Giving Context to Sounds through Mediation of Physical Objects

Shin-ya Sato
NTT Network Innovation Laboratories
3-9-11 Midori-cho
Musashino-shi, Tokyo
180-8585, JAPAN
shin-ya.sato@acm.org

Masami Takahashi
NTT Network Innovation Laboratories
3-9-11 Midori-cho
Musashino-shi, Tokyo
180-8585, JAPAN
t.masami@lab.ntt.co.jp

Masato Matsuo
NTT Network Innovation Laboratories
3-9-11 Midori-cho
Musashino-shi, Tokyo
180-8585, JAPAN
matsuo.masato@lab.ntt.co.jp

UbiComp'13 Adjunct, September 8–12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2215-7/13/09
DOI:http://dx.doi.org/10.1145/2494091.2494117

Abstract
We describe the concept of and approach for combining conceptual information produced by humans and data that convey situations of the real world without any modification or interpretation, which can be thought of as a method for bridging the Web and the real world. We conducted an experiment to validate our concept by making associations between everyday topics or situations and their characteristic sounds. We discuss the preliminary results obtained in the experiment.

Author Keywords
Context recognition, topic models, Web mining

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
Two types of information are usually used when people perform tasks that involve actions in the real world. One is abstract, conceptual information produced by humans and the other, more likely called data than information, conveys situations of the real world without any modification or interpretation such as visual and sound data. These two types of information become much more useful when they are used in combination. People actually

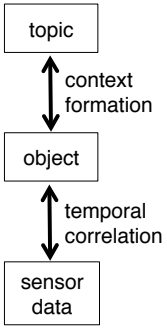

Figure 1: Schematic representation of our approach for bridging Web and real world through mediation of physical objects.

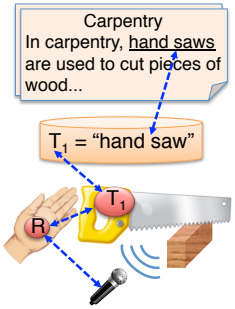

Figure 2: Experiment of associating topics and their characteristic sounds.

benefit from information combination when using an interactive navigation service such as Google Street View, where photos showing the surroundings of each spot are integrated into a map.

We have been developing an approach for bridging the Web and real world for bringing about the benefits of combining information extracted from the Web and sensor data, which is another pair of conceptual information and real world data. We describe the concept of combining conceptual information produced by humans and data that convey situations of the real world without any modification or interpretation and discuss preliminary results we obtained in an experiment.

Approach for building bridge

In our approach, physical objects in the real world play the role of mediators. Our approach for bridging the Web and real world through the mediation of objects is schematically summarized in Figure 1. A topic is equated with a group of documents written about a topic in our approach. We focus on documents on the Web that are written about things, events, and concepts related to our daily lives, which include blogs, social Q&A sites, and Wikipedia articles. An object is associated with a topic in which the object appears as one of the elements that construct the context of the topic. The object is also correlated with sensor data observed at the same time. By combining these associations, topics and sensor data are associated.

To help better understand our approach, let us discuss an experiment we conducted for validating our concept. In the experiment, we tried made associations between topics and their characteristic sounds as shown in Figure 2. For each object, a hand saw in this figure, a radio frequency identification (RFID) tag was attached (oval with ID T_1 in the figure). The tag ID could be translated into the name of the tagged object by looking it up in a manually prepared database. The object name was further associated with topics of Web documents in which the object name appeared as a keyword. Interactions between the tagged object and a participant of the experiment were detected using a wristwatch-type short-range RFID reader worn by the participant (oval with R in the figure). Furthermore, ambient sounds were captured in parallel throughout the experiment. The relevance between the object and a specific sound was then estimated based on the temporal correlation between detections of interaction involving the object and detections of the sound. Finally, by combining all relationships obtained so far, a concept and a sound, i.e., "carpentry" and the rasp of a saw on wood, were linked together.

Related work

Several studies have been conducted for understanding day-to-day human activities by making use of relationships between activities and physical objects involved in those activities. For example, in the study by Perkowitz et al. [3], directions (such as how-tos) written on Web pages are translated into models for activities, each of which is essentially a sequence of objects with probabilities of their involvement in the activity. Activities of a person in the real world are observed as interactions between the person and objects, or more precisely, contacts between an RFID tag on the object and an RFID reader worn by the person. Produced sensor readings are compared against the generated models to infer activities. We use this approach as a reference in linking topics and their characteristic objects. However, target activities, which correspond to topics in our approach, are not limited to those written in the form of directions, such as office meetings.

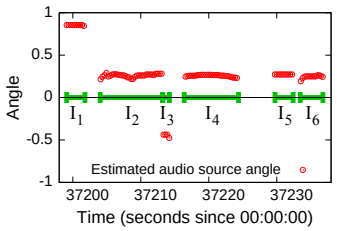

Figure 3: Temporal changes in sound source angles and contextually consistent time intervals.

Since what we hear in daily life is a mixture of sounds, we need to separate the target sound (e.g., rasp of a saw in the example in the previous section) from others to make links between topics and their characteristic sounds. We used a topic model, specifically Latent Dirichlet Allocation (LDA)[1], with the expectation that the target sound can be identified as a latent topic in audio signals (which can be called an "acoustic topic"). Similar approaches were adopted in a number of studies. For example, Raj et al. used LDA for speaker separation from mixed monaural recording [4], where each undisclosed speaker corresponds to a distinct acoustic topic. Our approach differs from the preceding one in that we intentionally discard low frequency components (lower than 3000 Hz) from the audio signals before applying LDA. This is because we would like to focus on sounds generated in conjunction with interactions between objects and people, rather than voice. Another novelty of our study is a new technique for audio signal segmentation that was developed for this experiment, which is explained later.

Experimental setting

In the experiment, we monitored sounds and interactions between objects and people that occurred within a compartment (about 7m x 5m) in our laboratory. The monitoring data were collected from 9:00 to 18:00. This data set is denoted as D in this paper. RFID tags were attached to 102 objects used in the office, which included computer peripherals, stationery, office furniture, and supplies. Three people participated in the experiment. Each participant wore a wristwatch-type RFID reader for detecting his/her contacts with the tagged objects. A Kinect sensor was used as an array microphone for capturing sounds and their source angles.

Identifying acoustic topics in audio signals

LDA is originally a model by which relationships between documents, words, and topics are explained. With this model, it is assumed that a document can be abstracted as an unordered collection of words (i.e., bag-of-words). An audio signal can be converted into a frequency spectrum by Fourier transform, which can be naturally regarded as bag-of-frequencies (BoF). By using this representation of audio signals, we extended the idea of LDA to a sound model that explains the relationships between audio signals, acoustic topics, and frequencies (frequency bins) in a similar way. That is, an acoustic topic is modeled as a probability distribution over the frequencies. In practice, a BoF for a given audio signal segment (equivalently a time interval) is obtained as follows: (1) calculate a power spectrum for each signal in the segment by discrete Fourier transform, (2) eliminate frequency components lower than 3000 Hz from the spectrum, and (3) obtain the sum of the obtained spectra. Acoustic topics were discovered by applying a commonly used learning algorithm [2] to the set of BoFs obtained from D.

For audio signal segmentation, we used sound source angles estimated using the Kinect sensor. Figure 3 shows actual temporal changes in the sound source angles. We observed time intervals $\mathrm{I}_1 \sim \mathrm{I}_6$ during each of which the range of fluctuation is small, which means sources of the sounds are spatially close to each other. Because of the temporal and spatial proximity of sources, it would be reasonable to expect that sounds occurring within such a single interval are generated by mutually related events. Based on this assumption, we chose intervals that could be identified in this way (contextually consistent intervals) for obtaining BoFs. Let $\mathbf{b}_I$ denote the BoF of a given time interval I.

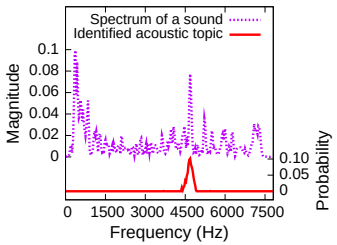

Figure 4: Spectrum of audio signal and identified latent acoustic topic.

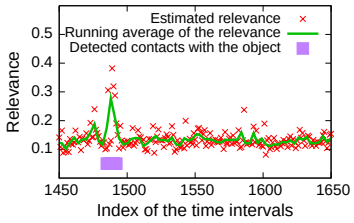

Figure 5: Relevance of contextually consistent time intervals to group of acoustic topics.

Associating acoustic topics with objects

Let $\{\mathbf{t}_i\}$ be the set of acoustic topics discovered from D. Then, $\mathbf{b}_I$ can be represented as the linear combination of acoustic topics $\mathbf{b}_I = \sum_i \alpha_{I,i}\mathbf{t}_i$. A physical object O and an acoustic topic $\mathbf{t}_i$ are considered to be correlated if $\alpha_{I,i}$ tends to be larger in response to detections of O in I. Figure 4 shows the acoustic topic $\mathbf{t}_\omega$ correlated with the object "white board marker" (solid lines) together with a spectrum of an actual sound (dashed lines). The acoustic topic $\mathbf{t}_\omega$ is distributed around 4700 Hz, which seems to correspond to the squeaky sound generated by the marker when it was in use. The graph also shows the correspondence between $\mathbf{t}_\omega$ and one of the distinctive peaks in the spectrum of the actual sound. It should be noted that we could find multiple acoustic topics associated with the object. Let $\mathbf{T}_\omega$ denote the set of acoustic topics associated with the marker.

We also confirmed that acoustic topics could be used as queries for searching relevant time intervals. By using elements in $\mathbf{T}_\omega$ as queries, we searched for contextually consistent intervals related to the white board marker in another data set D' captured on a different day. Let $\{\mathbf{t}'_i\}$ denote the set of acoustic topics discovered from D', and let $\{\mathrm{I}'_j\}$ be the temporally ordered set of contextually consistent intervals extracted from D'. Using a function $f(\mathbf{t}, \mathbf{t}'_i)$ that returns the degree of similarity between $\mathbf{t} \in \mathbf{T}_\omega$ and $\mathbf{t}'_i$, (e.g., a negative exponential of the Jensen Shannon divergence), we estimated the relevance between the marker and I'_j by $\mathcal{R}(\mathbf{T}_\omega, \mathrm{I}'_j) = \sum_{\mathbf{t}\in\mathbf{T}_\omega} \sum_i \alpha_{\mathrm{I}'_j,i} f(\mathbf{t}, \mathbf{t}'_i)$, where $\alpha_{\mathrm{I}'_j,i}$s are the coefficients of the linear combination $\mathbf{b}_{\mathrm{I}'_j} = \sum_i \alpha_{\mathrm{I}'_j,i}\mathbf{t}'_i$. Figure 5 plots $\mathcal{R}(\mathbf{T}_\omega, \mathrm{I}'_j)$ against j. A relevance peak around $j = 1490$ can be clearly observed. The figure also shows the points in time when the object was touched. We can see the temporal coincidence between the peak of $\mathcal{R}$ and the observed contacts with the object. These results indicate that intervals discovered based on acoustic topics are actually correlated with the object in question.

Linking topics with objects

Each physical object was also linked with topics of documents on the Web based on their names. First, relevant documents were located in predetermined information sources, which included a blog site, social Q&A site, and Wikipedia, by submitting the name of the object as a query to a search engine. Then, the located documents were divided into groups by using a clustering algorithm. We regarded the context shared by each group of documents as a targeted topic. In the case of white board markers, topics (situations), such as office meetings and school lectures, were discovered. Finally, these topics were linked with the squeaky sound identified as an acoustic topic through the mediation of the object.

References

[1] Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[2] Griths, T., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences 101* (2004), 5228–5235.

[3] Perkowitz, M., Philipose, M., Fishkin, K., and Patterson, D. J. Mining models of human activities from the web. In *Proc. of The 13th International World Wide Web Conference*, ACM Press (2004), 573–582.

[4] Raj, B., Shashanka, M., and Smaragdis, P. Latent dirichlet decomposition for single channel speaker separation. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (2006).