Journal of Advanced Computer Science & Technology, 4 (1) (2015) 53-59
www.sciencepubco.com/index.php/JACST

doi: 10.14419/jacst.v4i1.4012
Research Paper

A new approach for finding semantic similar scientific articles

Masumeh Islami Nasab [1], Reza Javidan* [2]

[1] *Department of Computer Engineering, Fars Sciences and Research Branch, Islamic Azad University, Fars, Iran*
[2] *Department of Computer Engineering and Information Technology, Shiraz University of Technology, Shiraz, Iran*
**Corresponding author E-mail: reza.javidan@gmail.com*

Abstract

Calculating article similarities enables users to find similar articles and documents in a collection of articles. Two similar documents are extremely helpful for text applications such as document-to-document similarity search, plagiarism checker, text mining for repetition, and text filtering. This paper proposes a new method for calculating the semantic similarities of articles. WordNet is used to find word semantic associations. The proposed technique first compares the similarity of each part two by two. The final results are then calculated based on weighted mean from different parts. Results are compared with human scores to find how it is close to Pearson's correlation coefficient. The correlation coefficient above 87 percent is the result of the proposed system. The system works precisely in identifying the similarities.

Keywords: *Similarities; Semantic Similarities; Text Preprocessing; WordNet.*

1. Introduction

With the increasing pace of advancement in information technology in all areas and applications, calculating the similarity of available articles and documents is now a subject of considerable debate, and the same is true of designing a system being able to effectively determine the semantic similarities of two documents. This is particularly more important about scientific articles. Two scientific articles may not necessarily be identical by context. For example, a paper discussing about identifying article similarities by fuzzy logic may not be similar by context to another paper explaining the general purpose and the performance of fuzzy logic. However, two papers are associated with each other, because the second elaborates a basis for implementing the first one. As a result, papers may be incomparable based on words and texts or fewer similarities may be found. And as papers and documents complied by different authors may not be contextually similar but semantically comparable, a system is needed to calculate the semantic similarities of articles.

The semantic resemblance refers to conceptual similarities which are not lexically identical. Semantic similarities are calculated based on word mapping by ontology and studying their associations in that ontology [2].

This paper presents a method for determining the article similarities. According to the proposed method, the weighted mean of similarities between titles, abstracts and keywords. WordNet was used to find the semantic relations. Articles were separated out into three parts of title, keywords and abstract, and a weight was assigned to each part. The similarities of parts were first compared two by two. The final results were calculated based on weighted means of different parts. Results were then compared with human scores and the proximity was estimated by Pearson correlation coefficient. The correlation coefficient was calculated at over 87 percent, showing that the system was highly precise at finding similarities.

The paper is structured as follows: In Section 2 related literatures in this regard are studied. The proposed method is explained in Section 3. In Section 4 the results are outlined and finally in Section 5 that the conclusions are presented.

2. Literature review

Evaluating the semantic similarities of texts and documents has been widely studies. Many scientific researchers have struggled for a long time to find a criterion for semantic similarities between two words or short texts and also between two documents. Works which have been carried out on semantic similarity are classified into three groups by purpose. The first group explains methods of evaluating semantic similarities of words. Ramprasath and Harry Haran [2] employed an adaptive constructive algorithm for measuring the similarities of each pair of question and answer. This paper studies all types of algorithms used for measuring the semantic similarities of words. And the results were compared by an adaptive constructive algorithm. According to results, this algorithm works better than web-based semantic similarity. The problem with web-based methods is that the occurrence statistical methods such as number-based enumeration are used to measure the semantic similarities of words. However, in the proposed method, the occurrence statistics is replaced by different adaptation levels for measuring semantic similarity of words. Sahami et al. [3] measured the semantic similarities between two concepts according to responses given to enquiries by the search engine. For any enquiry, some responses have been gathered by a search engine and displayed each response as a weighted TF-IDF word vector. The semantic similarity between two enquiries is defined by multiplying the vectors of the center of gravity of that enquiry. Madylova [4] offered a new method for calculating the semantic similarities of documents. This is based on calculating the cosine similarities of content vectors of documents including IS-A relations for categorization. Content vectors of documents were first considered by developing words in the vectors of documents titled IS-A. The cosine similarities of content vectors were then calculated. One problem with this method is that calculation of semantic similarities of work documents is time-consuming, the semantic similarity between each pair of words is found. To solve this problem, Mihalcea [5] proposed a method for measuring the similarity between two short texts by knowledge-based and literature-based measurement. This method offers a best adaptation and uses semantic similarity. The second group is those methods calculate semantic similarity by WordNet. Ghazizadeh et al. [6] measured the semantic similarity of words using harmonious hierarchical tree of WordNet. To calculate word similarities, weighting hierarchical in line edges is used. The weight of the edge of each hierarchy and the last edge associating with a leaf are both assumed to be 1. The weight of each tree edge is obtained by multiplying the tree weight by the leaf edge with constant coefficient γ. Results revealed that a low level of tree edges has a positive impact on continuity and similarity. Qasim et.al [7] offered a combined statistical method for measuring semantic similarity of Quran short suras, which have been translated into English. According to this method, three algorithms of cosine similarity combine the longest common subtail and N-Gram. The performance of the combined method varies with the applied algorithms. However, it does not work well for longer texts. In the third group, the semantic similarity is used for categorizing documents. Ana hong [8] used similarity measurement for categorizing documents. Accordingly, all types of measurements including the Euclidean distance, cosine similarity, relative entropy, and Jaccard coefficient were used. All algorithms were tested in different databases, and results confirmed that Euclidean distance did not work well. In [9], documents were implemented based on measuring semantic similarity by testing the genetic algorithm comparing with k criterion in similarity environment. The genetic algorithm had a better recall precision. In this paper, the semantic similarity mean of articles was calculated. Cosine method was also employed to estimate similarities.

3. The proposed method

Figure 1 illustrates the architecture of the proposed method and the subsequent sections are explained based on this architecture.

Texts supposed to be measured by similarities are the selected contexts of articles that have been classified in three parts (title, abstract, and keywords). All these parts should be assessed in all articles. The articles are compared two by two in terms of titles, keywords and abstract. A weighted mean of them is converted as the final result of the tool. Any parts of texts which are supposed to be compared with each other in terms of similarity are first preprocessed by above tools. The texts are standardized and equalized. The sentences are then identified and listed separately. Now they are labeled by a lexical component labeller and the type of each word including nouns, verbs, adjectives, adverbs, etc. is specified. As nouns, verbs and adjectives are most important in texts and other words do not have critical role in sentence similarities, words are divided into three groups of nouns, verbs and adjectives. Other words are overlooked. All words existing in any of above groups are compared two by two in titles, keywords, and abstracts. Similarity means of each group are then calculated. As previously mentioned, the weighted mean of the similarities is considered as the level of resemblance between two articles.

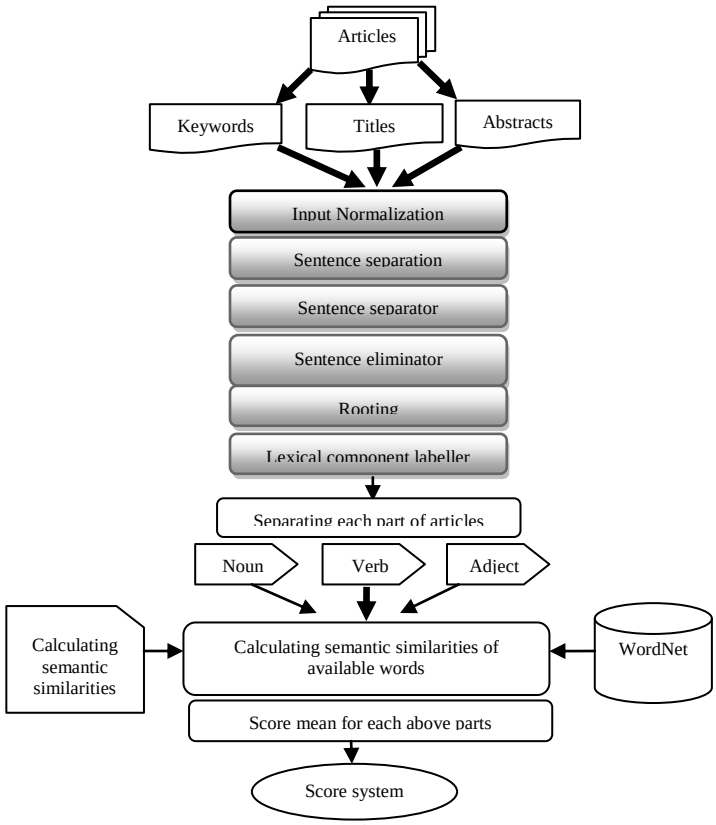

Fig. 1: The Architecture of the Proposed Method.

3.1. Text preprocessing tool

In the first step of designing and implementing the related tool for calculating the text similarities, most of basic tools for English processing should be designed and implemented according to a principal approach based on English grammar and writing rules. Among these tools, we can refer to normalizer or equalizer, word identifier, root finder, lexical component labeller, etc. normalizer substitute all text characters with a standard equivalent to equalize and normalize them. Sentences can be extracted from the text by the separating processor. When normalization is completed, the sentence identifier determines the sentence borders by marks ".", ",", "!", "?" in order to use them in subsequent steps. Separating processors helps with extracting text words. It also deletes less important words and/or stop words confirmed by users. Stop words, such as the, in, of, etc. are functional and less important words which are frequently seen when working with texts. At first sight, conjunction words and articles seem as stop words. Many verbs, auxiliary verbs, nouns, adverbs, and adjectives have been also identified as stop words. Root finder is responsible for finding word roots. Root findings means obtaining the root of words by eliminating their prefixes and affixes, as words with similar root have a similar appearance. For example, the words "computer", "compute", and "computing" are reduced to the word "compute" which is the main root. This process is so important in processing texts, because for example, car with two synonyms but with different appearance should not be considered as two unrelated words by root. Different algorithms have been proposed for word root finding. Porter algorithm [10] is the most prevalent algorithm in English language. The lexical component labeller is used to label lexical components.

3.2. WordNet

WordNet is a dictionary based on psycholinguistic theories and defines the models and meanings of words. WordNet more relies on word meanings than word forms. In morphology, however, verb drills have also been considered. WordNet consists of three databases: one for nouns, one for verbs, and one for both adjectives and adverbs. It includes a collection of word synonyms called "synsets" [1].

Each synset is composed of a meaning of a group of words. Synsets encompasses different semantic relations such as synonyms (similar), antonyms (opposite), hypernymy (super-concept), hyponymy (subconcept) (IS-A), meronymy (Part-of), holonymy (Has-A). According to grammatical categorizations, synsets have different syntactic relations [11]. WordNet provides also textual meanings of concepts (Glossary) including definitions and examples [12]. WordNet can be enumerated as a partial ordered collection of synonym resources

3.3. Calculating sentence semantic similarities between two texts

Having a look at different methods of calculating semantic similarities and given previous studies, we used cosine similarity for measuring sentence resemblances. A word vector is first made for each sentence (regarding the possibility of extending sentences by means of WordNet). The cosine distance is then calculated between the word vector of ith sentence Ci = [ci1, ci2… cim] and jth sentence Dj = [dj1, dj2… djm] according to formula (1) [8]:

$$\cos(C_i, D_j) = \frac{\sum_k c_{ik} d_{jk}}{\sqrt{\sum_k (c_{ik})^2} \times \sqrt{\sum_k (d_{jk})^2}} \quad (1)$$

3.4. Calculating word semantic similarities between two sentences

In last step, the semantic similarities between words are calculated by two by two WordNet. WordNet-based semantic similarity has been widely studied in Natural Language Processing (NLP) and Information Retrieval (IR). Many methods have been presented in this regard. Similarity criteria have been mostly applied to nouns, verbs and ARE-A relations in WordNet. This is because about 80 percent of relations and links between concepts are formed by hypernymy and hyponymy. When a semantic relation is studies in terms of concepts, however, synonyms, IS-A and Part-of have a bigger share in correlation of concepts. Two principal factors in determining the semantic similarities of two words are the relation and characteristics of WordNet nodes. Types of relation are the most important factor, which affect measuring word similarities by WordNet. Different links of WordNet would have different weights regarding the similarity weights. The most important relation is synonymy, which shows that two words are similar at the end of the link. In addition, IS-A link similarity weight is more than Part-Of link similarity weight. In applications of word similarities based on WordNet, the weight is generally defined as relation (2) [13].

$$sim_{wordnet}(w1, w2) \infty \begin{cases} 1 & type(w1, w2) = synonym \\ 0.95 & type(w1, w2) = Is - A \\ 0.85 & type(w1, w2) = part - of \end{cases} \quad (2)$$

Here, the definition of x-similarity in [14] is used to demonstrate the features of WordNet nodes. Features means sets of word definitions. It can be, thus, stated that if two words have identical definition sets, they are similar in terms of features. The commonality of two nodes is defined as relation (3):

$$sim_t(w1, w2) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

So, A and B are the explanatory words (after deleting stop words) and/or Synsets for two words w1 and w2.

3.5. Calculating semantic similarities between articles

As previously mentioned, articles were divided in three parts of titles, keywords, and abstracts. And a weight was allotted to each. The weights were considered based on empirical experiences and testing the tools for several articles. The final results are based on all three parts and the best tested weight relates to titles (0.35). Weights for keywords and abstracts were calculated at 0.40 and 0.25 respectively. They were shown by γ. Weighted means for article semantic similarities were calculated by the following formula (4):

$$Result = \gamma_1 * (sim_{title}) + \gamma_2 * (sim_{keyword}) + \gamma_3 * (sim_{abstract}) \quad (4)$$

4. Evaluation and results

1 This paper develops methods for calculating the semantic similarity (closeness)-relatedness of natural language words. The concept of semantic relatedness allows one to construct algorithmic models for the context-linguistic analysis with a view to solving problems such as word sense disambiguation, named entity recognition, natural language text analysis, etc. A new algorithm is proposed for estimating the semantic distance between natural language words. This method is a weighted modification of the well-known Lesk approach based on the lexical intersection of glossary entries.

2 Word similarity assessment is one of the most important elements in Natural Language Processing (NLP) and information retrieval. Evaluating semantic similarity of concepts is a problem that has been extensively investigated in the literature in different areas, such as artificial intelligence, cognitive science, databases and software engineering. Semantic similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Currently, its importance is growing in different settings, such as digital libraries, heterogeneous databases and in particular the Semantic Web. In this paper, authors present a search engine

Fig. 2: Data Sample Relating to Abstracts in Two Articles

According to the proposed method, the proposed system was implemented in C-Sharp. The articles were obtained by Thompson Reuter's dataset, downloaded from Web of Science. The dataset includes 100 articles with their semantic similarities have been scored by some individuals. Figures 2, 3, and 4 display the sample data used in this regard. The score means were taken into consideration as the semantic similarities between two articles. As it was mentioned before, the system results were saved in four output files. Results relating to the similarities of titles, keywords, and abstracts were presented. The input data were separately saved as a text file.

1 A method for the computation of the semantic similarity and relatednes between natural language words

2 A New Approach for Measuring Semantic Similarity in Ontology and Its Application in Information Retrieval

Fig. 3: Data Sample Relating to Articles

1 computer linguistics,semantic analysis of natural language texts,semantic similarity-relatedness of words,semantic ambiguity of words

2 Sense,Concept,Information content similarity

Fig. 4: Data Sample Relating to Keywords

After entering input files in each part, results are separately calculated and show a final file, which is a combination of three parts. Figures 5 to 7 display a number of output results in a diagram. The number of the interested article appears in the horizontal vector in which each part has been divided into several units. Each unit shows the number of another article which has been compared with that article. This means that these two articles have been compared with each other. The vertical vector indicates similarities. The similarity is a number between 0 and 1 and each article which has been compared with it is scored at 1.

Figure 5 depicts the results of article similarities of titles. Results are compared two by two and a number between 0 and 1 is saved as their semantic similarity in the output file as a matrix. Figures 6 to 7 present the similarities of keywords and abstracts. In figure 6, each word is compared with other words and the score is saved in a matrix. In figure 7, the score of comparisons are based on abstracts. For abstracts, sentence separation tool is employed and the sentences are separated out. This tool should be able to correctly identify sentences in the input text. In most verbal processing, the exact output of this tool is highly important. And use of abstract and keywords to have precise calculations. Figure 8 indicates the final results for all articles.

Pearson's correlation formula was used to find the correlation between system and human scores. The formula is stated as relation 5:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \quad (5)$$

Figure 9 depicts the comparison between system and human scores. As it is a huge figure, it has been presented at the end of the article. This figure shows the comparisons of ten articles. The horizontal diagram shows the article numbers. In other words, each article has been compared with ten other articles. The vertical diagram presents similarities. Finally, in figure 10, the proposed method has been compared with other methods. Results show the effectiveness of the proposed method.

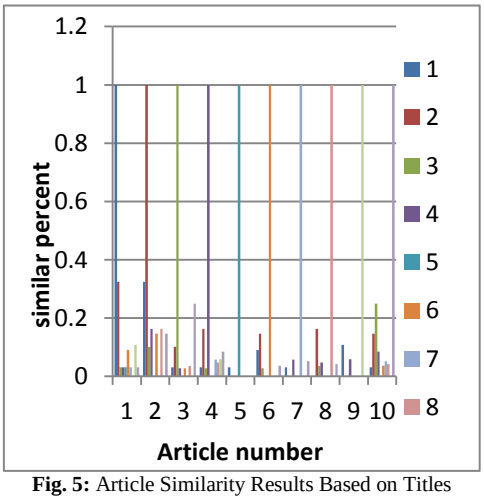

Fig. 5: Article Similarity Results Based on Titles

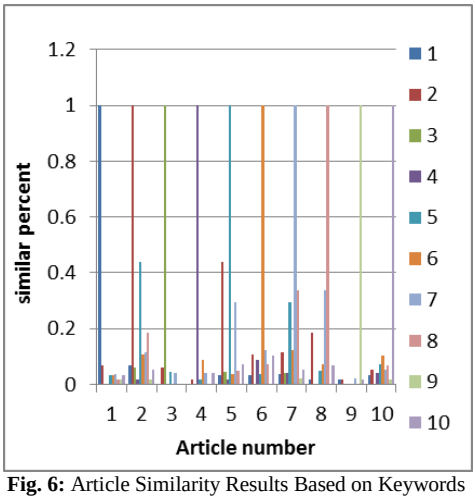

Fig. 6: Article Similarity Results Based on Keywords

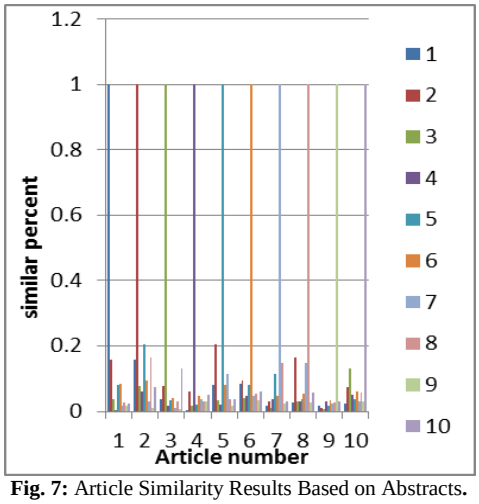

Fig. 7: Article Similarity Results Based on Abstracts.

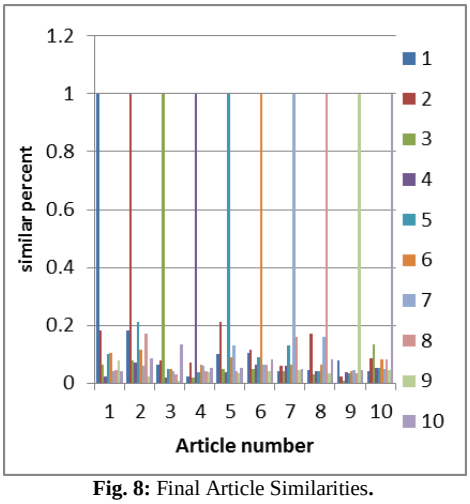

Fig. 8: Final Article Similarities.

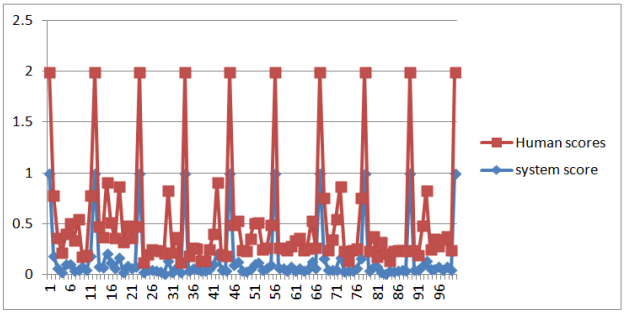

Fig. 9: Comparing System and Human Scorings

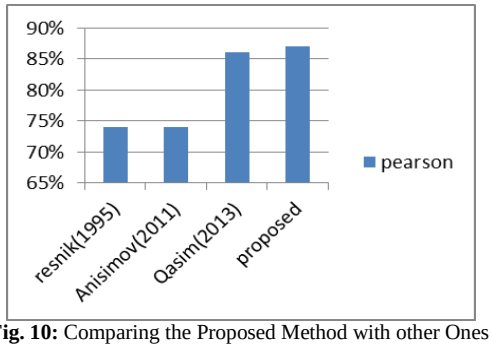

Fig. 10: Comparing the Proposed Method with other Ones

5. Conclusion

This paper presents a method for calculating article semantic similarities. According to the proposed method, similarities are stated based on the following points:

a) Separating out the article text into three parts of title, abstract, and keyword
b) Weighting titles, keywords and abstracts in terms of their contribution to the article
c) Calculating the weighted mean based on titles, abstracts and keywords
d) Using WordNet to achieve the semantic similarities of words

Pearson's correlation formula was used to find the correlation between system and human scores. The correlation coefficient of the proposed method was estimated at 87%, which is optimal comparing with methods proposed by Resnik [15] with correlation coefficient of 79%, by Anisimo et al. [16] with correlation coefficient of 74%, and by Qasim with correlation coefficient of 89%. It can also be focused on article similarities using a specialized WordNet. The proposed system can be used for other texts such as texts in Persian and in other language, which is need of a WordNet of that language.

References

[1] Sheth, A, Lytras M., "*Information Retrieval by Semantic Similarity*", int. journal on semantic web & information systems, 2(3), pp: 55-73. (2006).

[2] Ramprasath, M, Hariharan, Sh.,"Using ontology for Measuring Semantic Similarity for Question Answering System"IEEE International conference on Advanced Communication control and Computing Technologies(ICACCD), pp: 218-223. (2012).

[3] Sahami, M, Heilman, T., "*A Web-based Kernel Function for Measuring the Similarity of Short text Snippets*", Proceeding of 15th International Word Wide Web Conference. (2006). http://dx.doi.org/10.1145/1135777.1135834.

[4] Madylova, A., "*A Taxonomy based Semantic Similarity Documents Using Cosine Measure*", Computer an Information Sciences, IEEE,Iscis 2009.24th, International Symposium. (2009).

[5] Mihalcea, R., Corley, C, Strapparava, C., "*Corpus-based and Knowledge-based Measures of Text Semantic Similarity*", Proceeding of th National Conference on Artificial Intelligence ,pages:775-780. (2006).

[6] Ghazizadeh Ahsaee, M, Naghibzadeh, M, Yasrebi Naieni, S.E., "*Weighted Semantic Similarity Assesment Using Word Net* ", Dept. of Computer Engineering Ferdowsi University of Mashhad, Iran , International Conference on computer & Information Science(ICCIS), pp:66-71, (2012).

[7] Qasim, A, Omar, N, Albared, M., "*Combined Statistical Methods to Measure Semantic Text Similarity in Holy Qur'anic Translations*", Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, university Kebang Saan Mlaysia, 43600 Bangi Selangor, Malaysia, vol5(17), pp:1-7, (2013).

[8] Huang, A., "*Similarity Measure for Text Document Clustering*", Department of Computer Science The University of Waikato, Hamilton, New Zealand, pp:49-56, (2008).
[9] Song, W, Cheol Park, S., "*An Improved Genetic Algotithm for Document Clustering With Semantic Similarity Measure*", Division of Electronics and Information Engineering, Chonbuk National University, Jeonju, 561756, korea(IEEE), pp:536-540. (2006).
[10] Porter, M., "*An algorithm for suffix stripping. Program*".14(3), pp.130-137, (1980). http://dx.doi.org/10.1108/eb046814.
[11] Lin, F, Sandkuhl, K., "*A Survey of Exploiting WordNet in Ontology Matching*". In IFIP International Federation for Information Processing, Artificial Intelligence and Practice II; Max Bramer; (Boston: Springer), Vol 276, pages: 341–350, (2008).
[12] Cimiano, P., "*Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*", Springer, 2006.
[13] Lin, D., "*An information-theoretic definition of similarity*". In Proceeding of the15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, USA, pp. 296–304, (1998).
[14] Petrakis, E.G.M., Varelas, G., "*Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies*". In 4th Workshop on Multimedia Semantics (WMS'06), pp. 44–52, (2006).
[15] Resnic, P., "*Using Information content to evaluate semantic similarity in a taxonomy*", Proceedings of IJCAI-95, vol. 1, 448-453, (1995).
[16] Anisimov, A.V., Marchenko, O.O, and Kysenko .V.K., "*A Method for the Coputation of the Semantic Similarity and Relatedness between Natural Language Words*", Cybernetics and Systems Analysis, Vol 047, pp: 515-522, (2011). http://dx.doi.org/10.1007/s10559-011-9334-2.