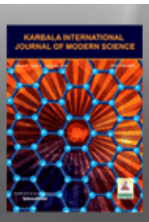

Karbala International Journal of Modern Science

Volume 5 | Issue 1 Article 9

About one approach to implementing semantic search

Murtadha Rasol
Southern Federal University, rasol@sfedu.ru

Jamal Challab
jamalalsadawy@gmail.com

Adnan Al-saeedi
adn.ak21@gmail.com

About one approach to implementing semantic search

Abstract

Traditional search mechanisms are based on the keyword search, which does not consider the semantic links between different concepts. This leads to the loss of relevant documents due to inaccurate query formulation or using contextually close words and concepts in the query. To solve the problems of formulating user queries and interdisciplinarity of concepts, it is suggested to use semantic search. The proposed method for implementing semantic search is applicable to large scopes of text data and is based on using a genetic algorithm. Unlike standard methods for information search, the suggested method allows us to consider the semantics of interrelationships between concepts and to handle interdisciplinary concepts correctly. By the aid of semantic tagging, documents contain concepts that are not present in the user's initial query but are semantically close to the requested concepts. Semantic tagging is performed for each document separately, which provides parallel tagging in several subject areas. By the time of the document ontological profile formation is completed, all semantic distances between pairs of distinguished concepts are calculated. Concepts are considered contextually close if their semantic proximity value is above a certain threshold value that is specified in the search parameters. Building a document ontological profile is a multicriteria task, since it depends on a lot of characteristics, so genetic algorithms can be used to solve it effectively. The developed genetic algorithm is intended for more accurate distribution of weight coefficients and estimation of semantic proximity of concepts.

Keywords

semantic search, information retrieval, ontology, genetic algorithm

1. Introduction

The standard mechanisms of the text analysis are mostly focused on the statistical analysis, which does not consider semantic relationships between the concepts. The syntactical methods work perfectly in the text analysis tasks but they have constraints in the contextual information search. The keyword-based search, which is used in the traditional search mechanisms can ignore the required document if the query contains contextually close words and concepts, which are not present in the desired document.

In Ref. [1] Amaral et al. describe two main problems in information search and retreival:

1) Query formulation. If the user defines the query using synonyms, generalizations or contextually close concepts, the result set will not contain many relevant documents.
2) Multi-disciplinarity. The same concepts in the different areas can have different meanings. For instance, the concept 'ontology' in philosophy and in computer science mean differently.

To solve the mentioned problems, the authors suggest developing the algorithm allowing us to consider ontological knowledge, which describes the subject area of contained document collections using semantic measures considering the query context through the example of scientific papers. The importance of using ontological knowledge in information search and retrieval is described by Huiping Jiang in Ref. [2] and Batra et al. in Ref. [3]. It is suggested to use the genetic algorithm to handle multi-disciplinary concepts correctly and to improve the mechanisms of query tagging. The state-of-art studies demonstrate the effectiveness and benefits of using the features of genetic algorithms in terms of information search. For instance, it is Kravchenko et al. in Ref. [4] presenting the genetic algorithm for semantic similarity estimation, or Yuxin Mao in Ref. [5] describing semantic-based genetic algorithm.

The proposed approach is implemented in terms of searching scientific papers. Regardless of different appearance requirements, scientific papers have a common pattern, which includes the title, the abstract, the keywords, the body and the conclusion. The developed algorithm considers this pattern for all papers. Usually the most important concepts are focused in the title, abstract or keywords. These concepts are rated higher than the concepts appeared in the document body only.

The rest of the paper is organized as follows. The next section is composed of five subsections describing the developed method of semantic search using genetic operators. The first one presents the existing approaches to implement the semantic search. It is noted that we can improve semantic search by using additional knowledge sources such as ontologies and novel mechanisms of semantic tagging. The second subsection describes the main steps of semantic tagging and the mathematical formula for calculating document dispersion. The next subsection presents the method of semantic closeness estimation to be used in the proposed model of semantic search. The last subsection is devoted to the developed genetic method: the model of the search space and chromosomes, genetic operators and fitness function, which demonstrates the accuracy of the weight coefficients allocation in terms of the considered document.

The third section presents the results of implementing the proposed method for semantic search using ontologies. The experiments show that the concepts appearing the most frequently are included in the title, abstract or keywords and have the greatest weight coefficients. The concepts, which are not in the text, but are semantically close with the most frequent concepts, have quite great weight coefficients. The last section summarizes and concludes the paper.

2. Materials and methods

2.1. Semantic search and building the document semantic profile

There are three main approaches to implement the semantic search:

1) to cluster the obtained results;
2) to analyze the syntax and semantics of the natural language;
3) to perform the search considering semantic relationships between the concepts.

Approaches 1 and 2 improve the search accuracy by using the correct semantic interpretation of the concepts and conflict resolution while interpreting the multi-disciplinarity. Their shortcomings are that these

https://doi.org/10.33640/2405-609X.1009

approaches only find the documents, which consist of at least one keyword from the query. Approach 3 provides decomposition of the concepts into triplets <subject, property, object>. It allows us to find relevant documents which do not include any keywords from the user's query but do have semantic relationships with them. The authors suggest using the ontologies to describe different concepts of the subject area and to establish the relationships between them [3–5].

To improve the quality of semantic search, it is needed to use the additional knowledge sources which describe the different concepts and relationships between them. For this purpose, the ontology warehouses are used. To consider the concepts appearing in the document as well as the concepts having the semantic relationship it is required to apply a new mechanism of tagging of the documents and users' queries.

The tagging mechanism mentioned above is to consider the following statistical characteristics [6]:

- the number of concept repetitions in a document;
- document frequency;
- concept dispersion;
- structure of the searched document.

It can be therefore assumed, that the task to be solved is a task of multicriteria optimization, which solution can be found through using of genetic algorithm (GA). The correct settings of the GA allow us to obtain the results, which are very close to the optimal ones in polynomial time. Genetic algorithms are highly effective in terms of solving optimization tasks without any clear requirements for the fitness function. To implement tagging of the documents and the user's query, it is needed to formulate the document ontological profile. The document ontological profile is defined by the list of pairs <concept, weight coefficient>, which are related to a certain subject area including the concepts found in the document or semantically related to them. The length of the semantic relationship (semantic distance) is determined in the GA parameters by using the threshold value. In that connection, the semantic tagging of the documents is the process of formulating the ontological profile considering the metadata of the available ontologies [6–10].

The ontological profile characteristics are [11]:

1) to depend on the ontology;
2) to include the multiword concepts;
3) to include all the concepts from the ontology, which are semantically related;
4) to include all the concepts, which meet the established criteria.

All the documents are tagged independently. Tagging of the documents is performed in a parallel way in terms of different subject areas, which provides the high speed of the GA processing the large collections of the documents. The algorithm of editing the queries include three steps [12]:

1) to pre-process the text;
2) to highlight the concepts from the ontology and the semantically related concepts in the document;
3) to estimate the weight coefficients using the GA.

2.2. Semantic tagging of the documents

The first step is to perform the morphological analysis of the document: to define the initial form of the words, part of speech, etc. It is defined if the concept appears in the title, abstract or keywords during the pre-processing stage.

At the second step, the concepts defined in the subject area ontology are to be determined in the text. The words remaining in the document are not considered serving as stop-words for the algorithm. The difficulty here includes the multi-word concepts in the chosen ontology, since several words in a row are processed rather than individual words during the text analysis. Let us formulate the following characteristics for all the concepts in a document:

1) F – denotes a number of concepts repetition in a document.
2) $(x_1, x_2,\ldots,x_F)$ – denote the position numbers, where the concept appears.
3) D – denotes the concept dispersion, which is calculated as follows:

$$D = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{F} \quad (1)$$

where $\overline{x}$ – is the average value of the positions $(x_1, x_2,\ldots,x_F)$.

The next step is to find the concepts, which are semantically close to those, that had already been highlighted in a document. For this purpose, it is

needed to build a chart containing all pairs <concept, weight coefficient>, where the weight coefficient denotes the semantic closeness (0; 1). The closer the concepts are, the higher is their semantic closeness. Semantic closeness can be estimated in several ways: considering the number of the paths between the concept pairs, changing the types of the paths, presence of the common parents, etc. The paths are represented as edges, and the concepts are represented as the vertexes in the ontology graph.

For each document, the semantic closeness is estimated only once in the context of each ontology of the subject area. By the time of the document ontological profile formation is completed, all semantic distances between pairs of distinguished concepts are calculated. The concepts are considered as close, if their semantic closeness is higher, than a certain threshold value, which is set in the search parameters. These concepts are added to the list of the concepts found in the document. Frequency and dispersion of the concepts, which do not appear in the document text, equal to 0 [15,16].

2.3. Methods of semantic closeness estimation

There are many methods for estimating semantic closeness between the concepts. Semantic closeness of the objects includes a lot of closeness aspects, thus, choosing the criteria of estimating is a rather difficult task depending on the purposes of the research in each individual case. Closeness measures of the ontological terms apply the different semantic characteristics of the compared concepts: their properties (different attributes and relationships with the other concepts), positions on the ontology, etc.

The hybrid measures of closeness compose the set of the measures of concepts closeness mentioned above. The more accurately the characteristics of the concepts are considered, the better quality of the closeness estimation is obtained. Thus, the hybrid measures of closeness are the most promising since they combine several approaches to estimate the closeness. In the developed algorithm the authors suggest using the hybrid measure of estimating the concepts closeness based on the additive convolution:

$$S(c_1, c_2) = \sum_{i=1}^{n} w_i S^i(c_1, c_2), \quad (2)$$

where S^i – denotes the closeness measure according to the chosen criterion, weight coefficient w_i denotes the relative importance of the criterion (total of the weight coefficients is 1), n – denotes the number of the used criteria. Weight coefficients can be determined by the subject area experts or users in an interactive way or by the genetic algorithm (GA) automatically [12–14].

Let us introduce $ONTO_i$ to denote the created list for the initial search, where i – is a number of the concepts in the list, $ONTO_i = (c_1,\ldots, c_i)$, where c_i – is a concept, which appears in the document or is contextually close.

After creating the initial list of the concepts it is needed to estimate the weight coefficients of the concepts in the text. Weight coefficient of the concept is a numerical characteristic, which evaluates the expression of the concept in the document text. It depends on the following criteria [11]:

1) statistical (such as dispersion and frequency);
2) ontological (presence or absence of the relationships between the concepts);
3) structural (the part of the document the concept appears: title, keywords or body).

At the beginning, GA generates the weight coefficients for all the concepts from the initial list. On the run of the genetic operators and during the evolutionary process those concepts, which meet the most criteria, are chosen. In the end, we obtain the optimal allocation of the weight coefficients meeting the mentioned criteria [10].

2.4. Search space and genetic operators

Search space is a set of vectors $(w_1,\ldots, w_i)$ of the length of i, where w_i – is the weight coefficient of the concept c_i from $ONTO_i$ taking on a value from 0 to 1, i – is a common number of the ontology concepts, which are semantically related with the document.

Chromosome is the vector $(w_1,\ldots, w_i)$ from the search space, genome is a weight coefficient of a concept. All chromosomes are generated randomly, thus, to estimate their fitness we use the fitness function and choose the classical genetic operators (GO) [9]:

1) to involve p percent of the species during the crossover.
2) to use a single-point crossover with a random choice of crossover point.
3) to change the value of one gene to a random value during the mutation.

Genetic algorithm (GA) is finished after stabilization of the population, when the best value does not change during several iterations.

2.5. Calculating of the fitness function

The value of the fitness function demonstrates the accuracy of the weight coefficients allocation in terms of the considered document. Due to the large amount of the criteria, they are represented as the heuristic rule in the notation of the predicate calculus language of the first order. Let us introduce the following designation to describe the rules:

1) Variable: *a,b* − denote the concepts.
2) Functions:
   a) *f(a)* − denotes the number of appearance of the concept in the document text.
   b) *d(a)* − dispersion of the concepts in the document.
   c) *c(a,b)* − semantic closeness between the concepts.
   d) *w(a)* − weight coefficient of the concept.
   e) *p(a)* − is a number of the paragraph, where the concept appears for the first time.
3) Predicates:
   a) *GT(X,Y), LT(X,Y), GE(X,Y), LE(X,Y), EQ(X,Y), NE(X,Y)* − are the standard predicates,
   b) $X > Y$, $X < Y$, $X \geq Y$, $X \leq Y$, $X = Y$, $X \neq Y$ are the predicates denoting the relationship.
   c) *Close(X,Y,Ɛ)* − X is close to Y or $|X-Y| < Ɛ$.
   d) *Title(a), Annotation(a), Keyword(a)* − concept *a* appears in the different parts of the document: title, abstract or keywords correspondingly.

Considering the assigned interpretations of the predicate and functional symbols, let us describe the heuristic rules to estimate the weight coefficients:

1) If the concept *a* appears in the document more frequently, than the concept *b*, the weight coefficient of the concept *a* is greater than or equal to the weight coefficient of the concept *b*:

$$ab(GT(f(a),f(b))) \Rightarrow (GE(w(a),w(b))).$$

2) If the dispersion value of the concept *a* is greater than the dispersion of the concept b, weight coefficient of the concept *a* is greater than or equal to the weight coefficient of the concept *b*:

$$ab(GT(d(a),d(b)\&GT(d(b),0)) \Rightarrow (GE(w(a),w(b))).$$

3) If the concept appears in the first paragraph, its weight coefficient is greater than or equal to the weight coefficients of the concepts, which appears in the other paragraphs for the first time:

$$ab(EQ(p(a),1)\&\&NE(p(b),1)) \Rightarrow (GE(w(a),w(b))).$$

4) The weight coefficients of the synonyms are close:

$$ab(GT(c(c,b),\beta) \Rightarrow (Close(w(a),w(b),Ɛ).$$

5) If the concepts *a* and *b* are synonyms, and the total of their frequencies is *f(a) + f(b),* the weight coefficient of these concepts is greater than or equal to the weight coefficient of any other concept *c*, which frequency is *f(c) < f(a) + f(b):*

$$abc(EQ(c(a,b),1))\&NE(c(b,c),1)\&EQ(c(a,c),1)\&LT(f©,f(a) + f(b))) \Rightarrow (GE(w(a),w(c))\&GE(w(b),w(c))).$$

6) If the concept appears in the title, abstract or keywords, its weight coefficient is greater than or equal to the weight coefficient of the concept, which appears in the document body only.

Natural language queries can be formulated arbitrarily: as a string in terms of the Internet search systems; as a fragment of the text document or the full text; as a list of concepts, which are highlighted in the hierarchical representation of the concepts from the ontology. Text analysis is performed in a similar way to analysis of the full text of the searched document. It represents the query as follows:

$$Q = ((cq_1, wq_1), \ldots, (cq_n, wq_n)), \quad (3)$$

where cq_i − denotes a concept *i-e* from the user's query, wq_i − denotes a weight coefficient of the concept cq_i, n − denotes the number of the highlighted concepts in the query. From the obtained resulting set of the documents, it is needed to calculate the scalar product $<Q_i, D_i>$ for

the ontological profile of each document D_i. The value of the scalar product represents the evaluation of the relevance of the document to the query. Thus, only those concepts, which scalar product value is greater, than the threshold value are put into the resulting set of the documents [14].

3. Results and discussion

The algorithm is represented through the example of semantic tagging of the document in the context of two subject ontologies. The first one describes the subject area for solving the optimization tasks, the second one is devoted to the artificial neural nets (Fig. 1).

In Fig. 1 the concepts, which do not appear in the document text are highlighted. There are the following relation types in the considered ontologies:

- *is-a* – is the relation class-subclass;
- *a-part-of* – is the relation whole-part;
- *syn* – is the relation of synonymy;
- *goal* – is the relation of goal;
- *mtd* – is the relation of method.

The ontological profiles of the document are built considering the context of the ontologies and represented in Tables 1 and 2.

During building the ontological profile of the document, it is assumed, that semantic closeness of the concepts is calculated on the basis of the hybrid measure of closeness described in the previous paragraphs. The distance between the synonyms is equal to 0, weight coefficient of the relations type *is-a* is equal to 0.1, *a-part-of* – 0.3, for all the other relations types – 0.6.

The example described above allows us to conclude, that:

Table 1
Ontological profile of the document in the context of the optimization methods.

Concept from ontology of the optimization methods	Weight coefficient
Minimum	0.94
Gradient descent	0.62
Iterative algorithm	0.45
Maximum	0.38
Optimization	0.27
Optimization task	0.15
Extremum	0.08

Table 2
Ontological profile of the document in the context of the artificial neural networks.

Concept from ontology of the artificial neural networks	Weight coefficient
Neuron	0.90
Artificial neural network	0.88
Neural network	0.75
Synapse	0.60
Perceptron	0.58
Artificial intelligence	0.55

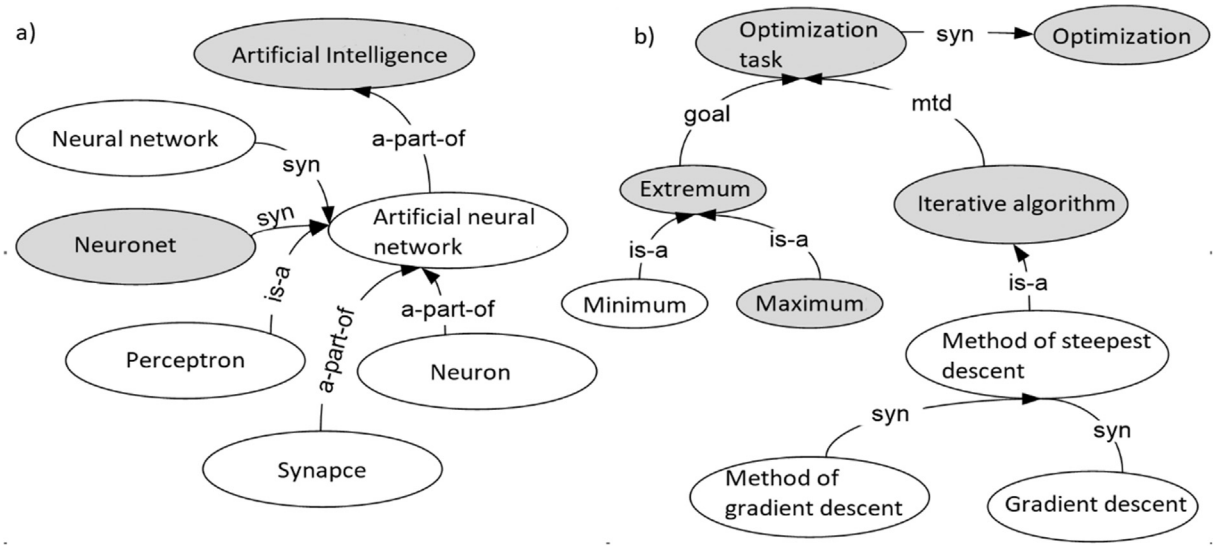

Fig. 1. Ontology fragments: a) – artificial neural networks, b) – optimization methods.

1) Ontological profile of the document represents the presence of interrelated concepts from the ontology, on the basis of which is has been built.
2) The concepts, which appear in the text the maximum number of times and are included in the title, abstract or keywords, have the greatest weight coefficients.
3) If the concept does not appear in the text, but there is a semantic relationship with one of the frequent concepts, its weight coefficient is rather great.

4. Conclusion

The paper describes the genetic algorithm, which provides semantic tagging of large collections of text documents according to the ontology context. The authors also demonstrate the algorithm of search in terms of the built semantic tagging. Using the genetic operators allows us to obtain more accurate results, semantic tagging in several ontologies simultaneously extends the search space. The calculated weight coefficients for estimating semantic closeness improve the accuracy of semantic search by semantic tagging the concepts, which are not used in the queries or documents, but are semantically close to them. The experiments show that the concepts appearing the most frequently are included in the title, abstract or keywords and have the greatest weight coefficients. The concepts, which are not in the text, but are semantically close with the most frequent concepts, have quite great weight coefficients.

References

[1] C. Amaral, D. Laurent, A. Martins, A. Mendes, C. Pinto, Design and implementation of a semantic search engine for Portuguese, in: LREC 2004 Proceedings of 4th International Conference on Language Resources and Evaluation, vol. I, 2004, pp. 247−250. Lisbon, Portugal.

[2] Jiang Huiping, Information retrieval and the semantic web, in: 2010 International Conference on Educational and Information Technology (ICEIT), vol. 3, IEEE, 2010.

[3] Mridula Batra, Sachin Sharma, Comparative study of page rank algorithm with different ranking algorithms adopted by search engine for website ranking, Int. J. Comput. Technol. App. 4 (1) (2013) 8.

[4] Y. Kravchenko, I. Kursitys, V. Bova, The development of genetic algorithm for semantic similarity estimation in terms of knowledge management problems, Adv. Intell. Syst. Comput. 573 (2017) 84−93.

[5] Yuxin Mao, A semantic-based genetic algorithm for sub-ontology evolution, Inf. Technol. J. 9 (2010) 609−620.

[6] D. Amerland, Google semantic search: search engine optimization (SEO) techniques that gets your company more traffic, in: Increases Brand Impact and Amplifies Your Online Presence, Que Publishing, 2013, p. 230.

[7] C. Daya, Wimalasuriya and Dejing Dou. Ontology-based information extraction: an introduction and a survey of current approaches, J. Inf. Sci. 36 (3) (June 2010) 306−323.

[8] Qing He, Xiu-Rong Zhao, Ping Luo, Zhong-Zhi Shi, Combination methodologies of multiagent hyper surface classifiers: design and implementation issues, in: Second International Workshop, Proceedings. − Springer Berlin Heidelberg, 2007, pp. 100−113. AIS-ADM 2007.

[9] H. Baazaoui Zghal, M.-A. Aufaure, N. Ben Mustapha, A model-Driven approach of ontological components for on-line semantic web information retrieval, Special Issue Eng. Semant. Web J. Web Eng. 6 (4) (2007) 309−336. Rinton Press.

[10] R. Srikant, R. Agrawal, Mining generalized association rules, Proc. VLDB 10 (2010) 407−419.

[11] A.F. Tuzovskij, S.V. Chirikov, V.Z. Jampol'skij, Sistemy upravlenija znanijami (metody i tehnologii)/pod obshh. red. V.Z. Jampol'skogo, Izd-vo NTL, Tomsk, 2005, 260 s.

[12] H. Peat, P. Willet, The limitations of term co-occurrence data from query expansion in document retrieval systems, J. Am. Soc. Inf. Sci. 42 (5) (2012) 378−383.

[13] J. Davies, R. Weeks, U. Krohn, QuizRDF: Search Technology for the Semantic Web. WWW2002 Workshop on RDF & Semantic Web Applications, Proc. WWW2008, 2008. Hawaii, USA.

[14] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, Y. Warke, Managing semantic content for the web, IEEE Internet Comput. 6 (4) (2010) 80−87.

[15] N. Stojanovic, R. Struder, L. Stojanovic, An Approach for the Ranking of Query Results in the Semantic Web. Proc. Of ISWC '03 (Sanibel Island, FL, October 2003), SpringerVerlag, 2013, pp. 500−516.

[16] B.N. Nguen, A.F. Tuzovskij, Obzor podhodov semanticheskogo poiska, Doklady Tomskogo gosudarstvennogo universiteta sistem upravlenija i radiojelektroniki 2 (2010) S. 234−S. 237. №. 2.