Technology Model to Support the Initiation of Innovation Artefacts

Maria-Iuliana Dascalu[1(✉)], Elisabeth Lazarou[1], and Victor Florin Constantin[2]

[1] Department of Engineering in Foreign Languages, University POLITEHNICA of Bucharest, Bucharest, Romania
{maria.dascalu, elisabeth.lazarou}@upb.ro
[2] Department of Mechatronics and Precision Mechanics, University POLITEHNICA of Bucharest, Bucharest, Romania
victor.constantin@upb.ro

Abstract. The current paper proposes a technology model to support the process of creating innovative artefacts, where artefact is any project proposal, business plan, business solution, article with a high degree of innovation. The model is based on an advanced technology stack, in which the central role is played by semantic high-performance computing. Several functionalities are available both for academic researchers and business consultants, from validating the innovation degree of an idea, to supporting its development with useful bibliographical recommendations or building research proposals based on that idea.

Keywords: Innovative artefact · Semantic search · High-performance computing · Technological model

1 Introduction

The current paper proposes a technology model – InnovRes, which can be used to implement a support product for shaping and validating innovative ideas by identifying and recommending valuable bibliographic resources (articles, patents, project descriptions, websites, etc.), and partially automating the writing process of research or statup projects. The model is useful both for academia researchers and business consultants, as they are trying to enter in a new cutting-edge domain. The model is based on semantic and Big Data processing technologies, as well as on the use of Application Programming Interfaces (APIs) for access the various scientific warehouses, thus providing a very current technological stack. The InnovRes model respects the digitization trend of the European Union: digital and large data platforms are becoming more widespread, impacting almost all industries [1]. Large data volumes generated by equipment, machines and people bring special opportunities for innovation, new business models, smart products and services, leading to industrial progress and adding value to the European society and beyond. There are digital tools for researchers and consultants - collaborative writing and visualization tools, search engines [2], but no tool combines both directions (search for relevant resources to create an innovative

A. G. Kravets et al. (Eds.): CIT&DS 2019, CCIS 1083, pp. 278–287, 2019.
https://doi.org/10.1007/978-3-030-29743-5_22

artefact and the creation itself), where artefact is any project proposal, business plan, business solution, article with a high degree innovation. The current papers presents the functions of the model, the technologies necessary to implement it, as well as the state-of-the-art related to its development. From the scientific point of view, the model has two directions of innovation: the semantic computing (as a new computational model) and the high performance computing, the real challenge being the realization of the interoperability between the semantic technologies and the high performance computations, thus optimizing the processing, searching and recommending relevant resources for innovative artefacts. In the context of the model, the semantic technologies are the ones that make the Big Data processed by the high-performance computing to be truly smart and useful.

2 Semantic Technologies, Potentiator of High-Performance Computing for Innovation Seekers

There are several types of tools which can be used by researchers or innovation seekers [2]:

- dedicated search engines (BibSonomy, Biohunter, DeepDyve etc.);
- article visualization tools (ACS ChemWorx, Colwiz, eLife Lens etc.);
- research data sharing instruments (BioLINCC, Code Ocean, ContentMine etc.);
- virtual communities (AcademicJoy, Addgene, AssayDepot etc.);
- crowdfunding (Benefunder, Consano, Experiment etc.) and so on.

Of course, dedicated tools to search for patents are developed or still in development [3, 4], but no tool offers integrated services for checking the validity of an innovative idea, supporting it with relevant resources, initiating the process of writing it, as our model proposes. In order for all those services to be successfully interconnected, a proper mix of advanced technologies is necessary. The context of these necessary technologies is further described.

High performance computing means the use of supercomputers and parallel processing techniques to solve complex computational problems [1]. In the case of our model, high-performance computation is necessary for large data collections of articles, project descriptions, scientific web resources, etc., needed to generate new ideas and innovation artefacts. In order to select the most relevant data, semantic models have to be applied to high-performance computing.

In a search engine, a keyword-based search returns documents, taking into account the greatest number of matching words in the query with the text of the documents. Semantic search seeks to improve search precision by understanding the intent of the seeker and contextual meaning of terms as they appear in the search data space, either on the Web or in a closed system, to generate more relevant results [5]. Semantic search represents the capacity of a search engine to determine what the user thought of in the moment of query and also to offer the user results that do not fully match the words that are typed in, but are equivalent in meaning. For a better understanding of the concept of semantic search, it must be placed, comparatively, in the context of another concept: keyword search. Semantic search intelligently understands the meaning of the words

that are typed in, to be more exact, it focuses on the context. Semantic search provides accurate and relevant results based on the queries. This approach is ontology-based, which is faster due to association between contents. Keyword search engines come in handy when the meaning of terms is not necessarily needed and the results are displayed within a decent amount of time. Surveys indicate that there are a lot of people not receiving accurate results in the first set of URLs returned, due to the fact that several words have the same meaning and one word could mean several thing, therefore it might lead to confusion. A clear comparison between semantic and keyword search is available in Fig. 1.

Keyword Search Engine	Semantic Web Search Engine
1. It is a traditional search engines that produce results of given query within the given context.	1. It works on Semantic based approach which is useful for having accurate and relevant information about the given query.
2. The information which is retrieved is dependent on keywords and page ranking algorithms that can produce spam results.	2. The information retrieved is independent of keywords and page rank algorithms that produce exact results rather than any irrelevant results.
3. It does not focus on stop words like is, or, and, how because it does not give accurate results what user is searching to get information.	3. It focuses on stop words and punctuation marks because it takes into account each and every small character as it affects search results.
4. It displays all web pages that may or may not satisfy user's query and to select relevant page from many pages is difficult task.	4. It will show only those results that will answer our query.
5. It does not highlight any words or phrases which are useful in answering getting accurate results.	5. It highlights the sentences or words that give answer to query asked by the user.
6. It makes use of keywords to expand query instead of using any methodology.	6. It uses ontology to get relations between the keywords.
7. It uses HTML, XML language for creation of metadata.	7. It uses Semantic Web languages like OWL, RDF for creation of metadata.

Fig. 1. Comparison between keyword search and semantic web search

To implement semantic searches, unstructured text needs to be transformed into a structured, easy-to-process computer form. Such form is the ontology, which models concepts and relationships within a domain, allowing an application to make automatic inferences, similar to the human way of thinking [6]. Although there are semantic search applications – OSSSE [7] or IBRI-CASONTO [8], these were developed only in the laboratory without being tested in an industrial environment: semantic search is not completely explored, thus an emergent direction of research nowadays. There are many criteria that classify approaches of semantic search proposals [9]: architecture

(stand-alone search engine or meta-search engine), coupling (tight coupling, loose coupling), transparency (transparent, interactive, hybrid), user context (learning, hard-coded), query modification (manual, query rewriting or graph-based), ontology structure (anonymous properties, standard properties, domain specific properties) and ontology technology. Some examples of semantic search approaches are the following ones:

- Simple HTML Ontology Extensions (SHOE) is a form based semantic search engine;
- Inquirus2;
- TAP;
- Hybrid spreading activation;
- Intelligent Semantic Web Retrieval Agent (ISRA);
- Librarian agent;
- Semantic Content Organization and Retrieval Engine (SCORE);
- TRUST;
- Audio data retrieval;
- Ontogator.

There are also some approaches proposed by researchers which are based on ontologies or XML and we will mention the most relevant ones. In [10], it is presented an engine for semantic search which would be used for tourism domain, which is able to provide precise and relevant results based on the input query. It is ontology based. The main modules are: Query Controller, Query Prototype, Query Similarity Mapper, State Parser, City-State Parser, Ontological Synset Parser, Distance Parser, Service Finder and Caller, Service Modules, Metaprocessor, and URL Generator. XSEarch [11] is a semantic search engine based on XML (eXtensive Markup Language) and the implementation of it was challenging due to the numerous steps that needed to be taken for the engine to return favorable and relevant results. SemSearch [12] is an ontology based search engine which distinguishes from the others with: low barrier to access for ordinary end users, dealing with complex queries, precise and self-explanatory results, quick response, along with the following layers: the Google-like User Interface Layer, the Text Search Layer, the Semantic Query Layer, the Formal Query Language Layer and the Semantic Data Layer.

Closely related to semantic search are recommender systems, which primarily aim to provide suggestions useful to the user [13]. There are three main types of such systems: (1) content-based systems - artefacts similar to previously-appreciated ones are suggested; (2) collaborative filtering based systems - are suggested artefacts which were appreciated by users with a similar profile; (3) hybrids - in which ontologies often play an important role in optimizing the performance of the recommendation model [13]. An extensive classification of recommender systems is available in Fig. 2. Creating suitable recommendation algorithms for an innovative artefacts will add value to this research direction. Another current trend to which the model can contribute is that of chatbots (conversational agents) - computer programs designed to simulate conversations with human users [14]. The InnovRes model requires the use of a conversational engine that will assist in identifying the right innovative resources.

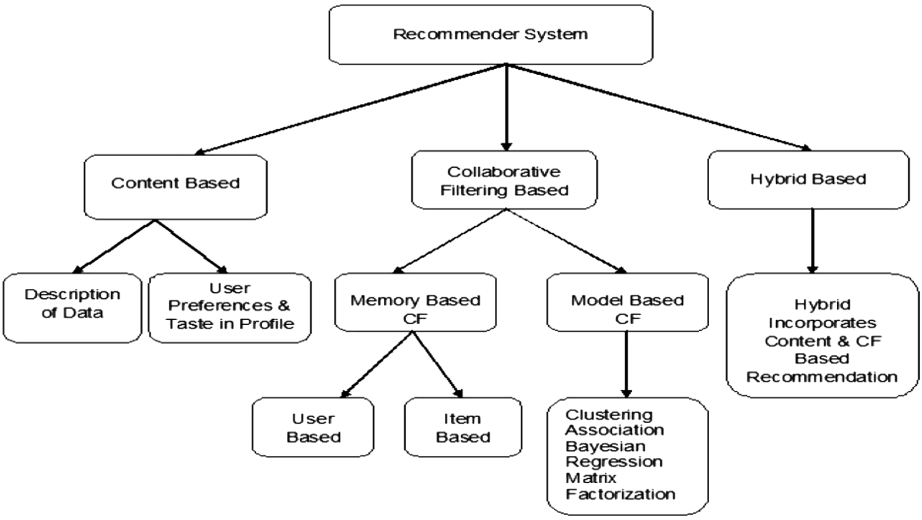

Fig. 2. Taxonomy of recommender systems

In terms of large data searches, high-performance computing will ensure the speed of response to the users. For this purpose, the Hadoop component for data storage – Hadoop Distributed File System (HDFS) [15] and the Apache Spark data processing framework [16] will be used. Unlike the MapReduce mechanism offered by Hadoop for data processing and used in a previous project [17], Apache Spark uses a resilient distributed data set that makes processing faster [18]. At the same time, Apache Spark is adapted to ontological data, which can be seen as a graph. The big data processing, as well as the graph/triplestores databases, are on the rise: the Neo4j Graph platform (ontological graph database) announced its collaboration with Apache Spark [19], so the interoperability between the two types of technologies (semantics, large data processing using high performance calculations) is doable.

3 Functional Description of the Model for Innovative Artefacts

The model will offer several functions via five services (see Fig. 3), from validating the innovation degree of an idea, supporting its development with useful bibliographical recommendations to building proposals for innovative artefacts based on that idea.

The first functionality is the **checking of the innovation degree of an idea**. This is possible using advanced data analytics: e.g., statistics of the number of similar reports, projects, patents, articles and the date of their publication. If none related references exist, then the idea is challenging: it might be very good, or not feasable. If old references exist, for sure the idea can't be the starting point to an innovative artefact. If a lot of recent references exist, then the idea respects a trending research direction and should be further exploited and so on.

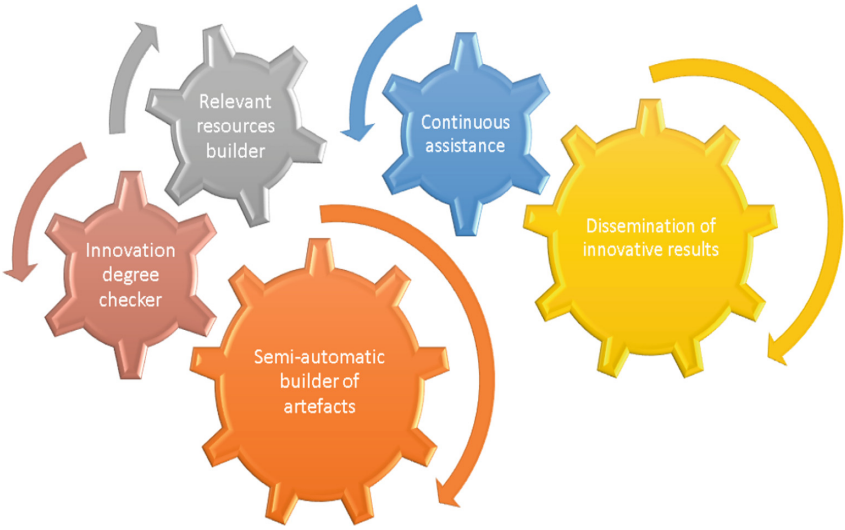

Fig. 3. Services offered by the technology model for innovative artefacts

The second functionality is the **building of relevant resources for the innovative artefact development**, through semantic search, recommendations and creation of resources, which let the users build personal innovation repositories. By combining semantic search and recommendation functionalities, our users will be able to obtain, in an integrated way, a relevant set of bibliographical resources, according to a specific research domain, abstract or keywords. Each resource could be evaluated for relevance by the user and thus further exploited in recommendations in next steps of the innovation process. The documents, links and other text-based artefacts will be transformed into data streams, then into ontologies (structured text), which can be interrogated by SPARQL questies processed by Apache Jena [20].

The third functionality is **the semi-automatic building of artefacts**, which will allow the user (researcher, business consultant, teacher, student) to customize some existent templates or to propose new ones, which will be available to the users after the administrator's acceptance.

The forth functionality is the **dissemination of own innovation results**: the users will be able to publish their own research, which will be searchable in the system after admin's validation.

The last functionality is **continuous assistance** via a trained chatbot and via a virtual forum community.

4 Technological Stack of the Model for Innovative Artefacts

The model can be implemented as a stand-alone system in different institutions (platform-as-a-service, difference instances of the same content management system) or as a cross-institutional system (software-as-a-service, single instance, joining the content of several organizations). The emergent technologies which are necessary for its implementation should be grouped in a closed layered architecture (as see in Fig. 4).

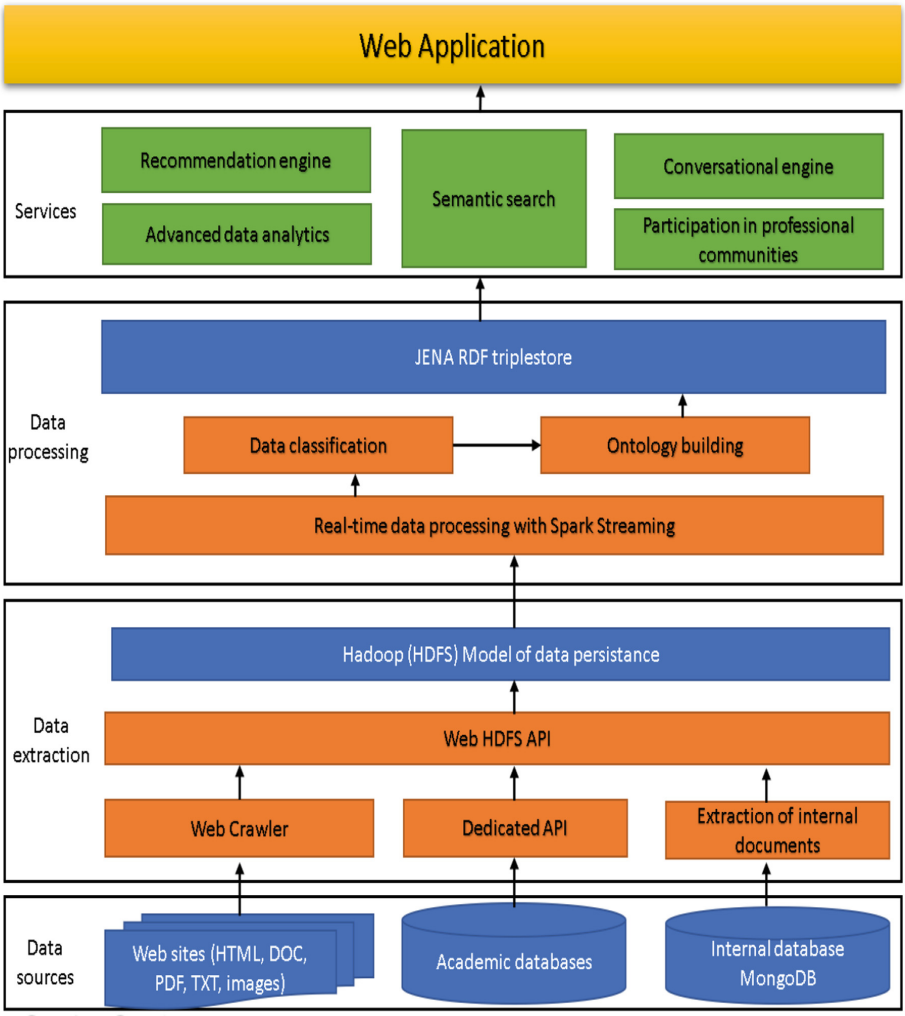

Fig. 4. Architecture of the technology model for innovative artefacts

The layer exposed to the users is given by the web application, which is in the form of a portal. This layer has access to the service layer, which allows the implementation

of the functionalities described in second section. The main components of the service layer are: the advanced analytics engine, the semantic search and recommendation engine, the conversational and the forum engine.

The service layer has access to the data processing layer, where the real time processing of data with Spark Streaming API [21] is done. Spark processing runs 10 times faster on disk and 100 times faster in memory than normal processing. For implementing recommendations and semantic search, ontologies will be built from extracted data, using several technologies: GATE, WordNet, Text2Onto, OpenNLP, Jena API [22]. The transformation of resources from plain text/natural language to structured text/ontologies, for them to be queried by computer-based applications, is not an easy task and advanced algorithms are needed [18], e.g. the ones for semantic similarity [23]. The ontologies will be saved in a semantic repository, e.g. JENA RDF triplestore [20]. Pre-processing of documents might be necessary and, for this purpose, the open-source Apache Kafka will be used.

The processing layer has access to the modules which deal with data extraction. Here, in data extraction layer, dedicated APIs for scientific databases are used, e.g. Springer APIs, IEEE Xplore API, Nature OpenSearch API [24] etc. All the data are saved in the files distributed system Hadoop – HDFS, using the Web HDFS API [25]. Various sources of data (as seen in Fig. 4) will be interrogated:

- websites;
- academic databases;
- an internal repository of innovative resources.

5 Conclusions

In this article, we propose a technological model for innovation seekers. Although there are many semantic search products [5], there are no tools which integrate the smart search of resources, the validation of an innovative idea and the initiation of its description, like our model proposes. Also, there are no search engines for business consultants, which makes our model a necessary documentation tool for them to make as many valuable project proposals as possible and to increase the absorption rate of European funds. We described a full technological stack and argument the interoperability of all proposes technologies, thus we claim that our model is feasible and implementable.

Acknowledgements. This work has been partially supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689/"Lib2Life - Revitalizarea bibliotecilor si a patrimoniului cultural prin tehnologii avansate"/"Revitalizing Libraries and Cultural Heritage through Advanced Technologies", within PNCDI II. Also, the work has partially received funding from the European Union's Erasmus+ Capacity Building in Higher Education program under grant agreement No. 586060-EPP-1-2017-1-RO-EPPKA2-CBHE-JP for the EXTEND project.

References

1. European Commission. https://ec.europa.eu/growth/industry/policy/digital-transformation_en. Accessed 2018
2. Digital tools for researchers. http://connectedresearchers.com/online-tools-for-researchers/. Accessed 2018
3. Alves, T., Rodrigues, R., Costa, H., Rocha, M.: Development of an information retrieval tool for biomedical patents. Comput. Methods Programs Biomed. **159**, 125–134 (2018)
4. Ribeiro Nogueira Ferraz, R., Quoniam, L., Reymond, D., Maccari, E.A.: Example of open-source OPS (Open Patent Services) for patent education and information using the computational tool Patent2Net. World Pat. Inf. **46**, 21–31 (2016)
5. Elbedweihy, K.M., Wrigley, S.N., Clough, P., Ciravegna, F.: An overview of semantic search evaluation initiatives. J. Web Semant. **30**, 82–105 (2015)
6. Bodea, C.N., Lipai, A., Dascalu, M.I.: An ontology-based search tool in the semantic web. In: Advancing Information Management through Semantic Web Concepts and Ontologies, pp. 221–249. IGI Global (2013)
7. Bošnjak, A., Podgorelec, V.: Upgrade of a current research information system with ontologically supported semantic search engine. Expert Syst. Appl. **66**, 189–202 (2016)
8. Sayed, A., Muqrishi, A.A.: IBRI-CASONTO: ontology-based semantic search engine. Egypt. Inform. J. **18**(3), 181–192 (2017)
9. Mangold, C.: A survey and classification of semantic search approaches. Int. J. Metadata Semant. Ontol. **2**(1), 23–34 (2007)
10. Laddha, S., Jawandhiya, P.M.: Semantic search engine. Indian J. Sci. Technol. **10**(23), 1–6 (2017)
11. Cohen, S., Mamou, J., Kanza, Y., Sagiv, Y.: XSEarch: a semantic search engine for XML. In: Proceedings of the 29th VLDB Conference, Berlin (2003)
12. Lei, Y., Uren, V., Motta, E.: SemSearch: a Search Engine for the Semantic Web. Knowledge Media Institute. Accessed 2018
13. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook. Springer, London (2011). https://doi.org/10.1007/978-0-387-85820-3
14. Stanica, I., Dascalu, M.I., Bodea, C.N., Moldoveanu, A.: VR job interview simulator: where virtual reality meets artificial intelligence for education. In: Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, pp. 9–12. IEEE (2018)
15. HADOOP – HDFS. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. Accessed 2018
16. Apache Spark. https://spark.apache.org/. Accessed 2018
17. Paraschiv, I.C., Dascalu, M., Banica, C., Trausan-Matu, S.: Designing a scalable technology hub for researchers. In: Proceedings of the 13th International Scientific Conference "eLearning and Software for Education", Bucharest, pp. 13–18 (2017)
18. Noyes, K.: Five things you need to know about Hadoop v. Apache Spark (2015). https://www.infoworld.com/article/3014440/big-data/five-things-you-need-to-know-about-hadoop-v-apache-spark.html
19. Burt, J.: Connecting The Dots With Graph Databases (2017). https://www.nextplatform.com/2017/10/24/connecting-dots-graph-databases/
20. Apache Jena. https://jena.apache.org/. Accessed 2019
21. Spark Data Sources. https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-data-sources.adoc. Accessed 2019

22. Dascalu, M.I., Bodea, C.N., Marin, I.: Semantic formative e-assessment for project management professionals. In: Proceedings of the 4th Eastern European Regional Conference on the Engineering of Computer Based Systems (ECBS-EERC), Brno, pp. 1–8. IEEE (2015)
23. Rus, V., Lintean, M., Banjade, R., Niraula, N., Stefanescu, D.: SEMILAR: the semantic similarity toolkit. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia (2013)
24. APIs for scholarly resources. https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/. Accessed 2019
25. HADOOP – HDFS. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. Accessed 2019
26. Apache Kafka. https://kafka.apache.org/. Accessed 2019