A newborn development insights mining and recommendation system from scientific literature and clinical guidelines*

Sergio Consoli[1,2], Kees Wouters[1], Renée Otte[1], and Adrienne Heinrich[1]

[1] Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands.
[2] European Commission, Joint Research Centre, Directorate A-Strategy, Work Programme and Resources, Scientific Development Unit, Via E. Fermi 2749, I-21027 Ispra (VA), Italy. sergio.consoli@ec.europa.eu

Abstract. In this short contribution we describe a method that is able to automatically retrieve relevant newborn development content from scientific documents, including scientific papers and standard guidelines, and then to recommend parents automatically the most relevant personal advices, also known as insights, from the extracted knowledge. The approach cannot replace specialist advice but it rather provides quick information from reliable sources with a certain degree of specificity for the parents and the child. The system builds on recent technological developments on big data, knowledge engineering, and cognitive computing, in particular related to the task of extracting relations between conceptual entities in the data sources.

Keywords: Generation and aggregation of health semantics; Ontologies; Recommendations for health data; Information Extraction; Insights.

1 Introduction

The availability of abundant computing and storage resources combined with the evolution of analytics has made affordable the use of cognitive computing technology to deliver industrial solutions of all kinds [11]. Cognitive computing systems depend on various aspects of artificial intelligence (AI), such as machine learning, reasoning, natural language processing, speech and vision, human-computer interaction, dialogue and narrative generation, and more. The machine learning algorithms learn and acquire knowledge from the massive amount of data fed into to them [10, 11].

Nowadays there is a lot of interest in adopting cognitive computing technologies in healthcare[3], which is particularly characterized by a vast amount of data coming from different sources [7, 4]. Through the application of natural

[3] https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/cognitive-technology-for-health-care.html

language processing (NLP), data mining, and advanced text analytics, cognitive systems can assist doctors in diagnosing and faster decision making [19, 4]. They optimize patient selection for clinical trials through intelligent matching. In oncology, these systems can assist in the creation of individualized treatment plans that enhance patient trust and experience [7, 19, 4].

The idea is that a machine can process more information than a doctor and potentially discover links and patterns not immediately visible at a first glance or that would require a complete overview of all possible interventions. An example of this is Watson[4], the popular question-answering (Q&A) machine by IBM, which has been recently employed to provide diagnosis and treatments to cancer patients, enabling faster and better care for patients[5] [10]. It can analyse the meaning and context of structured and unstructured data coming from a variety of inputs including handwritten documents [6, 8], and derive data from various sources including curated literature and rationales, as well as medical journals and textbooks[6].

Currently, there is an increasing number of new data-driven solutions in the market which provide pregnant women and new parents suggestions and personal advices, also known as insights, which address their needs and wishes, on the basis of behavioural and contextual data, scientific literature and clinical guidelines. It is important to underline that insights cannot replace specialist advice but they rather provide quick information from reliable sources with a certain degree of specificity for the parents and the child. Mor precisely, insights refer to small pieces of text that are suggested to parents on the basis of a technical rule [16]. This rule analyses parents-tracked data and the available scientific literature and guidelines, and defines when an insight text is presented to the user. For example, if a mother is keeping track of her newborn's breastfeeds, an insight could be the following: *"Recently, your tracked breast feeds with Sara have taken around 14 minutes. In general, newborns feed for 10-30 minutes at a time – occasionally even longer"* [16].
In this way, insights provide personal advice, tips, and information tailored to the unique situation of the parent and the newborn [16].

Currently these insights and articles are selected manually by curators, who need to grasp and exploit a large number of scientific documents and select the most relevant content from that to be selected as candidate insights or articles to users. However, manual generation of insights takes time and a lot of effort, because all the scientific content needs to be read and the most relevant insights or articles extracted. In addition, an optimal selection is hard to be manually established by a human who can only rely on his intuition, bringing in most cases non-optimal decisions. In addition, linking all the relationships among the

[4] https://www.ibm.com/watson/
[5] http://pulse.embs.org/may-2017/cognitive-computing-and-the-future-of-health-care/
[6] https://mihin.org/wp-content/uploads/2015/06/The-Impact-of-Cognitive-Computing-on-Healthcare-Final-Version-for-Handout.pdf

pregnancy or newborn concepts reported in the scientific documents still remains a challenge.

The aim of the system proposed in this contribution is to support the process of insights or article generation by recommending automatically a set of the most important facts coming from the scientific literature and clinical guidelines given in input. The system leverages on state-of-art technologies on Cognitive Computing, Natural Language Processing (NLP), Ontology Engineering, and Big-Data, producing an advanced AI system for semantically mining information from a scientific pregnancy and newborn development domains repository.

2 Description of the method

The algorithm leverages the recent AI developments in cognitive computing and applies them to automatically mining information from scientific literature and guidelines. In this way, new knowledge on the pregnancy or newborn development domains can be extracted, automatically providing applications with a recommendation of the most relevant insights or articles to give to users in a personalized way. The schematic workflow of the system is depicted in Figure 1.

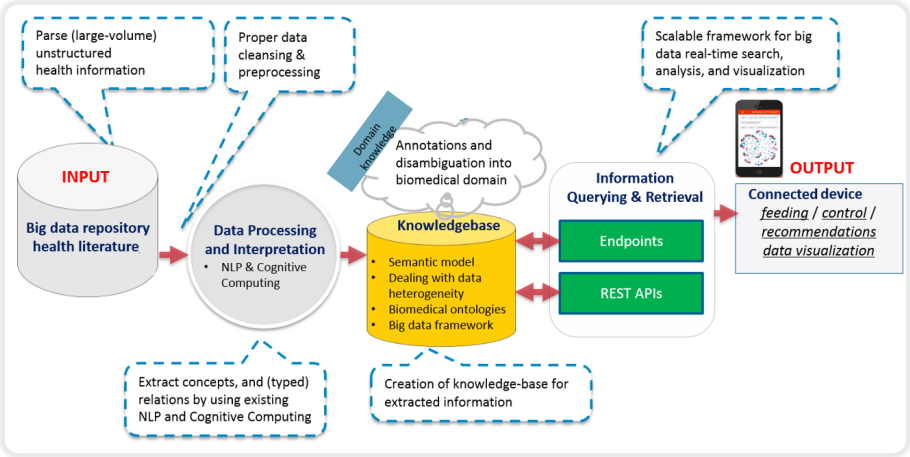

Fig. 1: Pipeline of the system illustrating its main elements.

The system is able to parse, extract, transform and load the unstructured information coming from clinical guidelines and scientific papers. It is able then to structure the free-format text using machine reading from natural language processing for extracting RDF[7]/OWL[8] graphs that are linked to the Linked Open

[7] RDF: Resource Description Framework
[8] OWL: Web Ontology Language

Data cloud and compliant to Semantic Web and Linked Data patterns [14]. In this way the information is translated into machine-readable semantic information in RDF/OWL format, which is a W3C standard for exchanging semantic data. Machine reading is typically much less accurate than human reading, but can process massive amounts of text in reasonable time, can detect regularities hardly noticeable by humans, and its results can be reused by machines for applied tasks [13].

The system recognizes and resolves named entities, links them to the existing knowledge base, and gives them a type by using different cognitive computing functionalities [1, 13, 14, 15, 18]: frame detection, topic extraction, named entity recognition, resolution and co-reference, terminology extraction, sense tagging, word-sense disambiguation, taxonomy induction, semantic-role labelling, and type induction.

In this way the algorithm recognizes the main entities and concepts, and most importantly, it performs relations extraction to derive the main relationships among them [17]. The main focus indeed is to extract the relationships among the obtained conceptual entities and link them together to allow interoperability among the information contained in the scientific documents.

Using named entity recognition (NER) and resolution (a.k.a. entity linking) with standard biomedical ontologies in the pregnancy or newborn development domains, the algorithm makes sure to restrict the extraction of the structured information within the specific pregnancy or newborn development domains. Biomedical ontologies ensure both syntactic and semantic interoperability among all heterogeneous data coming into the system [9].

The extracted big data leverage the pregnancy and newborn development knowledge-base, which is updated periodically by the system with new, updated information coming from its input sources. Each concept in the system is linked to other concepts in the knowledge-base by using ontologies and allowing interoperability. The stored relationships may include, for example, sensitive user status/conditions (e.g. pregnant, parent, infant, etc), and diseases (asthma, allergies, etc.) linked to other conditions and information. The knowledge-base is constantly maintained, updated, and integrated in the ontology model.

The knowledge-base contains the semantic model with the updated information coming from the scientific literature and clinical guidelines. In order to provide recommendation of the most relevant insights, a ranking is produced as following:

1. Consider each sentence with all its extracted relationships. Sum the absolute frequencies scores among all documents associated to the relationships. The result will be a score associated to the sentence.
2. Rank the extracted sentences with respect to the scores and identify the insight(s) having the largest score.
3. Put this sentence(s) set in the list of the insights to recommend.
4. Iteratively exclude from the remaining sentences the relationships that were already considered in the previous insights selection, and recalculate accordingly the scores of the sentences.

5. Re-order the sentence with the respect to the re-calculated scores, and identify the insight(s) having the largest re-calculated score.
6. Go to Step 3, and continue until no further sentences are remaining.
7. The output will be the final list of the insights to recommend.

The output of the system is a document with the list of top recommended insights that needs to be checked and validated by a curator, and successively the next step would be to automatically provide those top insights directly to end users. The recursive ranking re-calculation is aimed at increasing the diversification of the top insights that are finally recommended.

Based on a specific user profile the system may be also able to personalize the insights that are recommended[9]. In this way the algorithm could provide relevant insights to a pregnant woman or parent, linking validated information on pregnancy and newborn development to user-specific conditions. By storing the specific insights in the system for later usage, the method would be able to re-adopt this information by, and share with, different products of the user. The standardized RDF/OWL format of the produced information in the knowledge-base guarantees standard communication among the different products overcoming any incompatibility and interoperability issues.

Summarizing, the described system is able to identify novel insights that go beyond clinical guidelines, and provide relationships which can help parents and pregnant women.

3 A controlled experiment

In order to provide a business scenario as an example giving a detailed description of the system, a controlled prototype experiment of relations extraction in the pregnancy and newborn development domains from a small set of scientific paper and clinical guidelines as input (18 in total) has been carried out. In this experiment the system produced 38878 relationships among the extracted concepts from the 18 documents in the input repository.

Table 1 shows some examples of the extracted relationships among some concepts and their frequencies. Each entry in the subject and object columns represents the label associated with the related concept from a standard biomedical ontology. For example, "Child" is the label of the concept `http://purl.bioontology.org/ontology/HL7/C0008059` belonging to the *Health Level 7 (HL7)* ontology. Similarly, "Asthma" is the label of the concept `http://purl.bioontology.org/ontology/MESH/D001249` from the *Medical Subject Headings (MeSH)* ontology.

Each concept in the ontology is uniquely identified by its corresponding URI, which conveys other undirected information coming from the ontology, relations to other concepts and its position in the hierarchy, and most important, it allows disambiguation among terms and linking concepts together to ensure syntactic as well semantic interoperability.

[9] Functionality not yet implemented, under current development.

Table 1: Example of extracted relationships

subject	*relation*	*object*	*frequency*
"Infant"	"hasAttribute"	"Infection"	8
"Infant"	"have"	"Breastfeeding"	6
"Child"	"hasAttribute"	"Fever"	5
"Child"	"hasAttribute"	"Asthma"	5
"Physician"	"cure"	"Allergy"	5
"Night"	"related"	"Sleep"	4
"Milk"	"related"	"Metabolism"	4
"Parent"	"encourage"	"Infant"	4
"Individual"	"do"	"Feed"	3
"Mother"	"express"	"Milk"	3
"Breastfeeding"	"protect"	"Disease"	2
"Immunoglobulin E"	"protect"	"Asthma"	2
...	...	...	...

Figure 2 shows some relationships in the prototype example among the "Human milk" concept (http://purl.jp/bio/4/id/200906042293515949) and some of the other concepts extracted from the input data sources related to the pregnancy and newborn development domains of interest.

As an example of the process of extraction of relationships from the input source, consider the following sentence: *"Colostrum contains low concentrations of both lactose and fat in comparison to mature milk"*. The concepts detailed in Table 2 are extracted and related each other.

Table 2: Example of concepts and relations extracted for the sentence: *"Colostrum contains low concentrations of both lactose and fat in comparison to mature milk"*.

subject	*relation*	*object*
"Colostrum"	"contain"	"Decreased concentration"
"Colostrum"	"contain"	"Lactose"
"Decreased concentration"	"related"	"Lactose"
...	...	...

In particular, the reported concepts stand for:

- "Colostrum" is the concept: http://purl.obolibrary.org/obo/UBERON_0001914 coming from the *Uber-anatomy ontology (UBERON)* ontology;
- "Lactose" is the concept: http://purl.bioontology.org/ontology/HL7/C1696723 coming from the *Health Level 7 (HL7)* ontology;
- "Decreased concentration" is the concept: http://purl.obolibrary.org/obo/PATO_0001163 coming from the *Phenotype And Trait Ontology (PATO)* ontology;

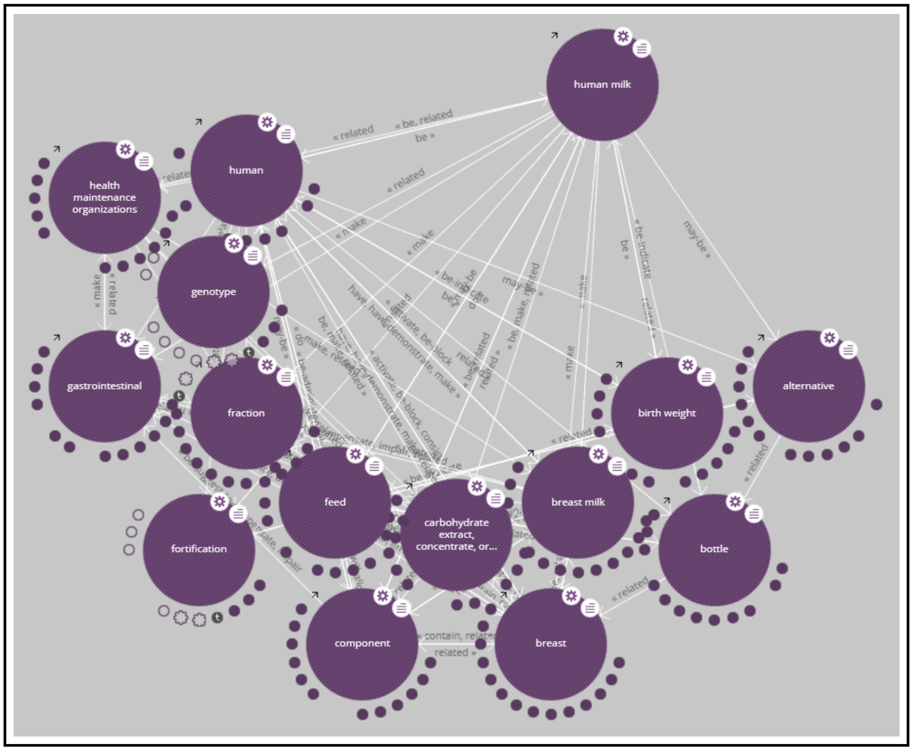

Fig. 2: View of the relationships among the "Human milk" concept (http://purl.jp/bio/4/id/200906042293515949) and other concepts extracted from the input data sources related to the pregnancy domain.

The relationships can be also explored by interactive chord diagrams for visualizations [12, 2]. A chord diagram is a graphical method of displaying the inter-relationships between data in a matrix. The data is arranged radially around a circle with the relationships between the points typically drawn as arcs connecting the data together. When a specific concept is selected interactively, only its relationships are visualized in the diagram, helping users to grasp and understand more intuitively the inter-relationships among the different entities. The format of chord diagrams is aesthetically pleasing, making it a popular choice in the world of data visualization [12, 3, 5]. For example, Figure 3 shows the relationships of the extracted concept "Breast", i.e. concept http://purl.obolibrary.org/obo/UBERON_0000310 from the *UBERON* ontology with the other concepts.

In the controlled prototype experiment with the 18 scientific paper and clinical guidelines as input, a total of 2573 different sentences were extracted. Figure 4 shows a chart with the ranked sentences extracted from the input pregnancy and newborn development repository for the controlled experiment.

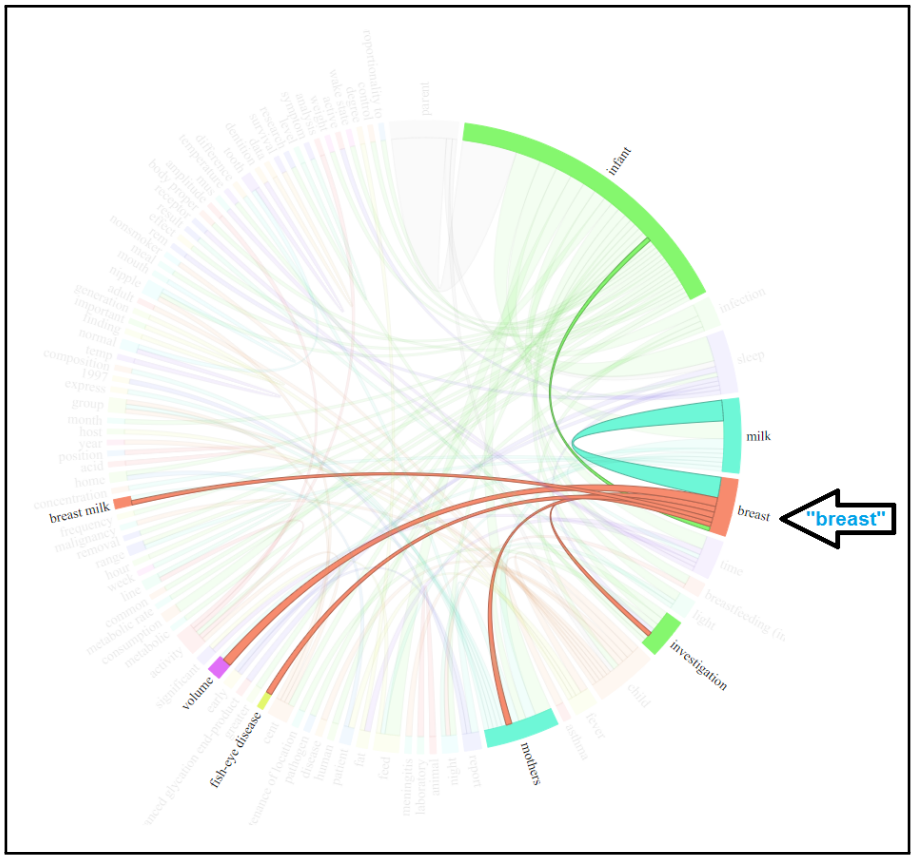

Fig. 3: Chord visualization of the relationships of the "Breast" (http://purl.obolibrary.org/obo/UBERON_0000310) concept with the other extracted concepts.

If in Figure 4 we consider a threshold to cut the tail of the curve, so that to have 95% of AUC for the given chart (corresponding to a minimum scoring threshold equal to 10), we can reduce the total number of insights that are recommended to 1523. If we aim for a more selective selection, if we cut at a score of 100, we have only 53 most relevant insights to look at. For example, the following are the top ones:

"Demonstrating the bioactivity of breast milk, a study on shed epithelial cells in the faeces of infants has shown that gene expression in the neonatal gastrointestinal tract is influenced by breastfeeding, with differential expression found between formula fed and breast fed infants in genes regulating intestinal cell proliferation, differentiation and barrier function."

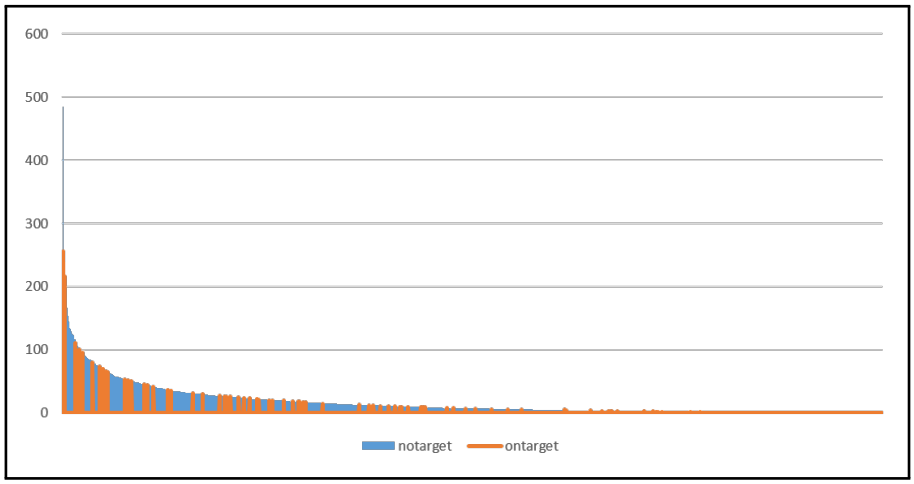

Fig. 4: Chart diagram showing the ranked sentences extracted from the input pregnancy repository for the controlled experiment.

"Protein concentration is highest in breast milk of mothers aged 20-30, however, maternal age does not seem to influence either lipid or lactose concentrations, and maternal age does not have a large impact on breast milk composition."

"Infants and young children have a higher resting metabolic rate and rate of oxygen consumption per unit body weight than adults because they have a larger surface area per unit body weight and because they are growing rapidly."

"The total protein content of human breast milk consists of 13% casein, the lowest casein concentration of any studied species, corresponding to the slow growth rate of human infants."

"Exposure to tobacco smoke in utero was associated with an increased risk of stillbirth (odds ratio = 2.0, 95% confidence interval: 1.4, 2.9), and infant mortality was almost doubled in children born to women who had smoked during pregnancy compared with children of nonsmokers (odds ratio = 1.8, 95% confidence interval: 1.3, 2.6)."

"Human milk oligosaccharides (HMO) also make up a significant fraction of breast milk carbohydrate, but are indigestible by the infant, their function instead is to nourish the gastrointestinal microbiota."

A preferred, ideal deployment of the system, not yet implemented but at a prototype stage, is shown in Figure 5, schematically depicting a recommendation device (e.g. a smartphone) comprising at least one (or more) communication unit(s) and a user interface, and a processing unit embedded in a remote server controlling the suggestions of the most relevant insights to the device. The processing unit comprises a cognitive system, of which the pipeline is described in Figure 5, and is connected to the recommendation device via the communication

unit. In various embodiments, the cognitive system in the processing unit maybe connected to different input data sources, including, but not limited to, a repository of scientific papers and clinical guidelines. The output information interface could be any device able to provide useful insights to a pregnant woman or a parent, for example a smartphone hosting an appropriate application. It might comprise a user interface for data input/output and a data display. Through a user interface, the user might input his profile details, e.g. sex, age, eventual health diseases (like allergies, asthma, etc.), and others, and then save these details for later usage. The recommendation device sends the user details to the processing unit via the communication unit.

Alternatively, the system may include also some automatic trackers, i.e. either devices embedded into the main system able to track and store automatically users' activities, or also special tracking devices that are external to the main system but directly linked to it.

The remote server is connected to the knowledge-base, which is a semantic triplestore containing the semantic information in the W3C standard format RDF/OWL, as described previously, and enabling semantic interoperability, reasoning and inferencing, containing the relationships among conceptual pregnancy and newborn development actors involved in specific situations, including health conditions, sensitivity information (manually set or derived by the system).

The cognitive system in the processing unit receives the user profile details and combines this information with the information in the knowledge-base. The system uses this knowledge to finally provide personalized insights to the recommendation device which is then displayed to the user.

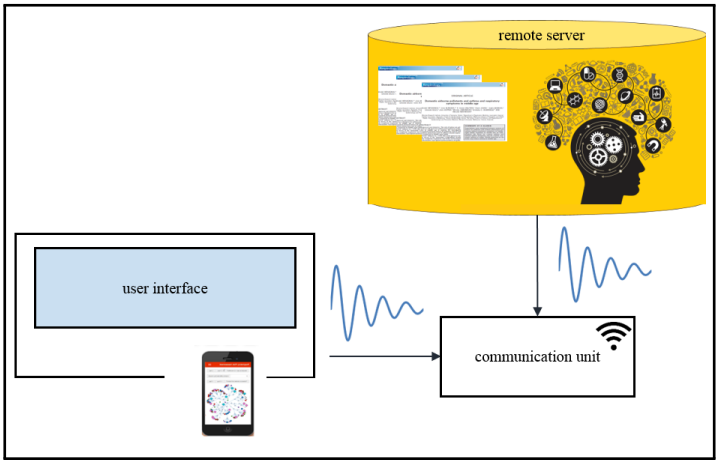

Fig. 5: Schematic illustration showing the ideal deployment of the system.

4 Conclusion

In this short contribution it has been described a method which is able to automatically retrieve relevant clinical content on newborn development from scientific papers and standard guidelines. The system may be used to feed in real-time a connected application to provide feedbacks and suggestion of useful insights to pregnant women and new parents derived from scientific literature and clinical guidelines. This method has the potential to be used in real pregnancy or newborn development recommendation systems. In addition, the method may be generalizable and applicable to other domains after choosing the relevant ontologies and information sources.

References

[1] S. Consoli and D. Reforgiato Recupero. Using FRED for named entity resolution, linking and typing for knowledge base population. In Gandon F., Cabrio E., Stankovic M., and Zimmermann A., editors, *Communications in Computer and Information Science*, volume 548, pages 40–50. Springer-Verlag, New York, 2015.

[2] S. Consoli and N.I. Stilianakis. A quartet method based on variable neighborhood search for biomedical literature extraction and clustering. *International Transactions in Operational Research*, 24(3):537–558, 2017.

[3] S. Consoli, K. Darby-Dowman, G. Geleijnse, J. Korst, and S. Pauws. Heuristic approaches for the quartet method of hierarchical clustering. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1428–1443, 2010.

[4] S. Consoli, D. Reforgiato Recupero, and M. Petkovic, editors. *Data Science for Healthcare: Methodologies and Applications*. Springer Nature, 2019.

[5] S. Consoli, J. Korst, S. Pauws, and G. Geleijnse. Improved metaheuristics for the quartet method of hierarchical clustering. *Journal of Global Optimization*, 78(2):241–270, 2020.

[6] D. Dessì, G. Fenu, D. Reforgiato Recupero, and S. Consoli. Exploration of IBM Watson for healthcare applications. Technical report, Philips Research Europe Technical Note, PR-TN 2017/00115, Eindhoven The Netherlands, 2017.

[7] D. Dessì, D. Reforgiato Recupero, G. Fenu, and S. Consoli. Exploiting cognitive computing and frame semantic features for biomedical document clustering. In *CEUR Workshop Proceedings*, volume 1948, pages 20–34, 2017.

[8] D. Dessì, D. Reforgiato Recupero, G. Fenu, and S. Consoli. A recommender system of medical reports leveraging cognitive computing and frame semantics. *Intelligent Systems Reference Library*, 149:7–30, 2019.

[9] A. Gangemi. Ontology design patterns for Semantic Web content. In *Lecture Notes in Computer Science*, volume 3729, pages 262–276, 2005.

[10] R.E. Gantenbein. Watson, come here! The role of intelligent systems in health care. In *World Automation Congress Proceedings*, art num 6935748, pages 165–168, 2014.

[11] J.O. Gutierrez-Garcia and E. López-Neri. Cognitive computing: A brief survey and open research challenges. In *3rd International Conference on Applied Computing and Information Technology and 2nd International Conference on Computational Science and Intelligence (ACIT-CSI)*, art num 7336083, pages 328–333, 2015.

[12] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[13] M. Mongiovì, D. Reforgiato Recupero, A. Gangemi, V. Presutti, A.G. Nuzzolese, and S. Consoli. Semantic reconciliation of knowledge extracted from text through a novel machine reader. In *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015*, 2015.

[14] M. Mongiovì, D. Reforgiato Recupero, A. Gangemi, V. Presutti, and S. Consoli. Merging open knowledge extracted from text with MERGILO. *Knowledge-Based Systems*, 108:155–167, 2016.

[15] A.G. Nuzzolese, A. Gangemi, and V. Presutti. Gathering lexical linked data and knowledge patterns from framenet. In *KCAP 2011 - Proceedings of the 2011 Knowledge Capture Conference*, pages 41–48, 2011.

[16] R.A. Otte, A.J.E. van Beukering, and L.-M. Boelens-Brockhuis. Tracker-based personal advice to support the baby's healthy development in a novel parenting app: Data-driven innovation. *JMIR mHealth and uHealth*, 7(7, art num e12666), 2019.

[17] V. Presutti, A.G. Nuzzolese, S. Consoli, A. Gangemi, and D. Reforgiato Recupero. From hyperlinks to Semantic Web properties using Open Knowledge Extraction. *Semantic Web*, 7(4):351–378, 2016.

[18] D. Reforgiato Recupero, A.G. Nuzzolese, S. Consoli, V. Presutti, S. Peroni, and M. Mongiovì. Extracting knowledge from text using SHELDON, a semantic holistic framEwork for LinkeD ONtology data. In *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, pages 235–238, 2015.

[19] M. Van Hartskamp, S. Consoli, W. Verhaegh, M. Petkovic, and A. Van De Stolpe. Artificial intelligence in clinical health care applications: Viewpoint. *Journal of Medical Internet Research*, 21(4), 2019.