Towards Semantic Web-Based Information Retrieval to solve Information Overload in an Applied Gaming Ecosystem

Philippe Tamla

FernUniversitaet in Hagen, Faculty for Multimedia and Computer Science, Hagen Germany
philippe.tamla@studium.fernuni-hagen.de

Abstract. The RAGE research project will provide access to a wide range of software assets for Applied Gaming (AG) enabling AG software developers to better collaborate, share their knowledge, and be able to react to new requirements and trends more efficiently. But, RAGE is facing a greater challenge which is Information Overload (IO) because of the permanent flow of new documents published daily on the Web including their different formats. To solve IO, existing contributors have used complicated mechanisms that index the Web and harvest large piles of documents to detect relevant materials. Others have used Semantic Web Technologies to enrich web resources with additional facts and meaning, but have let the system (and not the end-user) decide about the relevance of the search result failing to satisfy the users' requirements.
Thus, we propose a novel software architecture that combines NLP (by mean of Named Entity Recognition) and *Domain Ontology* to support searching and browsing (faceted-search), as well as reasoning about named entities on the Semantic Web. AG software developers will get only relevant information according to their specific need without much time spent on information harvesting.

Keywords: Social Networking · Applied Gaming · Information Retrieval · Named Entity Recognition · Semantic Web

1 Introduction

The market of non-leisure games, Applied Games (AG) or Serious Games [1], is an emerging business, but its growth potential is affected because its key actors (industries, developers, researchers) are not working well together. The AG market is characterized by weak interconnectedness, limited knowledge exchange, the absence of harmonising standards, limited specialisations, limited division of labour, and insufficient evidence of the products' efficacies [2]. The European project Realizing an Applied Gaming Ecosystem (RAGE) [3] will address this challenge by getting hold of advanced usable gaming assets (technology push) to get access to the associated business cases (commercial opportunity), innovative and co-creatively usable practice knowledge, and supplementary content

resources. It will support the integration with various Social Network Systems (SNSs) (like LinkedIn, Twitter, or Stack Exchange ("Hot Questions")), Scientific Publication and Presentation Platforms (SPPs) like Mendeley and SlideShare, Software Repositories (SRs) like GitHub ("Build software better"), and more [4, 5]. This whill allow its key participants to effectively and efficiently collaborate and access the associated innovative business cases and commercial opportunities [5]. To achieve this goal, RAGE will refer to Social Network Analysis (SNA) by mean of applying technologies for Natural Language Analysis (NLA) for discourse analysis, as well as Semantic Representation and Annotation (SRA) of its results [6]. This will help to extend the envisioned Ecosystem with features of a social mediation engine going beyond content syndication, i.e. it will serve as a social space that mediates collaboration partners, while content remains the main attractor [5].
But, RAGE is facing a greater challenge which is **Information Overload (IO)** due to the constant flow of web resources [7], the heterogeneous nature [6] and diversity of its contents [8], and the increasing number of tools and channels used to manage them [9]. This means for RAGE that new innovation requirements will generate new imports into Ecosystem which will saturate the target system with large piles of relevant materials, hence leading the user into overload. IO can decrease innovation because it has a negative impact on work productivity (*interruptions* [8], *distractions* [10]) and individual behaviour (like *anxiety* [11], *infobesity* [12]). Automated knowledge extraction can help tackle IO because it can retrieve useful knowledge (in form of named entities) from naturally written texts automatically, which can be used for efficiently searching and reasoning about existing documents in various domains on the Web [6]. Existing techniques dealing with IO have often focused on developing search tools [13, 14] that harvest the Web and/or construct complicated queries to find useful documents. But, they often fail because they do not really know more about the current search context, the goal of the user, and how to deal with heterogenous documents and contents in various domains on the Web [6]. Therefore, RAGE will apply NLA (by mean of Named Entity Recognition (NER)) and Semantic Web Technologies [15] to respectively extract useful knowledge (in form of NEs) from its Ecosystem and to construct and knowledge base (of NEs and their relationships) that will improve searching, accessing, and reasoning about NEs in different textual languages and across various technological domains of expertise [16, 17]. This paper is organized as follow: section 2 reviews the literature related to this work. Section 3 drafts our approach to solve IO by means of applying *NER* and *Semantic Web* and present the challenges related to it. We conclude in section 4.

2 Background and related works

2.1 The RAGE Ecosystem

As already outlined, the RAGE Ecosystem will support the integration with several external systems to help AG users get hold of advanced usable gam-

ing assets, best practice knowledge, and supplementary content resources and business cases [5]. However, due to the heterogeneous nature of RAGE, various knowledge is permanently integrated from different sources (SNSs, SPPs, SRs), different levels of scope (from single named entities (NEs) to entire web documents), across different relationships (e.g. static, dynamic relationships between NEs like users and documents [17]), and various technological domains of expertise. Thus, it is useful to extract and manage such knowledge efficiently in order to avoid IO. It is necessary to know which kind of content types and knowledge (NEs) can be captured from RAGE, how to represent, manage, further develop, and re-use such knowledge in different domains of expertise. This can help to better understand existing content items, their relationships with other items, to classify them properly (e.g. using a proper taxonomy), to depict the context in which these items are used, and finally, to retrieve useful documents for the user. For instance, such knowledge can be used to automatically answer questions like: Who can help me with my problem? What are the current hot topics, How to use a specific asset, Where are the bugs that need to be fixed? and many more.

2.2 Information Retrieval & Semantic Web Technologies

Information Retrieval (IR) techniques were introduced by many researchers to help users find relevant documents on the Web. Existing works mainly focused on building search engines [13, 19] that index Web documents to retrieve useful documents when searching with keywords. An index is often used instead of the Web document itself to retrieve and rank documents by their relevance, with the result that if a document is not yet indexed, useful information can be hidden to the user [20]. ***Named Entity Recognition (NER)*** is often used to assist the IR process because it can extract knowledge (or named entities) from natural language (NL) written texts [6] which can be used to optimize document search. But, since naturally written texts are often characterized by misspelling, synonyms, various word forms, and inflections making it hard to extract them [21], NER is often assisted by other NL preprocessing steps like regular expressions, tokenization, sentence splitting, part-of-speech (POS) tagging (etc.) [17]. Ananiadou et al. [22] developed an IR system based on NER and text-mining to retrieve documents from an education community. Deheng et al. [21] applied a linear chain Conditional Random Field (CRF) to crawl web documents and extract software related terms (such as programming language, API, frameworks) from the Java Programming Language. While NER is useful for extracting and classifying NEs, Semantic Web technologies are often introduced for facilitating the acquisition of extra related information about an existing NE, thus enriching the initial text according to a specific use case [23].

Semantic Web Technologies [15] help software agents easily access NEs and understand their relationships [23]. In the Semantic Web, NEs and their relationships are represented in a "knowledge graph" which is described using a formal specification (called *Domain Ontology*). XML, RDF (Resource Description Language) [20], and OWL (Web Ontology Language)[15] are the standard markup

languages used for describing domain-specific knowledge graph and making it available on the Web. Zhu et al. [16] combined NER and Semantic Web for Social Network Analysis. Domain adaptation was supported by observing the links between NEs represented in a domain ontology. Their tool, ESpotter, was able to understand the relationships between people, company, and projects by applying an adjacency matrix based on the triple (Document, Entity, and EntityType). NEs occurring in same documents were deemed to co-occur with each other. However, this approach is highly dependent on domain ontology and rules which are manually maintained. Witte et al. [17] applied Semantic Web technologies and NER to extract and link software related NEs (such as source code, requirements, and design documentation). They relied on the GATE NLP framework to build their NER pipeline. They constructed an ontology reasoner to query vulnerable software components, and to traceability link different NEs [24] such as pieces of code and API documentation. However, similar to the ESpotter tool, this approach is highly dependent on a specific domain ontology and requires experts' assistance to construct complex queries.

The drawbacks found in state-of-the-art techniques have motivated us to implement a focused application that can deal with individual aspects of resources in various domains of expertise, while still being able to integrate their results into a common knowledge base.

3 Towards Semantic Web-based Information Retrieval

Our assumption for creating a new SWB-IR is that traditional IR-systems which use indexing or rule-based NER mechanisms are not suitable approaches for accessing, exploring, and reasoning about existing knowledge in an heterogeneous-oriented domain like RAGE. Thus, we propose a new software architecture that can unify various types of knowledge and easily adapt to new domains. We rely on a layer-based architecture (as shown in Figure 1) for extracting various NEs in the RAGE Ecosystem as proposed by Nawroth et al. [6]. Our first component is the *RAGE Corpus Layer* which represents the various heterogeneous contents and knowledge sources available in Ecosystem. *The Adapter Layer* (with its sub-adapters) will be used to access existing knowledge sources. The *NLP Layer*, which contents a pipeline of several NLP steps (like POS, chunking, stop word removal etc.) will support the extraction of named entities (NE-chunks) from different textual inputs. The result of this extraction process is a vector whose features are NE-Chunks to be processed in the following layers. The *Knowledge Layer* contains two sub layers: The *SVM Classifier* (suitable for software assets classification [6]) for supporting faceted-search (browsing) within a domain-specific taxonomy, and an *Ontology Populator* for building a knowledge base which will allow semantic-based search and reasoning about NEs and their interrelationships in different domains of expertise. By using such a layered architecture, we hope to be able to ease the implementation of such a system (simplicity) by creating modular and independent sub-components (modularity) after inter-layer interface specification [6]. While the extracted knowledge and

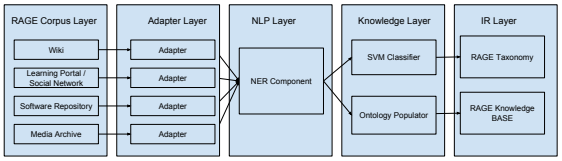

Fig. 1. Semantic Web-based Information Retrieval (SWB-IR)

the functionality may be covered by two main components, a NE extractor on one side, and an ontology reasoner on the other side, just a simple and user-friendly interface is necessary to provide a smart access to the existing materials.

4 Conclusion

Until now *Information Overload (IO)* has remained a great challenge event for Applied Gaming (AG) developers due to an ever increasing among web resources and the diversity of their different sources. Existing approaches have mainly relied on traditional *Information Retrieval (IR)* techniques that use complicated mechanisms (e.g. indexing or rules) to harvest the Web but often missing relevant documents. In this paper, we have presented a first architecture that combines NLP (like NER) and *Semantic Web technologies* to support searching and browsing (faceted-search) in the European-funded RAGE Ecosystem. In the next steps, we will analyse common practices of AG users and then use expert interviews to validate our proposed architecture on conceptual and technical levels.

References

1. Schmidt, R., Emmerich, K., Schmidt, B.: Applied games–in search of a new definition. In: International Conference on Entertainment Computing, Springer (2015) 100–111
2. Becker, J., Van Lankveld, G., Steiner, C., Hemmje, M.: Realizing an applied gaming ecosystem: Towards supporting service-based innovation knowledge management and transfer. In: International Conference on Games and Learning Alliance, Springer (2015) 540–549
3. : RAGE [Online]. ([accessed Juli 6, 2018])
4. Salman, M., Star, K., Nussbaumer, A., Fuchs, M., Brocks, H., Vu, D., Heutelbeck, D., Hemmje, M.: Towards social media platform integration with an applied gaming ecosystem. In: Submitted to: SOTICS, The Fifth International Conference on Social Media Technologies, Communication, and Informatics. (2015)
5. Salman, M., Fuchs, M., Vu, B., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M.: Integrating scientific publication into an applied gaming ecosystem. GSTF Journal on Computing (JoC) **5**(1) (2016) 45
6. Nawroth, C., Schmedding, M., Brocks, H., Kaufmann, M., Fuchs, M., Hemmje, M.: Towards cloud-based knowledge capturing based on natural language processing. Procedia Computer Science **68** (2015) 206–216

7. Hoq, K.M.G.: Information overload: Causes, consequences and remedies-a study. Philosophy and Progress **55**(1-2) (2016) 49–68
8. Gantz, J., Boyd, A., Dowling, S.: Tackling information overload at the source. IDC White Papers (2009)
9. Murphy, G.: Attacking information overload in software development. In: Visual Languages and Human-Centric Computing, 2009. VL/HCC 2009. IEEE Symposium on, IEEE (2009) 4–4
10. Spira, J.B., Burke, C.: Intels war on information overload: A case study. New York, Basex (2009)
11. Wurman, R.S.: Information anxiety. Doubleday (1989)
12. Case, D.O., Andrews, J.E., Johnson, J.D., Allard, S.L.: Avoiding versus seeking: the relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. Journal of the Medical Library Association **93**(3) (2005) 353
13. Li, J., Loo, B.T., Hellerstein, J.M., Kaashoek, M.F., Karger, D.R., Morris, R.: On the feasibility of peer-to-peer web indexing and search. In: International Workshop on Peer-to-Peer Systems, Springer (2003) 207–215
14. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Volume 283. Addison-Wesley Reading (2010)
15. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284**(5) (2001) 28–37
16. Zhu, J., Goncalves, A., Uren, V.: Adaptive named entity recognition for social network analysis and domain ontology maintenance. In: Proceedings of 3rd Professional Knowledge Management Conference, Springer, LNAI. (2005)
17. Witte, R., Zhang, Y., Rilling, J.: Empowering software maintainers with semantic web technologies. The Semantic Web: Research and Applications (2007) 37–52
18. Vu, D.B.: Realizing an applied gaming ecosystem - extending an education portal suite towards an ecosystem portal. Master's thesis, Technische Universitt Darmstadt, Germany (2015)
19. Khan, A., Martin, D., Tiropanis, T.: Using semantic indexing to improve searching performance in web archives. The First International Conference on Building and Exploring Web Based Environments-WEB2013 (2013)
20. Sibangiso, N., Khesani, Richard, C.: A semantic web solution for information overload. In: Research ana Intellectual expo (RIE), Harare, Zimbabwe. (2011)
21. Ye, D., Xing, Z., Foo, C.Y., Ang, Z.Q., Li, J., Kapre, N.: Software-specific named entity recognition in software engineering social content. In: Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on. Volume 1., IEEE (2016) 90–101
22. Ananiadou, S., Thompson, P., Thomas, J., Mu, T., Oliver, S., Rickinson, M., Sasaki, Y., Weissenbacher, D., McNaught, J.: Supporting the education evidence portal via text mining. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **368**(1925) (2010) 3829–3844
23. Velardi, P., Fabriani, P., Missikoff, M.: Using text processing techniques to automatically enrich a domain ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, ACM (2001) 270–284
24. Antoniol, G., Canfora, G., De Lucia, A., Casazza, G.: Information retrieval models for recovering traceability links between code and documentation. In: icsm, IEEE (2000) 40