications of the Association for Information Systems, V. 13(1), pp. 264-289 (2004).
16. Banerjee S., Ramanathan K., Gupta, A.: Clustering short texts using wikipedia. In: Proc.of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 787-788, ACM, (2007).
17. MediaWiki, https://www.mediawiki.org/wiki/MediaWiki.
18. Broekstra J., Klein M., Decker S., Fensel D., Van Harmelen F., Horrocks I..: Enabling knowledge representation on the web by extending RDF schema. In: Computer networks, 39(5), pp. 609-634, (2002).
19. McGuinness D. L., Van Harmelen F.: OWL web ontology language overview. In: W3C recommendation, 10(10), (2004).
20. Rogushina J.V.: Use of semantic properties of the Wiki resources for expansion of functional posibilities of “Great Ukrainian Encyclopedia”. In: Encyclopaedias in the modern information space/ Ed. Kirillon A.M., Kyiv, pp.104-115, (2017) [in Ukrainian]
21. Rogushina J.: Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies. In: International Journal of Mathematical Sciences and Computing (IJMSC), Vol.3, No.3, 2017, pp.50-58.
22. Rogushina J.V.: The Use of Ontological Knowledge for Semantic Search of Complex Information Objects. In: Proc. of OSTIS-2017, pp.127-132, (2017).