Towards Semantic Recommendation of Biodiversity Datasets based on Linked Open Data

Felicitas Löffler
Dept. of Mathematics
and Computer Science
Friedrich Schiller University
Jena, Germany

Bahar Sateli
Semantic Software Lab
Dept. of Computer Science
and Software Engineering
Concordia University
Montréal, Canada

René Witte
Semantic Software Lab
Dept. of Computer Science
and Software Engineering
Concordia University
Montréal, Canada

Birgitta König-Ries
Friedrich Schiller University
Jena, Germany and
German Centre for Integrative
Biodiversity Research (iDiv)
Halle-Jena-Leipzig, Germany

ABSTRACT

Conventional content-based filtering methods recommend documents based on extracted keywords. They calculate the similarity between keywords and user interests and return a list of matching documents. In the long run, this approach often leads to overspecialization and fewer new entries with respect to a user's preferences. Here, we propose a semantic recommender system using Linked Open Data for the user profile and adding semantic annotations to the index. Linked Open Data allows recommendations beyond the content domain and supports the detection of new information. One research area with a strong need for the discovery of new information is biodiversity. Due to their heterogeneity, the exploration of biodiversity data requires interdisciplinary collaboration. Personalization, in particular in recommender systems, can help to link the individual disciplines in biodiversity research and to discover relevant documents and datasets from various sources. We developed a first prototype for our semantic recommender system in this field, where a multitude of existing vocabularies facilitate our approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage And Retrieval**]: Online Information Services

General Terms

Design, Human Factors

Keywords

content filtering, diversity, Linked Open Data, recommender systems, semantic indexing, semantic recommendation

In: G. Specht, H. Gamper, F. Klan (eds.): Proceedings of the 26^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 21.10.2014 - 24.10.2014, Bozen, Italy, published at http://ceur-ws.org.

1. INTRODUCTION

Content-based recommender systems observe a user's browsing behaviour and record the interests [1]. By means of natural language processing and machine learning techniques, the user's preferences are extracted and stored in a user profile. The same methods are utilized to obtain suitable content keywords to establish a content profile. Based on previously seen documents, the system attempts to recommend similar content. Therefore, a mathematical representation of the user and content profile is needed. A widely used scheme are TF-IDF (term frequency-inverse document frequency) weights [19]. Computed from the frequency of keywords appearing in a document, these term vectors capture the influence of keywords in a document or preferences in a user profile. The angle between these vectors describes the distance or the closeness of the profiles and is calculated with similarity measures, like the cosine similarity. The recommendation lists of these traditional, keyword-based recommender systems often contain very similar results to those already seen, leading to overspecialization [11] and the "Filter-Bubble"-effect [17]: The user obtains only content according to the stored preferences, other related documents not perfectly matching the stored interests are not displayed. Thus, increasing diversity in recommendations has become an own research area [21, 25, 24, 18, 3, 6, 23], mainly used to improve the recommendation results in news or movie portals.

One field where content recommender systems could enhance daily work is research. Scientists need to be aware of relevant research in their own but also neighboring fields. Increasingly, in addition to literature, the underlying data itself and even data that has not been used in publications are being made publicly available. An important example for such a discipline is biodiversity research, which explores the variety of species and their genetic and characteristic diversity [12]. The morphological and genetic information of an organism, together with the ecological and geographical context, forms a highly diverse structure. Collected and stored in different data formats, the datasets often contain or link to spatial, temporal and environmental data [22]. Many important research questions cannot be answered by working with individual datasets or data collected by one group, but require meta-analysis across a wide range of data. Since the analysis of biodiversity data is quite time-consuming, there is a strong need for personalization and new filtering techniques in this research area. Ordinary search functions in relevant data portals or databases, e.g., the *Global Biodiversity In-*

formation Facility (GBIF)[1] and the *Catalog of Life,*[2] only return data that match the user's query exactly and fail at finding more diverse and semantically related content. Also, user interests are not taken into account in the result list. We believe our semantic-based content recommender system could facilitate the difficult and time-consuming research process in this domain.

Here, we propose a new semantic-based content recommender system that represents the user profile as Linked Open Data (LOD) [9] and incorporates semantic annotations into the recommendation process. Additionally, the search engine is connected to a terminology server and utilizes the provided vocabularies for a recommendation. The result list contains more diverse predictions and includes hierarchical concepts or individuals.

The structure of this paper is as follows: Next, we describe related work. Section 3 presents the architecture of our semantic recommender system and some implementation details. In Section 4, an application scenario is discussed. Finally, conclusions and future work are presented in Section 5.

2. RELATED WORK

The major goal of diversity research in recommender systems is to counteract overspecialization [11] and to recommend related products, articles or documents. More books of an author or different movies of a genre are the classical applications, mainly used in recommender systems based on collaborative filtering methods. In order to enhance the variety in book recommendations, Ziegler et al. [25] enrich user profiles with taxonomical super-topics. The recommendation list generated by this extended profile is merged with a rank in reverse order, called dissimilarity rank. Depending on a certain diversification factor, this merging process supports more or less diverse recommendations. Larger diversification factors lead to more diverse products beyond user interests. Zhang and Hurley [24] favor another mathematical solution and describe the balance between diversity and similarity as a constrained optimization problem. They compute a dissimilarity matrix according to applied criterias, e.g., movie genres, and assign a matching function to find a subset of products that are diverse as well as similar. One hybrid approach by van Setten [21] combines the results of several conventional algorithms, e.g., collaborative and case-based, to improve movie recommendations. Mainly focused on news or social media, approaches using content-based filtering methods try to present different viewpoints on an event to decrease the media bias in news portals [18, 3] or to facilitate the filtering of comments [6, 23].

Apart from Ziegler et al., none of the presented approaches have considered semantic technologies. However, utilizing ontologies and storing user or document profiles in triple stores represents a large potential for diversity research in recommender systems. Frasincar et al. [7] define semantically enhanced recommenders as systems with an underlying knowledge base. This can either be linguistic-based [8], where only linguistic relations (e.g., synonymy, hypernomy, meronymy, antonymy) are considered, or ontology-based. In the latter case, the content and the user profile are represented with concepts of an ontology. This has the advantage that several types of relations can be taken into account. For instance, for a user interested in "geology", the profile contains the concept "geology" that also permits the recommendation of inferred concepts, e.g., "fossil". The idea of recommending related concepts was first introduced by Middelton et al. [15]. They developed *Quickstep*, a recommender system for research papers with ontological terms in the user profile and for paper categories. The ontology only considers is-a relationships and omits other relation types (e.g., part-of). Another simple hierarchical approach from Shoval et al. [13] calculates the distance among concepts in a profile hierarchy. They distinguish between perfect, close and weak match. When the concept appears in both a user's and document's profile, it is called a perfect match. In a close match, the concept emerges only in one of the profiles and a child or parent concept appears in the other. The largest distance is called a weak match, where only one of the profiles contains a grandchild or grandparent concept. Finally, a weighted sum over all matching categories leads to the recommendation list. This ontological filtering method was integrated into the news recommender system *epaper*. Another semantically enhanced recommender system is *Athena* [10]. The underlying ontology is used to explore the semantic neighborhood in the news domain. The authors compared several ontology-based similarity measures with the traditional TF-IDF approach. However, this system lacks of a connection to a search engine that allows to query large datasets.

All presented systems use manually established vocabularies with a limited number of classes. None of them utilize a generic user profile to store the preferences in a semantic format (RDF/XML or OWL). The FOAF (Friend Of A Friend) project[3] provides a vocabulary for describing and connecting people, e.g., demographic information (name, address, age) or interests. As one of the first, in 2006 Celma [2] leveraged FOAF in his music recommender system to store users' preferences. Our approach goes beyond the FOAF interests, by incorporating another generic user model vocabulary, the Intelleo User Modelling Ontology (IUMO).[4] Besides user interests, IUMO offers elements to store learning goals, competences and recommendation preferences. This allows to adapt the results to a user's previous knowledge or to recommend only documents for a specific task.

3. DESIGN AND IMPLEMENTATION

In this section, we describe the architecture and some implementation details of our semantic-based recommender system (Figure 1). The user model component, described in Section 3.1, contains all user information. The source files, described in Section 3.2, are analyzed with GATE [5], as described in Section 3.3. Additionally, GATE is connected with a terminology server (Section 3.2) to annotate documents with concepts from the provided biodiversity vocabularies. In Section 3.4, we explain how the annotated documents are indexed with GATE Mímir [4]. The final recommendation list is generated in the recommender component (Section 3.5).

3.1 User profile

The user interests are stored in an RDF/XML format utilizing the FOAF vocabulary for general user information. In

[1]GBIF, http://www.gbif.org
[2]Catalog of Life, http://www.catalogueoflife.org/col/search/all/
[3]FOAF, http://xmlns.com/foaf/spec/
[4]IUMO, http://intelleo.eu/ontologies/user-model/spec/

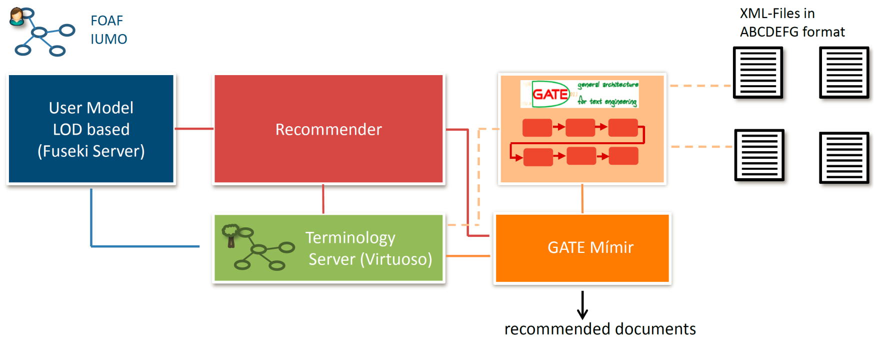

Figure 1: The architecture of our semantic content recommender system

order to improve the recommendations regarding a user's previous knowledge and to distinguish between learning goals, interests and recommendation preferences, we incorporate the Intelleo User Modelling Ontology for an extended profile description. Recommendation preferences will contain settings in respect of visualization, e.g., highlighting of interests, and recommender control options, e.g., keyword-search or more diverse results. Another adjustment will adapt the result set according to a user's previous knowledge. In order to enhance the comprehensibility for a beginner, the system could provide synonyms; and for an expert the recommender could include more specific documents.

The interests are stored in form of links to LOD resources. For instance, in our example profile in Listing 1, a user is interested in "biotic mesoscopic physical object", which is a concept from the ENVO[5] ontology. Note that the interest entry in the RDF file does not contain the textual description, but the link to the concept in the ontology, i.e., `http://purl.obolibrary.org/obo/ENVO_01000009`. Currently, we only support explicit user modelling. Thus, the user information has to be added manually to the RDF/XML file. Later, we intend to develop a user profiling component, which gathers a user's interests automatically. The profile is accessible via an Apache Fuseki[6] server.

Listing 1: User profile with interests stored as Linked Open Data URIs

```
<rdf:Description rdf:about="http://www.semanticsoftware.info/person
    /felicitasloeffler">
<rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
<foaf:firstName>Felicitas</foaf:firstName>
<foaf:lastName>Loeffler</foaf:lastName>
<foaf:name>Felicitas Loeffler</foaf:name>
<foaf:gender>Female</foaf:gender>
<foaf:workplaceHomepage rdf:resource="http://dbpedia.org/page/
    University_of_Jena"/>
<foaf:organization>Friedrich Schiller University Jena
</foaf:organization>
<foaf:mbox>felicitas.loeffler@uni-jena.de</foaf:mbox>
<um:TopicPreference rdf:resource="http://purl.obolibrary.org/obo/
    ENVO_01000009"/>
</rdf:Description>
```

3.2 Source files and terminology server

The content provided by our recommender comes from the biodiversity domain. This research area offers a wide range of existing vocabularies. Furthermore, biodiversity is an interdisciplinary field, where the results from several sources have to be linked to gain new knowledge. A recommender system for this domain needs to support scientists by improving this linking process and helping them finding relevant content in an acceptable time.

Researchers in the biodiversity domain are advised to store their datasets together with metadata, describing information about their collected data. A very common metadata format is ABCD.[7] This XML-based standard provides elements for general information (e.g., author, title, address), as well as additional biodiversity related metadata, like information about taxonomy, scientific name, units or gathering. Very often, each taxon needs specific ABCD fields, e.g., fossil datasets include data about the geological era. Therefore, several additional ABCD-related metadata standards have emerged (e.g., ABCDEFG[8], ABCDDNA[9]). One document may contain the metadata of one or more species observations in a textual description. This provides for annotation and indexing for a semantic search. For our prototype, we use the ABCDEFG metadata files provided by the GFBio[10] project; specifically, metadata files from the Museum für Naturkunde (MfN).[11] An example for an ABCDEFG metadata file is presented in Listing 2, containing the core ABCD structure as well as additional information about the geological era. The terminology server supplied by the GFBio project offers access to several biodiversity vocabularies, e.g., ENVO, BEFDATA, TDWGREGION. It also provides a SPARQL endpoint[12] for querying the ontologies.

3.3 Semantic annotation

The source documents are analyzed and annotated according to the vocabularies provided by the terminology server. For this process, we use GATE, an open source framework that offers several standard language engineering components [5]. We developed a custom GATE pipeline (Figure 2) that analyzes the documents: First, the documents are split into tokens and sentences, using the existing NLP components included in the GATE distribution. Afterwards, an 'Annotation Set Transfer' processing resource adds the original

[5]ENVO, `http://purl.obolibrary.org/obo/envo.owl`
[6]Apache Fuseki, `http://jena.apache.org/documentation/serving_data/`
[7]ABCD, `http://www.tdwg.org/standards/115/`
[8]ABCDEFG, `http://www.geocase.eu/efg`
[9]ABCDDNA, `http://www.tdwg.org/standards/640/`
[10]GFBio, `http://www.gfbio.org`
[11]MfN, `http://www.naturkundemuseum-berlin.de/`
[12]GFBio terminology server, `http://terminologies.gfbio.org/sparql/`

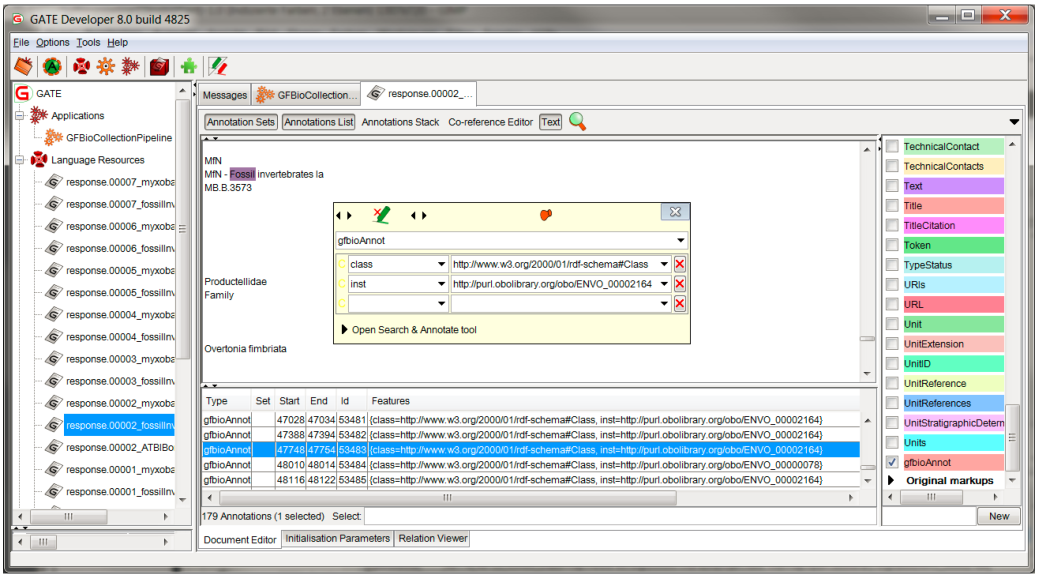

Figure 2: The GFBio pipeline in GATE presenting the GFBio annotations

markups of the ABCDEFG files to the annotation set, e.g., `abcd:HigherTaxon`. The following ontology-aware 'Large KB Gazetteer' is connected to the terminology server. For each document, all occurring ontology classes are added as specific "gfbioAnnot" annotations that have both instance (link to the concrete source document) and class URI. At the end, a 'GATE Mímir Processing Resource' submits the annotated documents to the semantic search engine.

3.4 Semantic indexing

For semantic indexing, we are using GATE Mímir:[13] "Mímir is a multi-paradigm information management index and repository which can be used to index and search over text, annotations, semantic schemas (ontologies), and semantic metadata (instance data)" [4]. Besides ordinary keyword-based search, Mímir incorporates the previously generated semantic annotations from GATE to the index. Additionally, it can be connected to the terminology server, allowing queries over the ontologies. All index relevant annotations and the connection to the terminology server are specified in an index template.

3.5 Content recommender

The Java-based content recommender sends a SPARQL query to the Fuseki Server and obtains the interests and preferred recommendation techniques from the user profile as a list of (LOD) URIs. This list is utilized for a second SPARQL query to the Mímir server. Presently, this query asks only for child nodes (Figure 3). The result set contains ABCDEFG metadata files related to a user's interests. We intend to experiment with further semantic relations in the future, e.g., object properties. Assuming that a specific fossil used to live in rocks, it might be interesting to know if other species, living in this geological era, occured in rocks. Another filtering method would be to use parent or grandparent nodes from the vocabularies to broaden the search. We will provide control options and feedback mechanisms to support the user in steering the recommendation process actively. The recommender component is still under development and has not been added to the implementation yet.

[13]GATE Mímir, https://gate.ac.uk/mimir/

Listing 2: Excerpt from a biodiversity metadata file in ABCDEFG format [20]

```
<abcd:DataSets xmlns:abcd="http://www.tdwg.org/schemas/abcd/2.06"
    xmlns:efg="http://www.synthesys.info/ABCDEFG/1.0">
<abcd:DataSet>
<abcd:Metadata>
<abcd:Description><abcd:Representation language="en">
<abcd:Title>MfN − Fossil invertebrates</abcd:Title>
<abcd:Details>Gastropods, bivalves, brachiopods, sponges</abcd:Details>
    </abcd:Representation></abcd:Description>
<abcd:Scope><abcd:TaxonomicTerms>
<abcd:TaxonomicTerm>Gastropods, Bivalves, Brachiopods, Sponges</
    abcd:TaxonomicTerm>
</abcd:TaxonomicTerms></abcd:Scope>
</abcd:Metadata>
<abcd:Units><abcd:Unit>
<abcd:SourceInstitutionID>MfN</abcd:SourceInstitutionID>
<abcd:SourceID>MfN − Fossil invertebrates Ia</abcd:SourceID>
<abcd:UnitID>MB.Ga.3895</abcd:UnitID>
<abcd:Identifications><abcd:Identification>
<abcd:Result><abcd:TaxonIdentified>
<abcd:HigherTaxa><abcd:HigherTaxon>
<abcd:HigherTaxonName>Euomphaloidea</abcd:HigherTaxonName>
<abcd:HigherTaxonRank>Family</abcd:HigherTaxonRank>
</abcd:HigherTaxon></abcd:HigherTaxa>
<abcd:ScientificName>
<abcd:FullScientificNameString>Euomphalus sp.</
    abcd:FullScientificNameString>
</abcd:ScientificName>
</abcd:TaxonIdentified></abcd:Result>
</abcd:Identification></abcd:Identifications>
<abcd:UnitExtension>
<efg:EarthScienceSpecimen><efg:UnitStratigraphicDetermination>
<efg:ChronostratigraphicAttributions>
<efg:ChronostratigraphicAttribution>
<efg:ChronoStratigraphicDivision>System</
    efg:ChronoStratigraphicDivision>
<efg:ChronostratigraphicName>Triassic</efg:ChronostratigraphicName>
</efg:ChronostratigraphicAttribution></
    efg:ChronostratigraphicAttributions>
</efg:UnitStratigraphicDetermination></efg:EarthScienceSpecimen>
</abcd:UnitExtension>
</abcd:Unit></abcd:Units></abcd:DataSet></abcd:DataSets>
```

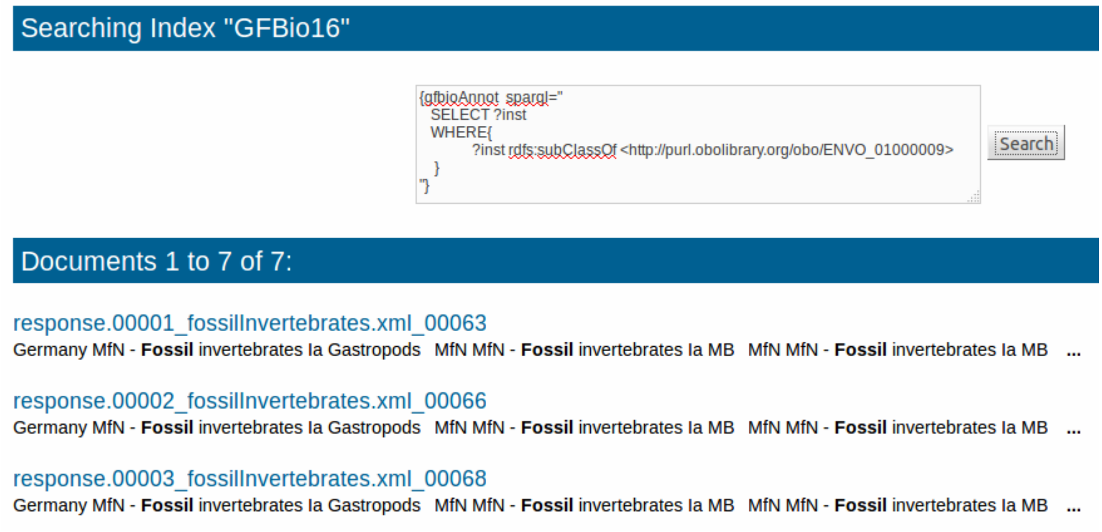

Figure 3: A search for "biotic mesoscopic physical object" returning documents about fossils (child concept)

4. APPLICATION

The semantic content recommender system allows the recommendation of more specific and diverse ABCDEFG metadata files with respect to the stored user interests. Listing 3 shows the query to obtain the interests from a user profile, introduced in Listing 1. The result contains a list of (LOD) URIs to concepts in an ontology.

Listing 3: SPARQL query to retrieve user interests

```
SELECT ?label ?interest  ?syn
WHERE
{
    ?s foaf:firstName  "Felicitas" .
    ?s um:TopicPreference ?interest  .
    ?interest  rdfs:label   ?label  .
    ?interest  oboInOwl:hasRelatedSynonym ?syn
}
```

In this example, the user would like to obtain biodiversity datasets about a "biotic mesoscopic physical object", which is the textual description of `http://purl.obolibrary.org/obo/ENVO_01000009`. This technical term might be incomprehensible for a beginner, e.g., a student, who would prefer a description like "organic material feature". Thus, for a later adjustment of the result according to a user's previous knowledge, the system additionally returns synonyms.

The returned interest (LOD) URI is utilized for a second query to the search engine (Figure 3). The connection to the terminology server allows Mímir to search within the ENVO ontology (Figure 4) and to include related child concepts as well as their children and individuals. Since there is no metadata file containing the exact term "biotic mesoscopic physical object", a simple keyword-based search would fail. However, Mímir can retrieve more specific information than stored in the user profile and is returning biodiversity metadata files about "fossil". That ontology class is a child node of "biotic mesoscopic physical object" and represents a semantic relation. Due to a high similarity regarding the content of the metadata files, the result set in Figure 3 contains only documents which closely resemble each other.

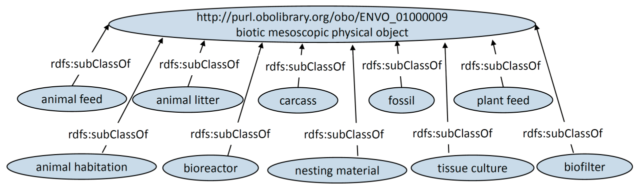

Figure 4: An excerpt from the ENVO ontology

5. CONCLUSIONS

We introduced our new semantically enhanced content recommender system for the biodiversity domain. Its main benefit lays in the connection to a search engine supporting integrated textual, linguistic and ontological queries. We are using existing vocabularies from the terminology server of the GFBio project. The recommendation list contains not only classical keyword-based results, but documents including semantically related concepts.

In future work, we intend to integrate semantic-based recommender algorithms to obtain further diverse results and to support the interdisciplinary linking process in biodiversity research. We will set up an experiment to evaluate the algorithms in large datasets with the established classification metrics *Precision* and *Recall* [14]. Additionally, we would like to extend the recommender component with control options for the user [1]. Integrated into a portal, the result list should be adapted according to a user's recommendation settings or adjusted to previous knowledge. These control functions allow the user to actively steer the recommendation process. We are planning to utilize the new layered evaluation approach for interactive adaptive systems from Paramythis, Weibelzahl and Masthoff [16]. Since adaptive systems present different results to each user, ordinary evaluation metrics are not appropriate. Thus, accuracy, validity, usability, scrutability and transparency will be assessed in several layers, e.g., the collection of input data and their interpretation or the decision upon the adaptation strategy. This should lead to an improved consideration of adaptivity in the evaluation process.

6. ACKNOWLEDGMENTS

This work was supported by DAAD (German Academic Exchange Service)[14] through the PPP Canada program and by DFG (German Research Foundation)[15] within the GFBio project.

7. REFERENCES

[1] F. Bakalov, M.-J. Meurs, B. König-Ries, B. Sateli, R. Witte, G. Butler, and A. Tsang. An approach to controlling user models and personalization effects in recommender systems. In *Proceedings of the 2013 international conference on Intelligent User Interfaces*, IUI '13, pages 49–56, New York, NY, USA, 2013. ACM.

[2] Ò. Celma. FOAFing the music: Bridging the semantic gap in music recommendation. In *Proceedings of 5th International Semantic Web Conference*, pages 927–934, Athens, GA, USA, 2006.

[3] S. Chhabra and P. Resnick. Cubethat: News article recommender. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 295–296, New York, NY, USA, 2012. ACM.

[4] H. Cunningham, V. Tablan, I. Roberts, M. Greenwood, and N. Aswani. Information extraction and semantic annotation for multi-paradigm information management. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 307–327. Springer Berlin Heidelberg, 2011.

[5] H. Cunningham et al. *Text Processing with GATE (Version 6)*. University of Sheffield, Dept. of Computer Science, 2011.

[6] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: A scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1175–1184, New York, NY, USA, 2010. ACM.

[7] F. Frasincar, W. IJntema, F. Goossen, and F. Hogenboom. A semantic approach for news recommendation. *Business Intelligence Applications and the Web: Models, Systems and Technologies, IGI Global*, pages 102–121, 2011.

[8] F. Getahun, J. Tekli, R. Chbeir, M. Viviani, and K. Yétongnon. Relating RSS News/Items. In M. Gaedke, M. Grossniklaus, and O. Díaz, editors, *ICWE*, volume 5648 of *Lecture Notes in Computer Science*, pages 442–452. Springer, 2009.

[9] T. Health and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.

[10] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, pages 16:1–16:6, New York, NY, USA, 2010. ACM.

[11] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

[12] M. Loreau. *Excellence in ecology*. International Ecology Institute, Oldendorf, Germany, 2010.

[13] V. Maidel, P. Shoval, B. Shapira, and M. Taieb-Maimon. Ontological content-based filtering for personalised newspapers: A method and its evaluation. *Online Information Review*, 34 Issue 5:729–756, 2010.

[14] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[15] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, Jan. 2004.

[16] A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: Framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, Dec. 2010.

[17] E. Pariser. *The Filter Bubble - What the internet is hiding from you*. Viking, 2011.

[18] S. Park, S. Kang, S. Chung, and J. Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 443–452, New York, NY, USA, 2009. ACM.

[19] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[20] Museum für Naturkunde Berlin. Fossil invertebrates, UnitID:MB.Ga.3895. http://coll.mfn-berlin.de/u/MB_Ga_3895.html.

[21] M. van Setten. *Supporting people in finding information: hybrid recommender systems and goal-based structuring*. PhD thesis, Telematica Instituut, University of Twente, The Netherlands, 2005.

[22] R. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, and et al. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE 9(3): e89606*, 2014.

[23] D. Wong, S. Faridani, E. Bitton, B. Hartmann, and K. Goldberg. The diversity donut: enabling participant control over the diversity of recommended responses. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 1471–1476, New York, NY, USA, 2011. ACM.

[24] M. Zhang and N. Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 123–130, New York, NY, USA, 2008. ACM.

[25] C.-N. Ziegler, G. Lausen, and L. Schmidt-Thieme. Taxonomy-driven computation of product recommendations. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 406–415, New York, NY, USA, 2004. ACM.

[14]DAAD, https://www.daad.de/de/
[15]DFG, http://www.dfg.de