

TOWARDS AUTOMATIC CONTENT TAGGING: ENHANCED WEB SERVICES IN DIGITAL LIBRARIES USING LEXICAL CHAINING

Ulli Waltinger¹, Alexander Mehler¹, Gerhard Heyer²

¹*Text Technology, University of Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Germany,*

²*Institute of Computer Science, NLP Department, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany*

{Ulli_Marc.Waltinger, Alexander.Mehler}@uni-bielefeld.de

heyer@informatik.uni-leipzig.de

Keywords: Topic Tracking, Topic Structuring, Topic Labelling, Social Tagging, Digital Library, Wikipedia, Lexical Network, Lexical Chaining.

Abstract: This paper proposes a web-based application which combines social tagging, enhanced visual representation of a document and the alignment to an open-ended social ontology. More precisely we introduce on the one hand an approach for automatic extraction of document related keywords for indexing and representing document content as an alternative to social tagging. On the other hand a proposal for automatic classification within a social ontology based on the German Wikipedia category taxonomy is proposed. This paper has two main goals: to describe the method of automatic tagging of digital documents and to provide an overview of the algorithmic patterns of lexical chaining that can be applied for topic tracking and –labelling of digital documents.

1 INTRODUCTION

Taxonomies and collaborative tagging

The phenomenon of the *web 2.0* can be directly associated to web technologies such as search engines, web mining, meta-standards but first and foremost with the socialisation and collaboration of internet users. An area which has grown in popularity particularly in the *blogsphere* and digital library services is collaborative tagging. In this scenario, weblogs, web-services and document repositories provide documents, bookmarks and multimedia content are organized by assigning keywords or tags by collaborating users. Interestingly, it turns out that this process is highly predictive showing that there are general principles of collective information organization. However the action of tagging content always is a process of a subjective decision. It is

neither exclusive nor necessarily hierarchical. One can introduce keywords without knowledge about whether and in which context that label has been used by others. Moreover, the new introduced tag might also be a reference for other users to describe their content. Clearly, collaborative tagging reflects the dedication of users in web communities, but common problems of natural language processing also appear in collaborative tagging. These are:

- wrong notation (keywords are written wrongly)
- polysemy (ambiguity of tags)
- synonymy (sense-related tags without being annotated that way)
- missing context views of the socially accepted usages of tags
- missing overviews of tag systems

There are many web technologies that do assist users in assigning related tags to content units.

Most commonly the representation of tag clouds, i.e., a weighted list of user generated content-tags, which indicate the most frequently, used classifiers. By means of such clouds, user's are not only inspired but also swayed to use already assigned terms. Moreover, several web services implement tag-recommendation systems which indicate previously assigned or shared tags by the user in sequence patterns during the action of tagging. On the other side, (Golder et al., 2006) have shown that the distribution of tags stabilises on base of a common denominator, that is, a shared vocabulary. Therefore, some users apply a wide range of different tags to their content, some introduce only a few, but it can be observed a stable pattern in tag proportions without global control.

Social tagging produces some sort of a tag-taxonomy. In contrast to existing ontologies, e.g., the tree-like Dewey decimal classification, social tagging induces *graphs* which are constantly changing. Furthermore, folksonomies do not force unambiguous categorizations, but realize multi-label classifications. A prototypical example is the category system of the *Wikipedia* (Voss, 2006) which is an open-ended social ontology enhanced by a community not only by publishing and interlinking of article, but also by enabling user to categorize documents (Gleim, Mehler, 2006).

This paper proposes a web-based application which combines social tagging, enhanced visual representation of a document and the alignment to an open-ended social ontology. More precisely we introduce an approach for automatic extraction of topic labels for indexing and content representation as an add-on to social ontologies. That is, we perform automatic document classification in the framework of a social ontology based on the *Wikipedia* category taxonomy. This paper has two main goals: to describe the method of automatic tagging of digital documents and to provide an overview of the algorithmic patterns of lexical chaining that can be applied for topic tracking and -labelling of digital documents. Thereby, we first explain the general architecture of the system in Section 2. Then we present a formal model of the used lexical chaining algorithm in Section 3. In Section 4, we outline the alignment with the *Wikipedia* category system. Finally, we give a conclusion and prospect future work.

2 RELATED WORK

The method proposed in this paper belongs to the domain of content classification in special the tagging of content though meta-information and the alignment of documents on a social ontology. (Braun et al., 2007) presented an application (*SOBOLEO*) on alignment of collaborative tagging to a light-weighted ontology. This approach enables users to add hyperlinks to an online-repository – so called ‘social bookmarks’ – by assigning tags to hyperlinks. Furthermore, each bookmark can be categorized by referring to a terminological ontology. The employed ontology can be specialised by assigning new concepts. In this case both, tagging and categorization of content has to be done manually. Contrary, our focus is set to an automatic – none manually - approach of tagging and categorization.

(Mika, 2005) presented an application for the extraction of community-base light-weighted ontologies from web-pages. In special creating actor-concept ontology by generating associations between an actor (e.g. person) and a concept (e.g. label). This is done by submitting a search query, combining the two terms, and measuring the resultant page count. This approach tends to be similar to the classical lexical chaining approach, using a lexical network (in this case a search engine) as a resource for generating associations between two terms. However an integrated structure and content-based text model is left out by using only already assigned tags from content.

3 ARCHITECTURE MODEL

The main concept towards automatic content tagging and topic tracking is an integrated structure and content-based text model approach. This means in first place the task of tracking semantically related tokens based on a lexical reference system is combined with a detailed structure analysis of text. The idea behind this is that each content element of a text (content and structure) is always semantically related to another segment in the same text. Therefore we can span associations between tokens, sentences, paragraphs and divisions based on their semantic relatedness. This is done by introducing a *Generic Lexical Network Model* exemplified by using a snapshot of the German *Wikipedia*-Project.

In addition an alignment to an existing ontology is computed by normalizing, labelling and categorizing

topic chains. Generally speaking, the application procedure can be subdivided into three coordinated main modules (see Figure 1) which provide an integrated structure- and content-based text model for topic tracking and automatic content tagging:

1. analysis of logical document structure
2. lexical content analysis and term extraction
3. ontology alignment and topic labelling

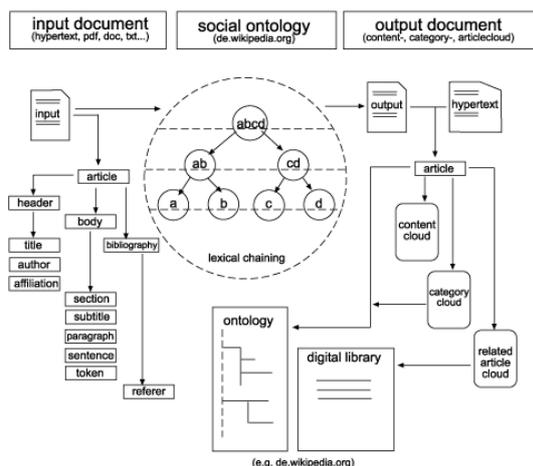


Figure 1: System Architecture.

3.1 Analysis of Logical Document Structure

A fundamental requirement of this module is to process a wide range of different input documents. Therefore *Plaintext*, *PDF*-, *Open Office*-, *Word*- and *(X) HTML* documents must be automatically analyzable. The possibility to process documents of a wide range of formats is indispensable from the point of view of digital libraries. We meet this demand by having integrated mapping routines for all these formats. Once having extracted the content of an input document a transformation to a XML-Format is deployed. All content is converted into the *Corpus Encoding Standard* (Ide et al., 1998) which has been designed for mapping *Logical Document Structures* (Power et al., 2003) of large corpora in language engineering. We provide this by extracting section (title, sub-title, header, body...), paragraph and sentence structures as well as images. As a result, each input document is mapped onto a tree-like representation which can be accessed for structure-oriented retrieval. Once the logical document structure has

been extracted, lemmatization of lexical content is deployed. The process of determining the lemma information for an extracted token is needed in order to retrieve information out of a lexical type network. Therefore, we developed an interoperable lemmatizer. It is based on the *Morphy* system (Lezius, 2000) which integrates a morphological analysis with part-of-speech tagging in a single package. We used a German edition of the *Wikipedia* as well as a ten years release of the German newspaper '*Süddeutsche Zeitung*' to extract the morphological information of *Morphy*. As a result, we generated a lexicon of more than 3.7 million word forms which are currently the basement of our tagging-application. The lemmatizer is used to annotate lexical information within input documents in the CESDOC-format. In addition token positions within sentences and paragraphs are annotated (see Figure 2). These so called corresponding 'c' attributes mark the position of the element in the XML DOM tree. As a result a hierarchical CESDOC-XML-Document is generated including logical document structure and lexical information.

```
<CESDOC>
<TEXT id='TEXT1'>
<BODY>
<H5>
<T c='1'>
    <O>Datum</O>
    <L p='NN'>Datum</L>
</T>
<T c='2'>
    <O>:</O>
    <L p='SZ'>:</L>
</T>
```

Figure 2: A Snapshot of a CESDOC-XML document.

3.2 Lexical Content Analysis

The second module (see Figure 1) of our application is concerned with lexical content analysis. The idea behind our lexical chain is the assumption that semantically related tokens of a document do occur within a restricted area of text segments (Halliday, Hassan, 1976). Following this idea, a token at position one in paragraph one tends to have a higher probability in being semantically related to a token in the same paragraph than with a token of the last paragraph of that document. Since we have a model of an ordered hierarchy of content objects, we are able to link any pair of tokens within instances of certain constituent types of the logical document structure. Thus, we can implement logical distances not only in terms of the numbers of tokens in-between, but also in terms of, e.g., intermediary

paragraphs, sentences etc. In order to classify a connection between a pair of token as a lexical edge, an external resource for lexical chaining is needed. This is provided by the usage of a type network as a model of a terminological ontology. Semantic taxonomies such as *WordNet* (Fellbaum, 1998) provide a rich source of lexical knowledge for text and web mining, but are limited in the sense that they do not cover special vocabularies as they are typical for scientific texts to be managed by digital libraries. Thus, we decided to use an open-ended social ontology as a resource for lexical chaining. In this case, the German release of Wikipedia.

More specifically, Wikipedia article, categories and portal documents have been used to induce vertices, whereas hyperlinks induce edges. In special, vertices are typed as articles, portals or categories and edges are labelled as, e.g., hyperonym of (in the case of a link from a superordinate to a subordinate category), article of (in the case of a link from an article to a portal) or as an association (in the case of a link between two articles). As a result we get a lexical network which spans the reference plane of lexical edges as the resource for computing lexical chains. More specifically, rating pairs of tokens on basis of their semantic relation equals to their minimal distance in the referred terminological ontology (Morris, Hirst, 1991). By that, lexical chains can be defined as graphs spreading over an inclusion hierarchy of text. Though lexical chains can be computed by the following algorithm:

```

foreach token T of paragraph P
{
  foreach token T' of paragraph P +/-
  paragraph distance parameter X
  {
    compute shortest-path as graph-
    distance D(T, T') within lexical
    network N;
  }
}

if ( pair D(T,T') < network distance Y )
{
  build lexical chain L;
}

```

In general, the time complexity of chaining algorithms is high as they rely on computing shortest paths which is of order $O(|V| |E| + |V|^2 \log |V|)$ (as, e.g., the Johnson all pairs shortest path algorithm (Johnson, 1977). There also exist proposals for a chaining algorithm in linear time (Silber, McCoy, 2002). However, this approach cannot be applied to the Wikipedia as it misses the rich type system of

WordNet utilized by Silber & McCoy. Thus, we alternatively explored the small world nature of the wiki graph (Zlatic et al., 2006; Mehler, 2006) and constrain the maximally allowed path length to a value < 3 where a distance of three links corresponds to the average geodesic distance in wiki graphs. As a consequence, shortest paths are efficiently computed as they are reduced to a simple look-up mechanism. More specifically, we reduce time complexity to an order of $O(|V||E|)$ supposed that the maximally allowed path distance in the terminological ontology is one. The reason is that in the worst case we have to consider all pairs (v,w) of lemma ($|V|^2$) where for each vertex v we have to examine on average $|E|/|V|$ edges. Next, all lexical chained pairs of tokens are clustered in order to get so called meta-chains describing the content of the input document. Depending on the used lexical-distance parameter, e.g. P, we get a snapshot of the content of the input document as in Figure 3.

So far, we have explored document structure, lemmatized all lexical content and have put all lexical items which are semantically related in terms of Wikipedia into the same lexical chain. As an output we get a set of such chains where the largest thereof represents the main document content. It can be accessed to further process the input document and to perform a semantic search, that is, a search by means of the most prominent lexical items of the main chain of the input document. This is described subsequently.

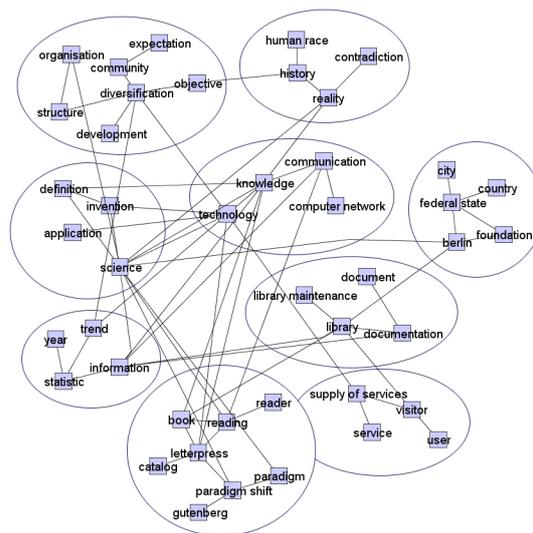


Figure 3: Lexical meta-chains of an input document (translated from German).

3.2 Ontology alignment / Topic Labelling

The third module of our applications is concerned with topic labelling and the categorization of a document. On base of the resultant meta-chains, representing the main document content, we are able to compute a topic label for each section of the input document. The first step in doing this is to determine the distribution of tags by employing again a lexical chaining limited to the entries of the meta-chain and ranking afterwards each returned keyword to its IDF/RIDF value in conjunction with the entropy of word frequencies³. As a result we gain a weighted list of tags out of which the topmost ranked units are selected. In order to classify and label a meta-chain we are going to align this information to the input taxonomy. In this case we are using again the social ontology of the Wikipedia Category System as a resource (See Gleim, Mehler, 2006). Therefore we explore the most probable categories and articles of the Wikipedia categorizing and relating to the input document. This is done by ‘firing’ search queries on the calculated index of the article-section of Wikipedia using the weighted tag list. The retrieved article weight is computed by frequency of tag occurrence within an article. This can be computed by the following algorithm:

```
nwt: number of weighted tags
rd:  retrieved article
rdw: retrieved article weight
ua:  used article
uc:  used category

while (rdw < 80%)
{
    submit search query with nwt(tags);
    nwt--;
}
add rd to ua

foreach(item of ua)
{
    parse article-site;
    retrieve category in site;
    add category to uc
}
foreach (item of uc)
{
    retrieve hypernym-category in category-graph;
    add new category to uc;
}
```

³ The IDF/RIDF-Index was computed on the base-ment of the German Wikipedia Project.

The explored categories are then used as topic-labels. As an outcome, three different weighted lists of tags are generated. Firstly, a content-list comprising the ‘classical’ content tags labelled with the category concepts. Secondly, a category-list as a subset of Wikipedia categories, tagging the input document. Thirdly, a hyperlink-list indicating the most likely connected Wikipedia articles. As a visual depiction all three weighted list are displayed as tag-clouds (Figure 4).



Figure 4: Tag Cloud-Representation

4 CONCLUSIONS

In summary, our system of topic labelling comprises classical text mining technologies which already have shown to produce reliable mining results with rising Web 2.0 technologies. Thus, a central outcome of the paper is to show a way to integrate text & web mining with social tagging systems that altogether provide semantic search as a future service of digital libraries. This paper presented the architecture of such integration. The evaluation of the usefulness of its ingredients has already been provided in the related literature. What remains to be done is a profound user study which shows the usefulness of our system from the point of user communities of digital libraries. Future work will focus on systematically evaluating this application, by using a hand-crafted tagged and categorized corpus of the German newspaper *Die Zeit*. The web-application is online accessible at:

<http://www.scientific-workplace.org/tagging/>

REFERENCES

- Allan J., 2002. Topic Detection and Tracking. Event-based Information Organization. Kluwer, Boston/Dordrecht/London.
- Barr M., Wells C., 1990. Category Theory for Computing Science. Prentice Hall, New York/London/Toronto.
- Barzilay R., Elhadad M., 1997. Using lexical chains for text summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain.
- Braun S., Schmidt A., Zacharias V., 2007. SO-BOLEO: vom kollaborativen Tagging zur leichte-wichtigen Ontologie. In Mensch&Computer 2007
- Budanitsky A., Hirst G., 2006. Evaluating WordNet-based measures of semantic distance. Computational Linguistics, 32(1):13-47.
- Fellbaum C., editor., 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge.
- Gleim R., Mehler A., Dehmer M., Pustyl'nikov O., 2007. Aisles through the category forest. In Proceedings Webist 2007.
- Golder S., Huberman B. (2006). Usage patterns of collaborative tagging systems. In Journal of Information Science, pages: 198—208.
- Heyer, G., Bordag, S., Quasthoff, U., 2003. Small worlds of concepts and other principles of semantic search, In Innovative Internet Community Systems, Proceedings of the Third International Workshop IICS 2003, June 2003 Leipzig, Lecture Notes in Computer Science, Springer Verlag: Berlin, Heidelberg, New York
- Hirst G., St-Onge D., 1997. Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambridge, MA: The MIT Press.
- Idea N., Pries-Dorman G., 1998. Corpus Encoding Standard. New York.
URL:<http://www.cs.vassar.edu/CES/>
- Leuf, B., Cunningham W., 2001. The Wiki way: quick collaboration on the Web. In Addison-Wesley.
- Lezius, W., 2000. Morphy - German Morphology, Part-of-Speech Tagging and Applications. In Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress pp. 619-623 Stuttgart, Germany
- Lossau N. (2004). Search Engine Technology and Digital Libraries, Libraries Need to Discover the Academic Internet. In: D-Lib Magazine, Bd. 10, Nr. 6, ISSN 1082-9873
- Mayr, W. (2005). Google Scholar - wie tief gräbt diese Suchmaschine? Bonn.
URL:http://www.ib.hu-berlin.de/~mayr/arbeiten/Mayr_Walter05-preprint.pdf.
- Mika P., 2005. Ontologies are us: A unified model of social networks and semantics. In: Proceedings of the Fourth International Semantic Web Conference (ISWC2005), Lecture Notes in Computer Science no. 3729, page 122-136, Galway, Ireland
- Morris J., Hirst G., 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics.
- O'Reilly, T., 2005: What Is Web 2.0. O'Reilly Media.
URL:<http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- Power, R., Scott, D., Bouayad-Agha N., 2003. Document structure. In: Computational Linguistics, 29(2), 211-260
- Silber H.G., McCoy K.F., 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics.
- Voss J., 2006. Collaborative thesaurus tagging the Wikipedia way.
URL: <http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0604036>.