Enhancing an Environment for Knowledge Acquisition based on Web Resources by Automatic Tag Type Identification

Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz
Technische Universität Darmstadt, Multimedia Communications Lab

Key words: resource based learning, tagging, semantic networks

Abstract:

Resource based learning is a form of demand-oriented learning where learners close knowledge gaps by self-directed knowledge acquisition. The WWW is commonly used as a major source for knowledge acquisition. Resource based learning requires a good resource management for persisting and reusing of research results. One possibility for resource management are tagging systems. This paper describes our research and learning environment supporting resource based learning on the basis of a semantic tagging concept. This concept enhances pure tagging by describing tags semantically by their type. From the evaluation, we conclude that there is a need for identifying the tag type automatically, thus we develop an algorithm for automatic tag type identification.

1 Motivation

Today, continuous learning is increasingly necessary because of fast changing circumstances and job-related requirements. When new challenges arise, knowledge has to be acquired on demand. This kind of learning is called resource based learning. It is described as interaction with a wide range of resources in order to acquire new knowledge [1]. The WWW is commonly used as a major source for knowledge acquisition because there is a huge amount of information which grows steadily and where up-to-date information can be found. However, resource based learning poses challenges for learners: Mostly, there is no teacher who structures the learning process and prepares the materials, and resources in the web are rarely created by the authors for serving as learning material. Hence, learners have to perform the entire learning process self-directed. Additionally, the huge amount of information in the web can lead to cognitive overload. Furthermore, learners have to organize and structure web pages they have found for later usage – in addition to the actual reading and learning process. Tagging systems provide an opportunity for managing and organizing web resources. The following section gives a short overview of existing semantic tagging applications. Then, in section 3, our semantic tagging concept is described, which is integrated into a research and learning environment. This concept enhances pure tagging by describing tags semantically by their type. In our research environment, users are enabled to create a knowledge structure and to organize their found web pages in this structure. In section 4, we present an evaluation, from which we conclude that there is a need for identifying the tag type automatically. Thus, in section 5, we show an algorithm for automatic tag type identification.
2 Semantic Tagging Applications

Users tag resources relevant for knowledge acquisition with keywords, e.g. in order to organize them in their resource collection. In general, tags have a special, often implicit, semantic meaning for the user who tags and thus, tags fulfil different functions [2]. They can express opinions like “interesting” and “relevant” or they can be used to relate topics or persons to the tagged resources. Some tags are only useful for the individual user if they refer to a very personal situation like tags which are used for personal goal management. These different meanings and functions show that users often have a particular concept in mind while tagging. In contrast to the simple tagging concept where only a plain keyword is saved, the semantic tagging concept aims to capture also the semantics of these keywords. Semantic tagging is applied in systems like Faviki1 and Fuzzzy2. These two selected examples will be analyzed here with regard to capturing the semantic of tags. The aspects considered are different types of tags, possible additional information, possible relations between tags and scope of the tags (unique to one user or valid to the whole community). Popular tagging applications like Delicious3 or Flickr4, although widely-used, are not addressed here because they do not support semantic tagging explicitly.

2.1 Faviki

Tags in Faviki differ only in their source of vocabulary. These are based either on Wikipedia articles or on Google’s search hits. While typing a keyword, Wikipedia articles are recommended in order to define the semantic of this keyword; i.e. if a user types “jaguar”, he can select either the article about the animal or about the car. The link between tag and Wikipedia article leads to a disambiguation because the meaning of a tag is related to the content of the article. If there is no appropriate article, users can create new tags in form of a Google’s search hit. The link to Wikipedia or Google is the only information which is added to the keyword manually by users. Further information like abstract and images are provided automatically. User-defined relations are not supported but tags are classified by the Wikipedia categories of the article. Tags have a personal scope but can be shared with the community. Faviki is not limited to the English language because there are a lot of Wikipedia articles in different languages.

2.2 Fuzzzy

Resources in Fuzzzy are differentiated between bookmarks (hyperlinks), books (articles, electronic books), audio, video, newsfeeds and events. Tags are classified implicitly by their relations to the resource, e.g. by the relation “is about” or “has contributor”. Short descriptions can be added to tags. Relations between tags are possible as well; synonyms, relationships and super-/subtopics can be created in order to describe the meaning of the tags more precisely. Tags and resources are globally valid for the whole community, thus users can vote for or against defined tags and relations.

---

1 http://www.faviki.com, Zugriff am 12.08.2010
2 http://www.fuzzzy.com/, Zugriff am 12.08.2010
3 http://www.delicious.com/, Zugriff am 20.08.2010
4 http://www.flickr.com/, Zugriff am 20.08.2010
3 Research and Learning Environment

A semantic tagging concept is integrated into our research and learning environment ELWMS.KOM which supports functionalities for resource management. In the first section, this environment in whole is presented briefly. Then the semantic tagging concept is explained and after that some benefits of the semantic tagging concept are shown.

3.1 Features of the Environment

In contrast to the above mentioned tagging applications, our system focuses on learning purposes, particularly with regard to goal management in the self-directed learning process of researching on the web and structuring knowledge resources.

![Figure 1: left: Information about currently opened web page; right: Goal management](image)

The environment is embedded directly as an add-on in the sidebar of the web browser Mozilla Firefox. Research takes place mostly in a web browser; thereby users can save resources without being interrupted by switching between applications. By tagging web pages, a user can externalize his knowledge and organize it in accordance with the knowledge structure in his mind, e.g. he can tag a web page with the topics covered or associated concepts while saving the page (like the fact that he searched for this web page in order to achieve the goal to prepare a specific talk). If a user visits a web page, which has already been saved by him or another user, the system notifies him (see figure 1-left) and the user gets further information in order to support remembering former researches on the one hand and on the other hand for getting recommendations to relevant resources as starting points for further researches. The evolving knowledge structure can be searched and browsed (graphically) later on in a web portal. Thereby, users can learn from the structure created by other users (e.g. associations between topics) and find more relevant web pages.

A user should not search unsystematically, thus an easy goal management tool is integrated. It serves for structuring research processes thematically and temporally in a goal hierarchy.
Goals can be subdivided and restructured during researches. The goal management view is shown in figure 1-right. A detailed explanation of this goal management tool is given in [3].

3.2 Semantic Tagging Concept

In our research environment, the semantic tagging concept is based on adding the type of a tag, expressing further information about this tag and defining relations between tags. Based on a requirement analysis [4] for the research environment in order to support resource based learning we identified the following different tag types: topic, goal, person, event and location. This basic set can be extended as required. The evaluation of the usage of these tag types is presented in section 4.

The basis for the implementation of our learning and research environment is a semantic net. A semantic network or net is defined as “a graphic notation for representing knowledge in patterns of interconnected nodes and arcs”[5]. Resources and tags form nodes in the network which are connected by relations. By tagging, resources are either linked with existing tag nodes or new tag nodes are created. Tags are saved in a personal scope. Beside relations from resources to tags, users can create relations between tags, e.g. a user can tag an online scientific paper with a person and tags this person with the name of a conference because he discussed this paper with this person at this conference. The probability is higher that the user remembers the person or the conference instead of the title or author of the searched-for paper.

Information which can be added to resources or tags is definable individually by each user. A user is not forced to fill in predefined metadata forms, in fact he can create his personal forms with information that are relevant to him, e.g. a user wants to add a deadline to goal tags so he extends the basic metadata form by a deadline input field. The semantic tagging concept leads to some benefits for the learning and research environment.

3.3 Benefits of the Tag Types

So-called folksonomies, which are generated by aggregating tags from different users, are often used for recommending tags and resources. As mentioned in the last section, tags can have different functions. Since users often use tags which are only useful for personal or organizational purposes which have no objective meaning (e.g. “to read”, “want to have”) or which are used for describing a personal situation (e.g. “my friends”), recommendations can be affected negatively. Recommendation algorithms, which recommend resources and tags based on frequency distribution, can benefit from ranking tag types differently. Tag types like goals with an individual relevance should not be taken into account and tag types like topics which are of value for others should be ranked higher.

Beside full-text searches in semantic networks, searching based on relations is possible, e.g. a resource can be found because of a location tag which is linked to a topic tag which again is linked to the resource although the location itself is not contained in this resource. By using different tag types, semantic searches can thus be narrowed down. Navigation in the research environment can be simplified as well by filtering particular tag types and the information view can be adapted to the type of tag. Special relations between tags serve to realize a hierarchical super-/subtopic structure and a goal management function.

4 Evaluation

The usage of the tag types in our research environment was evaluated in a long term user study with students of a seminar at the Technische Universität Darmstadt. Twelve students in the age range of 20-25 participated in this study. Self-directed researches in the web were
necessary for their seminar thesis. They used ELWMS.KOM for about three months. The evaluation was structured in a rollout phase, usage phase and closing phase, in which a presentation of their semantic networks was prepared.

In the rollout phase, the learning and research environment was introduced. During the main phase, the students on the one hand used ELWMS.KOM for organizing the research results for their written thesis and on the other hand for externalizing their knowledge in the semantic network. They could use their semantic network in an oral presentation about the proceeding and the structure of their seminar thesis at the end of the evaluation.

Table 1 itemizes the number of saved resources and the usage of each tag type per participant. A total of 258 resources were saved which were linked to 436 tags. The participant with the smallest number of resources saved ten; the one with the maximum had 43 resources. Two of the participants created only three tags and the most active user 134 tags. All in all, each tag type was used but not by every user. Tags without an assigned type were used more often than tags with a tag type. The usage of typed tags or rather the non-usage can possibly be traced back to the fact that none of the users had used a tagging system ever before and only two users even knew the term „tagging“ at all.

As a result, the aforementioned benefits of the semantic tagging concept could not be fully tapped. The goal management tool integrated in ELWMS.KOM most probably led to a high number of goal tags. With the miscellaneous type, we wanted to discover whether tag types are needed which were not identified in the requirements analysis. However, after reviewing all tags manually and interviewing the participants, it showed that there is no need for another tag type.

Table 1: Number of resources and tags itemized per participant

<table>
<thead>
<tr>
<th></th>
<th>Resources</th>
<th>Goals</th>
<th>Topics</th>
<th>Events</th>
<th>Locations</th>
<th>Persons</th>
<th>Miscellaneous</th>
<th>Tags without type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>28</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Participant 2</td>
<td>14</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Participant 3</td>
<td>43</td>
<td>15</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Participant 4</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>Participant 5</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Participant 6</td>
<td>21</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Participant 7</td>
<td>23</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Participant 8</td>
<td>12</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Participant 9</td>
<td>33</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Participant 10</td>
<td>10</td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Participant 11</td>
<td>25</td>
<td>16</td>
<td>38</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>Participant 12</td>
<td>26</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>100</td>
<td>74</td>
<td>2</td>
<td>11</td>
<td>32</td>
<td>20</td>
<td>197</td>
</tr>
</tbody>
</table>

The participants did not use all possible functionality of the research environment. This can probably be ascribed to the first-time usage of a tagging application and the missing training.
phase for using the tool. We guess that another reason is the missing possibility of using the learning environment in the seminar room, and as such, the participants were not aware of the more complex functions in the application. However, all students stated that they found the learning and research environment useful for researches and would use such a tool for their daily researches. This result corresponds to the results of a former user study with an older prototype of the research environment [6].

5 Automatically Identifying Tag Types

Supporting the process of assigning tag types seems to be useful in light of the performed user study in order to reduce the number of tags without an assigned tag type. One supporting possibility is to identify the tag type (semi-)automatically. This section discusses related work at first and then our approach is described which is implemented in ELWMS.KOM.

5.1 Related Work

Named entity recognition in natural language texts is part of the field of information extraction. This area of research aims to automatically identify named entities as well as to classify these entities into categories like persons, locations, organizations, etc. For classification, list and rule based methods and machine learning algorithms are often applied. List based approaches rely on dictionaries which have to be expensively created beforehand or manually revised for the particular usage scenario. For rule based approaches, patterns and heuristics have to be set up, e.g. the word “Mister” in front of a name refers to a person. Generally, these rules are language dependent and have to be adapted to new knowledge domains if the domain changes. Machine learning algorithms need labelled training data to extract rules automatically on how to classify an entity.

WALU [7] is a project example which aims to develop a tool with which domain experts without prior knowledge in computer linguistics can execute named entity recognition (semi-) automatically. In WALU, users can extract lists of named entities with which machine learning algorithms are trained in order to classify these names automatically in future. Beside machine learning algorithms, rule based methods are used, e.g. regular expressions for identifying date and time data.

There are only few approaches in the field of automatically classifying tag types. An overview of current approaches is given in [8]. These approaches try to classify tag types in the domains of pictures, web pages and music. Besides the keyword, some of them include additional metadata like the date when a photo was taken or the content of a tagged web page.

The approach developed in [8] covers all three mentioned domains but only specifics of the domain of web pages are addressed in the following because the other two are not relevant for the scenario in ELWMS.KOM. In [8] the following tag types are distinguished:

- Time (dates, holidays, “daily“, “current“...)
- Location (cities, countries, ...)
- Type/Medium (file formats, web genres like blog and wiki, ...)
- Author/Owner
- Self-references (“my friends“, “wish list“, ...)
- Topic
- Opinion (“interesting“, ”relevant“, ...)
- Usage context (“toRead“, “review“, ””later“, ...)

For the classification, all three methods of named entity recognition are used. Regular expressions and manually created lists, which contain e.g. holidays which are used for the time class. Predefined lists contain common file formats for type/medium and commonly used
terms for self-references. A list of locations is extracted from a geographic thesaurus. If the tag is contained in the URL of the tagged web page, it is classified as author/owner. Tags, which can not be assigned to one of these five classes, are tried to be classified as one of the other three classes by machine learning algorithms. The authors assume that topics are composed in most cases of nouns and opinion tags of adjectives and a higher number of words. The frequency of usage context tags is small in relation to the total amount of tags. Besides the URL, this approach does not use any other additional information or content of the tagged web page.

5.2 Approach in ELWMS.KOM

Tag types defined in our research and learning environment partially overlap with tag types in related work but there are some specifics in the scenario of resource based learning. Manually created, predefined word lists cannot be applied because the research area is not restricted to a particular knowledge domain. Likewise, users cannot be expected to create domain specific word lists. Therefore, the algorithm designed for the research environment is mainly based on querying external knowledge sources. Freebase and DBPedia are used as external knowledge sources. Freebase is a huge, open database which contains world knowledge and currently has entries covering more than twelve millions concepts in domains like movies, books, persons (2 mill.), locations (1.8 mill.), etc. DBPedia is a community, which tries to extract structured information from Wikipedia in order to make this information available in the course of the Linked Open Data initiative. The DBpedia database contains currently more than 3.4 million entries, including 0.3 million relating to persons and 0.4 million regarding locations. Both external sources have the advantage that they are maintained by a large community. Terms which become more important or are created because of world affairs and sciences, are added continuously to the database. In both sources, concepts are sorted into defined categories. These categories are mapped to tag types used in the research environment, see table 2. A tag typed by a user is used as search term to query the categories of the external sources. Then, mapping is performed for the tag types person, event and location. The sources do not only contain country-typical names (e.g. Köln for Cologne) but also translations into other languages, so that not only e.g. English tags can be classified.

If the tag cannot be found in one of the external sources, the term is looked up in dictionaries in order to identify the part of speech. Wordnet is used for the English language and

<table>
<thead>
<tr>
<th>Freebase</th>
<th>DBPedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person /people/person</td>
<td><a href="http://dbpedia.org/ontology/Person">http://dbpedia.org/ontology/Person</a></td>
</tr>
<tr>
<td>Location /location/location /geography</td>
<td><a href="http://dbpedia.org/ontology/Place">http://dbpedia.org/ontology/Place</a></td>
</tr>
<tr>
<td>Event /time/event /time/recurring_event /event/event /conferences/conference</td>
<td><a href="http://dbpedia.org/ontology/Event">http://dbpedia.org/ontology/Event</a></td>
</tr>
</tbody>
</table>

If the tag cannot be found in one of the external sources, the term is looked up in dictionaries in order to identify the part of speech. Wordnet is used for the English language and

---

5 http://www.freebase.com/, Zugriff am 25.08.2010
6 http://dbpedia.org/, Zugriff am 25.08.2010
7 http://linkeddata.org/see-also, Zugriff am 11.08.2010
8 http://wordnet.princeton.edu/, Zugriff am 22.07.2010
Beolingus⁹ for German. According to the assumptions in [8], nouns are classified as topic type and the other parts of speech as no type. For tags which consist of more than one word, the majority of the occurrence of a part of speech decides on the type. The tag type goal is not to be assigned automatically as it cannot be found in predefined lists because of its individual usage. Further, users are encouraged to set goals manually for planning their research process.

5.3 Integration in ELWMS.KOM

The presented approach is implemented and integrated into ELWMS.KOM. Figure 2 shows a screenshot of the dialog opened for saving a resource.

![Figure 2: Creating a new tag with a list of tag type suggestions](image)

Instead of one single tag, a list of tag types is suggested showing tags which could probably come into consideration as the user is not be limited in his tag decisions. From this list, the user selects the most appropriate tag. For example, if a user wants to tag with “Faust”, suggestions like “Goethe’s Faust (Topic)” and “Frederick Faust (Person)” are submitted. Consequently, users are made aware of different meanings for their input and they are inspired to choose a more exact term. All users of ELWMS.KOM can profit from more accurate terms. User inputs are saved in the system and in future tag type classifications they are recommended and presented prominently in the suggestion list.

The advantage of using external sources does not only lie in the identification of the tag types but also in the reduction of the effect of the so-called Cold Start problem which hampers the recommendations in the initial stages. Because of the lack of existing relevant tags, no adequate recommendations can be generated – primarily for the first users of a system.

⁹ http://quexit.sourceforge.net/, Zugriff am 22.07.2010
6 Summary

In this paper, a research and learning environment for supporting resource based learning and its integrated semantic tagging concept are presented. Related semantic tagging applications are discussed. The semantic tagging concept described here is based on the possibility to add semantic meaning to a tag in form of the tag type. These tag types enhance the functionality of the research and learning environment and may improve the resource recommendation algorithm.

The performed user study shows that the process of adding a tag type should be encouraged. This can be done by (semi-)automatically identifying the tag type. Based on existing methods in the area of named entity recognition, an approach is developed to identify the tag type (semi-)automatically and appropriate tag types are recommended to the user. Our approach, which uses mainly a list based method based on external sources, is implemented and integrated in the learning and research environment ELWMS.KOM. For future work, it is planned to extract additional information about the tags from the external sources, e.g. geo coordinates for locations. Saving additional information has already been realized in the system.

The knowledge about the selected tag type recommendation for tags with different meanings can be used in further research to reduce wrong search results in a search machine by removing web pages which refer to the wrong meaning.

References:

Authors:

Doreen Böhnstedt, M. Sc.
Technische Universität Darmstadt, Multimedia Communications Lab
Rundeturmstr. 10, 64283 Darmstadt
Doreen.Boehnstedt@kom.tu-darmstadt.de

Dipl.-Inf. Philipp Scholl
Technische Universität Darmstadt, Multimedia Communications Lab
Rundeturmstr. 10, 64283 Darmstadt
Philipp.Scholl@kom.tu-darmstadt.de

Dr.-Ing. Christoph Rensing
Technische Universität Darmstadt, Multimedia Communications Lab
Rundeturmstr. 10, 64283 Darmstadt
Christoph.Rensing@kom.tu-darmstadt.de

Prof. Dr.-Ing. Ralf Steinmetz
Technische Universität Darmstadt, Multimedia Communications Lab
Rundeturmstr. 10, 64283 Darmstadt
Ralf.Steinmetz@kom.tu-darmstadt.de