Temporal Tagging: Implicit Behavior Identifies Points of Interest in Complex Event Recordings

Steve Whittaker and Vaiva Kalnikaite

Information Studies Department, University of Sheffield, UK.
211 Portobello St., Sheffield, S1 4DP, UK.
s.whittaker/v.kalnikaite@sheffield.ac.uk

Abstract

We introduce *temporal tagging,* a general tagging principle that takes into account *when* tags were generated. We use this temporal information to identify points of interest in extended recordings of complex events. We empirically demonstrate the utility of the approach in a successful user trial of a system ChattyWeb (CW) that allows students to access lecture recordings. CW coindexes students' digital handwritten annotations and lecture recordings. In our trial, students were able to use CW to successfully answer questions related to recorded lectures, and to use others' digital notes to facilitate this. We describe how tags are generated at low cost, as a side-effect of current user practice, addressing a common limitation of tagging approaches.

Introduction

There has recently been huge interest in social tagging applications, in particular those that allow users to apply personal descriptors to web content, e.g. documents or web pages. Systems such as del.icio.us and citeulike began as methods for users to manage their personal web content (in particular bookmarks) across machines, but rapidly evolved; allowing tags to be shared across multiple users to find information of common interest. Despite initial concerns that such personal tags might be esoteric and hence of little communal use (Hammond et al., 2005), recent studies indicate that users tend to agree on common tags and are able to use others' tags to access new documents. There is also good evidence that tags tend to stabilize after initial periods of system usage (Golder and Huberman, 2006, Millen et al., 2007). The general utility of social tagging systems is indicated by their spread from the general internet into corporate domains (Millen et al, 2007, Damianos et al., 2007).

While early work largely focused on tagging textual content, social tagging techniques may have even greater benefits where content is non-textual, e.g. video, pictures or speech data. Retrieval algorithms for pictures and video are much less effective than textual search (the so called 'semantic gap' problem). Social tags can provide important metadata to supplement low level features that can be automatically extracted for these media types. Popular systems such as Youtube and Flickr rely on social tags to support search. And even for textual search, adding social metadata can improve retrieval over purely textual techniques (Millen et al., 2007).

This paper looks at the application of social tags in a new domain; identifying points of interest in recordings of complex everyday events. Advances in technology are now making it possible to capture everyday events. 'Lifelogging' tools are being built that automatically record everyday video, images and audio (Dickie et al., 2004, Sellen et al., 2007, Vermuri et al., 2004). Reductions in storage costs are making it possible to record huge amounts of data and databases are being built to organize such data (Gemmell et al., 2006).

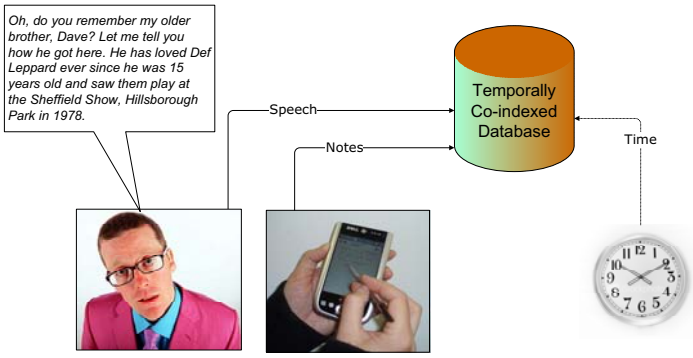

Figure 1. Temporal Tagging Principle

While there are many potential benefits to Lifelogging technologies in supporting fallible human memory, one major limitation is *retrieval.* How can we extract key *points of interest* from such complex media streams, where

current retrieval technologies are known to be weak? Even in constrained domains such as capture of meetings or lectures, it has proved difficult to generate adequate summaries or to identify key events such as actions using standard AI machine learning techniques (Purver et al., 2007).

Instead we take a different approach to analyzing complex event data, using people's *implicit behavior* as an indicator of points of interest. Our target application helps students to access recordings of lectures (although our technique can be applied to any multimedia recording). It is well known that people take notes about those parts of a lecture or meeting that they want to remember. Our application allows users to record speech data while simultaneously taking digital notes to annotate those recordings. It exploits the insight that knowing *when* tags are generated offers important clues about *points of interest* in an unfolding temporal event. Specifically, we use information about *when* users generate annotations, to indicate which parts of the lecture/meeting users think are important. In this way annotations can be seen as temporal markers on the media stream.

Figure 2. ChittyChatty Interface

Our research extends the approach taken by many social tagging systems, by exploiting temporal information about tags. We are not only interested in *what* tags are applied to data, but also *when* those tags are applied – a principle we call *temporal tagging*.

Figure 1 illustrates temporal tagging. Users take digital notes as they listen to the speech. These notes are then temporally co-indexed to the speech recording. Our interface allows users to browse back though these notes, and to use them as indices back into the underlying speech. By clicking on a note, users can replay exactly what was said at the time that the note was taken. In this way notes serve as a visual analogue to the underlying speech. This finesses many of the problems of accessing complex speech or video records (Whittaker et al., 2007) – allowing straightforward fine-grained access to points of interest in a long recording. We also address a key challenge to social tagging; a major benefit of this type of system is that there is *no additional annotation cost* compared with current user practice; people already take pen and paper notes to remember complex spoken information, and our application directly supports this. Temporal tagging annotations are therefore generated as a *side-effect* of people's usual behaviors and they require *no additional annotation effort*. This principle is not novel; various systems have been built that follow this approach (Landay,1993, Stifelman et al., 2001), although there has generally been little evaluation of the success of such systems.

In this paper we evaluate temporal tagging by testing a system we built to support access to student lectures. It is intended to work analogously to a paper notebook. Users take notes about a lecture and these notes are recorded digitally on a PDA application, ChittyChatty, which also records the lecture (see Fig 2). Notes and recordings are then uploaded to a web application, ChattyWeb (see Fig 3) that is made available to the class. Clicking on a given note replays exactly what was being said when that note was first taken.

Of course, there is a time lag between hearing something of interest and generating the relevant note about it. To control for this, we introduced a temporal offset. Pilot studies of user annotation behaviors, showed that offsetting notes by moving them forward 1.5 seconds, led them to be better aligned with relevant content. However it is important to note that absolute precision of alignment is not necessary for the successful functioning of the application. As we shall see, users can tolerate slight misalignments by for example clicking on an earlier note if they find that they have 'missed' the information they were looking for.

We first outline our evaluation demonstrating the utility of temporal tagging. We then describe an extension of the principle using pictures to annotate speech, and conclude with a discussion of the implications of our findings.

ChattyWeb: an evaluation of collaborative annotations to access lecture recordings.

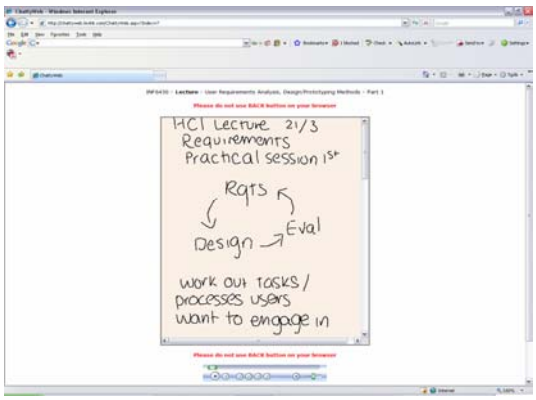

Figure 3. ChattyWeb Interface showing user notes

Fifteen students used Chattyweb (CW) during a 12 lecture course. We collected both controlled and naturalistic data about various aspects of its usage. Lecture recordings and student annotations were uploaded to the web. Students were encouraged to use these annotated recordings as a pedagogical aid to help them controlled access to prior lectures; allowing them to re-access complex discussions or parts of the lecture they had difficulty in understanding.

We logged student access to these recordings over the duration of the course to determine how and when students used annotations to access recordings. Did they browse an entire lecture or were they focused on specific aspects of that lecture? Did they make strategic use of the recordings to focus on lectures that were more relevant to their evaluation assignments? We also tested whether providing the annotated records improved students' performance on the course: did students who accessed the system more perform better overall on course assignments, than those who didn't use it?

We supplemented this naturalistic data with more controlled studies from class quizzes, where students were allowed to use ChattyWeb. We asked students questions about information provided in the lectures and compared how well they were able to answer these questions, with and without access to the temporally annotated recordings. We were also interested in the extent to which annotated recordings could substitute for lecture attendance. In principle, students with access to the recordings should be able to answer any question about the lectures, because these are a verbatim record. However our hypothesis was that socially tagged recordings would help non-attendees partially answer such questions, but they would not perform as well as students who had been present at the lectures themselves.

Finally we asked students a series of qualitative questions in a web survey intended to elicit their reactions to the accuracy, efficiency, usability and 'fun' of using the system.

Results

Students found the system extremely easy to master and a 5 minute tutorial was sufficient for them to learn how to use it. In a separate study, we showed that there was a strong overlap between handwritten annotations taken on ChittyChatty and those taken on standard pen and paper (Kalnikaite and Whittaker, 2007). We are therefore confident that the system supports natural handwriting annotation activities.

Quiz questions

Does an annotated recording improve quiz performance? We compared students' performance on quiz questions with and without access to the annotated digital record. Fig 4 shows that students performed better on quiz questions when they had access to the annotated record. A one way ANOVA with Accuracy as the dependent and Retrieval Strategy as the independent variable showed, as predicted, that retrieval was more accurate when CW was used ($F(1,121)=14.63$, $p<0.0001$). This clearly indicates that students are able to use handwritten annotations to retrieve information from complex speech recordings.

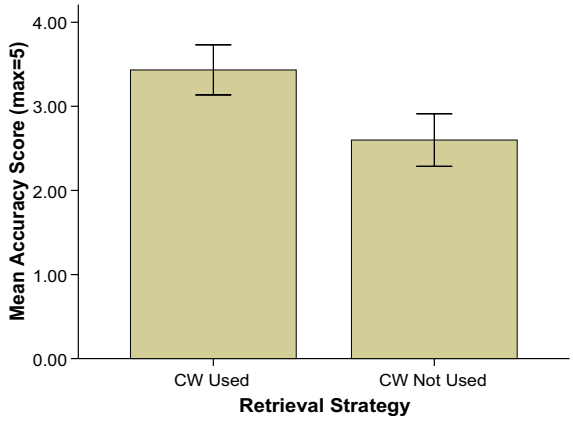

Figure 4. Retrieval Accuracy

Is the annotated recording a substitute for attending the lecture? Fig 5 compares performance on quiz questions for people who had been to the lecture compared with those who had not attended it. It shows that as we expected attendees are better able to exploit annotations to access relevant parts of the recordings. An ANOVA with Accuracy as the dependent variable and Attendance as the independent variable shows, as expected, that people perform better at answering questions when they attended the relevant lecture ($F(1,78)=4.9$, $p<0.03$). This result shows clear benefits to social tagging; non-attendees missed the original lecture and were thus reliant on others to tag material for them. Although non-attendees are not as effective as attendees at answering quiz questions, they are still able to answer over 40% of questions using the annotated record, showing the benefits of temporal annotations. This is an important result given concerns that students voiced about the quality and legibility of others' annotations.

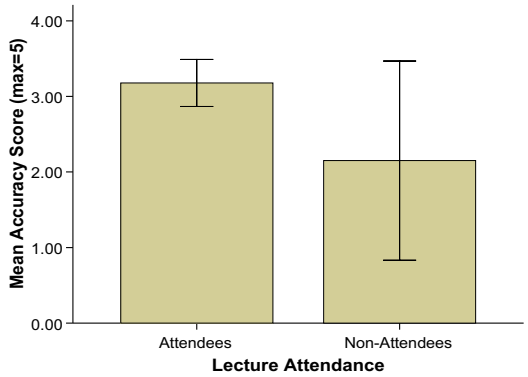

Figure 5. Attendance & Accuracy

Naturalistic Usage

We also looked at naturalistic access to the annotated record. An access session was defined as an unbroken interval spent accessing a specific lecture. If the student accessed another lecture or logged off, their session was deemed to have ended. There were large differences in access patterns between the students. Some accessed CW multiple times during the course, others only once or twice. However, a final group of six students never accessed the system. Those students who accessed the system had an average session lasting around 93 secs. There are also differences in the types of sessions we observed: 63% of sessions were short, lasting less than a minute, whereas for the remainder the average length of session was 244 secs.

How might we interpret these different patterns of usage? Shorter sessions are consistent with students accessing the system to address a specific question, whereas longer ones suggest that students want to revisit an extended segment of a lecture. Note too, that there were no instances of students using the system to revisit or replay large sections of a 120 minute lecture.

We looked at *when* students accessed the annotated record, finding that the 2 lectures that were most closely related to course assignments received many more accesses than those that were less relevant. Fig 6 shows this.

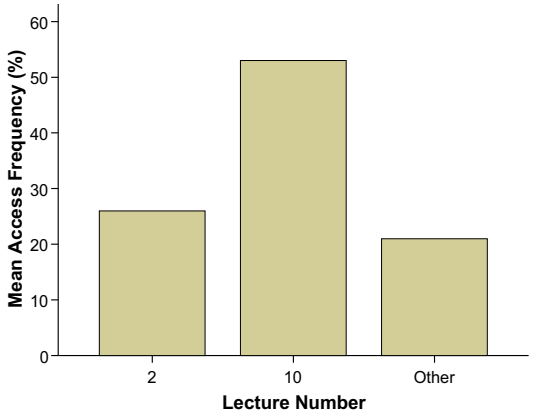

Figure 6. Frequency of Lecture Access

Finally we looked at the relationship between access behaviors and assignment scores. Somewhat to our surprise we found that students who accessed the system most often tended to perform worse overall on assignments, although this effect was just short of statistical significance ($r(14)=-.45$, $p = 0.09$). How can we interpret this? One possibility is that the system is used by weaker students who are concerned at their ability to effectively extract information from lectures. This interpretation is supported by the fact that students who accessed the system more, tended to perform worse on quizzes ($r(14)=-.54$, $p = 0.05$).

Subjective data

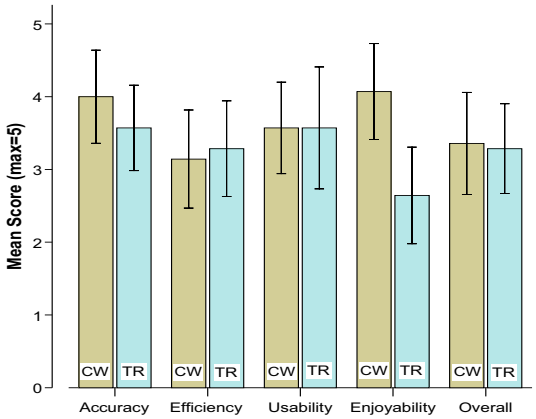

Figure 7. Subjective User Feedback

We also looked at students' reactions to the socially tagged recordings. These are shown in Fig. 7. Students judged that CW provided additional benefits over traditional tools ($t(13)=12.2$, $p<0.0001$), and it was perceived as being more Enjoyable to use ($t(13)=3.8$, $p<0.002$). However, there were no overall differences in perceived Accuracy ($t(13)=1.6$, $p>0.05$), Efficiency ($t(13)=-0.4$, $p>0.05$), and Usability ($t(13)=0.0001$, $p>0.05$).

Although students enjoyed using CW and found an added learning benefit, they did not perceive it to be more accurate, efficient or usable than traditional learning tools. We know that objectively, CW is less efficient than either referring to lecture handouts or relying on one's memory to answer questions, because speech is slow to access. The slight time off-set may also have led users to perceive CW to lack accuracy. To hear a current note, users sometimes had to select a previous note to compensate for the time delay during writing.

Students also made numerous comments about the tool which were consistent with our findings. These show that the main benefits of CW were jogging the students' memory about the lecture, sometimes, even when the notes themselves were not clear. The use of notes as memory joggers in combination with audio suggests that the tags themselves do not need to be of the highest quality.
"I believe ChattyWeb is very helpful because it provides the ability to record the lecturer. The notes that are taken are not very clear but with the [recording] of the lecture it's easy to remember the context of the lecture"

PiccyWeb: temporal image and audio co-indexing

We are also investigating the utility of temporal tagging for different *types* of tags, by extending our technique to pictorial tags. Instead of handwritten notes, we are using *images* as tags to access specific parts of a long recording.

Figure 8. Temporal Tagging with Images

Again we support a lecture environment, and students are given a picture taking device which records audio at the same time (Sony – Visual IC Recorder) see Fig 8. This not only provides a verbatim record of the lecture, but in addition an image tag temporally co-indexed to a specific part of the audio. This tool supports visual cueing to the underlying audio record of a lecture.

A complete recording, image tags and the time stamps are then placed in a custom built web application (PiccyWeb) for everyone in the class to share, see Fig 9.

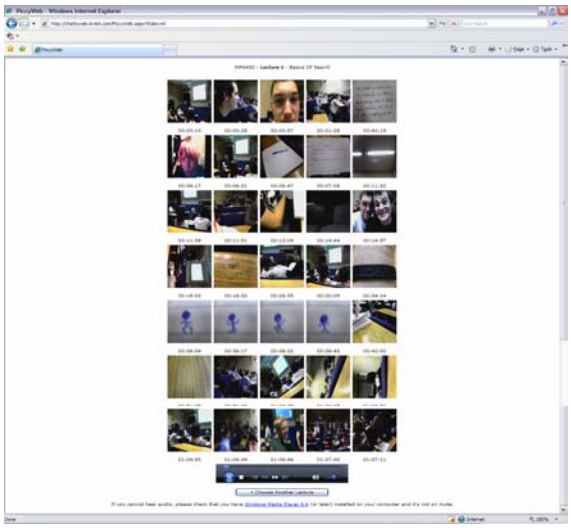

Figure 9. PiccyWeb Interface

When PiccyWeb is used for sharing the lecture recording, students simply click on the image tag that they want to hear more about and the audio that was captured at the same time as the image tag starts playing. Users can quickly navigate across different image tags by simply clicking on them. We are currently conducting a comparison of image versus handwritten annotations.

Conclusions

Our results build on and extend previous work on social tagging.

- We introduce a new social tagging concept – temporal tagging. This exploits information about *when* an annotation was generated, to provide useful information about points of interest within an extended temporal event. This contrasts with most prior tagging work which has looked at how tags are applied to *entire objects* (such as a document/video), rather than points of interest *within* that object. Temporal tagging promises to be valuable for recordings of complex events such as meetings, lectures or even everyday life recordings.

- We have also built and tested a system ChattyWeb which builds on this insight. One crucial design characteristic of the system is that it has *low annotation* costs. In most social tagging systems, adding tags is an additional activity, with an associated user cost, and it is well known that a shortage of tags can compromise their effectiveness and overall system viability. In contrast in CW, annotations are generated as a side-effect of an activity that users already engage in, namely taking notes about important parts of the lecture.

- Further we have evaluated the system in a naturalistic setting, showing the overall benefits of tags, specifically that:

  - Users provided with a temporally annotated record were better able to access information from a speech recording than those without such technology. These are striking results; despite the novelty of the system, and our concerns that students would not be able to read or understand the organization of others' notes, these data indicate that annotated recordings are effective in supplementing information already available from traditional tools; and that students perform better when they have such tools.

  - We also found when and why these recordings were most used; users exploited them to access specific facts about a lecture, rather than to extract the gist from the entire lecture.

  - The system was also positively evaluated by users who preferred it to traditional methods for accessing lectures.

Of course there are still many outstanding questions that we are beginning to address in this research. One immediate question is temporal consistency; the extent to which different users generate tags at the same time, and how we combine different annotations. Our current system makes no attempt to programmatically pool different user annotations, instead leaving it to the user to determine which set of annotations they want to attend to. However, an initial analysis of a controlled corpus of annotated

meetings recordings shows that there are indeed predictable points when multiple users simultaneously annotate, and furthermore that these events can be at least partially predicted from what is being said in the meeting.

We also plan to evaluate the utility of different tag types and are currently running a comparison of the utility of handwritten versus pictorial tags (PiccyWeb) for lecture recall. There are interesting trade-offs between pictures and notes as tags. Pictures are known to be easier to scan, but may not capture key lecture terms. Another possibility is to combine the two types of tags so that the same person can annotate lectures with both pictures and notes. We also plan to apply handwriting recognition to allow automatic processing of handwritten notes.

Another question concerns long term changes in annotation habits. Do users change their annotation habits as they become more experienced with the system? Do they take fewer notes, or change the type of notes that they generate? We have had 3 people use the CW system for almost 18 months – primarily in the context of meetings. An informal analysis of this data shows a shift over time in the types of annotations that users are taking. They suggest that long-term users are making more cryptic annotations as they gain experience with the system. Once users realize that they can retrieve underlying content effectively with CW, they no longer take notes constituted of complete coherent sentences, instead becoming more reliant on a few key words. In addition, they begin to develop conventions for temporally tagging significant content, e.g. using asterixes (with no other accompanying notes) to indicate when interesting things were said in the meeting. Overall this suggests that experience with the system leads to a shift from semantically oriented tagging (where users are reliant on the words in the annotation to retrieve what was said) to more purely temporal tagging, where the words in the tag are less significant, and users are tagging the mere fact that an important thing had been said. We are currently conducting a more systematic analysis of these long term tagging habits.

In conclusion then we have designed and demonstrated the value of an approach that exploits the temporal structure of implicit user actions to identify key points in complex event recordings. We believe temporal tagging has considerable general utility, not least because it addresses a major limitation of many social tagging systems, as tags are generated as a side effect of normal user practices.

References

Abowd G.D., Harvel L.D., Brotherton J.A. Building A Digital Library of Captured Educational Experiences. In Proc. ICDL 2000, Kyoto, Japan (2000), 395-402.

Damianos, L., et al. (2007). Exploring the Adoption, Utility, and Social Influences of Social Bookmarking in a Corporate Environment. 40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007. , Waikoloa, HI, USA

Dickie C., Vertegaal R., Chen D., Fono D., Cheng D., Sohn C. Augmenting and Sharing Memory with EyeBlog. In Proc. CARPE 2004. NYC, ACM Press (2004), 105-109.

Gemmell J., Bell G., Lueder R. MyLifeBits: a personal database for everything. In Comm. of the ACM, 49 (1) (2006), 88-95.

Hammond, T., T. Hannay, et al. (2005) "Social bookmarking tools (I) A General Review." D-Lib Magazine 11, DOI: 1082-9873

Landay J. A., Kaugmann T.R. User Interface Issues in Mobile Computing. In Proc. WWOS-IV 1993, IEEE Press (1993), 40-47.

Millen, D., Yeng., M., Whittaker, S., and Feinberg, J. (2007). Social Bookmarking and Exploratory Search. ECSCW 2007, Limerick, Ireland.

Purver, M., Dowding, J., Niekrasz, J., Ehlen, P., Noorbaloochi S., and Peters, S. Detecting and Summarizing Action Items in Multi-Party Dialogue. In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007.

Kalnikaite, V., and Whittaker, S. (2007). Does taking notes help you remember better? At HCI Workshop on Supporting Human Memory, Lancaster, UK.

Mann S., Fung J., Aimone C., Sehgal A., Chen, D. Designing EyeTap Digital Eyeglasses for Continuous Lifelong Capture and Sharing of Personal Experiences. In Proc. CHI 2005 Conference on Computer Human Interaction, ACM Press (2005).

Sellen A. J., Fogg A., Hodges S., Wood K. Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using SenseCam. In Press to appear in. CHI 2007 Conference on Computer Human Interaction, NY: ACM Press.

Stifelman L., Arons, B., and Schmandt, C. Augmented real-world objects: a paper-based audio notebook. In Proc. CHI 1996 Conference on Computer Human Interaction, Vancouver, Canada, 2001, ACM Press, New York, NY (2001), 182 – 189.

Vermuri, S., Schmandt, C., Bender, W., Tellex, S., & Lassey, B. (2004). An audio-based personal memory aid. Proc. Ubicomp 2004, 400-417.

Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. (in press). Design and Evaluation of Systems to Support Interaction Capture and Retrieval. To appear in Personal and Ubiquitous Computing. http://dx.doi.org/10.1007/s00779-007-0146-3