Improving Search in Tele-Lecturing: Using Folksonomies as Trigger to Query Semantic Datasets to Extract Additional Metadata

Franka Moritz, Maria Siebert and Christoph Meinel
Hasso-Plattner-Institute
University of Potsdam
Potsdam, Germany
franka.moritz|maria.siebert|christoph.meinel@hpi.uni-potsdam.de

ABSTRACT

Tele-teaching, where recorded lectures are streamed via the internet, was identified as the easiest adoptable method to produce large amounts of e-learning content. Due to the nature of the e-learning material produced, it is lacking searchable data and metadata. Social Web technologies have been identified as one way to overcome this problem, because metadata will be generated by users. But still, the amount of metadata generated that way is not sufficient and the contextual information is missing. This information can be extracted from the Semantic Web, especially from Linked Data initiatives. This paper describes a workflow that utilizes the metadata generated by users to trigger the query of semantic datasets. These datasets are providing additional metadata, that can be extracted via an interface, which is publicly available. Thereby it is possible to supply new search strategies and achieve an extension of the search space for multimedia e-learning data. An extension of an existing search functionality and the similarity detection is finally described.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Information Storage and Retrieval—*On-line Information Services*; K.3.1 [**Computing Milieux**]: Computer in Education—*Computer Uses in Education*

General Terms

Community Tags, Folksonomies, Linked Data, Semantic Web, Metadata, Tele-Teaching

1. INTRODUCTION

The internet era has opened up a whole new world of knowledge. For learners this is an enormous opportunity. But at the same time it is also a burden, because the choice of information is too large and the learner faces the problem, that he must decide which source to choose, which material to look at in detail and simply which topics really belong in the field he would like to study. This problem not only exists within the internet, but also in limited e-learning environments. Especially multimedia e-learning data is a challenge, because it is not yet easily possible to search through audio and video files. So far, textual information in different forms is required to be able to search amongst content. One approach to overcome this issue for multimedia files is the enrichment with metadata. Metadata is information that is connected with a content item, explaining details the content itself does not reveal.

Tele-teaching is the most widely adopted method of e-learning. This is the case, because large amounts of content can be created without much effort. But still, the lecture recordings are created manually and metadata is just as well inserted manually. Because time and money are mostly an issue, the quality of the metadata is usually not very good and not in the least sufficient for proper search activities.

Since the era of Web 2.0, it was found out that Social Web technologies are one method to improve the metadata base. In many fields users are happily participating in the creation of adequate metadata. But this only solves that a single content item can be found more easily. It is still not obvious, which topics belong together and what is the context of the content item. Semantic Web activities have been researched for quite some time in order to overcome this restriction. Wikipedia and its semantic counterpart DBpedia are one example how semi-structured user-generated data can be transformed to structured semantic data. By crawling the text published on Wikipedia, connections between content items and the metadata to a content items are gathered and stored in the DBpedia in a structured manner by using Semantic Web technologies. Therefore the Wikipedia also becomes machine-readable.

This paper will first give an introduction into tele-teaching and Social Web activities in tele-teaching. Why those technologies can be utilized to query further contextual information is explained. Second, the idea of the Semantic Web is introduced in more detail. Our idea how to utilize the Linked Data initiative within the Semantic Web in order to query contextual information for user-generated keywords is described afterwards. Finally, the utility of this contextual information for application in the tele-teaching context is explicated.

WIMS'11 May 25-27, 2011 Sogndal, Norway

2. TELE-LECTURING TODAY

Within the long history of e-learning, several approaches to create content have been tried. Because tele-lecturing facilitates the production and distribution of large amounts of content, this approach is one of the most widespread methods for e-learning. The Hasso-Plattner-Institute (HPI) in Potsdam has developed its own tele-lecturing solution, tele-TASK [13], which will be explained in the next paragraph. Current perspectives and drawbacks of this approach will also be described.

2.1 Tele-Lecturing with tele-TASK

The tele-Teaching Anywhere Solution Kit [13], short tele-TASK, is an e-learning project at the chair Internet-Technologies and -Systems at the HPI. The tele-TASK project was started in 2002 at the university of Trier by developing a hardware system for lecture recording. An all-in-one solution was developed including hard- and software for lecture recording. It is a plug-and-play solution. Two video steams (a video of the lecturer and screen capturing of his laptop or a smart-board) and one audio stream can be recorded at once.

More than 3500 lectures and 8800 podcasts of the tele-TASK archive can be accessed free of charge via the tele-TASK web portal (www.tele-task.de) or portable device. The large video archive and the web-platform tele-TASK are the basis for further research and development at the HPI. Perspectives and drawbacks that could be experienced with this project will be explained now.

2.2 Perspectives and Drawbacks

The problem, how to easily develop and distribute large amounts of e-learning content, is in fact solved with the help of tele-lecturing. Nevertheless, manpower is still required to record the lectures and work on the post production process including the creation of applicable metadata. This manual work is still a very time-consuming process. But, without being enriched with additional metadata, like the title, the date of recording, the lecturer and a description, it will not be possible to find the recorded lecture within a large number of other recordings. Other sources for metadata in a tele-lecturing environment will be explained in this paragraph.

2.2.1 *Sources for Keywords in a Tele-Lecturing Environment*

There are different modules of the sample tele-teaching application tele-TASK that can be the source for content-related keywords. Also the core plug-in can provide keywords, for example those keywords retrieved from the manually inserted title or description information. Another module, that has knowledge about important keywords in the portal, is the search module. It knows about keywords that users are searching for and can therefore supply these keywords to the proposed *dbpedia* module in order to find out the semantic context about the keywords.

Several approaches to automatically harvest metadata from the lecture recording have been explored in the past and are still being developed. The basis for the automatic metadata harvesting are the audio and video data of the recording. Optical character recognition and audio transcription are two approaches to gather metadata with the help of machines and not manpower.

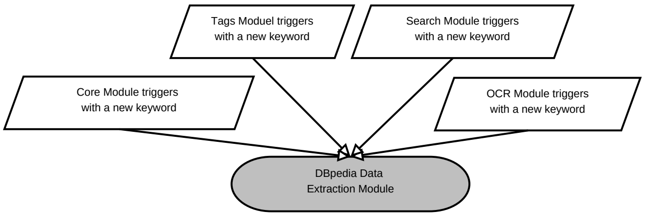
Tags Moduel triggers with a new keyword
Search Module triggers with a new keyword
Core Module triggers with a new keyword
OCR Module triggers with a new keyword
DBpedia Data Extraction Module

Figure 1: Different Modules May Trigger The Extraction of Further Metadata

The OCR plug-in can extract and aggregate keywords directly from the presenters slides and the audio plugin-in will gather spoken words and aggregate keywords from the audio transcript of the lecturers speech. All these keywords, collected by different modules in the tele-teaching portal, will be provided to the *dbpedia* module like proposed in this paper (see figure 1).

In order to receive valuable content-related keywords from the modules described in this paragraph, algorithms to aggregate the most essential information from OCR and audio data have to be applied and actual keywords need to be extracted from the aggregated data at first. The same process needs to be gone through for keywords that should be extracted from the title or description information. The keywords received from the search module are not related to any specific content item. Therefore those are less interesting for the improvement of the search and browsing options.

Next to automatic metadata harvesting and the content-unrelated keywords researchers are also exploring ways to include users in the metadata creation process. This approach will be explained in the following paragraph. The module we are focusing on as metadata source in this paper specifically is the tagging plug-in, which will be described in more detail afterwards.

2.2.2 *Community and Social Web Functionalities in Tele-Lecturing Scenarios*

Since the beginning of the Web 2.0 era [11] numerous Social Web portals, whose main motivation is fostered around the user participation, have evolved and grew very quickly. A number of Social Web and community features have been found to be useful to the users. These include blogging, the collaborate creation of wikis, social annotating and tagging, evaluating (eg. rating and commenting), recommending, content sharing and linking of content items [11]. The key success factor for these features is the interaction amongst users [8].

That community functionalities are not only useful for networking, but also for learning contexts was found out at the beginning of the e-learning era around 2000 already [10, 12]. But only recently research started on joining tele-lecturing with community functionalities. During the workshop *eLectures 2009* at the conference DeLFI 2009 [17] an approach of integrating tele-lecturing applications into facebook, a combination of wikis and tele-lecturing and other social e-learning approaches were shown. This paper will evolve around utilizing a tagging functionality to improve the metadata base in the tele-lecturing scenario. Therefore the tagging functionality will be introduced separately in the next chapter in order to fully describe its specific utility as trigger for the extraction of semantic data.

2.2.3 Collaborative Tagging

Tagging is the assignment of keywords, the so called tags, to a resource. In collaborative tagging, where people are taking over the role of assigning tags to resources, this keyword can be chosen according to the experiences with the resource and the perception of the person choosing the keyword, the tagger. The new aspect thereby is, that not a librarian or an administrator chooses those tags, but all users of the community are allowed to participate in the tagging activity.

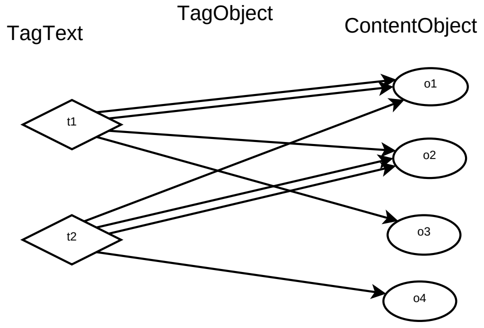

Figure 2: Relations between tags and content objects

Each keyword is meant to be one aspect of the resource. Tags therefore help to summarize and classify the content of a resource. [1, 5] In collaborative tagging, the users can tag freely without limitations. That means that they can choose any vocabulary they can think of including compound words and a combination of several words. This enables the user to categorize content according to his personal strategy. The compound of all tags collected via collaborative tagging is finally called folksonomy, although there is some discussion about the correctness of the term [5]. A folksonomy is formally defined as a quadruple $F := (U, T, R, Y)$, whereby U, T and R are finite sets, whose elements are users, tags and resources. Y represents the tag assignment, a relation between the three other sets, which can formally be described as $Y \subseteq U \times T \times R$ [9]. This also means that one content object might have several tags assigned to it and one tag may be allocated to multiple content objects (see figure 2).

Tagging is a community feature, that has been vastly deployed and is very popular amongst social web sites. One category of tags are content-related tags. Those are the most interesting ones for an extension of the metadata base of the tele-lecturing portal. This is the case, because these really create added value by creating keywords that are related to the content of the video and therefore help contextualizing it.

Because tagging is the community feature that creates content-related keywords, and the only module in the telelecturing portal where keywords are directly created (as explained in section 2.2.1), it is the ideal candidate to be further utilized for browsing, search and recommendation purposes. But even tags lack one very important characteristic for the just mentioned tasks: the context. Yes, they are connected to one content item, but they are lacking the greater context, the relation to other tags and a global coherence. A research area, that has been in the focus of researchers for quite a while, and that deals with this topic of contextualization, is the Semantic Web. The next chapter will give an introduction into the topic and explain our approach to combine the three fields Semantic Web, Social Web and e-learning .

3. RETRIEVING CONTEXTUAL AND METADATA INFORMATION FOR COMMUNITY TAGS

The connection of Social and Semantic Web has been researched in papers already [6]. It was especially identified, that tagging data provides valuable information that serves to bridge the gap between unordered social data and structured data [7]. The following paragraph will give an introduction into the idea of the Semantic Web, the Linked Data initiative and one of its projects - the DBpedia. Afterwards, an overview of current research, to combine the Social and Semantic Web with e-learning, will be given. Finally, our own solution, to use the tagging data acquired in an e-learning portal in order to trigger the retrieval of additional metadata and semantic information, will be described.

3.1 Semantic Web and Linked Data

There has been and still is the vision that all the knowledge existing in the world can be united and joined together. This, so the vision, would lead to machines being able to answer questions for mankind that human knowledge alone would not be able to answer, simply by combining common knowledge from different fields and thereby creating new knowledge. Ultimately machines should be able to answer semantically comprehensive queries. This inspiration has lead to decades of research fostered around that problem and finally the idea of the Semantic Web was born [2, 3].

In order to be able to have many researchers following up on the vision, experimenting with semantic data and trying out different methods to extract and reason over the data provided, a dataset to start with was required. Several research groups together found the Linked Data initiative, that has to goal to provide exactly that start dataset. Linked Data is a technique to publish resources and their descriptions on the web. It is based on two main technologies: the resource description framework (RDF) and uniform resource identifiers (URIs). The URI is the address under which the descriptions of the resource, that is identified by that URI, can be found. The description itself is presented in the RDF format and contains information about the resource as well as links to related resources. URIs will deliver the required format; when accessed by a Semantic Web agent they will return RDF data and when approached by a standard web browser an HTML representation of the same information will be returned. [3]

3.2 From Community-generated to Semantic Community-generated Metadata

One project within the Linked Data initiative is DBpedia. Their goal is to convert the content of the very popular user-generated online-encyclopaedia Wikipedia into structured knowledge in order to enable the data to be accessed with Semantic Web methods.

Two different techniques are applied by the DBpedia operators in order to transfer the user-generated metadata to

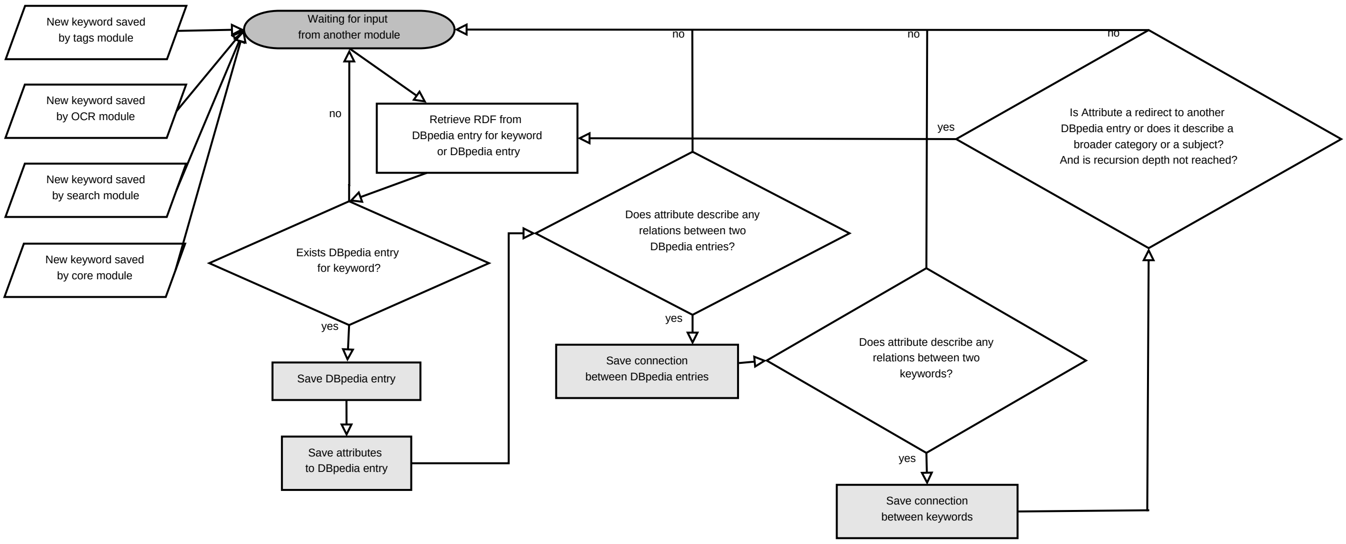
New keyword saved by tags module
New keyword saved by OCR module
New keyword saved by search module
New keyword saved by core module
Waiting for input from another module
Retrieve RDF from DBpedia entry for keyword or DBpedia entry
no
Exists DBpedia entry for keyword?
yes
Save DBpedia entry
Save attributes to DBpedia entry
Does attribute describe any relations between two DBpedia entries?
yes
Save connection between DBpedia entries
Does attribute describe any relations between two keywords?
yes
Save connection between keywords
Is Attribute a redirect to another DBpedia entry or does it describe a broader category or a subject? And is recursion depth not reached?
yes
no
no
no

Figure 3: Flow diagram of the *dbpedia* module that retrieves data from DBpedia for a given keyword and looks up connections between keywords

semantic user-generated metadata. Both are based on regular dumps of the database tables from Wikipedia. First of all, relationships between different entries from the database are mapped onto RDF. Second, further information, that is displayed in the article and info boxes, is withdrawn directly from the text on the site. [3]

With the help of this initiative the community effort that was put into creating the online encyclopaedia Wikipedia becomes accessible by machines. So far only users were able to navigate through the vast knowledge base of Wikipedia. Now that all the data exists in a structured manner, other applications can access this knowledge base via the interfaces provided. The raw data can be retrieved in the formats CSV, RDF or OData.

One area of application, where the utilization of the semantic community-generated metadata is really beneficial, is e-learning. The next chapter will therefore give an introduction into some related work connecting the Social and the Semantic Web with e-learning.

3.3 Related Work

To connect the Social and Semantic Web with e-learning was also the center of some research attention in the last years. Several different ideas fostered around this area were tried out. This section will give an overview about related work before the next section will go into more detail explaining our approach to the field.

One method that was tried was to engage the users into creating a semantic context to existing content. This was achieved by asking users to also add the relations between tags they assign, which is part of the semantic context, manually. The reason for the users to take on this activity is similar as for tagging itself - enable oneself, by contributing to the metadata of the content, to receive more exact search results. The evaluation showed however, that only a very minor percentage of users participated in this activity. Therefore this approach was not considered to be successful. [4]

Other researchers described a technique to suggest related tags to an ontology concept. Thereby teachers should be supported in creating domain ontologies for their subject. The tags, that students attach to content items, serve as basis for a content-based relatedness algorithm that computes, if the new tag would fit somewhere in the domain ontology. The approach was validated as being very promising. [16]

A semantic search interface for an academic video search engine is proposed in [18]. Thereby the search query is forwarded to DBpedia in order to build up an exploratory search interface. In the interface semantic connections of the search term are suggested to the user in order to enable free browsing and serendipitous search results.

Our aim however is not to build on user or administrator activity to create additional metadata and semantic information, like the first two approaches. Like the third approach, we would like to trigger an automatic extraction of semantic information. But we suggest to convert this information for storage in our own database, mapped to our own data structure, so that different modules in our portal may utilize it. Furthermore we want to make the most of the metadata that we already have available in our system. Therefore, we are suggesting a method that leverages keywords that are inserted into our system, either by information retrieval or user participation. Those keywords will trigger an automatic extraction of supplementary data from a Linked Data set. The next paragraph will explain our approach in detail.

3.4 A Semantic Information Retrieval Module - Its Triggers and Workflow

Our semantic information retrieval module in the tele-TASK portal is based on two things. First, the knowledge, that there are about 2 million resources described in the DBpedia dataset. Each of them is identified by a unique uniform resource identifier (URI), which is presented in the form http://dbpedia.org/resource/Keyword [3]. Second, the foundation that we regularly do have new keywords that are generated from diverse sources.

Once a new keyword was saved in the system, this keyword will trigger the first function within the *dbpedia* module that will extract related metadata from the DBpedia. The first function tries to call http://dbpedia.org/data/Keyword.rdf, the URI, by inserting the new keyword at the end of the URI. For compound words the space will be replaced by underscore. This URI will directly return all information existing for that entity in RDF format. If this URI is available and valid data is returned from it, the next step is the parsing of all rdf triples and the extraction of required data. The flow diagram in figure 3 visualizes those and the following steps, that are processed in the *dbpedia* module.

First, the new DBpedia entity with its keyword, URI and resource type will be stored in the database. There are two varying entity types that need to be distinguished - resources and categories. They can be differentiated by their URI. The resource entities have the before mentioned URIs whereas for the category types the keyword in the URI is replaced by *Category:Keyword*. It is reasonable to distinguish between those two, because both URIs can exist for one keyword, the resource that includes all properties of the entity and the category that includes connections between the content items. The first step in further recursions of this module is therefore to differentiate between those different entity types and store them in the database. In the first recursion the standard URI is called, therefore one already knows that the entity type is a resource.

Next, the different attributes of the entity will be extracted and stored in the database. The attributes may have the properties *language*, *type* and *value*. It is of special interest for us to retrieve connections between the keywords we are extracting from DBpedia. This information is revealed by the attribute *type*. We built the system extensible. At the moment we only obtain predefined attribute types. The source entity is the entity whose attributes we are currently extracting and the target entity is represented in the value of the attribute. Those connections between DBpedia entities are stored in the database separately in order to be able to reuse them later on.

At this stage the *dbpedia* module can actually return information to the module that triggered the start of the whole extraction process and handed over the keyword to the *dbpedia* module. In our example, the tagging application delivered the keyword to the *dbpedia* module. For searching it is also interesting to see relations between tags. Therefore the *dbpedia* module looks if the connections between two DBpedia entities can be conveyed to connections between tags. Due to the recursivity of the *dbpedia* module this will not always be possible, as a lot more DBpedia entities will have been extracted than tags are available. Furthermore, it is also not desirable to ascribe the keywords extracted from DBpedia back to the tags, as tags are by definition solely produced by users. But still, it is possible to store connections between existing tags. The same process is imaginable for any other module triggering the start of the *dbpedia* module.

There are three attribute types that are of special interest for us: *redirect*, that is an indicator for synonyms, or *subject* and *broader*, that are signs for a term generalization. These attribute types include references to other related DBpedia entities, like synonyms and categories. When one of these types is found, the value of the attribute will include the URI representing the related entity. This is when the before mentioned recursion is started and this new entity is retrieved from DBpedia as well.

4. UTILITY OF SEMANTIC TAG INFORMATION

Collecting data never ends in itself. There is a purpose in generating more metadata. In the tele-teaching context, collecting data normally has the purpose in providing more knowledge to the user. It could also help finding appropriate information faster or showing connections between different teaching units.

In the following parts, we will show the points of interest concerning an enhancement of tele-teaching metadata and give some examples, how semantic information will supplement these functions.

4.1 How Users Retrieve Information

In the internet most user utilize search engines, like Google or Bing, for finding the content they are looking for. On a web page more possibilities are available for retrieving the desired information.

The first entry point to a web page is the navigation via menu and lists. Normally the menu provides the possibility to browse the content by predefined categories. This navigation is determined by the administrated metadata and cannot be influenced by the users themselves.

Problems occur for the user, when the number of items in a list for a specific category becomes too large. No user likes to scan long lists for the desired content, many of them will stop after a few pages without reaching their goal. In the beginning this problem can be solved with using subcategories, but this will only help for a short time. With a growing amount of data, navigation through menus and lists becomes impossible. So menus and lists are only suitable for navigation on smaller web pages.

Alternatively, user generated metadata, like tags, can be used for navigation as well. Browsing by tags provides two possibilities. On the one hand it is often used, when the user found content he is interested in and wants to find other content for an aspect of it. One of the tags can be used for navigation in this case. On the other hand a tag cloud can be used as an entry point for finding content for a specific keyword. This navigation is influenced by the users, because the users provide the tags and classify the content.

The available metadata is not only used for direct navigation, but also for the search function, which is an important entry point for big archives. In our portal we provide two search possibilities, a simple ajax search, which combines as much metadata as possible, and a more specified search, where the user can define more search criteria.

When a user searches for a special tag without semantic metadata, he will only find the same results he would get, when he uses the link in the tag cloud. With semantic information more results are presented, by also displaying items with synonym tags, more generalized or more specific tags.

Another possibility for navigation are suggestions to the user. As known from big web shops, like amazon, on a webpage dedicated to a special item other items are recommended. This recommendation is based on the similarity of the items as well as the preferences of the user.

The last three types of navigation, which depend on the metadata available, can be enhanced using semantic information. In the following paragraphs we will show, how the search function and the recommendation system were enhanced through semantic metadata.

4.2 Extending the Metadata Base for Search

Searching data is one of the most important functionalities of a tele-teaching portal. Especially when the amount of content becomes huge, it is nearly impossible to find the required content without a search function. Therefore we implemented a pluggable search function in the portal, to be able to enhance the basic search function with more specialized search options, as it is described in [14].

One of these enhancements is the search for tags. As described before, tags are terms which have a connection with the tag object. Therefore when searching for a tag, this means, that we are searching for these objects, which are connected with this term.

A problem is, that the number of user generated tags is normally small. It also happens often, that similar terms are searched, but they are not used as tags in the lectures or series with appropriate content. For example it is possible, that the tag CSS3 is used in a lecture about Cascading Stylesheets Level 3. The user searches for CSS. Without the knowledge of the generalisation relation between the terms CSS and CSS3 the lecture would not be found, even though it is appropriate.

Of course it would be possible to administrate these relations between tags manually, but when the number of tags and relations between them is increasing, this would result in a huge amount of work. It would also not help to find terms, which are not used in tags so far. For example the tag CSS is used in our portal, but the tag Cascading Stylesheet is not available. So it would be necessary to add all synonyms to get a good search experience for the users.

Other enhancements of the search function, like searching for spoken words in the content, are also possible, if the metadata is available. Instead of having too few data for search, the audio transcription function will for example provide a huge amount of data. Therefore the semantic knowledge is used for finding the most appropriate data, which can be used to classify the lectures. It can also be used to recognize the topic of a lecture, even if homonyms are used. For example if a lecture deals with safari it can have different topics. If the lecturer talks about some animals, like elephants or giraffes, you can conclude that it is a lecture about an African safari, but if the lecturer talks about web browsers or internet, it will be about the web browser. With this knowledge, it is easier to collect the appropriate search results for the user.

4.3 Using information from Semantically Enhanced Metadata for the Detection of Content Similarity

In huge archives, it is hard for the user to find the content, he is interested in. Therefore different approaches are possible. One is to suggest content to the user, which could be interesting for him. Therefore the content has to be clustered. One possibility for this clustering is the detection of content similarity. Therefore a number is calculated, which describes the similarity of two content objects.

For the detection of content similarity, the metadata of the content has to be compared. Therefore we built a flexible architecture, which allows to combine different results of similarity calculations in an overall result, like we described in [15]. In this work the similarity measure is described as a function $s : O^2 \mapsto [0, 100]$, where O is the set of objects, which should be compared. In the paper a basic algorithm which bases on the connection between similarity and distance of two objects was described. This claims the following three characteristics for a content similarity measure function:

- Symmetry: The values of $s(o_i, o_k)$ and $s(o_k, o_i)$ should be the same for all objects.
- Maximum of $s(o_i, o_k)$: The similarity of an object to itself should be 100. In [15] it is described, how every similarity function can be extended to fulfil this requirement. Therefore it has not be fulfilled in the basic calculation algorithm, but the algorithm has to be adapted before implementation.
- Transitivity of the distance: Because of the connection between distance and similarity ($d(o_i, o_k) = 100 - s(o_i.o_k)$, the similarity function has to follow the transitivity of the distance function. This means, that the distance of two objects should be less or equal to the sum of the distance of both objects to another one.

We started with the comparison of the title, the description and the lecturer of a lecture. But enhancing the similarity comparison with more metadata produced better results.

4.3.1 Basic Algorithm for Similarity Calculation

The solution for embedding tags into the similarity calculation is to define objects as similar, if they have many tags in common. Therefore we defined a set of tags of an object as a subset of the set of all Tags T like follows:

$$T(o) = \{\text{tags of object } o\} \subseteq T \qquad (1)$$

Furthermore we defined the value $w_M(t)$ as the number of connections between the objects of the set M and the tag t, where $c_i(o, t)$ is the i-th connection between the object o and the tag t.

$$w_M(t) = \#\{c_i(o, t) | o \in M\} \in \mathbb{N} \qquad (2)$$

With these definitions we can define the similarity of two objects concerning their tags using the Jaccard index as follows:

$$s_{\text{Tags}}(o_i, o_k) = 100 \cdot \frac{\#(T(o_i) \cap T(o_k))}{\#(T(o_i) \cup T(o_k))} \qquad (3)$$

When considering the weight of tags, so that globally often used tags are less important and locally often used tags are more important, the calculation is extended as follows:

$$s_{\text{Tags}}(o_i, o_k) = 100 \cdot \frac{\sum\limits_{t \in (T(o_i) \cap T(o_k))} w_{\{o_i, o_k\}}(t)}{\sum\limits_{t \in (T(o_i) \cup T(o_k))} w_{\{o_i, o_k\}}(t)} \qquad (4)$$

This function should be expanded using semantic information. Therefore we analyse the different types of semantic data provided by DBPedia.

4.3.2 Using Synonyms

Using semantic information we can compare more tags than without it. The first step is to consider the synonyms. Obviously the synonym relation between the tags should create an equivalence relation. Therefore the synonym relation divides the set of tags in equivalence classes (see 5), which can be used instead of the tags and the set of equivalence classes of tags (see 6) instead of the set of tags.

$$[t]\text{syn} = \{t^* \in T \mid t^* \text{ is a synonym of } t\} \tag{5}$$

$$T/\text{syn} = \{[t]\text{syn} \mid t \in T\} \tag{6}$$

$$T/\text{syn}(o) = \{[t]\text{syn} \mid t \in T(o)\} \tag{7}$$

$$T\text{syn}(o) = \bigcup_{S \in T/\text{syn}(o)} S \tag{8}$$

Using these equivalence classes instead of the tags, more similarities between the different objects can be found.

$$\omega_{\{o_i,o_k\}}(t) = \frac{w_{T\text{syn}(o_i) \cup T\text{syn}(o_k)}(t)}{\#\left(T\text{syn}(o_i) \cup T\text{syn}(o_k)\right)} \tag{9}$$

$$s_{\text{Tags}_s}(o_i,o_k) = 100 \cdot \frac{\sum\limits_{t \in (T\text{syn}(o_i) \cap T\text{syn}(o_k))} \omega_{\{o_i,o_k\}}(t)}{\sum\limits_{t \in (T\text{syn}(o_i) \cup T\text{syn}(o_k))} \omega_{\{o_i,o_k\}}(t)} \tag{10}$$

4.3.3 Using Generalization

But synonyms are not the only information, which can be found using semantic data. The relations between tags are more complex. The generalization is a partial order instead of an equivalence relation, which is more complicate to handle. Furthermore in this partial order it is not defined, how much a term is generalised by another tag.

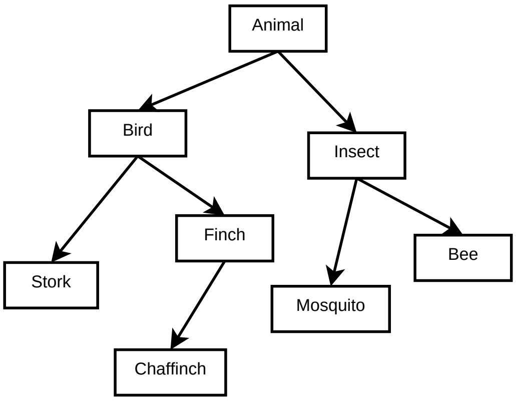

Figure 4: Example for generalization of terms as partial order

For example birds are animals, a finch is a special bird and a chaffinch is a special finch (see figure 4). A stork is also a bird. Therefore, if one object is tagged as a stork and another one is tagged as chaffinch, a similarity should be detected, because both are birds and animals.

But the tag finch should get another influence by the tag animal as the tag bird. It also has to be discussed, if the influence of the term finch on the term chaffinch is higher, than the influence of the term animal on the term bird.

$$G(t) = \{g \in T \mid g \text{ is a general term of } t\} \cup \{t\} \tag{11}$$

$$G(M) = \bigcup_{t \in M} G(t) \tag{12}$$

$$S(t) = \{s \in T \mid t \text{ is a general term of } s\} \cup \{t\} \tag{13}$$

If the full relation between all terms is known, some calculation algorithms to decide the value of the generalization are thinkable. For the length of the longest distance between two terms $dist(t_i, t_k)$ could be considered. If this distance is high, the influence of a term should be small.

$$md(t_i,t_k) = \#(S(t_i) \cap G(t_k)) \tag{14}$$

$$dist(t_i,t_k) = \begin{cases} md(t_i,t_k) & \text{if } t_i \in G(t_k) \\ md(t_k,t_i) & \text{if } t_k \in G(t_i) \\ \infty & \text{else} \end{cases} \tag{15}$$

$$\Omega_{\{o_i,o_k\}}(t) = \frac{\sum\limits_{x \in S(t)} \frac{w_{\{o_i,o_k\}}(x)}{dist(t,x)}}{0,5 \cdot \left(\min\limits_{x \in T(o_i)} dist(t,x) + \min\limits_{x \in T(o_k)} dist(t,x)\right)} \tag{16}$$

$$s_{\text{Tags}_g}(o_i,o_k) = 100 \cdot \frac{\sum\limits_{t \in G(T(o_i)) \cap G(T(o_k))} \Omega_{\{o_i,o_k\}}(t)}{\sum\limits_{t \in G(T(o_i)) \cup G(T(o_k))} \Omega_{\{o_i,o_k\}}(t)} \tag{17}$$

Furthermore, the number of relations starting from a term could be a measure. A term, which is a generalization for a high number of other terms, is less important than a term which generalizes only a few words.

The problem is, that the full relation is not known. Only a small part of the whole relation is visible when parsing the information from DBPedia. Also some connections might not have been parsed, because intermediate objects were not parsed. For example if the term bird is not parsed, there is no knowledge about a chaffinch being an animal, even if the term animal is parsed.

So the calculation becomes better, the more data is available. Until now, we don't have enough data in our database to get good results by using the generalization. When more data is available it can be checked, if this approach is also promising.

5. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the combination of Web 2.0 with the Semantic Web for the tele-teaching field of application. We argued, that the metadata base of multimedia tele-teaching content can be immensely enhanced by including Social Web functionalities, like tagging. Tagging generates valuable content-related keywords, but the semantics and the context of those keywords are missing. We therefore suggested an algorithm that utilizes the user-generated keywords to retrieve semantic data for them from a Linked Data source like DBpedia.

With the help of this newly gained semantic content-related data many algorithms can be expanded and improved. One example that we explicated further is the similarity detection between different objects in the context of tags. The similarity detection is one approach to be able to provide recommendation for related tele-teaching content to the users. Our similarity detection algorithm is able to combine different data sources into the similarity calculation. In order to enrich the recommendation function further, other data sources need to be included.

The next step is the implementation of additional triggers to fetch more semantic information. We are planning to use keywords from the speech transcript and optical character recognition. Also the titles and descriptions of the objects or search terms, that were entered by the users numerous times, can provide more keywords. We want to detect more sources for keywords and use the collected data to enhance the search function and similarity detection.

Also the tagging functionality itself can be improved using the newly retrieved semantic data related to the existing tags. As described in this paper, users do not only try to retrieve the relevant information by using the navigation, search or recommendation functionality. In some cases browsing through visible connections between the content can be a real benefit. By offering a browsing opportunity via the semantically related tags of a tag, a new approach to navigate through the content can be offered to students. This can for example be visualized in a topic map. The benefit for the students is that they can gain an overview of a whole topic without having to know details of that topic.

The problematic side of leveraging the Social Web for information enhancement is, that not a lot of users are participating in the content creation so far. Therefore we for example only have a small number of different tags we can utilize for the two applications mentioned so far. That is why another big focus in our research will be to find out how more users can be activated to contribute in the user-generated content.

Once all the metadata available in the portal is included in the extraction of semantic data and the user participation is increasing, the aggregated metadata can be used to create a global overview of the content of the whole tele-teaching portal. This can be done by creating an ontology for the special topic of the portal from the Linked Data that was extracted. These ontologies can also be used as alternative navigation. They can furthermore serve as part of a recommendation system, because logical connections between topics are included. In the e-learning area of application this can ultimately lead to a recommendation of a course schedule or related topics based on previous knowledge of the learner.

6. REFERENCES

[1] S. Bateman, C. Brooks, G. Mccalla, and P. Brusilovsky. Applying Collaborative Tagging to E-Learning. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization (WWW'07)*, Banff, Canada, 2007. ACM.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American Magazine*, 284(5):34–43, 2001.

[3] C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. DBpedia : A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, 2007.

[4] C. Brooks, S. Bateman, J. Greer, and G. Mccalla. *Lessons Learned using Social and Semantic Web Technologies for E-Learning*, chapter 14, pages 260–278. IOS Press, 2009.

[5] S. A. Golder and B. A. Huberman. The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208, 2005.

[6] T. Gruber. Collective knowledge systems: Where the Social Web meets the Semantic Web. *World Wide Web Internet And Web Information Systems*, 6:4–13, 2007.

[7] T. Gruber. Ontology of Folksonomy: A Mash-up of Apples and Oranges. *International Journal on Semantic Web & Information Systems*, 3(2):1–11, 2007.

[8] D. Mican and N. Tomai. Web 2.0 and Collaborative Tagging. *2010 Fifth International Conference on Internet and Web Applications and Services*, pages 519–524, May 2010.

[9] M. G. Noll. *Understanding and Leveraging the Social Web for Information Retrieval Dissertation.* Phd thesis, Hasso-Plattner-Institut für Softwaresystemtechnik, 2010.

[10] R. M. Palloff and K. Pratt. *Building Learning Communities in Cyberspace: Effective Strategies for the Online Classroom (The Jossey-Bass Higher and Adult Education Series) Share your own customer images Search inside this book Building Learning Communities in Cyberspace: Effective St.* Jossey-Bass, 1 edition, 1999.

[11] T. O. Reilly. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, No. 1(65):17–37, 2007.

[12] M. J. Rosenberg. *E-learning: strategies for delivering knowledge in the digital age.* McGraw-Hill, 2001.

[13] V. Schillings and C. Meinel. Tele-TASK – tele-teaching anywhere solution kit. In *Proceedings of ACM SIGUCCS*, Providence, USA, 2002.

[14] M. Siebert and C. Meinel. Realization of an Expandable Search Function for an E-LearningWeb Portal. In *Workshop on e-Activity at the Ninth IEEE/ACIS International Conference on Computer and Information Science Article*, page 6, Yamagata/Japan, 2010.

[15] M. Siebert, F. Moritz, and C. Meinel. Distributed Recognition of Content Similarity in a Tele-Teaching Portal. In *2nd International Conference on Information and Multimedia Technology (to appear)*, Hong-Kong, 2010.

[16] C. Torniai, J. Jovanovic, S. Bateman, D. Gasevic, and M. Hatala. Leveraging Folksonomies for Ontology Evolution in E-learning Environments. In *2008 IEEE International Conference on Semantic Computing*, pages 206–213, Los Alamitos, CA, USA, Aug. 2008. IEEE Computer Society.

[17] S. Trahasch, S. Linckels, and W. Hürst. Vorlesungsaufzeichnungen - Anwendungen, Erfahrungen und Forschungsperspektiven. Beobachtungen vom GI-Workshop "eLectures 2009". *i-com*, 8(3 Social Semantic Web):62–64, 2009.

[18] J. Waitelonis, H. Sack, J. Hercher, and Z. Kramer. Semantically Enabled Exploratory Video Search. In *Proc. of Semantic Search Workshop at the 19th Int. World Wide Web Conference*, Raleigh, NC, USA, 2010.