# Tag Recommendation in Photo Sharing Services based on Multi-Granular Context Dependency

Xingtian SHI†, Fei CHEN††, Masashi TOYODA†, Min WANG††, and Masaru KITSUREGAWA†

† Institute of Industrial Science, the University of Tokyo
4–6–1 Komaba, Meguro–ku, Tokyo 153–8505 Japan
†† HP Labs, China
A505 No. 1, Zhongguancun East Rd, Haidian District, Beijing, 100084 P.R. China
E-mail: †{xingtian,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp, ††{fei.chen4,min.wang6}@hp.com

**Abstract** Contextual information, such as time and location, is now easy to access in many online photo sharing services thanks to Web 2.0 and the wide use of mobile devices. While context-aware recommendation system is developed to improve user's satisfaction on recommendations by tailoring some particular contexts, the effect of multiple contexts with granularity structure is of critical importance to the recommendation performance. Furthermore, how to detect the best context combination for recommendation with regard to the given input still remains a problem. In this paper, a generic framework to exploit time and location contexts is proposed for tag recommendation in photo sharing services which aims at enhancing user experiences and makes rich annotation of photos possible. A tag-sensitive method is also presented to automatically detect the context dependency for recommendations. Experiments conducted on real data set show that the proposed approach has significant performance improvement compared to competitive baselines.

**Key words** context-aware, tag, recommendation

## 1. Introduction

Recent years we have witnessed the fast growth of mobile devices and the explosion of online user-generated content (UGC) such as pictures and videos. More and more users take photos or record videos by their smart phones or other mobile devices and upload the multimedia information onto the sharing websites such as Flickr[1] and Youtube[2]. For later browsing, indexing, and searching, people add some word annotations to the uploaded files. This process is called tagging. In order to facilitate users' demand for more accurate and valuable tags, a tag recommendation system is needed. On the other hand, the small screen size of mobile devices also requires that the tag recommendations be more precise with only a few high quality candidates to display.

Meanwhile, mobile devices and Web 2.0 allow an easier access to rich contextual information. For the picture tag recommendation task, time and location the pictures were taken, camera models, and user profiles (name, personal interest, and social network) are all important contexts which we can utilize for better recommendation. For example, a tourist takes a picture of Eiffel Tower in Paris. If we know that at this specific location, many previous pictures in the log data are annotated with tag " Eiffel tower ", then we can easily recommend " Eiffel tower " to the user. If another user takes a picture on the night of 14 July at the same place, not only " Eiffel tower " will be recommended, tags such as " Bastille Day " and " fireworks " will also be good recommendations. As illustrated by the examples, an ideal tag recommendation system should utilize rich contextual information for better recommendation.

While a substantial amount of research has been already conducted on picture tag recommendation [2], [3], [5], such as picture content-based method or tag co-occurrence-based method, there is little work that takes contextual information into account. On the other hand, it is also a popular research problem discussing how to integrate contextual information in the recommendation model [1]. However, most works ignored the context structure and only static method is proposed to select the best context for recommendation. In this paper, we propose a generic framework that can exploit a rich set of contexts in recommendation models. A tag-sensitive method to detect the best context combination

1 : http://www.flickr.com/
2 : http://www.youtube.com/

and granularity is developed. Experiments show that our approach can benefit more from rich contextual information compared with competitive baselines.

The remainder of this paper is organized as follows. In Sec.2., related works are briefly introduced. We give the definition of context and formulate the Flickr tag recommendation task in Sec.3.. Sec.4. provides the evaluation results of our approach on a real Flickr dataset. We conclude and discuss future steps in Sec.5..

## 2. Related work

The two research areas that are close to this paper are picture tag recommendation and context-aware recommendation. In the research literature, different methods have been proposed for tag recommendation, such as using tag quality, tag co-occurrence, and object features as summarized in [6]. Context-aware recommendation has also been studied [1], but the previous work either focuses on a few contexts regardless of context granularity or just sets up simple rules to decide how to use the contextual information. To the best of our knowledge, there is no previous work on how to integrate rich context information into the tag recommendation system.

## 3. Proposed model

We first introduce the expression of multi-granular context and then give the pre-filtering paradigm of the Flickr tag recommendation task. Then the method to select the best context combination and granularity is proposed and a running sample is given to show how the approach works.

### 3.1 Multi-granular context

In the tag recommendation setting, the context refers to the information that characterizes the picture and its tags. The contextual information reveals the user's intent for taking the picture and has an impact on the user's tagging behavior. Time and location the pictures were taken, camera models, and user profiles (name, personal interest, and social network) are all important contexts which we can utilize for better recommendation.

Context has a hierarchical structure and the multiple-granular organization can be employed to help context-aware recommendation. For example, location context may have several granularities such as neighborhood, district, city, region, country, etc. If a photo is taken at <latitude=35.713573, longitude=139.774085>, the location context granularity might be $< Ueno - park(neighborhood), Taito - ku(district), Tokyo(city) >$. We call this expression of the multi-granular context as the **context granularity path** (abbr. cgp). Time context also has such property. 2012-02-08-14:30 may have context granularity path as $< daytime, Feb., winter >$. The context structure can be obtained by a pre-defined ontology. We define **context combination** (abbr. cc) as the cartesian product of all context granularity paths for one picture. For example. Table 1 shows all 9 possible context combinations for the above mentioned two granularity paths.

| cgp | **daytime** | **Feb** | **winter** |
|---|---|---|---|
| **Ueno-park** | <Ueno-park, daytime > | <Ueno-park, Feb. > | <Ueno-park, winter > |
| **Taito-ku** | <Taito-ku, daytime > | <Taito-ku, Feb. > | <Taito-ku, winter > |
| **Tokyo** | <Tokyo, daytime > | <Tokyo, Feb. > | <Tokyo, winter > |

Table 1 Example of possible context combinations

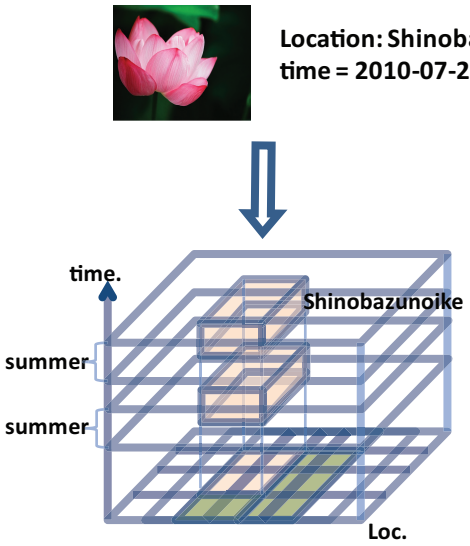



Fig. 1 Example of the pre-filtering paradigm

### 3.2 Pre-filtering recommendation paradigm

Follow the recommendation paradigm in [1], we employ pre-filtering data model in tag recommendation. We will first select the pictures according to certain context combination and based on the selected data, the recommendation is made to return tag candidates. Fig. 1 shows an example how the pre-filtering paradigm works. One picture is taken in *Shinobazunoike* at $2010-07-24\ 14:10$. We can select the pictures previously taken in *summer* and at *Shinobazunoike* and base on the selected data, recommendation is made.

### 3.3 Framework of the recommendation model

The recommendation task can be divided into two parts, **trainer** and **recommender**.

The trainer is an offline process to discover potentially predictive relationships between tag recommendation and contexts. The input of the trainer includes:

(1) Ranking function

$$R(t_i, t_j) :\rightarrow \mathbb{R}$$

where $R$ represents any method used to rank the candidate tags $t_j$ according to the given tag $t_i$. $R$ can be picture content based or tag co-occurrence based model.

(2) Data repository $D$ comprising

$$pic =< image, \{tags\}, cgp_{loc}, cgp_{time} >$$

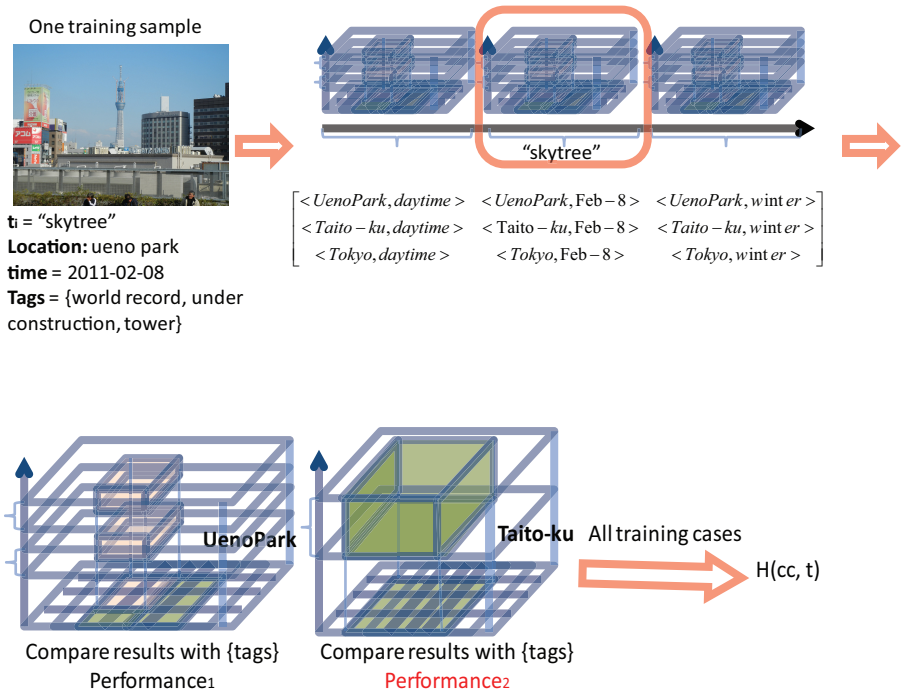



Fig. 2 Training procedure for one sample picture

where $pic$ is one picture and $\{tags\}$ are the tags that the user uses to annotate the picture.

The trainer outputs the **Detect function**, denoted as

$$H(pic) :\rightarrow best\ cc$$

which decides the appropriate context combination for $pic$. The trainer is built offline. How to generate the detect function is the main focus of this paper.

During the online phase, the recommender gets $pic'$ as input. $pic'$ includes image information, one given tag [3], and location & time information. $H(pic')$ will first detect the best context combination ($bcc$) and then based on the selected data on $D$ over $bcc$ as described in Sec. 3.2 (i.e., the selection operation over $D$ according to $bcc$), $R$ returns a set of tags as recommendation.

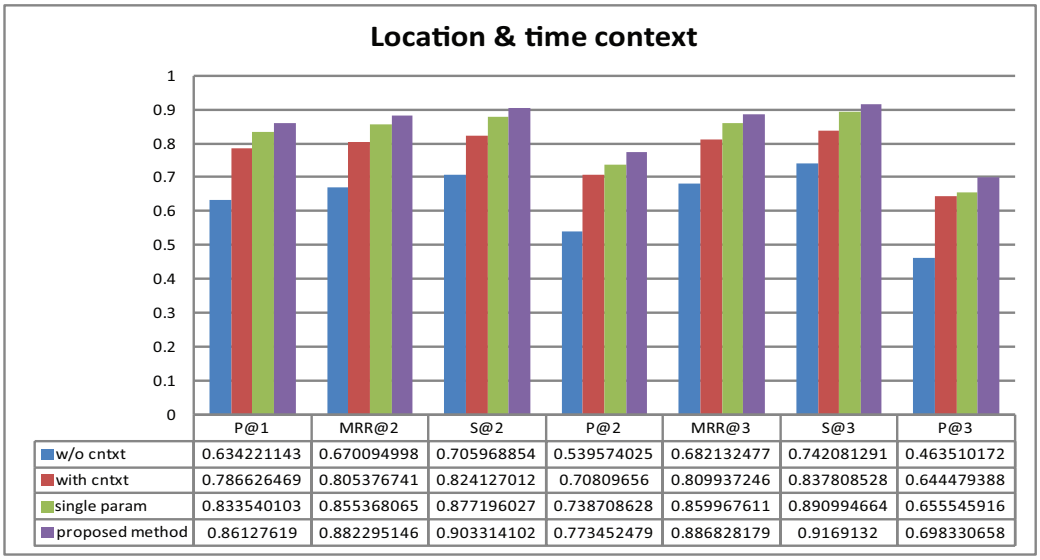
### 3.4 Design of $H(pic)$

We describe how to build $H$ in this section. There are some training samples to build the model. The main idea is that for each training sample, we select data over different context granularities and based on the select data, recommendation is made by $R$ and we test the recommendation performance compared with the ground truth of the training sample under some measurement. After all training samples are run, we can infer from the results the best context combination for each $cc$.

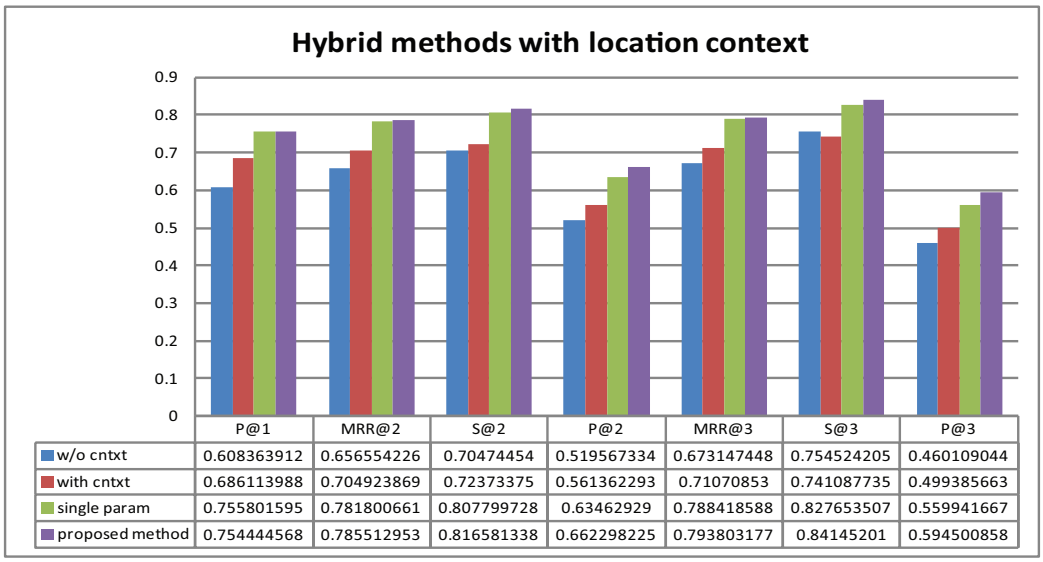
We observe that how to identify the appropriate context combination is not only dependent on the current contextual information. The given tag should also be taken into consideration since the given tag provides more information of the picture and reveals the user's intent of the tagging behavior. For example, when only location is considered, for pictures taken at $Ueno$ with tags such as $panda$ or $sakura$, the best location granularity for recommendation might be at $neighborhood$ level. However, if it is a picture taken also at $Ueno$ and the user tags the picture with $skytree$, the best location granularity might be $district$ level since Skytree is not actually located at $Ueno$. From the examples, we can learn that even for the pictures in the same context, the appropriate context granularly is different since the picture semantics are different. Therefore, we propose to find the best context combination for each $< tag, cc >$ pair with the context structure. For $< tag, cc >$, we select from the training samples which contain $tag$ and calculate the recommendation performance over different context combinations. We choose the context combination that has the maximum performance as the best combination for $< tag, cc >$.

Fig. 2 demonstrates the training procedure for one sample picture. The sample picture is about the $Skytree$ taken in $Ueno$. The given tag is $skytree$. To decide the best context combination for $< skytree, Uneo >$, we pre-filter the data according the the 9 possible context combinations and based on different selected data, we calculate the recommendation performance (e.g., precision) compared with the real tags the user input. After all applicable training samples are run, the training process will return the context combination for $< skytree, Uneo >$ that has the best performance.

3: We require user input at least one tag as the hint for recommendation.




| | P@1 | MRR@2 | S@2 | P@2 | MRR@3 | S@3 | P@3 |
|---|---|---|---|---|---|---|---|
| w/o cntxt | 0.634221143 | 0.670094998 | 0.705968854 | 0.539574025 | 0.682132477 | 0.742081291 | 0.463510172 |
| with cntxt | 0.786626469 | 0.805376741 | 0.824127012 | 0.70809656 | 0.809937246 | 0.837808528 | 0.644479388 |
| single param | 0.833540103 | 0.855368065 | 0.877196027 | 0.738708628 | 0.859967611 | 0.890994664 | 0.655545916 |
| proposed method | 0.86127619 | 0.882295146 | 0.903314102 | 0.773452479 | 0.886828179 | 0.9169132 | 0.698330658 |




| | P@1 | MRR@2 | S@2 | P@2 | MRR@3 | S@3 | P@3 |
|---|---|---|---|---|---|---|---|
| w/o cntxt | 0.608363912 | 0.656554226 | 0.70474454 | 0.519567334 | 0.673147448 | 0.754524205 | 0.460109044 |
| with cntxt | 0.686113988 | 0.704923869 | 0.72373375 | 0.561362293 | 0.71070853 | 0.741087735 | 0.499385663 |
| single param | 0.755801595 | 0.781800661 | 0.807799728 | 0.63462929 | 0.788418588 | 0.827653507 | 0.559941667 |
| proposed method | 0.754444568 | 0.785512953 | 0.816581338 | 0.662298225 | 0.793803177 | 0.84145201 | 0.594500858 |

Fig. 3 Comparison between baselines and proposed method

## 4. Experiments

### 4.1 Dataset and experiment setting

The experiments are conducted on a pre-processed data set from [4] comprising meta-data (locations & time and tags) of 230,532 images taken in more than 20 European cities. We have two contexts considered in the experiment. For location context, there are 5 granularities $\{neighborhood \rightarrow locality \rightarrow county \rightarrow region \rightarrow country\}$, which are predefined according to the map and for time context, one granularity *season* is considered. For testing, we treat the words that real users used for tagging as ground truth. The reason why human labeling is not performed in the experiment is that human labeling is cost-consuming and not scalable. In [5], it is also argued that using real data as ground truth for tag recommendation gives an underestimation of the system performance.

We first choose a probabilistic model for $R$, for one given tag $t_i$, recommendation model will recommend $t_j$ according to

$$R_1(t_i, t_j) = p(t_j|t_i) = \frac{p(t_i, t_j)}{p(t_i)}$$

$p(t_i, t_j)$ means the probability that $t_i$ and $t_j$ co-occur in one picture in the selected data over the best context. $R_1$ captures how often the tag $t_i$ co-occurs with tag $t_j$ normalized by the total frequency of tag $t_i$. This can be interpreted as the probability of a photo being annotated with tag $t_j$ given that it is at the same time annotated with tag $t_i$.

We use three standard IR metrics for evaluation:

(1) Mean Reciprocal Rank (**MRR@k**), if the first relevant tag among top k returned tags is at rank $r$, then MRR is $1/r$ (if there is no relevant tag, MRR = 0).

(2) Success at rank k (**S@k**), probability of finding a good descriptive tag among top k returned tags.

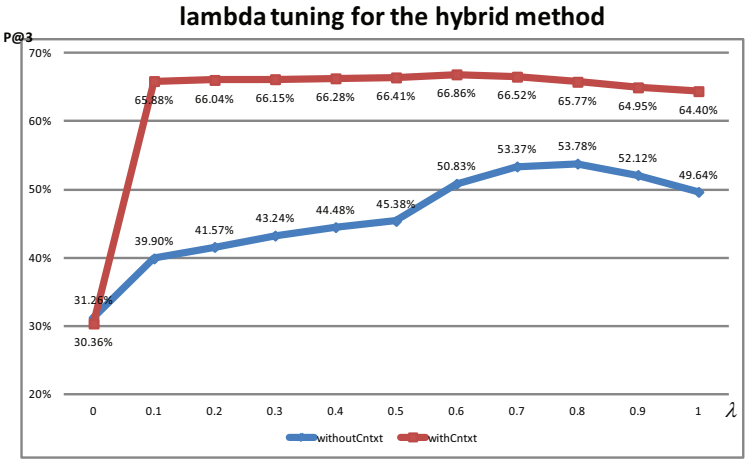



Fig. 4 Tuning of $\lambda$

(3) Precision at rank k (**P@k**), the percentage of relevant tags among the top k returned tags.

We set $k = 1, 2, 3$ since the screen size of mobile devices is only available for few candidates to display.

For comparison, three baselines are designed.

(1) **w/o cntxt**, that recommends tags without any contextual information.

(2) **with cntxt**, that always uses the finest contextual information to make recommendations.

(3) **single param**, that implements the method introduced in [1]. This method detects whether to use the context only based on the current context, regardless of the given tag that the user inputs. Compared to our proposed method, *single param* is a static decision model and cannot dynamically change the recommendation strategy when the input tag is different.

### 4.2 Results and observations

The upper chart of Fig.3 shows the 10-fold cross-validation results of the three baselines and the proposed method. As can be concluded from the figure, using contextual information greatly improves the recommendation results (18% for $P$@3), which reveals that contextual does matter in tag recommendation. The second observation is that when context granularity is considered (*single param*), the performance increases by about 5% ($S$@3). This indicates that the finest context does not necessarily produce best results. The third observation is that our proposed method outperforms the best baseline by 4.3% ($P$@3), which shows that considering the input tag can adapt to the appropriate context granularity and benefit more from rich contextual information.

As the task is picture tag recommendation, we add the picture content feature into the ranking function $R$. If two pictures are similar, the tags that are attached to the pictures are also similar. 100 SIFT features are extracted from each picture, and we compare the similarity between two pictures based on their matched points.

$$similarity(img_i, img_j) = \frac{\#matched\_points}{(\#points_{img_i} + \#points_{img_j})/2}$$

Then ranking function based on picture content between two tags can be designed as

$$R_2(t_i, t_j) = \max(similarity(img_i, img_j))$$

where $t_i$ is the tag for $img_i$ and $t_j$ for $imge_j$. $R_2$ means the ranking value of $t_i$ and $t_j$ are determined by the most similar two images which contain the two tags respectively. We combine $R_1$ and $R_2$ linearly to form a hybrid ranking function as

$$R_3(t_i, t_j) = \lambda \times R_1(t_i, t_j) + (1 - \lambda) \times \frac{R_2(t_i, t_j)}{\mu}, \lambda \in [0, 1]$$

$\mu$ is a smooth factor and the parameter $\lambda$ needs to be tuned. Figure 4 shows the performance ($P$@3) of "w/o cntxt" and "with cntxt" when $\lambda$ changes from 0 to 1 increased by 0.1. We can discover that when $\lambda = 0.7$, the combination of the two rank functions reaches the highest performance ("w/o cntxt" + "with cntxt"). In the following experiment, we set $\lambda = 0.7$.

The lower chart of Fig.3 shows the 5-fold cross-validation results of the hybrid method with location context considered. A subset of the data set with 101,168 images is chosen. These pictures are taken in Italy or the United Kingdom. As can be seen from the chart, the proposed method still outperforms the baselines, which shows that the tag-sensitive detect function can be applied to different ranking functions with consistent improvement.

### 4.3 A case study

We choose one typical picture (Fig. 5) from the data set and compare the four recommendation made by the three

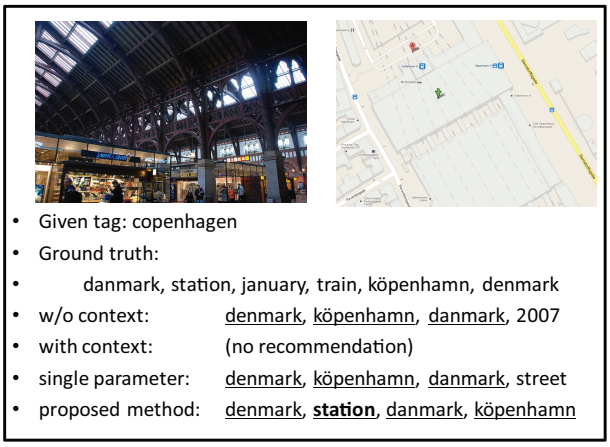



Fig. 5 Example to show that the proposed method understands the user intent better

baselines and the proposed method. For simplification, we only consider the location context and the co-occurrence feature to explain the results. The underlined tags are the correct tags compared with the ground truth. Figure 5 is taken near the railway station in Copenhagen, capital of Denmark. Always using the finest granularity is not a good choice. As can be seen from the example, "with context" baseline cannot generate any recommendation since in the finest granularity there is no tag that co-occurs with "copenhagen". The "single param" method selects a high granularity and recommends some general tags such as "denmark" while the proposed method is able to detect a finer granularity than the "single param" method and recommend "station", which is more relevant to the user. This shows that the proposed method can select the granularity more accurately with the contextual information and the given tag.

## 5. Conclusion and future works

In this paper, the tag recommendation based on multi-granular context dependency for online photo sharing services is formally defined. We build the recommendation model with the new concepts of context granularity, granularity path, and context combination. The model is a generic framework to exploit a rich set of contexts with multiple granularity which can be applied to different ranking functions, whether it is co-occurrence based or picture-content-based. A tag-sensitie method which takes into account both contextual information and given tags is presented to automatically detect the best context combination for recommendations.

We have implemented our system over a real Flickr dataset where after pre-filtering, there are 0.2 million images with meta-data. Extensive experiments have been conducted to validate our rich context-aware model. The proposed method is verified as a dynamic approach and is able to generate consistently better recommendation compared to other competitive state-of-the-art baselines.

For the next steps, we plan to do the following:

• Each context represents one dimension. When more contexts and context granularities are available, this will lead to the sparsity problem when select data on $D$. We will focus on this issue and investigate the tradeoff between fine granularity and sparse data problem.

• Context-aware problem is not only restricted to tag recommendation. We would like to extend the model to other tasks such as search engine to see how multi-granular contexts cast effect in these problems.

## References

[1] Adomavicius, Gediminas and Sankaranarayanan, Ramesh and Sen, Shahana and Tuzhilin, Alexander. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. ACM Trans. Information Systems, vol. 23, no. 1. *TOIS'05*, 2005.

[2] Sigurbjörnsson, Börkur and van Zwol, Roelof. Flickr tag recommendation based on collective knowledge. Proceedings of the 17th international conference on World Wide Web. *WWW'08*, 2008.

[3] Liu, Dong and Hua, Xian-Sheng and Yang, Linjun and Wang, Meng and Zhang, Hong-Jiang. Tag ranking. in Proceedings of the 18th international conference on World wide web. *WWW'09*, 2009.

[4] Avrithis, Yannis and Kalantidis, Yannis and Tolias, Giorgos and Spyrou, Evaggelos. Retrieving landmark and non-landmark images from community photo collections. in Proceedings of the international conference on Multimedia. *MM'10*, 2010.

[5] Garg, Nikhil and Weber, Ingmar. Personalized, interactive tag recommendation for flickr. Proceedings of the 2008 ACM conference on Recommender systems. *RecSys'08*, 2008.

[6] Gupta, Manish and Li, Rui and Yin, Zhijun and Han, Jiawei. Survey on social tagging techniques. SIGKDD Explor. Newsl. 12, 1 (November 2010).