Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data

Harald Sack
Institut für Informatik
Friedrich-Schiller-Universität Jena
Germany
sack@minet.uni-jena.de

Jörg Waitelonis
Institut für Informatik
Friedrich-Schiller-Universität Jena
Germany
joerg@minet.uni-jena.de

ABSTRACT

Collaborative tagging systems have become rather popular for annotating any kind of resources ranging from electronic documents to real world objects. In current tagging systems resources as a whole are annotated with and referenced by user defined tags. For multimedia data, as e. g. for video data, single scenes can be identified and annotated by using MPEG-7 metadata. We propose a collaborative tagging system that is combined with an automated annotation system for synchronized multimedia presentations. MPEG-7 metadata are used for the annotation of single scenes with user compiled tagging information in combination with metadata provided directly by the author or by other annotation systems. Thus, we propose a system being able to search within multimedia data that can further be extended to search within any kind of (partial) document to achieve a more tightly focused and personalized search.

1. INTRODUCTION

Online Social Networking enables collaboration relationships and allows exploiting these relationships for automated information distribution and classification. In particular, *collaborative tagging systems* (CTS) have become increasingly popular for annotating any kind of electronic documents (e. g. web pages, images, videos) or even real world objects (e. g. books, consumer goods, people). In a CTS the users assign freely chosen terms (i. e. *tags*) to specific resources with the purpose of referencing those resources later on with the help of the assigned tags.

By considering also other users' tags serendipitous discovery of new, previously unknown resources is possible via so called *tag browsing*, i. e. all resources that are annotated with the same tag(s) as a decisive resource will be referenced. For an overview of CTS see [8, 6]. Current CTS usually consider the resources being tagged as a whole. Thus, tag based search produces a hit list that contains entire resources, although the tags describing these resources might refer only to specific parts of that resources. In case of electronic documents, as e. g., HTML encoded documents, single parts or fractions of the document can be referenced, if the document author – and not the document reader – has provided anchors encoded within the document for the identification of those document parts. In case of multimedia data, as e. g., recorded video, specific document parts – i. e. single video scenes – can be identified and annotated by using MPEG-7 metadata.

We propose the combination of a CTS with an automated annotation system for synchronized multimedia presentations that is able to annotate single parts of multimedia data with user defined tags. We have developed a system for automated annotation of synchronized multimedia documents that is focused on lecture recordings. The video recording of the lecturer is synchronized with a recorded desktop presentation [11], which serves as a basis for an automated creation of MPEG-7 metadata and enables content-based annotation of single scenes within the video recording. This MPEG-7 annotation is endorsed with user defined tags to enable a personalized search that can be performed on a large multimedia database as well as within a single multimedia file.

The paper is structured as follows: Section 2 gives a short overview on related work concerning video annotation systems and CTS. Section 3 illustrates our approach that combines tagging information with MPEG-7 metadata and shows how to apply this combined information for content based search within multimedia data. Section 4 concludes the paper with an outlook on how to apply our concept of partial document tagging to the processing of large text documents.

2. RELATED WORK

In this section we give a short overview on current video annotation systems and CTS. The service that we are focusing on in this paper combines collaborative tagging and traditional video annotation. MPEG-7 [4, 9] is an XML based markup language for the description and annotation of multimedia data. We have developed an MPEG-7 based annotation service that is focused on the automated annotation of lecture video recordings. The recorded video is synchronized with a desktop presentation given by the lecturer. The textual content of this presentation is used to annotate single sections of the video with weighted descriptors. A keyword based search can be performed on the annotated video recordings resulting in a list of video sections related to the search term (see [11] for a more detailed description). Repp and Meinel have proposed a similar video annotation system based on the transcription of spoken language in the audio part of the video data [10]. In a similar way Hauptmann et al. extracted textual annotation from recorded video by optical character recognition (OCR) and speech recognition [7]. The major difference between our new approach and the just mentioned video annotation systems is that there, the annotation is conducted in a centralized way either by the author or producer of the video or by an independent automated system. The user of the video data does not have the possibility to add his own annotations and to make them available for the system's search facilities. Furthermore, reliability of speech recognition itself depends on training data and it is difficult to identify context and semantically connected sequences [11].

On the other hand, there are manual multimedia data annotation systems that enable the user to connect personal annotations to single scenes of a video recording [12, 1, 3]. In difference to our approach, those multimedia annotation systems are focused on personal annotations only. Indeed, they enable personalized search facilities, but without simultaneously providing a platform that is able to use annotations from different users in a collaborative way.

CTS enable personalized annotation of resources that can be utilized collaboratively by all users. YouTube [2] is a rather popular system for the collaborative annotation of video data. But, YouTube only allows the annotation of the video data as a whole and not the annotation of single parts of a video document. The majority of available video clips in YouTube is rather short and most times those clips only cover a single subject. Thus, for YouTube it is probably not necessary to provide a possibility for partial document annotation. Our system is focused on lecture recordings, where most lectures cover a variety of different topics. By providing partial document annotation facilities the user is able to annotate single video scenes that are related to a specific topic according to his own interests. By considering also those annotations that have been provided by other users, the system enables the discovery of related (similar) video scenes by tag browsing.

3. INTEGRATING COLLABORATIVE TAGGING INFORMATION AND MPEG-7

3.1 MPEG-7 Encoding

This section describes how MPEG-7 metadata can be used to maintain collaborative tagging information. MPEG-7 is an XML based markup language for the description of multimedia metadata. Besides various standard metadata information MPEG-7 enables the identification and annotation of distinct spatial and temporal segments within multimedia data. For our purpose, the description of temporal decomposition of video data is essential. Thereby, MPEG-7 allows the identification and annotation of overlapping temporal segments, which is a prerequisite for storing collaborative tagging information that is provided by different users.

Video segments can be annotated with various information by utilizing the <TemporalDecomposition> tag of the MPEG-7 metadata description scheme. Each video segment is identified and annotated with the <VideoSegment> element (see Fig. 1). Within each <VideoSegment> the elements <MediaTimePoint> and <MediaDuration> specify the segment's temporal location within the video stream (see Fig. 2). For textual annotation MPEG-7 provides the tags <KeywordAnnotation>, <FreeTextAnnotation>, and <StructuredAnnotation>. The information connected to these tags can be utilized for a keyword based search within the video data facilitating a fine-grained access.

```
<Mpeg7 xmlns="...">
 <Description xsi:type="ContentEntityType">
   ...
  <MultimediaContent xsi:type="VideoType">
   <Video>
    <MediaInformation>
     ...
    <TemporalDecomposition>
     <VideoSegment>...</VideoSegment>
     <VideoSegment>...</VideoSegment>
      ...
    </TemporalDecomposition>
   </Video>
  </MultimediaContent>
 </Description>
</Mpeg7>
```

Figure 1: Simplified MPEG-7 basic elements.

```
<VideoSegment>
 <CreationInformation>...</CreationInformation>
  ...
 <TextAnnotation>
  <KeywordAnnotation>
   <Keyword>cat</Keyword>
   <Keyword>mouse</Keyword>
  </KeywordAnnotation>
  <FreeTextAnnotation>
   billy the cat is catching a mouse
  </FreeTextAnnotation>
 </TextAnnotation>
 <MediaTime>
  <MediaTimePoint>T00:05:05:0F25</MediaTimePoint>
  <MediaDuration>PT00H00M31S0N25F</MediaDuration>
 </MediaTime>
</VideoSegment>
```

Figure 2: Simplified <VideoSegment> element.

For the integration of collaborative tagging information into the MPEG-7 metadata description schema an obvious approach would be to use the <Keyword> element associated with each video segment. But, for each set of tags additional user dependent information has to be stored to facilitate a personalized search. Collaborative tagging information can be encoded as a tupel

$$(\{tagset\}, username, date, [rating]),$$

where a set of tags is supplemented by user, date, and auxiliary (optional) rating information. Therefore, instead of the <Keyword> element we use the <MediaReview> element, which allows a video segment to be annotated with user specific textual information including also a rating indicator (see Fig. 3). The tagset denotes the set of all tags that a distinct user has employed to annotate a video segment. It is

```
<CreationInformation>
 <Classification>
  <MediaReview>
   <Rating>
     <RatingValue>9.1</RatingValue>
     <RatingScheme style=”higherBetter”/>
   </Rating>
   <FreeTextReview>
     tag1, tag2, tag3
   </FreeTextReview>
   <ReviewReference>
    <CreationInformation>
      <Date>...</Date>
    </CreationInformation>
   </ReviewReference>
   <Reviewer xsi:type=”PersonType” >
     <Name>Harald Sack</Name>
   </Reviewer>
  </MediaReview>
  <MediaReview>...</MediaReview>
 </Classification>
</CreationInformation>
```

Figure 3: Simplified `<MediaReview>` element.

represented as comma-separated list of tags and is encoded in the `<FreeTextReview>` element. The date of the last modification of the tagset is encoded with the `<CreationInformation>` element. The user identification is encoded in the `<Reviewer>` element, which is derived from the MPEG-7 agent type. Furthermore, an optional rating indicator can be included to enable the ranking of video content. Thus, the `<MediaReview>` element provides the possibility to store all necessary collaborative tagging information. The `<MediaReview>` element is embedded inside the `<CreationInformation>` and `<Classification>` elements of a video segment. Within the `<Classification>` element several different `<MediaReview>` elements can be combined that each represent annotations from different users.

3.2 Browser-Based User Interface

For collaborative tagging of video segments the design of an efficient user interface is mandatory. Thus, we define three distinguished areas in the browser's user interface: the video display area (1), the tag display area (2), and the tag/segment definition area (3) (see Fig. 4 for an overview of the user interface). The tag display is organized as a tag cloud (2). The single tags are ordered alphabetically while their font size indicates additional information that can refer to frequency of usage or tag rating (according to the relevance indicator). We consider different display modes: either personal or popular tags can be displayed, while a *static view* includes all tags for the entire video in difference to a *dynamic view* that refers to tags used at a distinct point in time within the video. By pointing at a tag with the mouse device a list of video segments annotated with that tag will be displayed in a separate window (4). There, the video segments are represented by a miniature screen shot and by their starting time and end time. The user can select a particular video segment from the list for playback. On the other hand, the user has to get an overview of all (non disjunctive) segments that have already been annotated in the video. This information is displayed within a coordinate system with the x-axis representing the timeline and the y-axis representing overlapping sequences (5). By pointing at a video sequence within the coordinate system all tags referring to that segment are displayed. Besides user annotation, we also consider annotations provided by the author of a video resource. These annotations can include structural informations (cut points) as well as semantic information (tags, headings, comments). The interface provides the possibility to use the annotation given by the author as a default starting point for user dependent annotation. Alternatively, the video can be pre-cut at fixed time intervals that can be fine-tuned by the user. For selecting a new *video sequence* to be annotated, the user is able to mark starting time and end time simply by clicking special buttons in the video display during playback or/and by adjusting those cut points in a separate timeline display (6). After selecting a video sequence the user is able to add his tags in a separate tag definition window (7). For faster processing it is possible to place tags just at a specific *point in time* during the video playback without denoting an entire segment. Then, starting point and end point of a sequence being annotated with that tag is chosen using predefined or author-given cut-points. To consider the most important parts of a video a rating index is displayed along a separate timeline (8).

3.3 Searching Tagged MPEG-7 Metadata

CTS enable different ways of searching the system's resources that can be adapted to our multimedia search:

Personalized Search By utilizing his own set of tags the user is able to perform a search based on his personal information needs. These tags can be descriptive or functional by nature, i. e. they either describe a resource in general – and thus, are also useful for other users – or they draw the focus on a certain aspect that (most times) is only relevant for the user who supplied it. Esp. the functional tags are suitable to extend a general search according to personal information needs. As e.g., the user might tag several sequences of a lecture video that are relevant for an examination with the tag *exam*.

General Search By considering the (descriptive) tags of all users in combination with the original MPEG-7 annotations of the resource's author, a general keyword-based search can be performed.

Tag Browsing Here, we refer to the retrieval of all resources that are annotated with the same tags as a specific resource under current consideration. Now, esp. those resources become important that have been annotated with the same tags, but by other users. In that way the user is able to discover new resources that are considered to be similar to the original resource.

Social Networking Additionally, in CTS the inherent social network of users can be considered. To participate in a CTS the user has to register which often includes the delivery of a personal profile. Thus, a social network can be defined connecting users that are considered to be similar according to their profiles. On the other hand, users that have annotated the same resource (probably even with the same tags) can be considered to be similar. Thus, by browsing resources that have been annotated by similar users, new relevant resources can be discovered.

4. CONCLUSIONS AND OUTLOOK

We have shown how to integrate collaborative tagging information within a MPEG-7 framework to facilitate a search

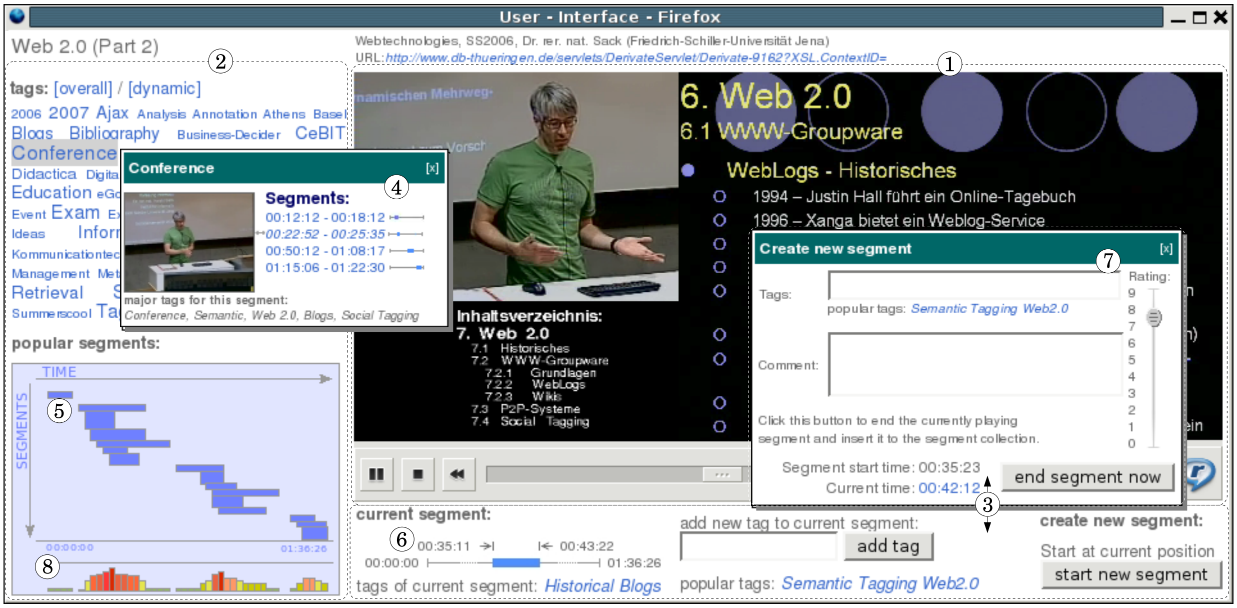
User - Interface - Firefox
Web 2.0 (Part 2)
Webtechnologies, SS2006, Dr. rer. nat. Sack (Friedrich-Schiller-Universität Jena)
URL: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-9162?XSL.ContextID=
1
2
tags: [overall] / [dynamic]
2006 2007 Ajax Analysis Annotation Athens Basel
Blogs Bibliography Business-Decider CeBIT
Conference
Didactica Digita
Education eGo
Event Exam Ex
Ideas Infor
Kommunicationtec
Management Met
Retrieval S
Summerscool Ta
Conference [x]
4
Segments:
00:12:12 - 00:18:12
00:22:52 - 00:25:35
00:50:12 - 01:08:17
01:15:06 - 01:22:30
major tags for this segment:
Conference, Semantic, Web 2.0, Blogs, Social Tagging
popular segments:
TIME
SEGMENTS
5
00:00:00
01:36:26
8
6. Web 2.0
6.1 WWW-Groupware
WebLogs - Historisches
1994 – Justin Hall führt ein Online-Tagebuch
1996 – Xanga bietet ein Weblog-Service
Inhaltsverzeichnis:
7. Web 2.0
7.1 Historisches
7.2 WWW-Groupware
7.2.1 Grundlagen
7.2.2 WebLogs
7.2.3 Wikis
7.3 P2P-Systeme
7.4 Social Tagging
Create new segment [x]
7
Tags:
popular tags: Semantic Tagging Web2.0
Comment:
Rating:
9 8 7 6 5 4 3 2 1 0
Click this button to end the currently playing segment and insert it to the segment collection.
Segment start time: 00:35:23
Current time: 00:42:12
end segment now
3
current segment:
6
00:35:11
00:43:22
00:00:00
01:36:26
tags of current segment: Historical Blogs
add new tag to current segment:
add tag
popular tags: Semantic Tagging Web2.0
create new segment:
Start at current position
start new segment

Figure 4: User interface combining collaborative tagging and MPEG-7 annotation.

function on multimedia data that is able to deliver distinct parts of interest within a multimedia document. In difference to current CTS our approach allows the annotation of partial documents which is important esp. for time-dependent media, as e. g., video data. A prototype of the proposed system for collaborative video scene tagging and retrieval is under current development.

The concept of collaboratively annotating partial video documents can be extended for other types of media, as e. g., for large text documents (textbooks). There, the users (document readers) should have the possibility to annotate distinct sections of the text document and to benefit from these annotations in a personal or collaborative way. One way to facilitate the identification of distinct sections within any type of document can be realized with the help of the document object model (DOM) [5]. The DOM representation of a document is a rooted graph (document tree), where different sections (at different levels within the document's hierarchy) are represented by nodes that can be linked with user annotations. Thus, with the collaborative annotation of partial documents a more focused and personalized search can be achieved for any type of document.

5. REFERENCES

[1] Ricoh movie tool, `http://m7itb.nist.gov/M7Validation.html`.

[2] YouTube - video sharing and tagging system, `http://www.youtube.com/`.

[3] D. Bargeron, A. Gupta, J. Grudin, and E. Sanocki. Annotations for streaming video on the web: System design and usage studies. *Computer Networks*, 31(11-16), 1999.

[4] S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 Standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

[5] Document Object Model Level 1 specification. http://www.w3.org/TR/REC-DOM-Level-1/.

[6] S. Golder and B. A. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208, 2006.

[7] A. G. Hauptmann, R. Jin, and T. D. Ng. Multi-modal information retrieval from broadcast video using OCR and speech recognition. In *JCDL'02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, Video and multimedia digital libraries, pages 160–161, 2002.

[8] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006.

[9] National Institute of Standards and Technology. NIST MPEG-7 Validation Service and MPEG-7 XML-schema specifications, `http://m7itb.nist.gov/M7Validation.html`.

[10] S. Repp and C. Meinel. Semantic indexing for recorded educational lecture videos. In *4th Annual IEEE Int. Conference on Pervasive Computing and Communications Workshops (PERCOMW'06)*, 2006.

[11] H. Sack and J. Waitelonis. Automated annotations of synchronized multimedia presentations. In *In Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, CEUR Workshop Proceedings*, June 2006.

[12] J. R. Smith and B. Lugeon. A visual annotation tool for multimedia content description. In *Proc. SPIE Photonics East*, Internet Multimedia Management Systems, pages 160–161, 2000.