BUILDING OF HUMAN ACTIVITY CORRELATION MAP FROM WEBLOGS

Takahiro Kawamura, Nguyen Minh The and Akihiko Ohsuga
Graduate School of Information Systems, University of Electro-Communications, Japan

Keywords: Ontology, Mining, Weblog.

Abstract: Recently, context-dependent recommendation services for smart phone users have been publicly available, which provide any of useful information related to the users' current situation. These services are based on a technique called Human Activity Mining. However, a method collecting every event all the day, which is seen in some projects on ubiquitous computing has some problems. Therefore, this paper proposes an approach to build human Activity Correlation Map using ontologies from users' behavior logs on weblogs. Then, we show its efficiency through a preliminary evaluation.

1 INTRODUCTION

Recently, information recommendation service for a smart phone depending mainly on its location has been publicly available. In such a service, several kinds of daily events like web access logs from the phone, and informations from GPS and sensors like MEMS accelerometer are recorded as personal activity logs, then any of regular rules, that is, frequent event sequences are discovered in the logs, and the related information to the user's current situation are provided (weather, train schedule, traffic jam, daily schedule, etc), further the next possible activities are recommend (have dinner in nearest restaurant?), if any.

For example, "My Life Assist Service" (Kitsuregawa, 2008) which is a service of Japanese national project "Information Grand Voyage Project", and "Life Log" (Ohashi, 2008) which is a joint research project organized by a telecom carrier, KDDI are providing such services. Furthermore, one of those are already commercialized. Another telecom carrier, NTT DoCoMo just started an activity assist service called "i concier" (Yamada, 2008) on November 2008, which is sort of an agent service for mobile users.

However, a background technique of those services, Human Activity Mining currently has the following problems.

A. Because the events have lots of noise data, it's difficult to find meaningful event sequence.

B. The personal logs are not enough to find elements of surprise for him/her. However, Using other users' logs has difficulties of data amount, security, and privacy, etc.

C. Further, to find regularities from enormous size of the event data and make them rules, it needs substantial cost for building and maintenance.

Therefore, this paper proposes a way to build the human Activity Correlation Map by first extracting users' activity metadata from CGM (Consumer Generated Media) like weblogs and SNS, and combining the similar activities referring activity ontology. The actual activity recommendation, that is, the agent service would be realized by searching on the map from a certain topic node and selecting some adjacent information and activity nodes. We note that, however, this paper discusses part of building of the map.

2 HOW TO BUILD ACTIVITY CORRELATION MAP

First of all, the Activity Correlation Map in this paper is like Figure 3, 4. There is no definition of this term, but we consider it as a graph which expresses the relationship and its strength among a target activity and other activities most users tend to do with/after the target activity. Our proposed map has a unique feature which shows a modification relation between a subject of the activity and its object.

Kawamura T., Minh The N. and Ohsuga A. (2009).
BUILDING OF HUMAN ACTIVITY CORRELATION MAP FROM WEBLOGS.
In *Proceedings of the 4th International Conference on Software and Data Technologies*, pages 346-352
DOI: 10.5220/0002241203460352

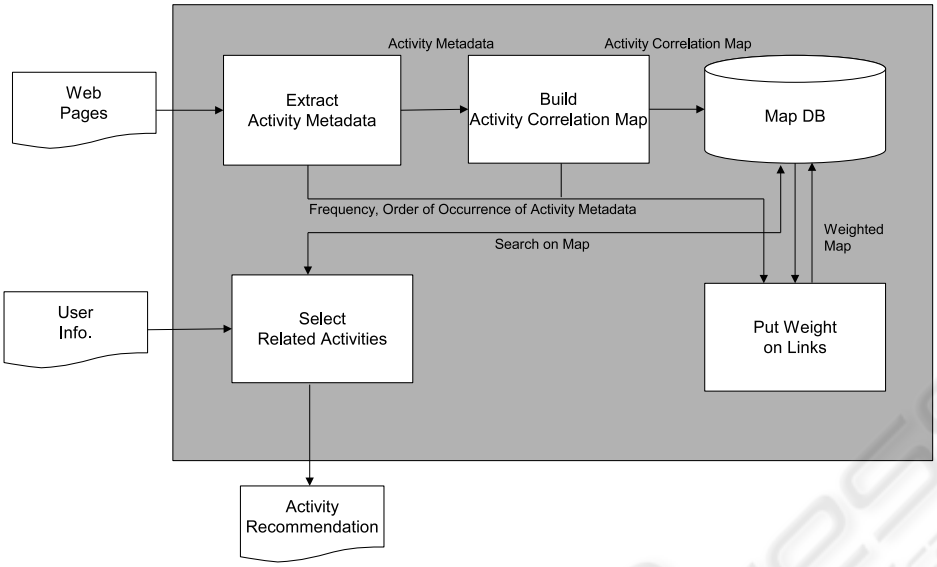


Figure 1: System Architecture.

Figure 1 shows an overall architecture to build the Activity Correlation Map. In our approach,

1. Firstly, we extract sentences include a certain topic key (search key word) from CGM like weblogs and SNS.
2. Next, by morphological analysis and syntactic parsing we find modification relations among the topic key and words (instances) of the activity ontology and product ontologies from the sentences, then organize them the activity metadata which are triples of S, V, {O,C}.

   The activity ontology is a graph representing types and names of human activities (see Figure 2). Then, the product ontology is a graph representing category and names of products (see Figure 2). Both of ontologies represent a type of activity and a category of product as a class, and its actual names (includes paraphrases) as instances of the class. They are based on SUMO (Pease, 2008), and their main classes are retrieved from SUMO, then the Japanese instances are added to them. As a result, those ontologies become relatively instance rich. The extraction of ontology from the sentence uses a word match with the instance.
3. Then, we combine those metadata referring the classes of words (instances) found in the metadata, and create the 2-dimensional Activity Correlation Map (Figure 3). We also combine the words in the case that one of which is included in the other one, even if the corresponding word is not found in the instances of the ontologies. Here, the activity and its subject, object are represented as nodes, and their modification relations are links between them.
4. Finally, we put weights on the links according to co-occurrence function considering consecutiveness and occurrence order of the activity metadata in the original CGM.

Additionally, when we recommend the possible information and activity to the user, we are now considering to select the related activities by searching around the topic node and some nodes representing the user's current situation (station name, etc), if any. On the other word, we would handle the information recommendation as a graph search problem.

3 MAP BUILDING EXPERIMENT AND EVALUATION

3.1 Experimental Result

This section shows an example to build an Activity Correlation Map for a topic key "Harry Potter". Firstly, we collect blog pages, etc. which contain the term "Harry Potter" using any of keyword search engines.

Secondly, we extract the activity metadata, that is, lists of words which have the modification relations from sentences in the above pages. The modification relations are extracted by the syntactic parsing using triggers like: the term itself "Harry Potter", the

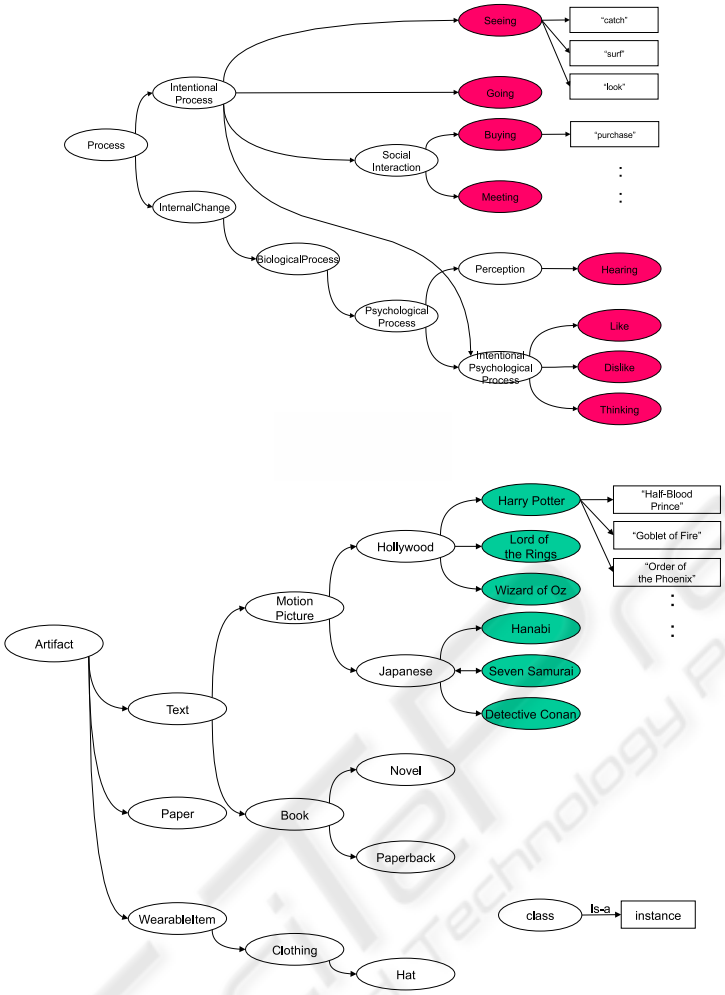


Figure 2: Part of Activity Ontology (above) and Product Ontology (below).

classes and their instances of the activity ontology like *look, read, rent*, etc., and those of the product ontology like subtitles of "Harry Potter". The extracted metadata is as follows:

Original Sentence: "The LEGO: Harry Potter game is also rumored to be published by Warner Bros."

Extracted Metadata: {Subject: Warner Bros, Verb: publish, Object: LEGO}

Then, by looking at the ontology classes of words (instances) in the metadata, we arrange synonyms among the metadata as one node. Further, we build the map by representing the modification relations between the words as the links between the nodes.

Finally, we put the weights on the links based on frequency and order of the word occurrence, and the words semantic closeness in the ontologies.

Figure 3 shows the actual output of the Activity Correlation Map. But, currently the system can only take Japanese sentences, because the morphological analysis and syntactic parsing engines depend on language. For illustrative purposes, therefore, Figure 4 shows a part of the map for the topic "Harry Potter".

This map was composed of the activity metadata extracted from 149 sentences of 26 blogs about "Harry Potter" collected by Google Blog Search on November, 2007 by using our reputation extraction engine from blogs (Kawamura et al., 2007). We should note that the size of the activity ontology created for this experiment is at most 85 nodes.

Currently, the weight on a link between two nodes are calculated as co-occurrence ratio of two words. The co-occurrence ratio of word A and B Co_{ab} is:

$$Co_{ab}\ (\%) = \frac{O_{ab}}{O_a + O_b - O_{ab}} \times 100$$

,where O_a and O_b are the number of sentences with occurrence of word A and B respectively, and

O_{ab} is the number of sentences with simultaneous occurrence of word A and B, that is, the modification relation from word A(B) to B(A).

3.2 Preliminary Evaluation

3.2.1 Map Analysis

Although we definitely need more accurate evaluation, we can at first qualitatively confirm that the following points are contained in the map.

- Practical activities the user would like to do such as seeing the movie, renting the DVD, reading the book, etc.
- Activities or topics of surprise that the user might feel interested such that:
  - dubbed-in voices in the movie attract lots of attentions (a dubbed version of this movie is more popular than a closed-captioned one, in spite of the fact that close-captioned movies are usually more popular in Japan),
  - there are many talks on "Detective Conan" (a cartoon program in TV) in the same blogs (we imagine that their audiences and readers would overlap with each other),
  - there are also many talks and even EC sites on hat coming up in the movie (the audiences and readers of that age seem to yearn for such an item),
  - people tend to be getting tired and sleep just after seeing the program (we guess that's because the screen time of the "Harry Potter" movies are relatively long).

3.2.2 User Test

Furthermore, we conducted simple user tests for quantitative evaluation, since it's difficult to have comparison with other systems [1].

In this experiment, we measured precision and recallconcerning the activity correlation maps for certain topics like "Harry Potter" from 10 users. The precision means here "how many nodes within a given distance from the topic node are useful activities (verb) or informations (subject, object) for the users". Also, we made a list of the related activities and informations to the topic which are interesting for the users before the experiment by questionnaire. Then, we measured as the recall "how many of those are contained in the nodes within a given distance from the topic node". The distance is here product of weights (co-occurrence) on links from the topic node, and we conducted two cases of 5.0% and 3.0%. In the experiments, we took the average of the precision and the recall of 10 users regarding three maps for three topics. We note here that the number of nodes within 5.0% distance was 38 as average of three topics and its link length was average 3 links. The number of nodes within 3.0% distance was 177 as average of three topics and its link length was average 6 links. Also, the number of nodes listed by the users in advance was 9 as average of 10 users. The result is shown in Table 1.

Table 1: Precision and Recall.

Topics	Topic A		Topic B		Topic C	
Distance (%)	5.0	3.0	5.0	3.0	5.0	3.0
Precision (%)	18.4	8.4	14.3	10.9	18.9	14.0
Recall (%)	77.8	96.2	62.5	87.5	75.0	87.5

Table 2: Precisions according to distance from 2 topics.

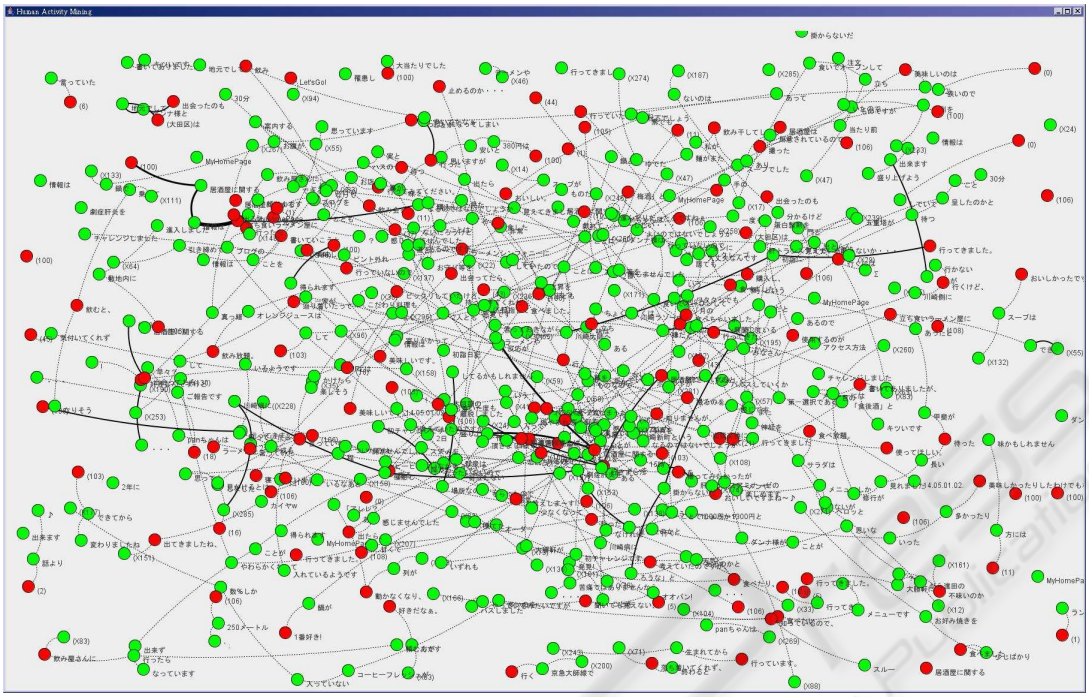
Topics	Topic A	Topic B	Topic C
Precision (%)	40.0	40.0	36.4

As a result, we found that both of 3.0% and 5.0% distance have high recalls, but remain low precisions. A reason is that sort of common activities and information which are not necessarily related to the topic are frequently revealed in the map, since a mass of general words are also co-occurred to the topic in the modification relation extracted from the ordinary blog sentences. So, we are considering to exclude them from the map by adding the general words to a filiter which is currently used to take out adverbs, etc.

Additionally, it is not preferable to show lots of choices with low accuracy to the user in the actual information recommendation service. Therefore, we should try to put more useful activities and information for the user in less choices as possible by narrowing range from the topic node. An approach is, for example, to obtain the second topic which is interesting for the user from a search history, etc., then select the nodes close to both of the first and second topic node. So, we made the users specify the second topic for each map in Table 1, and measured the precision for the nodes within the same distance (co-occurrence) from the first and second nodes. The result is shown in Table 2.

As a result, it was found that this approach can raise the precision in comparison with the case of a topic node. Thus, we can confirm that it is possible to extract practical activities and information related to

[1] Although we considered the comparison with collaborative filtering systems like Amazon.com, Amazon can neither output products not available on Amazon, nor the human activity such as other systems.



Note: Line width means link weight (co-occurrence ratio).
Verb is a red node, and subject and object is a green node as well as other figures.

Figure 3: System Output of the Map (in Japanese).

a certain topic from the Activity Correlation Map. In future, we will plan to realize a highly accurate recommendation service applying this approach.

In addition to the above, it is known difficult to set any metrics to measure the elements of surprise in recommendation technique, and to have objective experiment on this issue. So that, we simply measured by a questionnaire the number of nodes of activities and informations which are unexpected for the users. However, it is expected that the unexpected nodes for the users would increase because co-occurrence falls according to the distance from the topic. Then, we measured the average value of 10 users in the case of 5.0% distance. The result is shown in Table 3.

Table 3: Ratio of nodes of surprise.

Topics	Topic A	Topic B	Topic C
Surprise (%)	5.3	4.8	10.8

As a result, we found that the nodes within relatively close distance from the topic include a certain amount of the elements of surprise. In future, we will consider how to present the above practical nodes and the node of surprise to the users in the recommendation service.

4 DISCUSSION

The proposed method in this paper is trying to solve the three problems mentioned in section 1 by the following approach.

A. We are handling CGM, which has less noise than the daily event data from the sensors.

B. There is a possibility to find elements of surprise from any other users' blogs.

C. Since the Activity Correlation Map represents the links and its strength among the topic and other activities,searching on the links makes writing the rules unnecessary.

Also, we can make a group node which contains semantically similar nodes in the map by referring more abstract class of the product ontologies and the activity ontology. Then, if there is no exact node for a topic, we can recommend any activities or informations from similar topics (grouped nodes). It's a handling to a problem that we will not able to collect lots of the activity metadata for minor movies, etc. from CGM.

On the other hand, the Activity Correlation Map can be useful as a visualization tool for BI (Business Intelligence) which is a IT system supporting

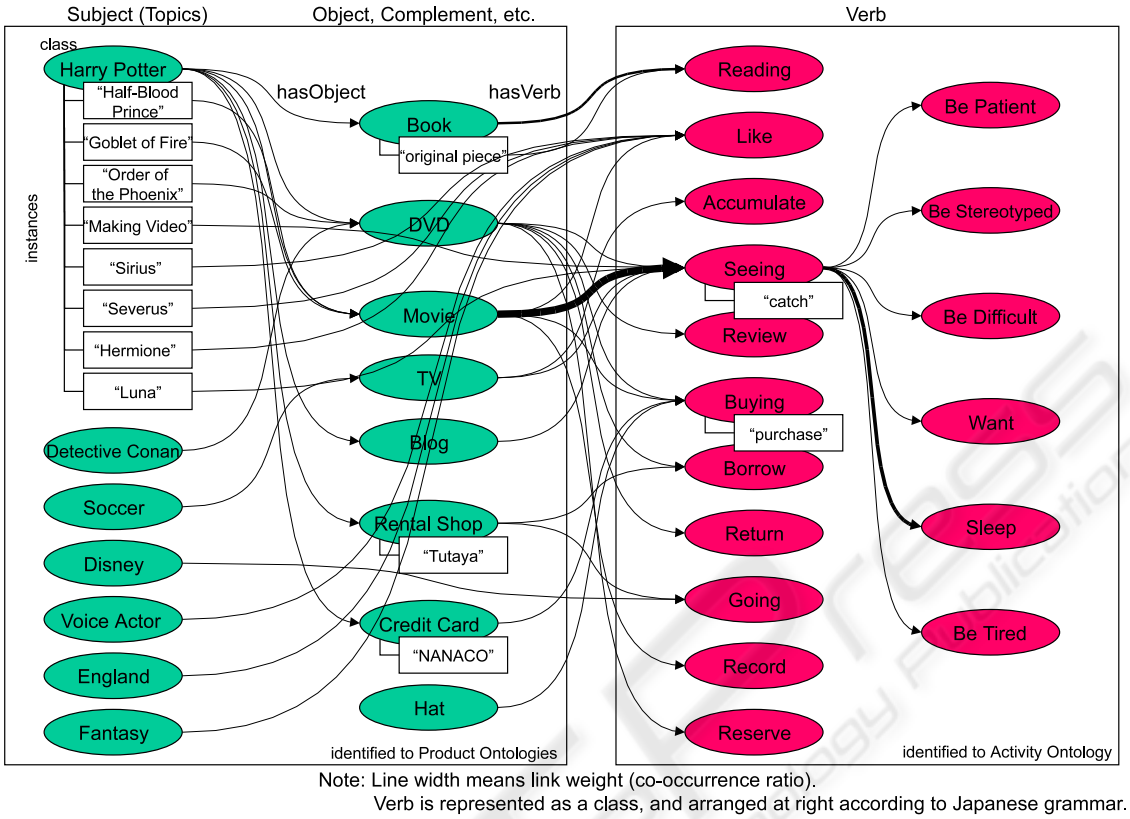


Figure 4: Part of Map for "Harry Potter".

corporate management. Here, data mining is attracting some attention as a core technique, and already many package vendors commercialized the mining tools as add-ons of DWH (Data Ware House), ERP (Enterprise Resource Planning). Now we are considering this map can be applicable as the visualization tool for customers' activity, which would help product marketing and merchandising. One similar tool is a product recommendation service like the collaborative filtering system of Amazon.com. Another application is BTA (Behavioral Targeting Advertisement). Recently, big companies have increased ads to BTA, especially combination to social media like SNS commands considerable attention.

5 RELATED WORK

In addition to the projects mentioned in section 1, there are related researches in the area of ubiquitous computing and web. For example, (Kamei et al., 2007) is trying to combine the real-space events, that is, the user's activity history taken from sensors which are embedded all over the space like other many ubiquitous researches, and some general knowledge gotten from the web.

Moreover, (Perkowitz et al., 2004) is also focusing on the combination of fine-grained sensor data and high-level structured information like time and space, and conducting an experiment on a certain data model.

In our approach, on the other hand, the granularity of data model (Activity Correlation Map in this paper) are predefined by the activity ontology. We would discuss the design of the activity ontology in the other paper.

In addition to the above, our Activity Correlation Map would be similar to ConceptNet (Arnold, 2008), a common-sense project of MIT with respect that its representation is a semantic network of human activity. However, it's different that ConceptNet is based on collective intelligence, that is, relying on human volunteer via the web. Therefore, it would be difficult to incorporate unconscious correlation of activities, and the element of surprise which is important in the future activity recommendation service can not be found in the network. So that we believe that our approach which is mining from CGM is complementary to ConceptNet approach.

6 CONCLUSIONS

As a future work, we are considering to develop an application to recommend the activities to the users related to his/her current situation and page views, based on the pre-build map for several topics. As mentioned before, a marketing application as part of BI is also under review.

In technical aspect, the development of the original co-occurrence function considering consecutiveness and occurrence order of the activities, and how to search on the map, i.e., how to determine the next possible activities to recommend when a topic is given are scheduled to be studied.

REFERENCES

Arnold, K. (2008). Conceptnet 3 - a semantic network representation of the open mind common sense project. http://conceptnet.media.mit.edu/.

Kamei, K., Yanagisawa, Y., Maekawa, T., Kishino, Y., Sakurai, Y., and Okadome, T. (2007). Tagging strategies for extracting real-world events with networked sensors. In *Proceedings of 1st International Workshop Tagging, Mining and Retrieval of Human-Related Activity Information*.

Kawamura, T., Nagano, S., Inaba, M., and Mizoguchi, Y. (2007). Mobile service for reputation extraction from weblogs - public experiment and evaluation. In *Proceedings of Twenty-Second Conference on Artificial Intelligence (AAAI-07)*.

Kitsuregawa, M. (2008). Outline of the "information grand voyage project" (my life assist, p.8). http://www.eurojapan-ict.org/ppts_forum_march/MasaruKitsuregawa.pdf.

Ohashi, M. (2008). Ubiquitous network - next generation context aware network - (lifelog, p.7). http://www.eurojapan-ict.org/ppts_forum_march/MasayoshiOhashi.pdf.

Pease, A. (2008). Suggested upper merged ontology (sumo). http://www.ontologyportal.org/.

Perkowitz, M., Philipose, M., Fishkin, K., and Patterson, D. J. (2004). Mining models of human activities from the web. In *Proceedings of the 13th International Conference on World Wide Web*.

Yamada, R. (2008). Docomo's change and challenge to achieve new growth (i concier, p.16). http://www.nttdocomo.co.jp/english/corporate/ir/ binary/pdf/library/presentation/081204/all_e.pdf.