Collaborative Semantic Management and Automated Analysis of Scientific Literature

Bahar Sateli and René Witte

Semantic Software Lab
Department of Computer Science and Software Engineering
Concordia University, Montréal, QC, Canada

Abstract. The overabundance of literature available in online repositories is an ongoing challenge for scientists that have to efficiently manage and analyze content for their information needs. Most of the existing literature management systems merely provide support for storing bibliographical metadata, tagging, and simple annotation capabilities. In this demo paper, we go beyond these approaches by demonstrating how an innovative combination of semantic web technologies with natural language processing can mitigate the information overload by helping in curating and organizing scientific literature. We present the *Zeeva* system as a first prototype that demonstrates how we can turn existing papers into a queryable knowledge base.

1 Introduction

Every research group faces the task of managing research literature pertinent to ongoing projects. This includes storing and indexing publications that are required as background or foundation for a specific topic, finding and discussing related work, as well as sharing and linking research techniques, data, and software. Existing bibliographical tools – whether online or locally installed – mainly focus on managing bibliographic metadata, together with some limited form of social support, like tagging or free-text comments. But they all lack further support when it comes to explicitly model, store, and query a paper's *content*, such as *goals, claims, methods,* or *results*. While strategies for the semantic markup of newly created publications have been proposed before [1], no tools exist that would help researchers dealing with existing papers, in particular by integrating automated text analysis workflows.

Our overall research goal is to improve the management of scientific literature, in particular for individual researchers and research groups. Our hypothesis is that semantic technologies, including semantic wikis and text mining, can improve several tasks that users are facing on a daily basis. To investigate the feasibility and impact of semantic literature management support, we have been developing *Zeeva*, a first prototype that integrates wiki-based collaboration with semantic knowledge representation and text mining in a coherent, user-friendly interface.

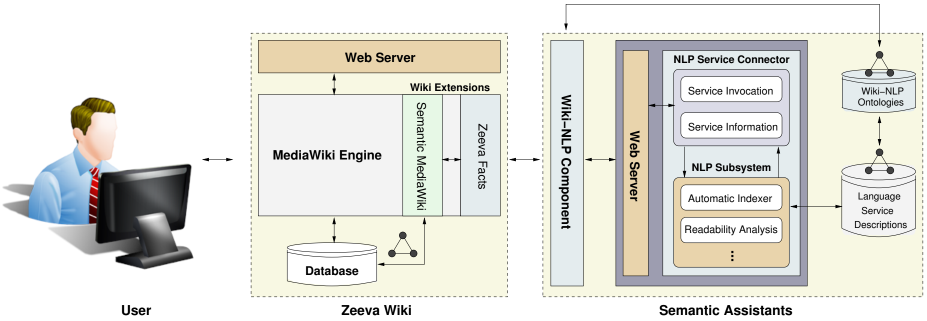


Fig. 1. A high-level overview of the Zeeva system architecture

2 Zeeva System Architecture

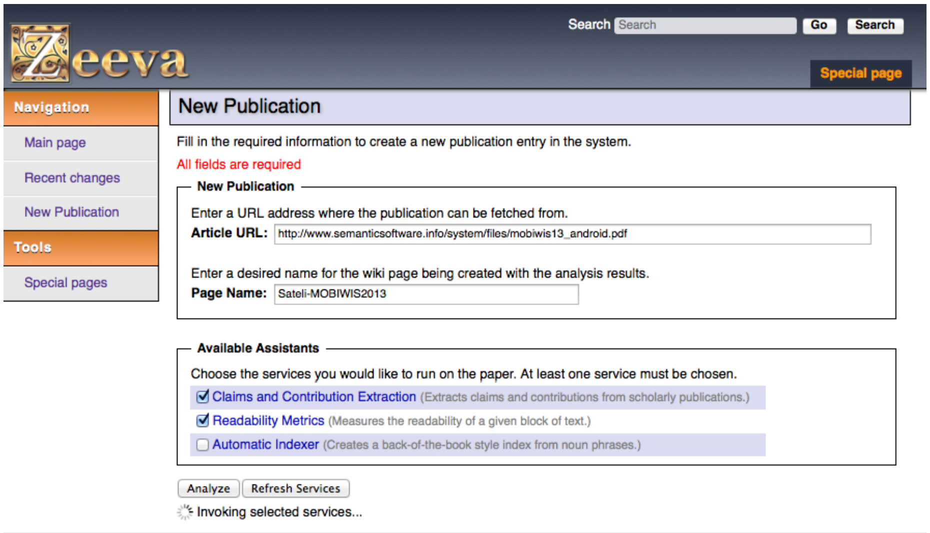
The Zeeva system (Fig. 1) features a wiki as its front-end. Powered by the highly scalable MediaWiki[1] engine, users interact with the Zeeva wiki using their Web browser. They can view and edit the wiki content using a simple markup language, called *wiki markup*. Semantic capabilities are provided through the Semantic MediaWiki (SMW) [2] extension. SMW allows special markup to be inserted into wiki pages in order to embed metadata about the page's content. The metadata is subsequently transformed internally into RDF[2] triples. In addition, the Zeeva wiki has a special extension, called *Zeeva Facts*, which allows wiki users to seamlessly interact with natural language processing (NLP) pipelines directly within the wiki environment to automatically analyze scientific publications.

The NLP services in Zeeva are provided by the Semantic Assistants [3], an open source framework that can publish various NLP pipelines, implemented based on the General Architecture for Text Engineering (GATE) [4], as W3C standard web services. The service-oriented architecture of the Semantic Assistants framework allows us to add or remove arbitrary NLP pipelines from the Zeeva wiki to experiment different use cases without any modifications to its wiki engine.

When users invoke NLP services through the wiki interface, a RESTful request is sent to the Semantic Assistants server via the Zeeva Facts extension. The Semantic Assistants server then fetches the content of the paper from the provided URL and executes the user-selected NLP pipelines on the retrieved text. The NLP results are passed on to the Semantic Assistants Wiki-NLP connector [5], which transforms them into wiki-friendly markup and stores them in the wiki database. Each semantic wiki markup internally translates into a semantic triple, with the wiki page as the subject, the declared property as the predicate, and the given value as the object. SMW stores the generated triples in the wiki repository that can be later queried both within the wiki and from external applications through an RDF feed. Currently, the Zeeva system does not employ

[1] MediaWiki, http://www.mediawiki.org

[2] Resource Description Framework, http://www.w3.org/RDF/



New Publication

Fill in the required information to create a new publication entry in the system.

All fields are required

New Publication

Enter a URL address where the publication can be fetched from.

Article URL: http://www.semanticsoftware.info/system/files/mobiwis13_android.pdf

Enter a desired name for the wiki page being created with the analysis results.

Page Name: Sateli-MOBIWIS2013

Available Assistants

Choose the services you would like to run on the paper. At least one service must be chosen.

☑ Claims and Contribution Extraction (Extracts claims and contributions from scholarly publications.)

☑ Readability Metrics (Measures the readability of a given block of text.)

☐ Automatic Indexer (Creates a back-of-the-book style index from noun phrases.)

Invoking selected services...

Fig. 2. Invoking integrated text mining assistants in the Zeeva wiki

any ontology specific to the literature analysis domain on its backend, rather it uses the SWIVT[3] ontology provided by the SMW extension.

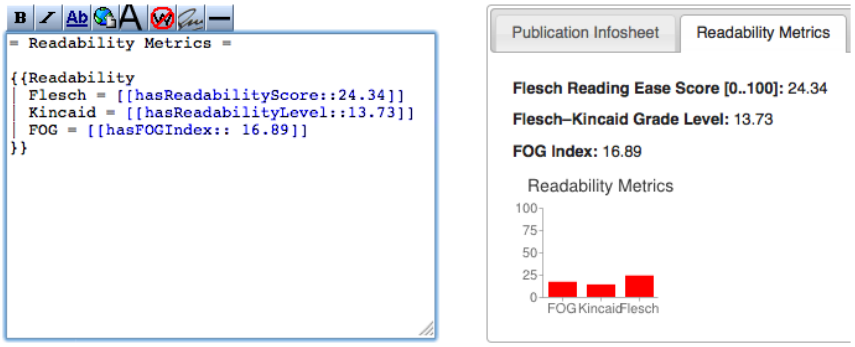
3 Demonstration

In this demo,[4] we show how a number of concrete tasks, like finding contributions of an author over a set of gathered papers, are supported in Zeeva. In particular, we explain how a researcher can use the Zeeva system to interact with the NLP services (Fig. 2) in order to automatically extract structural and rhetorical entities from scientific publications. During the demonstration, visitors can see how we make use of MediaWiki's *templating* mechanism to transform NLP pipelines output to semantic triples in real-time. Zeeva's pre-defined templates, like the one illustrated in Fig. 3, define *(i)* the look and feel of the results when embedded in wiki pages, and *(ii)* the semantic metadata that should be attached to each pipeline output.

In real-world scenarios, research projects are typically collaborative work between two or more researchers. Therefore, one key requirement in the design of the Zeeva system has been the creation of a shared space, where all researchers of a team have access to the most up-to-date information and can easily keep track of content modifications. During the demo, we will show how multiple researchers can interact through the wiki, while always having an up-to-date view of the knowledge created by other users of the system using the SMW inline queries.

[3] Semantic Wiki Vocabulary and Terminology, http://semantic-mediawiki.org/swivt/

[4] Please see http://www.semanticsoftware.info/eswc2014 for demo screencasts.



```
= Readability Metrics =

{{Readability
| Flesch = [[hasReadabilityScore::24.34]]
| Kincaid = [[hasReadabilityLevel::13.73]]
| FOG = [[hasFOGIndex:: 16.89]]
}}
```

Flesch Reading Ease Score [0..100]: 24.34

Flesch–Kincaid Grade Level: 13.73

FOG Index: 16.89

```
<rdf:RDF>
  <swivt:Subject rdf:about="http://localhost/.../Android-2DMOBIWIS2013">
    <rdf:type rdf:resource="http://localhost/.../Category-3APublication"/>
    <rdfs:label>Android-MOBIWIS2013</rdfs:label>
    <swivt:page rdf:resource="http://localhost/.../Android-MOBIWIS2013"/>
    <property:HasFOGIndex rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
       16.89
    </property:HasFOGIndex>
    <property:HasReadabilityLevel rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
       13.73
    </property:HasReadabilityLevel>
    <property:HasReadabilityScore rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
       24.34
    </property:HasReadabilityScore>
    <!-- additional properties not shown in this excerpt -->
  </swivt:Subject>
  </property:HasTitle>
</rdf:RDF>
```

Fig. 3. Wiki template with semantic properties (left), preview in browser (right) and the RDF document (bottom) generated by Semantic MediaWiki

Finally, we will show how the Zeeva wiki can be transformed from an analysis platform to a queryable knowledge base. We will show how the results of human-AI collaboration on the Zeeva wiki can be exported to standalone RDF documents (Fig. 3) that can be directly queried with SPARQL queries.

4 Related Work

A large body of research exists that deals with improving access to the ever-increasing amount of scientific literature. Within the scope of this demo paper, we focus on collaborative solutions and semantic web ontologies.

Wiki-based systems, such as WikiPapers[5] and AcaWiki,[6] are recent efforts for collaborative literature analysis. Any user can register on these websites and submit summaries or reviews of peer-reviewed articles to the wiki. The goal of these systems is to collect a community-driven, comprehensive compilation of

[5] WikiPapers, http://wikipapers.referata.com

[6] AcaWiki, http://www.acawiki.org

bibliographical and semantical metadata and make them available to the general public.

Within the semantic web framework, researchers like Groza et al. [1] are envisioning an approach where authors can explicitly encode their bibliographical and rhetorical metadata in their publications prior to publishing the documents, i.e., using special markup in text as they are writing their content. This special markup can then be automatically extracted and mapped onto formal descriptions in pre-defined ontologies, such as SALT [1], when accessed by machines.

Our work is complementary to these efforts. While we also aim at formalizing the body of knowledge contained in scientific publications within a collaborative (wiki-based) space, our approach offers an innovative way of generating bibliographical and semantical metadata from a collaboration between human users and 'intelligent' natural language processing agents. This way, scientific publications can be enriched with metadata that is generated automatically, hence, transforming them into queryable artifacts, while remaining amenable to human-created semantic annotations within the wiki.

5 Conclusion and Future Work

Currently existing literature management tools provide limited support for research groups when dealing with knowledge-intensive tasks, like literature surveys. We propose Zeeva, a proof-of-concept system that demonstrates how the next generation of literature management tools can go beyond simply storing bibliographical data and support research groups by transforming publications into an active knowledge base. Zeeva's embedded "Semantic Assistants" play the role of intelligent agents that collaboratively work with human users on scientific publications text analysis. Future work includes the addition of further text mining pipelines, as well as designing and integrating an ontology specific to the scientific literature analysis domain. In addition, we plan to perform user studies to measure the impact of the semantic support in real-world scenarios.

References

1. Groza, T., Handschuh, S., Mller, K., Decker, S.: SALT - Semantically Annotated LaTeX for Scientific Publications. In: The Semantic Web: Research and Applications. Volume 4519 of LNCS. Springer Berlin Heidelberg (2007) 518–532
2. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: The Semantic Web (ISWC 2006). Volume 4273 of LNCS. Springer Berlin Heidelberg (2006) 935–942
3. Witte, R., Gitzinger, T.: Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In: 3rd Asian Semantic Web Conference (ASWC 2008). Volume 5367 of LNCS., Bangkok, Thailand, Springer (2008) 360–374
4. Cunningham, H., et al.: Text Processing with GATE (Version 6). University of Sheffield, Department of Computer Science (2011)
5. Sateli, B., Witte, R.: Natural Language Processing for MediaWiki: The Semantic Assistants Approach. In: The 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012), Linz, Austria, ACM (2012)