ISPRS Int. J. Geo-Inf. **2013**, *2*, 256-275; doi:10.3390/ijgi2020256

OPEN ACCESS

ISPRS International
Journal of
Geo-Information
ISSN 2220-9964
www.mdpi.com/journal/ijgi/

Article

From Geoportals to Geographic Knowledge Portals

Bernhard Vockner [1],* Andreas Richter [2] and Manfred Mittlböck [1]

[1] Research Studios Austria—Studio iSPACE, Schillerstr. 25, Salzburg 5020, Austria;
E-Mail: manfred.mittlboeck@researchstudio.at

[2] OMV Aktiengesellschaft, Trabrennstraße 6-8, Vienna 1020, Austria;
E-Mail: andreas.richter1@omv.com

* Author to whom correspondence should be addressed; E-Mail: bernhard.vockner@researchstudio.at;
Tel.: +43-662-908585-222; Fax: +43-662-908585-299.

Received: 24 January 2013; in revised form: 21 March 2013 / Accepted: 21 March 2013 / Published: 28 March 2013

Abstract: We present the application of Latent Semantic Analysis (LSA) in combination with recommender systems, in order to enhance discovery in geoportals. As a basis for discovery, metadata of spatial data and services, as well as of non-spatial resources, such as documents and scientific papers, is created and registered in the catalogue of the geoportal (semi-)automatically. Links that are not inherent in the data itself are established based on the semantic similarity of its textual content using LSA. This leads to the transition from unstructured data to structured (metadata) information, serving as a basis for the generation of knowledge. The metadata information is integrated into a recommendation system that provides a ranked list showing (1) what other users viewed and (2) the related resources discovered by the LSA workflow as a result. Based on the assumptions that similar texts have something in common and that users are likely to be interested in what other users viewed, recommendations provide a broader, but also more precise, search result; on the one hand, the recommender engine considers additional information; on the other hand, it ranks resources based on the discovery experience of other users and the likeliness of the documents being related to each other.

Keywords: latent semantic analysis; LSA; recommender; matching; similarity

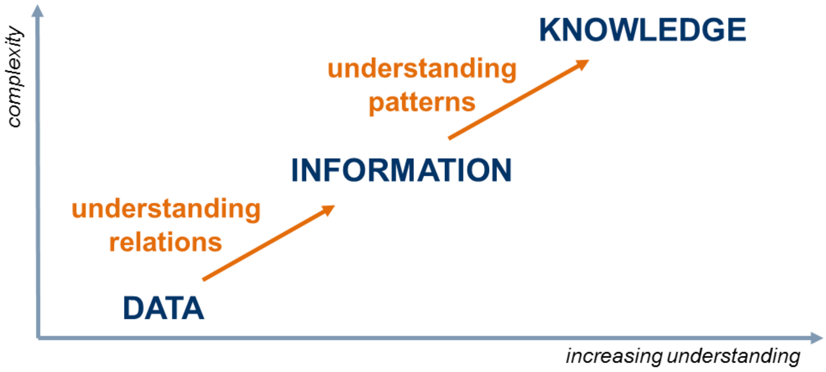
1. Introduction

With the emergence of smart devices, distributed and mobile computing, as well as the broadening of application areas for information systems, the transition of our society into an information society is long complete. In 1997, IBM envisioned an information society as a "society characterized by a high level of information intensity in everyday life of most citizens, in most organizations and workplaces" with the ability to "transmit, receive and exchange digital data rapidly between places irrespective of distance" [1]. In our society, the creation, distribution, manipulation and interpretation of information largely influences both our working environments and our everyday lives. Ever new requirements and ideas have led to a rapid development of technology that helps us to improve the speed, amount and accuracy of gaining what has become the soil of our economy and lives: data.

In this context, it has become a major challenge to find relevant data in that vast and ever-increasing digital universe that in 2011 exceeded 1.8 zettabytes (1.8 × 10^{12} gigabytes) contained in 500 quadrillion files [2]. Although it has not been verified, widespread use in literature, as well as the variety of mainstream technologies using location information, prove it plausible that more than 80% of all data has a spatial component.

This big amount of data, however, is of little value until meaningful information is derived from it. As Bellinger [3] puts it, data alone is "just a meaningless point in space and time, without reference to either space or time". For gaining information from flat data entities, it is necessary to put them into context or relations (*cf*. Figure 1). If not only relations between data are identified, but also patterns can be extracted and autonomously re-applied, knowledge is generated. This leads to increasing complexity, on the one hand, but also increased understanding, on the other hand.

Figure 1. From data to knowledge.



Considering the means of effectively turning data into information and knowledge as a basis for decision-making is obligatory for the optimization of data usage [4], due to the rapidly increasing amounts of data produced by man, as well as automated systems, such as sensors, effectively turning data into information has become a major challenge for individuals and businesses today. The difficulty, however, is not only to extract useful value from data and information, but also to find pieces of information that seem to be relevant in the first place [4]. To make things even more challenging, 70% of all data in the digital universe is unstructured [5] and often without a yet important context that allows meaningful discovery. Examples for such unstructured data are e-mails, text documents, presentations, images, videos and any other data that is not supplied with metadata, but also holds true for spatial data.

The growth of unstructured data amounts, therefore, challenges search mechanisms and algorithms, which are designed to discharge users in their quest for relevant information. According to Croisier [6], information overload, missing contextual filtering methods in unconnected information pools, as well as search engines that create flat result lists based on simple keyword queries have negative effects on efficiency and productivity.

In this context, an important task is the reduction of the percentage of unstructured data by developing new methods for discovery, which address the challenge of structuring flat data entities. According to Croisier [6], there are two different approaches of developing a semantic capacity in information systems: "First, the bottom-up approach is problematic, as it assumes metadata will be added to each piece of content to include additional information about its context. […] Second, the top-down approach might have more success for the rest of the data, as it focuses on developing automated natural language-based text annotation capabilities".

We consider semantic text matching in combination with recommendations as a solution to overcome challenges in information discovery within the domain of spatial data infrastructures (SDI). In its simplest form, semantic text matching deals with the relatedness of two entities or texts. While semantic text matching works fine for documents containing text, it can also be applied to structured metadata, thereby mixing the bottom-up and top-down approaches of Croisier [6], while enhancing the capabilities of information retrieval through information extraction and promoting knowledge discovery of structured content. Those techniques are not only used in active discovery processes ("What do I want to know?"), but also in passive recommendation processes ("Based on what I want to know, it is likely I also want to know that…"). This meets a major challenge of search engines operating on vast amounts of data, often making it difficult for users to conduct searches on things whose existence they are not yet aware.

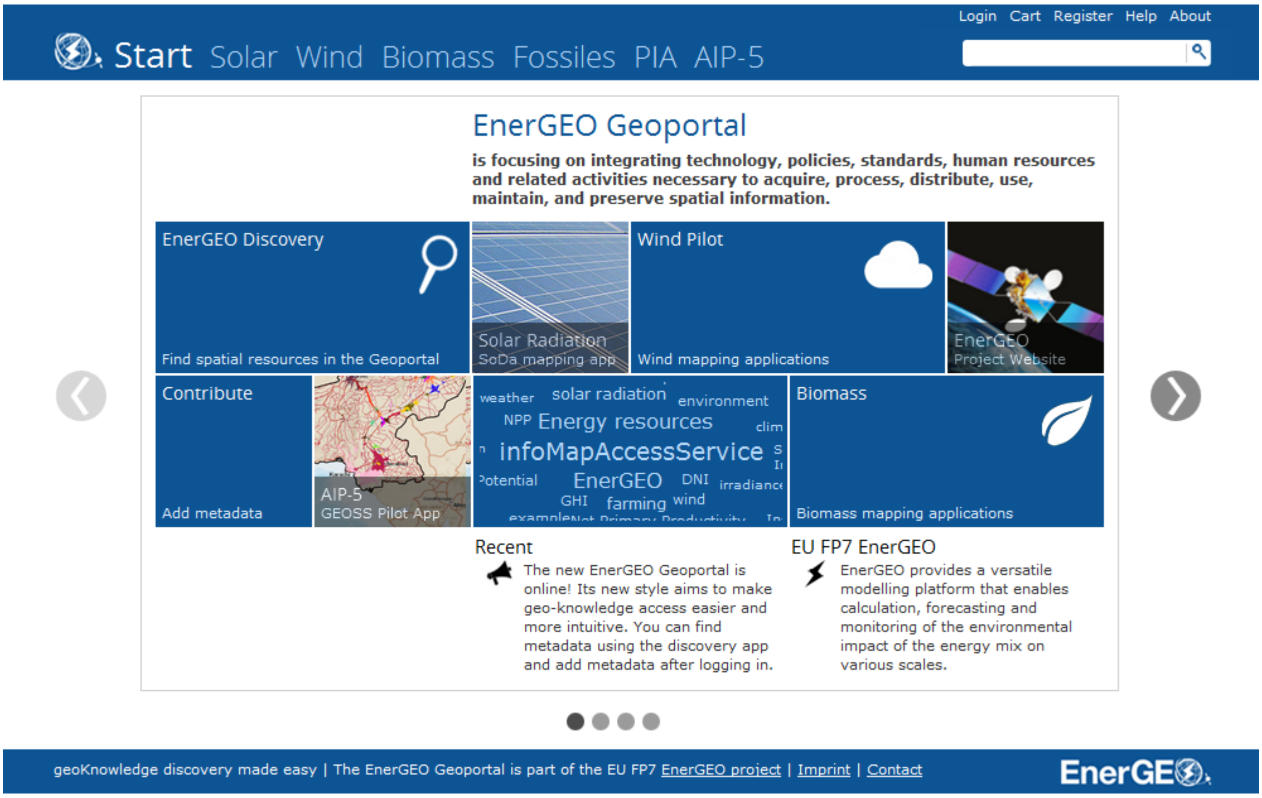
2. Spatial Portals

In many information environments, like spatial data infrastructures (SDIs), so-called portals have become important gateways as single points for accessing the information pools of companies, organizations or joint ventures. The portal is a website that acts as the final, critical bridge that brings together users and content [7]. It is a central platform for discovering, publishing, accessing and sharing information and knowledge [8]. While common examples, like Google or Bing, may also be considered as portals, in the geospatial domain, the terms "spatial portal" or "geoportal" were. Therefore, a geoportal is the information broker between geospatial resources and their potential users [8]. The crucial factor for the usefulness and eventual acceptance of a portal, however, is the provision of relevant resources that can be searched in a structured and documented manner. In any environment, a portal that offers little or no documented resources does not offer any benefits in terms of information retrieval and knowledge derivation [4].

The current state of the art SDIs use computer-readable, XML-structured metadata for discovery purposes. Metadata—that often follows acknowledged standards in order to ensure interoperability and comparability—has to be entered by the resource owners manually. The elements partially consist of structured information, but some elements, such as abstracts or lineage, are unstructured free text [9]. The metadata records themselves are kept in either central or distributed databases that form so-called catalogues [10]. As registries for resources, they act as the broker between users and providers. In a basic discovery mechanism, keyword-based searches of users are syntactically matched against words contained in metadata files [9]. Only metadata that contains exactly the keywords entered by the user is returned. Beside keyword-based search, tag clouds, auto-complete lists of content, as well as spatial (bounding box) and temporal filters are mostly applied in today's search interfaces within SDI [11]. The keyword-based search methods, however, reveal challenges that are trying to be met through the consideration of thesauri, taxonomies and ontologies [12,13]. In contrast to these approaches, we use recommendations in combination with semantic text matching methods to overcome these challenges.

As a geoportal must serve users' needs, we built the basic structure of the geoportal for the EU FP7 Project EnerGEO (*cf*. Figure 2) based on the task-oriented design paradigm, as described by Scholz and Mittlböck [14]. It therefore fits the "one question-one answer" principle for applications, as well as for discovery. This means that a portal has to be easy to use, providing meaningful information and not just plain data to users. Beside this, it serves as the demonstrator for the concept of semantic text matching and recommendations presented in this paper.

Figure 2. EnerGEO Geoportal.


Login Cart Register Help About
Start Solar Wind Biomass Fossiles PIA AIP-5

EnerGEO Geoportal

is focusing on integrating technology, policies, standards, human resources and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial information.

EnerGEO Discovery
Find spatial resources in the Geoportal
Solar Radiation
SoDa mapping app
Wind Pilot
Wind mapping applications
EnerGEO
Project Website
Contribute
Add metadata
AIP-5
GEOSS Pilot App
weather solar radiation environment
NPP Energy resources clim
infoMapAccessService
Potential EnerGEO DNI irradiance
GHI farming wind
Biomass
Biomass mapping applications

Recent

The new EnerGEO Geoportal is online! Its new style aims to make geo-knowledge access easier and more intuitive. You can find metadata using the discovery app and add metadata after logging in.

EU FP7 EnerGEO

EnerGEO provides a versatile modelling platform that enables calculation, forecasting and monitoring of the environmental impact of the energy mix on various scales.

geoKnowledge discovery made easy | The EnerGEO Geoportal is part of the EU FP7 EnerGEO project | Imprint | Contact

In this context, search methodologies were enhanced by semantic matching tools that help to exploit the large amounts of unstructured data by revealing semantic structures [6] and leading the way from simple keyword-matching data portals to contextualized knowledge portals. Smart search methods match and discover hidden relationships between semantic structures with similar meaning and link raw content with trusted data and information sources [6]. Such "content enrichment services" [6] are capable of suggesting keywords, extracting topics and entities and performing other forms of automated classifications, such as sentiment analysis.

3. Methods of Semantic Text Matching

In order to establish links between resources in a geoportal, semantic similarity algorithms need to be applied. In the following, only a small subset of the most common and prominent methods for semantic text matching in literature will be presented and evaluated. For a more complete overview of semantic text matching algorithms, we refer to Islam and Inkpen and Mihalcea *et al.* [15,16].

3.1. Method Overview

The Vector Space Model (VSM) is one of the simplest methods for computing semantic similarity. The VSM automatically extracts knowledge and, therefore, requires less manual work compared to other semantic approaches, such as ontologies [17]. VSM is based on the exact matching of terms that can be found in documents. It converts texts to n-dimensional vectors for measuring distances among them. As a distance measure, the cosine of the angle between two vectors is used in most cases. The result is a value of similarity ranging from 0 to 1, where 1 indicates an exact/high match between terms and 0 indicates that there is no match. This means the higher the value of the cosine, the higher the likelihood that two terms are equal.

However, the fact that the method is solely based on the exact matching of words raises problems, such as synonymy and polysemy. Synonymy deals with different words having the same meaning. For example, car and automobile are synonyms, which are not considered to be equal in the VSM. This can lead to poor recall, meaning that not all relevant information sources are discovered. Polysemy refers to words having more than one distinct meaning. For example, the term "model" can be a scaled representation of a real world object, a person employed to display clothing merchandise or a design type of a car. This can lead to poor precision, meaning that the "accuracy" of the retrieval is not sufficient for the user, who gets a high number of search results that are not relevant for the questions he or she posed. Another problem is that common words, such as "the" and "is", and correlating high similarity measures result in a high match, which does not represent the actual desired result.

To alleviate the drawbacks of the VSM, Dumais *et al.* [18] presented Latent Semantic Analysis (LSA), a statistical, corpus-based text comparison method. Throughout the literature, it is sometimes also referred to as Latent Semantic Indexing (LSI), which is primarily used in the field of information retrieval, whereas LSA is used in other application areas [19]. The process of learning words that are related to each other is based on their statistical co-occurrence together in a context [20].

LSA is a method of unsupervised document analysis. Unsupervised means that no direct human input is needed for conducting the analysis. Wiemer-Hastings [21] claims that LSA is even able to learn words at a rate similar to human beings. Landauer *et al.* [22] proved that LSA is capable of estimating the knowledge level of students by examining short essays they wrote. The results show that there is little difference between human judges and the model. This is related to the fact that the meaning of text passages can be carried by words only [22]. In contrast to human beings, LSA does not require word order or syntax to extract the essential meaning stored in documents. The key principle of LSA is that it does not use any manually created resources, such as thesauri or dictionaries.

It only depends on large amounts of texts to induce knowledge about the meanings of documents and words [23]. LSA assumes that the meaning of a text can be extracted as the sum of the meaning of its words [24].

LSA uses a weighted term-document matrix that is created out of a large collection of documents. For weighting or transformation purposes, several constellations can be used. tf-idf (term frequency-inverse document frequency) and log-entropy are the most common methods, although. Overall. 20 different combinations of local and global weighting schemes exist [25]. In the tf-idf weighting method, terms occurring less are upweighted to reflect their relative importance [24]. The application of a log-entropy weight reduces the effect of words occurring across a wide variety of contexts [21]. The goal of the transformation is to discover relationships between words and to use such relationships to describe the documents.

LSA uses the Singular Value Decomposition for dimension reduction. During compression, semantic information that is latent (~"hidden") in the corpus itself is captured. In contrast to VSM, it extracts concepts instead of words.

Summing up, LSA consists of four steps: (1) creation of a term-document matrix out of a collection of texts, (2) application of a transformation (e.g., tf-idf, log-entropy), (3) dimension reduction using Singular Value Decomposition (SVD) and (4) retrieval in reduced space by cosine similarity.

Although it was created for large collections of texts, Terzi *et al.* [26] consider LSA a state-of-the-art similarity measure for comparison of short texts, such as user reviews or abstracts, both of which can also be found in geoportals. However, Terzi *et al.* [26] think that LSA underperforms in the case of short user-generated reviews in recommender systems. As LSA does not make use of any syntactic information, it is better suited for longer texts than very short abstracts consisting of two or three sentences only [15]. This is especially true for noisy, unstructured information that contains spelling mistakes [26]. This disadvantage is also related to the fact that word co-occurrence may be rare in short texts [27]. In contrast to spatial resources, documents containing the URL to the file provide a short abstract version, as well as the full text of the document. Thus, one challenge is that information of different lengths is compared using LSA within our research.

Alongside with LSA, Pointwise Mutual Information in combination with Information Retrieval techniques (PMI-IR) [28] is a common semantic text matching method. It is an unsupervised measure that is based on co-occurrences of words, like LSA [16]. It uses a large corpus of statistical data collected by Information Retrieval (IR) processes from the Web [16,28].

Besides LSA and PMI-IR, Gabrilovich and Markovitch [29] proposed Explicit Semantic Analysis (ESA) as a method to analyze similarity in texts. It uses Wikipedia as a knowledge base to derive predetermined sets of natural concepts [29]. ESA creates a feature vector for each text. Each feature is related to a Wikipedia article. It determines the extent to which each word is related to words used within Wikipedia. ESA can be considered as being "explicit", because it uses external categories coming from Wikipedia, in contrast to LSA, which uses latent topics [30]. Gabrilovich and Markovitch [29] state that the advantage of ESA is that it can make use of "human knowledge encoded in Wikipedia".

Latent Dirichlet Allocation (LDA) [31] assumes that each document is a mixture of latent topics [20]. L'Huillier *et al.* [32] state that "[…] every topic is modeled as a probability distribution over the set of words represented by the vocabulary and every document as a probability distribution over a set of topics". As can be seen from this statement, the focus of LDA is on topic modeling rather than the meaning of words [20]. In contrast to LSA, LDA uses a probabilistic background instead of SVD [33]. Blei *et al.* [31] consider LDA a simple model and, therefore, a competitor to LSA in the future.

Newer approaches of semantic text matching methods comprise STASIS, STS and OMIOTIS. STASIS [27] takes information of the lexical database, WordNet, in order to calculate similarities between texts [26]. STS [15] uses string similarity in combination with corpus-based word similarity in shorter texts [26]. It is a modified version of the Longest Common Subsequence (LCS) string matching algorithm [15], which deals with finding the longest common subsequence of two sequences. The main difference to other approaches is that Islam and Inkpen [15] focus on the similarity between two sentences or short paragraphs, but not full texts. OMIOTIS [34] extends a word-to-word measure to be able to deal with texts and to establish links between concepts from WordNet [26]. For weighting purposes, it uses the semantic path length, the node depth in the thesaurus' hierarchy and the types of the semantic edges that compose the path [35].

3.2. Comparison of Methods for Semantic Text Matching

Throughout the literature, we are not aware of any all-encompassing evaluation of all methods previously presented. However, there are a few scientific papers comparing some of these methods in specific contexts. For example, Ramage *et al.* [36] evaluated the ability of a keyword-based, n-gram vector space method and LSA to model human judgments. The LSA model was consistent with those of human judgments with correlations of 0.6 [36]. PMI-IR, as another member of corpus-based techniques, nearly produces the same results as LSA [16,37]. In some cases, such as the test for English as a foreign language (TOEFL), PMI-IR achieved 10 percent better results than LSA [28]. However, Turney [28] states that this may be due to the fact that different amounts of data were used in the analyses. Tsatsaronis *et al.* [35] showed that their method, OMIOTIS, performed best with the Microsoft Paraphrase Corpus [38]. OMIOTIS [34] had the highest Spearman’s correlation ($p = 0.8905$) in contrast to LSA, STASIS [27] and STS [15], with LSA having the second best Spearman’s correlation ($p = 0.8714$) compared to human beings [35].

Mohler and Mihalcea [39] discovered that a small, domain-specific corpus performs better than a generic corpus, like one coming from Wikipedia. In that case, LSA ($r = 0.4628$) shows a higher Pearson’s correlation than ESA ($r = 0.4385$). Thus, for LSA, the quality of texts is more important than their quantity [39]. For domain-related information, LSA outperforms ESA, whereas ESA is better suited for generic corpora [39]. Cimiano *et al.* [33] showed that LSA performed better than LDA, no matter if LSA was trained on domain-specific documents or a general source, like Wikipedia.

As we mainly deal with energy-related information in EnerGEO, and all other results we discovered in the literature showed that LSA is among the best algorithms compared to human beings, we chose LSA as the most appropriate method to calculate the semantic similarity of texts.

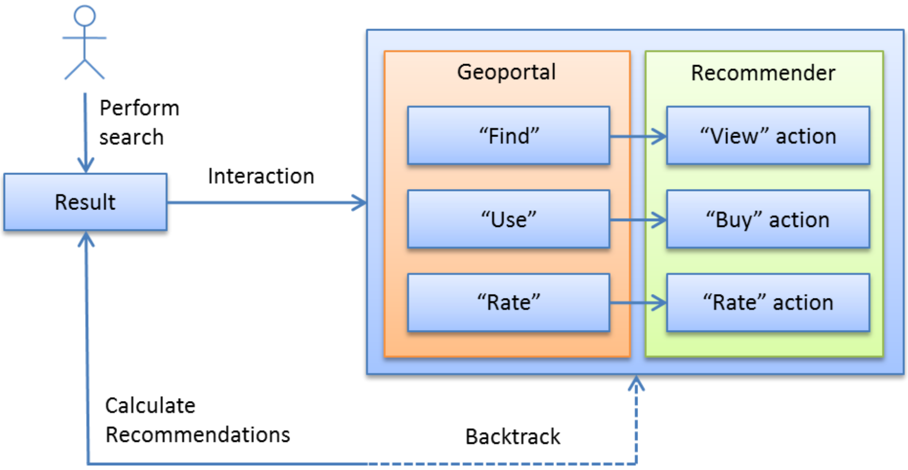
4. Recommendation Approaches

Online stores, such as Amazon.com, have established recommendations in the WWW in order to offer related “items”, which their customers might not have thought of when looking for a specific item. The recommendations themselves are based on calculations of user interactions in background processes. In general, a recommendation engine takes into account what other users viewed, bought or rated by "tracking" user clicks. Within the context of online stores, "items" are books or CDs. In the domain of SDIs, we propose spatial (vector-, raster- and service-data), as well as non-spatial content, such as scientific papers or project reports, to be part of a recommender engine, providing information to end-users in order to assist them in their task-gaining knowledge.

Recommender systems use different types of algorithms for calculating recommendations. For example, collaborative filtering approaches [40] use previous user interactions with items to provide recommendations to the user. They require a big amount of information and users in order to predict user preferences by comparison with other users. Content-based filtering [41] is based on previous sessions of a single user and the user profile. Thus, it requires a smaller amount of users to provide useful recommendations.

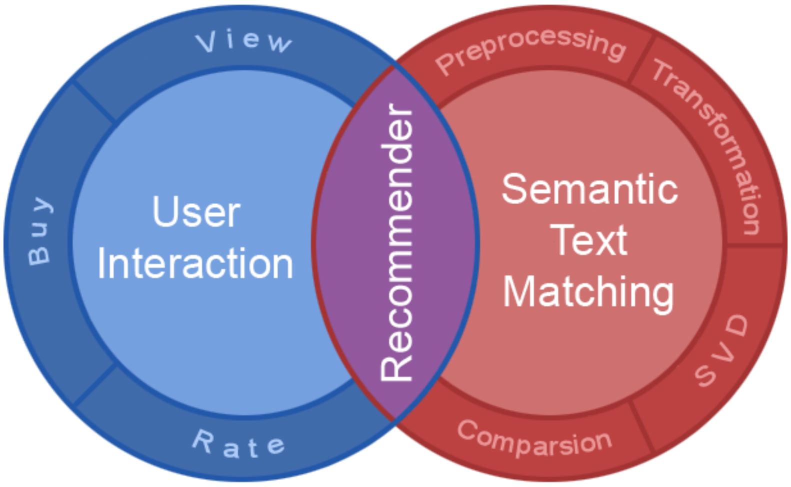
For the integration of recommendation engines in geoportals, we created an analogy to interrelate the differing concepts. Figure 3 shows the workflow of a typical user interaction in the geoportal. A user performs a search, retrieves a list of results and interacts with these. Thus, the task of "finding" information and having a look at a single result in a geoportal can be considered a "view" action, while "using" a resource (such as having a look at the preview) is interlinked with a "buy" action in the recommender engine. "Rating" exists in both the world of SDI as in online stores. Another important aspect in recommender systems is the so-called backtracking capability. This means that if a user clicks on the recommendation results, this interaction is submitted to the recommender engine as well, showing that the provided recommendations really were useful.

Figure 3. User interaction recommendation workflow.



As we face the situation that the spatial domain and geoportals do not have the same mass of users as online stores, we propose to extend the concept of recommendations based on user interactions with semantic text matching results as additional input for the calculation of recommendations. Therefore, the proposed architecture consists of two interacting components to provide meaningful recommendations: (1) "tracking" of user interactions in the geoportal and (2) "semantic text matching" (see also Figure 4).

Figure 4. The two interacting components to provide recommendations.



5. Implementation

We implemented the concepts of integrating automatic metadata extraction tools, semantic text matching algorithms and recommender systems in the EnerGEO Geoportal. The EnerGEO Geoportal is a spatial portal that contains information resources from the energy domain. The major compartments of the proposed system are presented in the following.

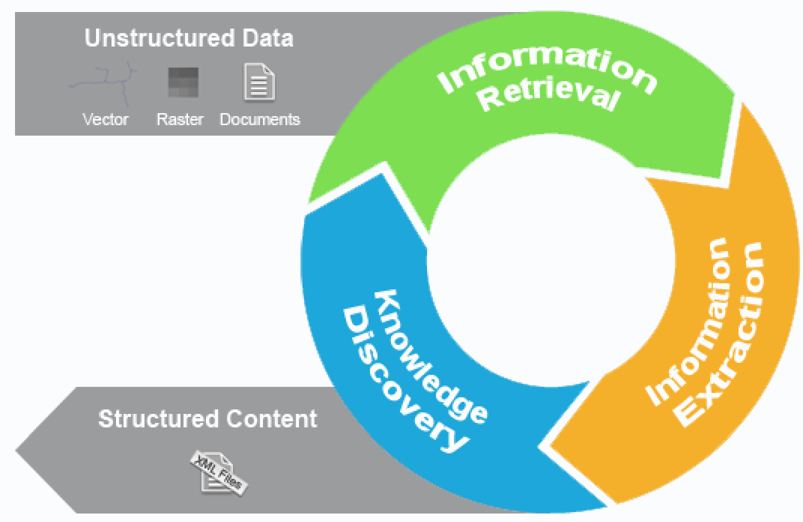
5.1. Geoportal Framework

A geoportal serves as the framework for the approach of integrating semantic text matching tools, as well as recommendations. One example for such a framework is the ESRI Geoportal Server [42], an open source implementation of a catalogue service with a highly customizable user interface. The ESRI portal enables management of spatial and non-spatial resources, as well as basic discovery mechanisms based on the Lucene index. Using the JavaServer Faces (JSF) framework, the ESRI Geoportal allows the integration of advanced discovery mechanisms using JavaScript. The layout of the standard Geoportal Server can be adapted using the Apache Struts Tiles framework, where a web application page is split into fragments that are assembled into the complete page at runtime [43]. The Geoportal Server is the technological basis for the EnerGEO Geoportal and further extensions that are presented in this paper.

5.2. (Semi-)Automatic Metadata Extraction Tool

As a basis for semantic similarity matching of texts, standardized and structured metadata content is needed. Since users in geoportals often tended to enter only a few datasets or services if they had to create their entries manually, a (semi-)automatic metadata extraction tool was developed. This tool extracts relevant information not only from spatial resources, but also from documents, such as scientific papers. Therefore, it transforms unstructured data to structured content as a basis for the generation of knowledge (Figure 5).

Figure 5. Data-knowledge circle.

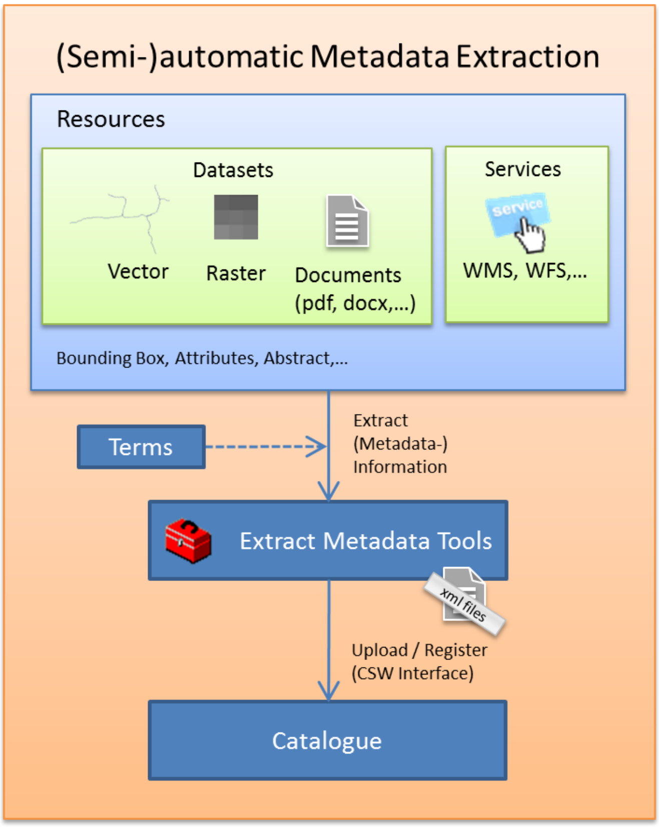


The current tool was written in the Python programming language and makes it possible to extract information from Portable Document Formats (pdf), Microsoft Word documents (doc, docx) and text documents (txt). For supporting pdf documents, we use the additional Python library, gfx [44]. For all other types of documents, we utilize the standard modules of Python, together with win32com [45]. If the documents are already tagged with metadata (such as author, abstract or creation date), this information is extracted and integrated into a standardized Dublin Core (DC) XML document.

For spatial resources, the following metadata standards are used: ISO 19110 (Feature Catalogue), ISO 19115, ISO 19119 and ISO 19139. Currently, metadata from all vector and raster data formats, as well as services of ESRI ArcGIS, can be automatically extracted by the ArcGIS Python module, ArcPy. It is based on the metadata managed in the ESRI ArcCatalog. Among these are folders that contain Shapefiles, Feature Classes (File Geodatabase, Personal Geodatabase), SDE Feature Classes, GRID and TIFF files. The tool allows the automatic deduction of bounding boxes, feature attribute lists (ISO 19110), an automatic linkage of ISO 19115 and ISO 19110 documents, abstracts (if present in the item description of ArcGIS), keywords (if present) and the path or link to the actual spatial dataset or service.

After extraction of the metadata, the structured information is automatically registered in the geoportal (see also Figure 6). We use the standardized OGC CSW (Catalogue Service Web 2.0.2) interface in order to automatically upload/register information. The whole process itself is considered as semi-automatic, since it cannot derive all information automatically that is needed to fulfill the ISO standards, as well as the complete EnerGEO metadata profile. In the case of quality information, the user must manually enter the data. As for non-spatial textual information, some content (e.g., terms) may be extracted by looking at word frequencies in combination with dictionaries of common words. In the EnerGEO Geoportal metadata extraction tool, the Python module, Topia Termextract [46], is therefore used.

Figure 6. (Semi-)automatic metadata extraction workflow.

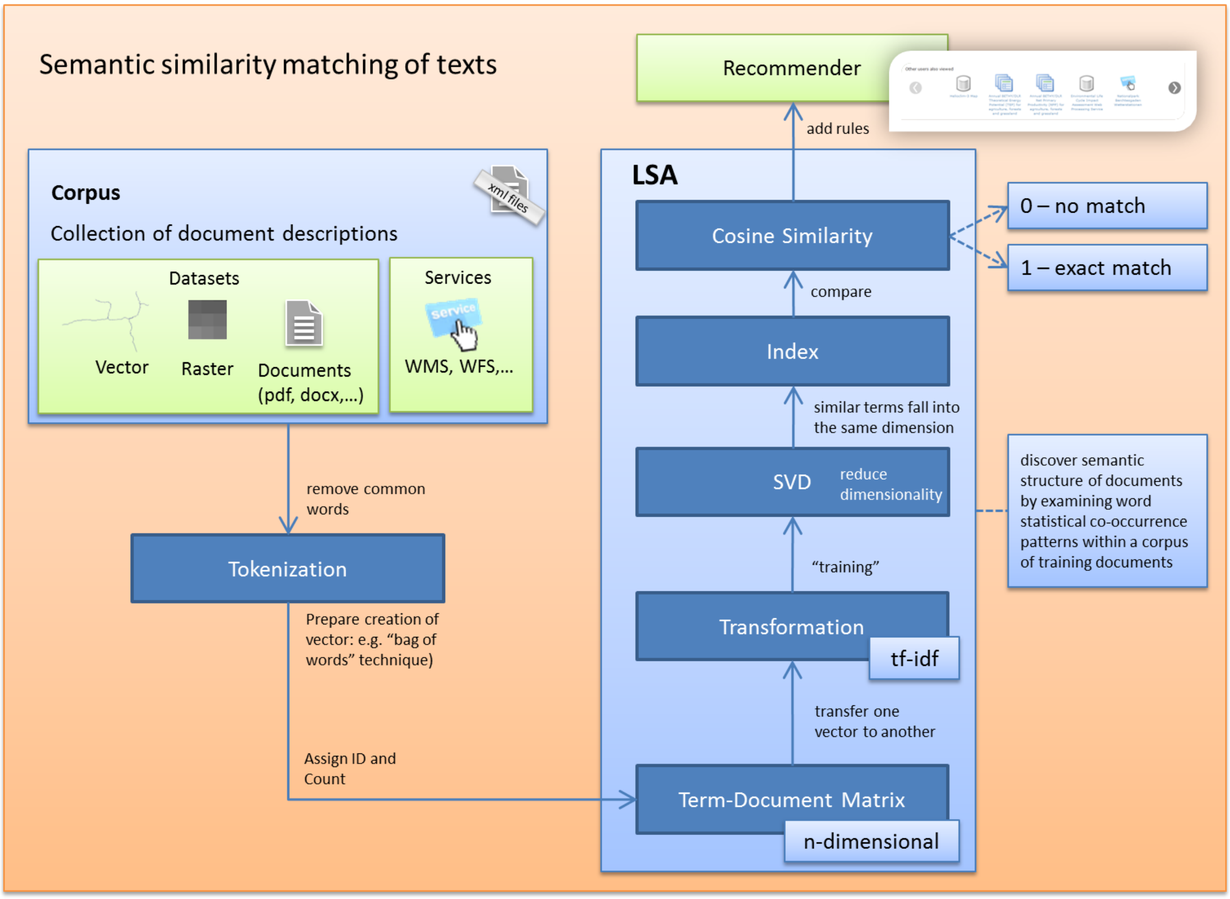

(Semi-)automatic Metadata Extraction
Resources
Datasets
Vector
Raster
Documents (pdf, docx,...)
Services
service
WMS, WFS,...
Bounding Box, Attributes, Abstract,...
Terms
Extract (Metadata-) Information
Extract Metadata Tools
xml files
Upload / Register (CSW Interface)
Catalogue

Putting resources into standardized formats offers many benefits. On the one hand, it makes content transparent and interchangeable between different metadata collections. On the other hand, it provides sections that can be compared with each other. For example, abstracts of documents can be matched with abstracts of spatial resources or data quality information of one item with the same kind of information of the other. This can be used to interlink the different resources.

5.3. Semantic Text Matching Tool

For the semantic text matching software tool, we use the Python modules gensim and simserver [47]. The reason for choosing these two modules is mainly due to the performance of the approach. Most calculations occur in the RAM of the computer. [48] states that the creation of the LSA model for the complete English Wikipedia took approximately four hours on a MacBook Pro (Intel Core i7 2.3 GHz, 16GB DDR3 RAM, OS X). Thus, gensim is able to process about 16,000 documents per minute (including all I/O) [48]. Figure 7 shows the overall implementation of the semantic similarity text matching method. It starts with a collection of documents (the so-called corpus), which need to be converted into a vector representation. The documents comprise both structured (e.g., spatial metadata for web services) and non-structured information (e.g., scientific papers). Within our work, the subsequent processes to be applied are the same for all kinds of information.

Figure 7. Workflow of semantic similarity matching of texts.


Semantic similarity matching of texts
Recommender
add rules
Corpus
Collection of document descriptions
xml files
Datasets
Vector
Raster
Documents (pdf, docx,...)
Services
service
WMS, WFS,...
remove common words
Tokenization
Prepare creation of vector: e.g. "bag of words" technique)
Assign ID and Count
LSA
Cosine Similarity
0 – no match
1 – exact match
compare
Index
similar terms fall into the same dimension
SVD
reduce dimensionality
discover semantic structure of documents by examining word statistical co-occurrence patterns within a corpus of training documents
"training"
Transformation
tf-idf
transfer one vector to another
Term-Document Matrix
n-dimensional

Before, a process called tokenization is required. This means that a full text is broken up into single words or meaningful concepts. It is also useful for removing common words, such as articles or prepositions, by using a list of stop-words. The vector itself is created through various techniques. One of the simplest is the so-called "bag of words" method. It is made up of question-answer-pairs. As an example, the question: "How many times does the word … appear in the document?" could be answered with "two times". Afterwards, each word is assigned an ID, as well as the count. It can be assumed that if the numbers in two vectors are similar, the documents are also likely to be similar, because the questions are the same for each document.

The result of the previous step is an n-dimensional vector space. To transfer one vector to another, a transformation, such as tf-idf or log-entropy, needs to be applied. Based on the amount of documents considered for the calculation, either tf-idf or log-entropy is used by the tool. As already stated, the goal of the transformation is to discover semantic relationships between words and to use them for describing the documents. Transformation is sometimes also referred to as "training of documents". For training purposes, either the documents that need to be compared with each other or a set of common documents coming, e.g., from Wikipedia can be used. Within the scope of this work, only the documents registered in the geoportal were used for training. This is because the used resources are restricted to the energy domain, where training with a large set of documents coming, e.g., from Wikipedia does not lead to better results [33].

Singular Value Decomposition (SVD) afterwards reduces the n-dimensional vector space to a lower dimensionality. This is necessary to discover the semantic structure of the documents by examining the statistical co-occurrence patterns of words, within a corpus of training documents [47]. It leads to similar terms falling into the same dimension. As a last step before applying the cosine similarity measure, the documents coming from the catalogue are indexed. Finally, the cosine similarity value indicates if there is an exact/high match (1) between two vectors or no match (0) at all, with possible degrees in between.

5.4. Recommender System

As a recommender system, we implemented the open source software product, easyrec [49]. Like the ESRI Geoportal Server, easyrec is a Java servlet. easyrec is mainly based on two algorithms: the Apriori algorithm R [50] and SlopeOne [51]. Both form the basis for the shopping cart analyzer called "Association Rule Miner (ARM)". Apriori is a learning algorithm for association rules between specific items. SlopeOne is a member of item-based collaborative filtering techniques. Collaborative filtering methods predict preferences of users based on the behavior of other users.

easyrec distinguishes between three different methods of user interaction: "view", "buy" and "rate". Within the implementation of easyrec in the ESRI Geoportal Server, clicks on the Apache Lucene-based search result list ("find" action in the domain of SDI) are considered as view actions (*cf*. Figure 3). A more detailed look at the full metadata document, as well as any preview or download ("use") of a resource is considered a buy action. A click on the "thumbs up" or "thumbs down", buttons are assumed to be rating actions. A more detailed description of the algorithms utilized by easyrec and its implementation for logging user clicks in the form of "view", "buy" and "rate" actions in the ESRI Geoportal Server can be found in Vockner *et al.* [11].

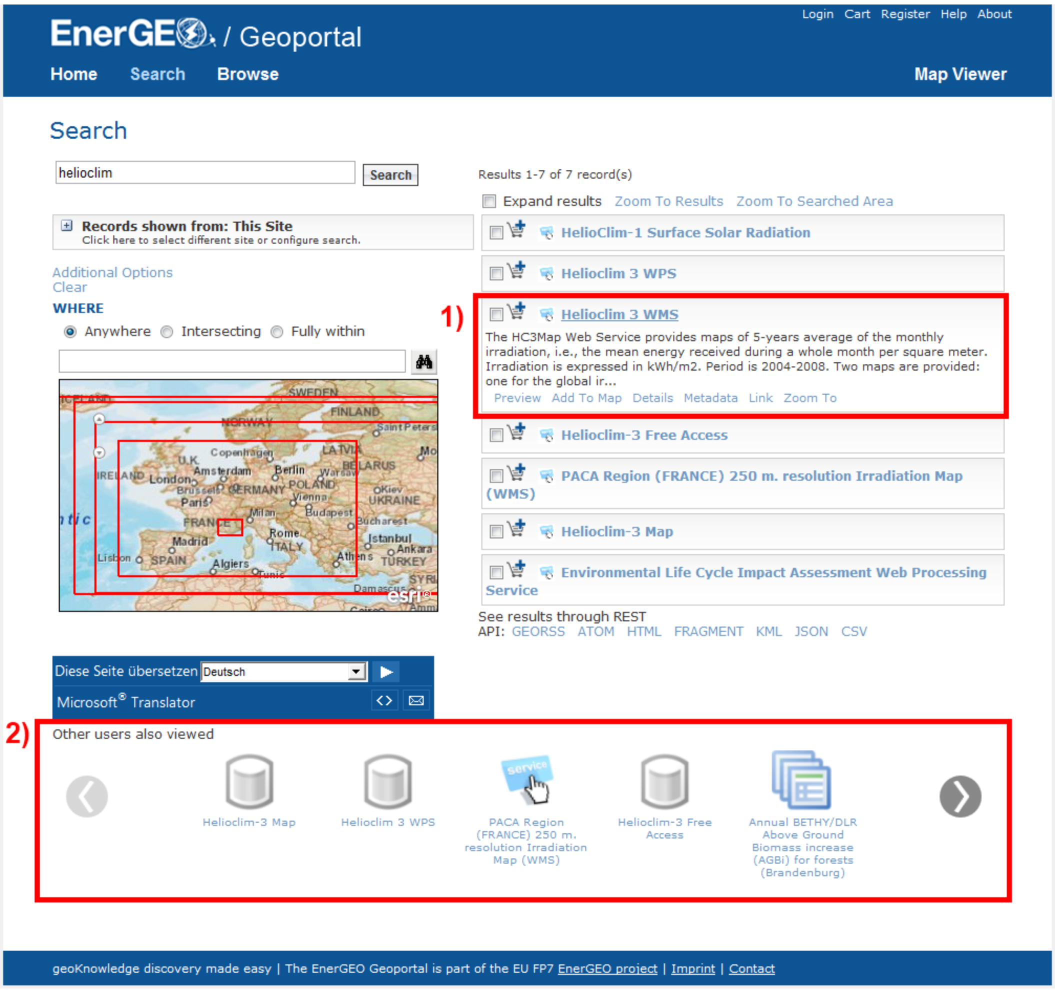
The major advantage of using easyrec within the context of semantic text matching algorithms is because it provides an API that is capable of receiving additional input for rule generation. The text matching values of two resources calculated by our tool are converted to percentage values and sent to the easyrec servlet. There, they are used for calculating recommendations.

6. Results

The final implementation result is a combination of the two compartments presented in the previous chapters. Figure 8 shows the result list of the recommendation engine integrated in the search page of the EnerGEO Geoportal in the form of an image carousel. If a user clicks on an item in the results list on the right (1), the recommendations linked with this item are presented in the section below (2). As already mentioned, the recommendations rely on calculations of user interactions, as well as the percentage compliance between the resources derived from the semantic text-matching tool. The links between various spatial and non-spatial resources of the energy domain are based on structured metadata information that is either manually entered or automatically extracted. The tool for automatically extracting content addresses the issue of having huge amounts of data in unstructured forms that contain valuable information, which is a fundamental input for processing and further analysis of the resources.

As the percentage of matches between sections of textual metadata (e.g., abstracts) is used to establish links between non-interrelated resources, a new method for resource discovery was added to the EnerGEO Geoportal.

Figure 8. Recommender-enhanced discovery in the EnerGEO Geoportal.



EnerGEO / Geoportal

Login Cart Register Help About

Home Search Browse Map Viewer

Search

helioclim [Search]

⊞ Records shown from: This Site
Click here to select different site or configure search.

Additional Options
Clear

WHERE

● Anywhere ○ Intersecting ○ Fully within

Results 1-7 of 7 record(s)

□ Expand results Zoom To Results Zoom To Searched Area

□ HelioClim-1 Surface Solar Radiation

□ Helioclim 3 WPS

1)

□ Helioclim 3 WMS
The HC3Map Web Service provides maps of 5-years average of the monthly irradiation, i.e., the mean energy received during a whole month per square meter. Irradiation is expressed in kWh/m2. Period is 2004-2008. Two maps are provided: one for the global ir...
Preview Add To Map Details Metadata Link Zoom To

□ Helioclim-3 Free Access

□ PACA Region (FRANCE) 250 m. resolution Irradiation Map (WMS)

□ Helioclim-3 Map

□ Environmental Life Cycle Impact Assessment Web Processing Service

See results through REST
API: GEORSS ATOM HTML FRAGMENT KML JSON CSV

Diese Seite übersetzen Deutsch ▶

Microsoft® Translator

2)

Other users also viewed

Helioclim-3 Map

Helioclim 3 WPS

PACA Region (FRANCE) 250 m. resolution Irradiation Map (WMS)

Helioclim-3 Free Access

Annual BETHY/DLR Above Ground Biomass increase (AGBi) for forests (Brandenburg)

geoKnowledge discovery made easy | The EnerGEO Geoportal is part of the EU FP7 EnerGEO project | Imprint | Contact

7. Outlook and Discussion

Consistent research activities have been dedicated to the improvement of information discovery within geoportals [9,13,52,53]. Various approaches make use of thesauri and ontologies. In our approach, we propose the usage of semantic text matching algorithms in combination with recommender systems to overcome problems arising from different meanings and usages of terms, especially due to the heterogeneous scientific backgrounds of the user groups within the energy domain. The semantic text-matching tool in combination with the recommender system, easyrec, is not a standalone solution to replace a general keyword-based search, but, rather, an approach to provide additional ordered results based on their similarity and other users' contexts. To validate our approach, we implemented these components in the EnerGEO Geoportal. Input data contains energy resources in spatial and non-spatial formats. Metadata of these resources can be (semi-)automatically extracted and used for further analysis using LSA.

All related work that we discovered either focused on the development of new algorithms or the application of these algorithms in other scientific domains. Within the domain of spatial data infrastructures and geoportals, we do not know of any implementation of a vector-based text matching method for improving information discovery. Thus, we present related work found in different scientific disciplines in the following.

Related work, such as *Omiotis* [34] coming from the field of bibliography, uses a thesaurus-based measure of text-relatedness. It is primarily an extension of the VSM with the *WordNet* Thesaurus. In the case of this paper, however, WordNet may not improve the VSM much, as it is not primarily constructed for energy related data. Thus, our preference is the LSA.

Nevertheless, we consider extending the current approach with thesauri or ontologies. For the latter, Ankolekar *et al.* [54] state that extension of text similarity measures with ontologies may lead to issues, for instance, that the semantic knowledge encoded in ontologies does not correspond to the concepts significant for text classification. Another question is how to integrate the relatively strict concept of ontologies in the fuzzy-oriented semantic text matching approach used by us.

Mihalcea *et al.* [16] present a method for corpus-based and knowledge-based measures of semantic text similarity. It is especially suited for short texts. Compared to vector-based similarity metrics, experiments show that their method reduces the error rate by up to 13 percent [16]. The focus of this paper is not on the development or improvement of the semantic text matching algorithms themselves, but on the application of algorithms in order to show the possibilities of semantic text matching as a means to generate knowledge out of information by interlinking resources. Therefore, the most appropriate solution for our use case, namely LSA, was chosen to be implemented in the EnerGEO Geoportal as a first step. In further stages of the implementation, algorithms better suited for relatively short texts will be used und evaluated.

A major advantage of the application of LSA is that it addresses the cross-language information retrieval problem that occurs if the search queries are in a different language than the resource language. As all documents and paraphrases are transformed to vectors, they can be compared by application of the specialized extension called Cross-Language LSA (CL-LSA) [23]. Cimiano *et al.* [33] have evaluated cross-language text matching algorithms in the English and French language. As already presented, their results show that ESA outperforms LSA or LDA, except for domain-specific training documents, like we have in our case [33]. LSA is clearly superior to both LDA and ESA when it is trained on the retrieval documents themselves [33].

A possible drawback of the proposed method is that recommendations may not coincide with the users' expectations. Issues causing this are usually small amounts of users or small amounts of resources registered in the catalogue. The first issue relates to the fact that a recommender system is based on user actions to calculate rules for recommendations. If only a small amount of resources exists, too few items are present to provide recommendations to the users. In that case, the recommendation list remains empty, leading to dissatisfaction of the user.

In addition, users may receive inappropriate recommendations when they come from domains other than the one domain-specific portal that was created. As recommendations are calculated based on what other users clicked, interests might differ.

Another issue may arise when having texts of different lengths. In the cases of very short texts, matching is only possible based on a few words. The longer the texts, the better the algorithms can be used to derive related documents.

To validate the quality of text matching within recommendations, we use backtracking mechanisms of user interactions in the recommendation list at first sight. Based on the experience of [16,33,35,37,39], we discovered the suitability of the LSA algorithm regarding our specific needs of comparing the semantic content of metadata. In addition, we conducted different user tests to evaluate the quality of discovery and recommendations. The recommendations revealed promising results to our internal experts group. This outcome encourages us to start with enhanced user experience tests as a next research step.

Hence, a survey will be conducted, where the users' experience with the combination of LSA and recommender systems in Geoportals will be quantitatively evaluated, leveraging enhanced web analytics tools, such as Piwik [55] and mouse-tracking analysis (e.g., creating mouse-pointer heatmaps). This test will be accompanied by an online survey, querying the qualitative users' experience.

8. Conclusion

As the main result of the presented work, we propose the integration of the semantic text matching algorithms and recommender systems for enhancing metadata discovery quality and user experience in geoportals. Therefore, we developed a tool for (semi-)automatic metadata extraction from spatial and non-spatial content to generate location-aware knowledge. Structured, standardized metadata serves as the input for semantic text matching of content using LSA. With this new approach, links can be automatically created between resources that were not interrelated before. These quantitatively established links are presented using recommendations on the contextual similarity of texts. Additionally analyzed user interactions on the geoportal discovery interface further enhance the recommendation rankings. Thus, the EU FP7 EnerGEO Geoportal shows the results of our research as proof-of-concept. It presents discovery results that are not inherent in the data itself, but rather derived from the context in form of textual similarity and what other users viewed.

Acknowledgments

We implemented the recommender and semantic text matching extension as part of the EnerGEO Geoportal. The EU FP7 project EnerGEO is the European contribution to the energy theme of GEOSS. The aim of this project is to develop a strategy for a global assessment of the current and future impact of the exploitation of energy resources on the environment and ecosystems and to demonstrate this strategy for a variety of energy resources worldwide. Therefore, the goal is to provide resources in order to generate domain-specific knowledge by integrating resources into the common architecture, as well as making these resources easily discoverable and accessible.

References

1. Rouse, M. *Information Society. IT Standards and Organizations Glossary*. Available online: http://whatis.techtarget.com/definition/Information-Society (accessed on 19 January 2013).

2. Gantz, J.; Reinsel, D. *Extracting Value from Chaos*. Available online: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (accessed on 17 January 2011).
3. Bellinger, G. *Knowledge Management—Emerging Perspectives*. Available online: http://www.systems-thinking.org/kmgmt/kmgmt.htm (accessed on 11 January 2013).
4. Richter, A. Enterprise Spatial Information & Knowledge Infrastructures—Concepts and Technologies for the Oil & Gas Business Using the Example of OMV. M.S. Thesis, University of Salzburg, Salzburg, Austria, 2012.
5. SAS. *Text Analytics—Contextual Intelligence*. Available online: http://www.eu.gov.hk/sc_chi/cmps/files/cmps_20100125_1210_sas.pdf (accessed on 1 May 2012).
6. Croisier, S. The Rise of Semantic-Aware Applications. In *Semantic Technologies in Content Management Systems*; Maass, W., Kowatsch, T., Eds.; Springer: Berlin, Germany, 2012; pp. 23–33.
7. Tang, W.; Selwood, J. *Spatial Portals. Gateways to Geographic Information*; ESRI Press: Redlands, CA, USA, 2005.
8. Tang, W.; Selwood, J. *Spatial Portals. Adding Value to Spatial Data Infrastructures*. Available online: http://www.isprs.org/proceedings/XXXVI/4-W6/papers/35-40WinnieTang-A022.pdf (accessed on 21 July 2012).
9. Fugazza, C.; Luraschi, G. Semantics-aware indexing of geospatial resources based on multilingual thesauri: Methodology and preliminary results. *Int. J. Spat. Data Infrastruct. Res.* **2012**, *7*, 16–37.
10. Smits, P.C.; Friis-Christensen, A. Resource discovery in a European spatial data infrastructure. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 85–95.
11. Vockner, B.; Belgiu, M.; Mittlböck, M. Recommender-based enhancement of discovery in Geoportals. *Int. J. Spat. Data Infrastruct. Res.* **2012**, *7*, 441–463.
12. Latre, M.A.; Hofer, B.; Lacasta, J.; Nogueras-Iso, J. The Development and interlinkage of a drought vocabulary in the EuroGEOSS interoperable catalogue infrastructure. *Int. J. Spat. Data Infrastruct. Res.* **2012**, *7*, 225–248.
13. Janowicz, K.; Schwarz, M.; Wilkes, M. Implementation and Evaluation of a Semantics-Based User Interface for Web Gazetteers. In Proceedings of Visual Interfaces to the Social and the Semantic Web (VISSW 2009) Workshop in Conjunction with the International Conference on Intelligent User Interfaces (IUI 2009), Sanibel Island, FL, USA, 8–11 February 2009.
14. Scholz, J.; Mittlböck, M. Spatio-Temporal Visualization of Simulation Results Using a Task-Oriented Tile-Based Design-Metaphor. In *Service Oriented Mapping 2012*; Jobst, M., Ed.; Jobsstmedia Management Verlag: Vienna, Austria, 2012; pp. 369–382.
15. Islam, A.; Inkpen, D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data* **2008**, *2*, 1–25.
16. Mihalcea, R.; Corley, C.; Strapparava, C. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. In Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 July 2006; Volume 1, pp. 775–780.
17. Turney, P.D.; Pantel, P. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.* **2010**, *37*, 141–188.

18. Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Deerwester, S.; Harshman, R. Using Latent Semantic Analysis to Improve Access to Textual Information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Washington, DC, USA, 15–19 May 1988; pp. 281–285.
19. Deerwester, S.; Dumais, S.; Landauer, T.; Furnas, G.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407.
20. Maas, A.; Daly, R.; Pham, P.; Huang, D.; Ng, A.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
21. Wiemer-Hastings, P. How Latent is Latent Semantic Analysis? In Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 31 July–6 August 1999; Volume 2, pp. 932–937.
22. Landauer, T.; Laham, D.; Rehder, B.; Schreiner, M.E. How Well can Passage Meaning be Derived without Using Word Order. In Proceedings of the 19th Annual Meeting of the Cognitive Science Society, Palo Alto, CA, USA, 7–10 August 1997.
23. Dumais, S. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2004**, *38*, 188–230.
24. Wicijowski, J.; Ziolko, B. Extracting Semantic Knowledge from Wikipedia. In *Intelligent Information Systems: New Approaches*; Klopotek, M.A., Ed.; Publishing House of University of Podlasie: Podlasie, Poland, 2011; pp. 91–98.
25. Nakov, P.; Popova, A.; Mateev, P. Weight Functions Impact on LSA Performance. In Proceedings of the EuroConference Recent Advances in Natural Language Processing (RANLP'01), Tzigov Chark, Bulgaria, 5–7 September 2001; pp. 187–193.
26. Terzi, M.; Ferrario, M.-A.; Whittle, J. Free Text in User Reviews: Their Role in Recommender Systems. In Proceedings of the 3rd ACM RecSys'10 Workshop on Recommender Systems and the Social Web, Chicago, IL, USA, 23–27 October 2011; pp. 45–48.
27. Li, Y.; McLean, D.; Bandar, Z.; O'Shea, J.; Crockett, K. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1138–1150.
28. Turney, P.D. Mining the Web for Synonyms: PMI-IR *versus* LSA on TOEFL. In Proceedings of the 12th European Conference on Machine Learning, Freiburg, Germany, 5–7 September 2001; pp. 491–502.
29. Gabrilovich, E.; Markovitch, S. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hydrabad, India, 6–12 January 2007; pp. 1606–1611.
30. Sorg, P.; Cimiano, P. Cross-Lingual Information Retrieval with Explicit Semantic Analysis. In Proceedings of Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, 17–19 September 2008.
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. L'Huillier, G.; Hevia, A.; Weber, R.; R ós, S.A. Latent Semantic Analysis and Keyword Extraction for Phishing Classification. In Proceedings of 2010 IEEE International Conference on Intelligence and Security Informatics (ISI), Vancouver, BC, Canada, 23–26 May 2010; pp. 129–131.

33. Cimiano, P.; Schultz, A.; Sizov, S.; Sorg, P.; Staab, S. Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. In Proceedings of the 21st International Joint Conference on Artifical Intelligence, Pasadena, CA, USA, 11–17 July 2009; pp. 1513–1518.
34. Tsatsaronis, G.; Varlamis, I.; Vazirgiannis, M.; Norvag, K. Omiotis: A Thesaurus-Based Measure of Text Relatedness. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, Bled, Slovenia, 7–11 September 2009; pp. 742–745.
35. Tsatsaronis, G.; Varlamis, I.; Vazirgiannis, M. Text relatedness based on a word thesaurus. *J. Artif. Int. Res.* **2010**, *37*, 1–40.
36. Lee, M.D.; Pincombe, B.M.; Welsh, M.B. An Empirical Evaluation of Models of Text Document Similarity. In Proceedings of the XXVII Annual Conference of the Cognitive Science Society, Stresa, Italy, 21–23 July 2005; pp. 1254–1259.
37. Ramage, D.; Rafferty, A.N.; Manning, C.D. Random Walks for Text Semantic Similarity. In Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing, Suntec, Singapore, 7 August 2009; pp. 23–31.
38. Dolan, B.; Quirk, C.; Brockett, C. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004; p. 350.
39. Mohler, M.; Mihalcea, R. Text-to-Text Semantic Similarity for Automatic Short Answer Grading. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 20 March–3 April 2009; pp. 567–575.
40. Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; Riedl, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, Chapel Hill, NC, USA, 22–26 October 1994; pp. 175–186.
41. Pazzani, M.J. A framework for collaborative, content-based and demographic filtering. *Artif. Intell. Rev.* **1999**, *13*, 393–408.
42. *ESRI Geoportal Server*. Available online: http://www.esri.com/software/arcgis/geoportal (accessed on 16 March 2013).
43. *Apache Software Foundation. Apache Tiles*. Available online: http://tiles.apache.org (accessed on 20 August 2012).
44. *gfx*. Available online: http://www.swftools.org/gfx_tutorial.html (accessed on 10 September 2012).
45. *Win32com*. Available online: http://starship.python.net/~skippy/win32/Downloads.html (accessed on 10 September 2012).
46. *Topia Termextract*. Available online: http://pypi.python.org/pypi/topia.termextract/ (accessed on 12 September 2012).
47. Rehurek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 17–23 May 2010; pp. 45–50.
48. Rehurek, R. *Experiments with the English Wikipedia*. Available online: http://radimrehurek.com/gensim/wiki.html (accessed on 15 February 2013).
49. *easyrec*. Available online: http://www.easyrec.org (accessed on 15 March 2012).

50. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, 12–15 September 1994; pp. 487–499.
51. Lemire, D.; Maclachlan, A. Slope One Predictors for Online Rating-Based Collaborative Filtering. In Proceedings of the 2005 SIAM International Conference on Data Mining (SDM'05), Newport Beach, CA, USA, 21–23 April 2007.
52. Abargues, C.; Granell, C.; Díaz, L.; Huerta, J.; Beltran, A. Discovery of User-Generated Geographic Data Using Web Search Engines. In *Advances in Geoscience and Remote Sensing*; Jedlovec, G., Ed.; InTech: Rijeka, Croatia, 2009.
53. Pearlman, J.; Craglia, M.; Bertrand, F.; Nativi, S.; Gaigalas, G.; Dubois, G.; Niemeyer, S.; Fritz, S. EuroGEOSS: An Interdisciplinary Approach to Research and Applications for Forestry, Biodiversity and Drought. In Proceedings of the 34th International Symposium on Remote Sensing of Environment, Sydney, Australia, 10–15 April 2011; pp. 1–4.
54. Ankolekar, A.; Seo, Y.W.; Sycara, K. Investigating Semantic Knowledge for Text Learning. In Proceedings of ACM SIGIR Workshop on Semantic Web, Toronto, ON, Canada, 28 July–1 August 2003.
55. *Piwik*. Available online: http://piwik.org/ (accessed on 10 February 2013).