

A Bidirectional-Based Spreading Activation Method for Human Diseases Relatedness Detection Using Disease Ontology

Said Fathalla^{1,2(✉)} and Yaman Kannot³

¹ Enterprise Information Systems (EIS), University of Bonn, Bonn, Germany
fathalla@cs.uni-bonn.de

² Faculty of Science, Alexandria University, Alexandria, Egypt

³ Software Engineer, Alexandria, Egypt
yaman.kce@gmail.com

Abstract. There is a numerous demand for a standard representation of the ubiquitous available information on the web. Developing an efficient algorithm for traversing large ontologies is a key challenge for many semantic web applications. This paper proposes spreading activation over ontology method based on bidirectional search technique in order to detect the relatedness between two human diseases. The aim of our work is to detect disease relatedness by considering semantic domain knowledge and description logic rules to identify diseases relatedness. The proposed method is divided into two phases: Semantic Matching and Disease Relatedness Detection. In Semantic matching phase, diseases in submitted query are semantically identified in the ontology graph. In Disease relatedness detection phase, disease relatedness is detected by running a bidirectional-based spreading activation algorithm and return the related path (set of diseases) if so. In addition, the classification of these diseases is provided as well.

Keywords: Bidirectional search · Disease ontology · Semantic web · Spreading activation

1 Introduction

The use of ontologies in the field of health informatics has become a mainstream activity within bioinformatics due to the vast growing of healthcare system. In bioinformatics, ontology is used for representing and organizing medical vocabularies. Spreading Activation (SA) could be run on semantic networks and could also be used for information retrieval process [2]. Spreading activation is appropriate to run on incomplete and large graphs. It runs on a graph structure that comprises a set of nodes connected by edges in which concepts are nodes with an activation value and the relations between them are represented by edges. An activation value is assigned to each node in the graph and then the algorithm spreads to the nodes with the higher activation value. The algorithm runs in a set of iterations and terminates when a stopping condition is reached. The output is a list of

activated nodes within each iteration. For each iteration or cycle, there are three substantial actions: (1) The list of nodes is expanded by adding adjacent nodes (all nodes which have links to the nodes in the list), (2) The activation value at each node in the list is recomputed based on the activation value of the node itself and the weight of links which exist between other nodes, and (3) The list is filtered by excluding the nodes with activation values less than a given threshold. Below a group of definitions related to the proposed methodology:

- *Semantic Relation*: A semantic relation between any two concepts in the ontology is one of the relations among the set of semantic relations $\Sigma = \{\text{Hypernym, Hyponymy, Synonymy}\}$,
- *Hyponym, Hypernym*: a hyponym is a word whose semantics is a specific meaning of another word which called its Hyperonym or its Hypernym. For instance, vaccinia and smallpox are all Hyponyms of viral infectious disease (their Hypernym),
- *Co-hyponyms*: Given a concept C has two hyponyms A and B, then A and B are identified as co-hyponyms.

This research investigates the question of whether two human diseases are related to each other by any means and what is the relation between them if exists. For instance, is there a relatedness between *Vasculogenic Impotence* and *Transvestism*? If so, what is the relatedness? A striking feature of finding a relatedness between diseases is that physicians can treat patients not only based on symptoms they suffer but they can treat the real cause of this disease which may be related to another disease that causes these symptoms. Therefore, physicians can treat the real cause disease, not the symptoms. For instance, Gallstones disease (Cholelithiasis) may be caused by *Hemolytic Anemia* disease so crushing gallstones is not a solution or treatment because the stones will develop again [24]. Therefore, the objectives of detecting the relatedness between diseases are¹:

- *Causality*: A disease may occur due to the existence of another disease. For example, *Hereditary Spherocytosis* diseases is an autosomal preponderant anomaly of erythrocytes that causes gallstones. Pigmented gallstones occur in approximately half of untreated patients,
- *Complications of diseases*: one disease may increase the complications of another. For example, Diabetes and HCV hepatitis,
- *Treatments prescription*: Treatments may differ when there is a relation between two diseases.

The remainder of the article is structured as follows: We present an overview on related work in Sect. 2. The disease ontology which is used as a semantic knowledge base for the proposed method is described in Sect. 3. We present the proposed method in Sect. 4. The workflow of the proposed method is presented in Sect. 5. The proposed method is illustrated by using a running example in Sect. 6. The conclusion and the directions for future work are outlined in Sect. 7.

¹ Thank you Dr. Diaa Elsayed, Gastroenterology and Hepatology specialist, for the Counseling.

2 Related Work

There are a numerous amount of literature on biomedical knowledge management and medical decision making due to the explosion of biomedical knowledge over the last recent years. Therefore, biomedical knowledge available on the web is growing considerably as most of the biomedical research papers are published online. Semantic matching is used to expose information which is semantically related to structured data based matching concepts not keyword-based [12]. Abundant assorted frameworks and algorithms of semantic matching have been proposed so far such as [13, 21, 25]. Some examples of individual approaches addressing the matching problem can be found in [6, 7]. In their cutting edge paper of 2010, Ngo, Cao, and Le [20] proposed an ontology-based vector space model for semantic annotation and semantic search by combines different ontologies. It takes advantage of ontological features of both named entities and WordNet vocabularies and develops a spreading activation algorithm for query expansion. As anticipated, their experiments evince that their model is better than the solely keyword-based model and also better than the ones using only WordNet or named entities. In addition, it merges various ontologies to improve the semantic search process. De Maio et al. [5] proposed a project named ODINO which uses a fuzzy knowledge approach for disease diagnosis that supports medical decision-making. ODINO has three main features which are a faceted search of diseases through taxonomy constraints, disease catalog browsing, and preliminary medical diagnosis. In the field of bioinformatics, there are a lot of research has been made to represent medical information as a semantic knowledgebase for further processing by semantic applications. Due to the existence of many medical ontologies, it is important to reuse and integrate ontologies to establish suitable mappings between their concepts. Therefore, a lot of research in ontology matching and integration [8, 23] have been done in the recent years. Shvaiko and Euzenat [23] surveyed the state of the art of ontology matching and addressed some worthy challenges for ontology matching techniques. In addition, they analyzed the results of recent ontology matching evaluations. Some of the famous medical-related ontologies are:

- *Human Disease Ontology (DO)*² is a standardized biomedical ontology which contains a considerable number of disease terminologies,
- *Vaccine Ontology (VO)*³ is a biomedical ontology which contains more than 2000 terms and relationships for vaccines and vaccinations,
- *Infectious Disease Ontology (IDO)*⁴ comprises a set of ontologies which represent infectious diseases,
- *Ontology for Biomedical Investigations (OBI)*⁵ is an integrated ontology for the concepts which belong to life-science and clinical practice,

² <http://www.disease-ontology.org>.

³ <http://www.violinet.org/vaccineontology>.

⁴ http://infectiousdiseaseontology.org/page/Main_Page.

⁵ http://obi-ontology.org/page/Main_Page.

3 The Disease Ontology

The main purpose of developing ontology is to use it as a semantic knowledge base for identifying concepts in a specific domain and to share a data semantics among software agents so that it becomes machine understandable [10]. Köhler et al. [17] and Croft et al. [4] discuss in their research that the human disease data is a cornerstone of biomedical research. Therefore, there's an enormous need for a consistent representation of human disease for robust data analysis [18]. Creating a biomedical knowledgebase in the form of ontologies creates a rigid knowledgebase for semantic annotation of biomedical data through defined concepts and relations connecting them [14]. The disease ontology (DO) has been selected to be used as the semantic knowledgebase for the representation of human diseases. The objective of using DO is to provide the biomedical community with a convenient, reusable and robust knowledgebase of human disease concepts [16]. Major enhancements to the DO database since 2012 has been made including: the content of DO has had several revisions, including the addition of 32% of all terms. The Disease Ontology database has been updated to the latest ontology as of March 2, 2017. The DO project has had a considerable influence on the development of biomedical resources, as evidenced by 307 Google Scholar citations (as of April 14, 2017) to DO's paper [22] published in 2012. We have used the disease hierarchy in DO to infer disease relatedness. Furthermore, synonyms of diseases are also used in semantic matching phase. For instance, *Carotenemia* disease has exact synonym *Hypercarotinemia*.

4 Methodology

The methodology of the proposed work is divided into two phases: Semantic matching and Disease relatedness detection.

In Semantic matching phase, diseases in submitted query are semantically identified in the ontology graph. The output of this phase is whether these diseases are found in the ontology or not. If the disease has been identified then, the Uniform Resource Identifier (URI) of both diseases is retrieved. In the Disease relatedness detection phase, the URI of each disease is passed to the relatedness detector to find whether they are related or not. If they are related, the algorithm returns the set of diseases that connect them in the path from the first to the second and the classification of both diseases as well.

4.1 Semantic Matching

One of the most common approaches to perform semantic matching for determining the semantic similarity between concepts in an ontology is the node-based approach [3] which we used in this work. Semantic matching technique is used to identify candidate diseases in the disease ontology. Concept disambiguation is performed by querying WordNet [11] and DO. Each disease name in the query is first disambiguated into concepts using vector space models [3],

representing concepts as vectors of features in a k -dimensional space where k is the number of pertinent keywords for each disease. In other words, each disease in the query is represented by a vector of pertinent keywords found through WordNet and DO. Pertinent keywords are *hyponyms*, *direct-hypernyms*, *co-hyponyms*, and *synonyms* of the disease. One-level Hyponyms and direct-hypernyms are retrieved from DO and synonyms are retrieved from both DO and WordNet. Direct-hypernyms are one-level up of a disease node in DO hierarchy and Hyponyms are its siblings. After sense disambiguation, the proposed matching algorithm returns the URIs of matched diseases. It is assumed that, when searching for a concept, it is also important to match synonyms of that concept. For example, the synonyms of the disease *Hyperuricemia* are *Hyperuricaemia* and *Uricacidemia*. Therefore, diseases describing these concepts are retrieved as well. Figure 1 shows the semantic matching process cycle.

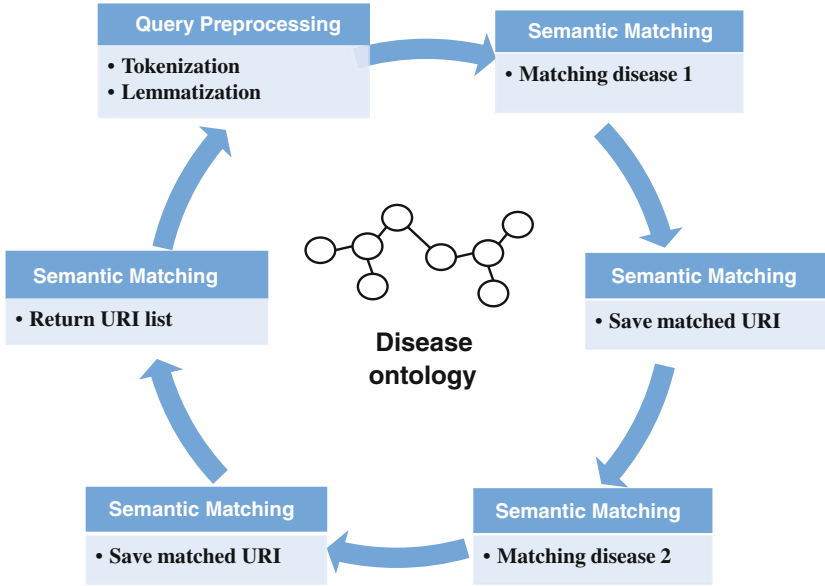


Fig. 1. The semantic matching process cycle.

4.2 Disease Relatedness Detection

In disease relatedness detection phase, diseases relatedness is detected by using bidirectional spreading activation on ontology graph which consists of a set of finite cycles/iterations. Checking for termination conditions is performed in each cycle. We are considering only hierarchical relations for nodes activation. The worthiness of bidirectional search is the speed and it requires less memory [19]. Figure 2 depicts the relatedness detector components. The algorithm starts with two initial nodes which are the two diseases and the follows the following steps:

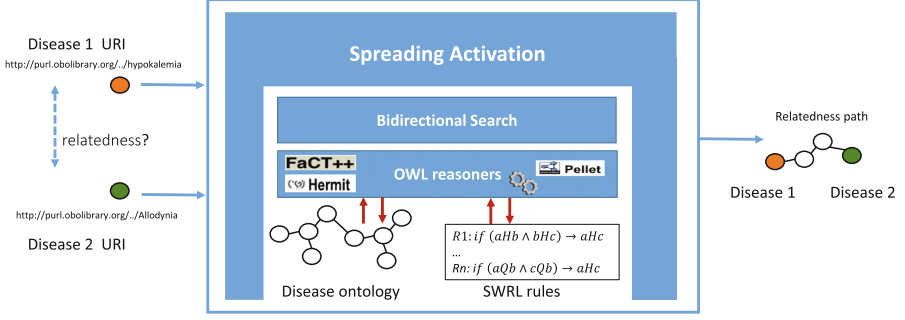


Fig. 2. Relatedness detector components.

1. Assume all nodes in the graph have activation values of zero and starting nodes have activation value of 1.0,
2. For each Link l_{ij} connecting the source node n_i with target node n_j , compute $a_j = \sum_{i=1}^n a_i w_{i,j}$ where n is the number of nodes connected to n_j and $w_{i,j}$ is the relatedness weight,
3. If a candidate node takes an activation value exceeds 1.0, then set its new activation value is set to 1.0. Likewise, set the activation value of the candidate node to 0.0 if it takes a value below 0.0. Nodes in the disease ontology receive the highest value of $w_{i,j}$ if they are one of the pertinent keywords of the current activated node ($w_{i,j} = 0.8$ for hyponyms and $w_{i,j} = 0.1$ for direct-hypernyms) because they build the taxonomy using “is-a” relation. A low value of $w_{i,j} = 0.0$ is assigned to any other relation which is not effective in this case,
4. A significance threshold ($F = 0.8$) determines whether to include the activated node to the output list,
5. Activated node will not be considered in the next cycles,
6. Nodes with activation value exceeds the threshold F are marked as activated on the next cycle,
7. The procedure terminates when a node is reached from more than one path (sexual disorder in the running example shown in Fig. 3).

An ambiguity may arise if two homonyms are used in the search query but there is no relatedness between them. For example, *Stewart* in *Stewart-Treves syndrome*, (a chronic lymphedema disease), is different from the one in *Stewart-Bluefarb syndrome*. The latter is a type of *acro angiodermatitis* which was described independently by Stewart as well as by Bluefarb and Adams on the legs of patients with Arterio-venous malformations [1]. Semantically matched diseases could be found using Jena reasoners. One feature of Jena is the support of different reasoners, which infer additional knowledge. Jena inference subsystem contains various inference engines or reasoners. These reasoners are used to check ontology consistency and allow additional facts to be inferred from instance data and class descriptions. The predefined reasoners included in the Jena distribution

are [15]: *Transitive reasoner*, *RDFS rule reasoner*, and *Generic rule reasoner*. The idea of using bidirectional search is to cut back the search time by looking out forward from the beginning and backward from the goal at the same time. When the two search frontiers meet, the algorithm will reconstruct one path that extends from the beginning state through the frontier intersection to the goal.

5 Workflow of the Proposed Method

In this section, we will illustrate how the proposed method is used over the underlying disease ontology. User submitted query is automatically processed in the following steps:

1. User submits the two diseases as a string using the system interface.
2. Perform pre-processing on input, a detailed discussion of pre-processing falls outside the scope of this paper:
 - (a) *Intelligent Tokenization and Stop list elimination*: This step includes tokenizing the query stream by breaking it down into understandable segments. Then, eliminate all stop words. The intelligence here is that the method does not blindly remove all stop words like traditional techniques but using a concept tokenization technique in which the keyword is taken with the preposition after it as a concept. If that concept has a match in the ontology, then this proposition will not be removed otherwise it will be removed,
 - (b) *Stemming*: Stemming removes word suffixes: both inflectional suffixes (-s, -es, -ed) and derivational suffixes (-able, -ability) are stemmed [9].
3. Formulate semantic query using SPARQL query language and executes it using Jena-embedded query engine (ARQ) against the ontology which performs the semantic matching process described in the methodology section,
4. Diseases relatedness is detected by using bidirectional-based spreading activation on ontology graph if the diseases found in the matching process. The output of this process is a set of diseases that builds the path between the two diseases. In addition, the classification of these diseases could be detected by performing one more cycle of the spreading process which could be the parent disease (if available) in the hierarchy.

As soon as these steps have been carried out, a set of diseases that may connect the two initially submitted diseases are displayed to the user and the classification of them is displayed as well.

6 Running Example

As a running example in this paper, suppose a patient already has *Vasculogenic impotence* disease and the physician discovered that he/she got recently *Transvestism* so the question is does these two diseases is related to each other by any means? The answer to this question will support the medical decision of the

physician for treatment prescription and whether there are any complications caused by one because of the other. In this case, the physician should submit a query includes two diseases: *Vasculogenic impotence* and *Transvestism*. As shown in Fig. 3, the relatedness between *Vasculogenic impotence* and *Transvestism* diseases are detected by the intersection node *sexual disorder*. The algorithm runs in a set of cycles. When the algorithm detects an intersection node, it stops. One more cycle can be performed to get one more up-level disease connecting both which could be considered as a classification of both diseases.

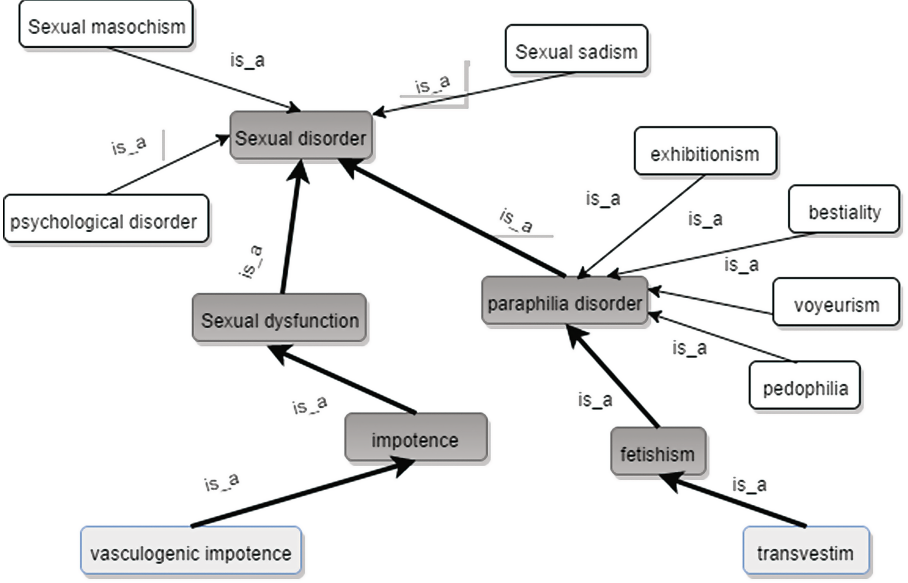


Fig. 3. Partial view of the disease ontology shows the relatedness path (marked in bold) between *Vasculogenic Impotence* and *Transvestism* diseases.

7 Conclusion and Future Work

In this paper, we proposed a Bidirectional-Based Spreading Activation Method in order to detect the relatedness between two human diseases. This relatedness can be detected by running spreading activation algorithm using bidirectional search methodology on a large disease ontology. One key feature of the proposed method is to identify whether two human diseases are related to each other. Moreover, identify related diseases that may connect them to a common path. As a result, detecting the relatedness between diseases helps in finding out the real cause of a disease (in case it is caused by another disease), decrease the prognosis of a disease by treating the other disease which increases the complications of that disease and helps in treatments prescriptions. Therefore, physicians can treat the main cause, not just the symptoms. Consequently, we believe that

the proposed method will assist physicians and will support medical decision-making by considering semantic domain knowledge to infer diseases relatedness. This idea could also be applied to VO ontology to find out relations between vaccines and also in Gene ontology to find out relations between genes. Our main line of future research involves:

- Extending our approach to integrate different biomedical ontologies using and ontology matching and integration services,
- Detecting the relatedness between more than two diseases using multiple goal search algorithms and provide the relatedness graph between them not only a simple path,
- A relatedness set is provided which is a subset of the power set of the set of input diseases. In other words, which of these diseases is related to the others?
- Finally, a bilingual bidirectional-based spreading activation method will be proposed and implemented.

References

1. Al Aboud, A., Al Aboud, K.: Similar names and terms in dermatology; an appraisal. *Our Dermatol Online* **3**, 367–368 (2012)
2. Anderson, J.R.: A spreading activation theory of memory. *J. Verbal Learn. Verbal Behav.* **22**(3), 261–295 (1983)
3. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? towards personalized similarity measures in ontologies. In: Ferstl, O.K., Sinz, E.J., Eckert, S., Isselhorst, T. (eds.) *Wirtschaftsinformatik 2005*, pp. 1347–1366. Physica-Verlag HD, Heidelberg (2005). doi:[10.1007/3-7908-1624-8_71](https://doi.org/10.1007/3-7908-1624-8_71)
4. Croft, D., et al.: The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**(D1), D472–D477 (2014)
5. De Maio, C., et al.: Fuzzy knowledge approach to automatic disease diagnosis. In: 2011 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 2088–2095. IEEE (2011)
6. Dhamankar, R., et al.: iMAP: discovering complex semantic matches between database schemas. In: *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pp. 383–394. ACM (2004)
7. Do, H.-H., Rahm, E.: COMA: a system for flexible combination of schema matching approaches. In: *Proceedings of the 28th International Conference on Very Large Data Bases*, pp. 610–621. VLDB Endowment (2002)
8. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: six years of experience. In: Spaccapietra, S., et al. (eds.) *Journal on Data Semantics XV*. LNCS, vol. 6720, pp. 158–192. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-22630-4_6](https://doi.org/10.1007/978-3-642-22630-4_6)
9. Fan, Y., Huang, X., An, A.: York university at TREC 2006: enterprise email discussion search. In: *TREC 2006* (2006)
10. Fathalla, S.M., Hassan, Y.F., El-Sayed, M.: A hybrid method for user query reformation and classification. In: *2012 22nd International Conference on Computer Theory and Applications (ICCTA)*, pp. 132–138. IEEE (2012)
11. Fellbaum, C.: *WordNet*. Wiley Online Library, New York (1998)

12. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: algorithms and implementation. In: Spaccapietra, S., et al. (eds.) *Journal on Data Semantics IX*. LNCS, vol. 4601, pp. 1–38. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-74987-5_1](https://doi.org/10.1007/978-3-540-74987-5_1)
13. Guo, J., et al.: Semantic matching by non-linear word transportation for information retrieval. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 701–710. ACM (2016)
14. Hoehndorf, R., Dumontier, M., Gkoutos, G.V.: Evaluation of research in biomedical ontologies. *Brief. Bioinform.* **14**(6), 696–712 (2013)
15. Jena, A.: Reasoners and rule engines: jena inference support. The Apache Software Foundation (2013)
16. Kibbe, W.A., et al.: Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**(D1), D1071–D1078 (2015)
17. Köhler, S., et al.: The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**(D1), D966–D974 (2014)
18. LePendur, P., Musen, M.A., Shah, N.H.: Enabling enrichment analysis with the human disease ontology. *J. Biomed. Inform.* **44**, S31–S38 (2011)
19. Li, H., Xu, J., et al.: Semantic matching in search. *Found. Trends R Inf. Retrieval.* **7**(5), 343–469 (2014)
20. Ngo, V.M., Cao, T.H., Le, T.M.: Combining named entities with wordnet and using query-oriented spreading activation for semantic text search. In: *2010 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pp. 1–6. IEEE (2010)
21. Qin, Y., Yao, L., Sheng, Q.Z.: Approximate semantic matching over linked data streams. In: Hartmann, S., Ma, H. (eds.) *DEXA 2016*. LNCS, vol. 9828, pp. 37–51. Springer, Cham (2016). doi:[10.1007/978-3-319-44406-2_5](https://doi.org/10.1007/978-3-319-44406-2_5)
22. Schriml, L.M., et al.: Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* **40**(D1), D940–D946 (2012)
23. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. *IEEE Trans. Knowl. Data Eng.* **25**(1), 158–176 (2013)
24. Trotman, B.W., et al.: Studies on the pathogenesis of pigment gallstones in hemolytic anemia: description and characteristics of a mouse model. *J. Clin. Invest.* **65**(6), 1301 (1980)
25. Wu, Z., et al.: An efficient Wikipedia semantic matching approach to text document classification. *Inf. Sci.* **393**, 15–28 (2017)

Computational Collective Intelligence

9th International Conference, ICCCI 2017, Nicosia,

Cyprus, September 27-29, 2017, Proceedings, Part I

Nguyen, N.T.; Papadopoulos, G.A.; Jędrzejowicz, P.;

Trawiński, B.; Vossen, G. (Eds.)

2017, XXVIII, 592 p. 168 illus., Softcover

ISBN: 978-3-319-67073-7