

# Tag Transformer

Yicheng Song<sup>1,2</sup>, Juan Cao<sup>1</sup>, Zhineng Chen<sup>1,2</sup>, Yongdong Zhang<sup>1</sup>, Jintao Li<sup>1</sup>

<sup>1</sup>Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China

<sup>2</sup>Graduate School of the Chinese Academy of Science, Beijing 100039, China

{songyicheng, caojuan, chenzhineng, zhyd, jtli}@ict.ac.cn

## ABSTRACT

Human annotations (titles and tags) of web videos facilitate most web video applications. However, the raw tags are noisy, sparse and structureless, which limit the effectiveness of tags. In this paper, we propose a tag transformer schema to solve these problems. We first eliminate those imprecise and meaningless tags with Wikipedia, and then transform the remaining tags to the Wikipedia category set to gather a precise, complete and structural description of the tags. Our experimental results on web video categorization demonstrate the superiority of the transformed space. We also apply tag transformer into the first study of using Wikipedia category system to structurally recommend the related videos. The online user study of the demo system suggests that our method could bring fantastic experience to the web users.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

## General Terms

Algorithms, Design, Performance, Experimentation

## Keywords

Tag Transformer, Wikipedia Category Tree, Tag Cleaning, Structural Web Video Recommendation, Online User Study

## 1. INTRODUCTION

Coming with the prosperity of community-contributed multimedia, nowadays internet can provide billions of videos along with their annotations. Due to the unsatisfactory performance of multimedia understanding, currently, most of web video applications take advantage of tags to optimize their performance [6]. Although a lot of encouraging results have been reported, there still remain four main challenges:

**Noisiness:** There are two kinds of noisy tags. One is the kind of imprecise and meaningless tags which are unrelated to the video content. The other is the kind of ambiguous tags which will misguide the video understanding without the further confirmation.

**Sparseness:** Existing studies [1] reveal that the average number of tag per video is 11.4, which still include a great part of meaningless and imprecise tags. These sparse tags are far from completely describing the rich video content.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'10, October 25–29, 2010, Firenze, Italy.

Copyright 2010 ACM 978-1-60558-933-6/10/10...\$10.00.

**Structurelessness:** There exist structural relationships between tags. For example, “Huston Rockets” is subordinate to “NBA” and neighbor to “Los Angeles Lakers”. But current tag related applications cannot make good use of this important relationship because the raw tags are structureless

**Scalability:** Facing with the huge amount of tags on the web, it is impossible for the current tag related applications to recognize all of them. Usually, a space consisted of certain number of tags is required in these applications. But this finite space lacks the ability to cover all existing web video tags as well as emerging tags.



Figure 1. The framework of Tag Transformer.

In this paper, we leverage Wikipedia to meet above challenges. As the largest online encyclopedia, Wikipedia covers most concepts in the world and develops a category system to structural organized them: Each concept in Wikipedia is placed in the existing categories it logically belongs to, and the category system is organized hierarchically into an ontology. Motivated by these characteristics of Wikipedia [3], we propose a method named Tag Transformer (see Figure 1) to address above four challenges:

- We first adopt Wikipedia to validate and disambiguate each tag. Then we rank tags according to their relevance with the video content. By this way we eliminate those imprecise and meaningless tags to address the noisiness challenge.
- Then, we transform the tag to Wikipedia category space to gather a Wikipedia category set as a precise, enriched description of the raw tag to overcome the sparseness challenge.
- After transforming the tag to Wikipedia categories, we utilize Wikipedia category system to structurally organize the tags. By this way we can easily tell the tag relationship.

- The fourth challenge is the lack of scalability to handle the massive and ever-increasing tags on the web. By utilizing Wikipedia to recognize each of them and transform these massive tags to the finite and relatively stable Wikipedia category space, we solve the fourth challenge.

In conclusion, our contribution can be summarized as follows: 1) Tag transformer offers an effective representation of web videos to overcome the limitation of original tags. 2) This precise, enriched and structural description of web video can introduce novel opportunity to many related applications.

The rest of this paper is organized as follows: We define the proposed tag transformer framework in section 2. In section 3, we demonstrated the superiority of tag transformer than the original tags. Then, inspired by the structural result of tag transformer, we propose a novel web video recommendation approach in section 4.

## 2. TAG TRANSFORMER

Tag Transformer is a two-stage process which is illustrated in Figure 1. The first stage named *tag cleaning* is designed to face noisiness challenge. The second stage utilized Wikipedia category system to provide a precise, enriched and structural description of the tag to overcome the rest challenges.

### 2.1 Tag Cleaning

Tag cleaning aims to remove those imprecise and meaningless tags and resolve ambiguous tags to gather those *clean tags*.

We first use the longest match principle to validate each tag in Wikipedia. For example, a video tagged with “New, York”, “New” and “New York” are both detected as Wikipedia concept, only “New York” is selected. We employ Wikipedia to validate the tag due to following reasons: 1) As the largest online encyclopedia, Wikipedia covers the vast majority of current English concepts. Recent study [3] has shown that Wikipedia accuracy to rival that of Encyclopedia Britannica and contains much more concepts. These characteristics ensure that those tags not recruited in Wikipedia are mainly spelling errors or the too specific nouns. 2) Different from those fixed dictionary such as WordNet, Wikipedia is an online encyclopedia which continuously gathering knowledge from over 160,000 volunteer editors [3] all over the world. This ensures Wikipedia will recruit new concepts as soon as possible.

When facing with ambiguous tag  $T$ , our disambiguous method is different from the prior one [7] which is base on analyzing the distribution of tag combination in the finite dataset. We utilize the precise disambiguous system in Wikipedia to solve this problem. Wikipedia can recognize ambiguous tags and use a disambiguation page to list all its possible meanings [9]. As illustrated in Figure 1, “apple” is identified as an ambiguous tag in Wikipedia; its disambiguous page lists its 25 candidate meaning. We build TFIDF VSM models ( $M_1...M_p$ ) for each meaning based on its abstraction in Wikipedia, and another TFIDF VSM model  $M_T$  for the video base on its tags. Then, we calculate the similarity of  $M_T$  with each of  $M_1...M_p$  based on the classical *Cosine* similarity measurement. The one with the maximum similarity with  $M_T$  is regard to be de disambiguous result of tag  $T$ . As illustrated in Figure 1, “Apple” is identified as ambiguous tag and replaced by the concept “Apple Inc.”.

After that, we rank each tag according to their relevance with the video content. Analogous to that in [2], our approach is

formulated as a random walk problem along the complete graph, where video tags are nodes and the edges between them are weighted by tag co-occurrence and video visual similarities. This process score each tag based on its relevance to video content and other tags, which will promote the tags that have many close neighbors and weaken the isolated ones. Finally, we eliminate those low scored tags. Figure 1 shows a perfect tag cleaning result.

### 2.2 Transform Tag to Wikipedia Categories

Web video tag cleaning select those *clean tags* to address the first challenge, then we transform these tags to Wikipedia category space to overcome the rest challenges.

As mentioned above, each concept in Wikipedia is placed in several categories it logically belongs to. By this way, we transform the tags to Wikipedia categories. Considering most concepts in Wikipedia have few categories which will also cause the sparseness problem. We navigate through Wikipedia category system to gather the father and grandfather categories of the concept to form a category tree. The category tree comprehensively describes the concept from different angles. Similar method is adopted in [5], they utilize the precise Wikipedia categories and template system to conduct tag classification. The difference is that: we focus on supplying a new web video description to overcome the limitations of original tags and use it to trigger more novel applications.

The transform from the tag “Apple Inc.” to its category tree is depicted in Figure 1. The category tree  $Tree = \{C_1, C_2, \dots, C_N\}$  supplies a precise and enriched description of the tag  $T$  to address the first challenge. To evaluate the probability of video  $V$  belong to the category  $C_i$  based on tag  $T$ , following criterion is illustrated:

$$\Pr(C_i | T, V) = \frac{d(C_i)}{Level(C_i)Rank(T, V)} \quad (1)$$

where  $d(C_i)$  and  $Level(C_i)$  respectively denotes the in-degree and level of category  $C_i$  in the category tree,  $Rank(T, V)$  represents the rank of tag  $T$  in the video  $V$  after tag cleaning.

After transforming the tags to structural Wikipedia categories, we can utilize the category mechanism of Wikipedia to structurally organize the tags. The related application is illustrated in section 4.

## 3. EXPERIMENTS

We begin with the introduction of dataset used in this paper and then show two experiments in Section 3.2 and 3.3.

### 3.1 Dataset

**Wikipedia Dataset:** Wikipedia release its downloadable data at <http://download.wikipedia.org> in the form of MYSQL database dumps. Moreover, Wikipedia dump is published periodically, from several days to several weeks apart. In this paper, we use the Wikipedia dump released on 09 September, 2009.

**Web Video Dataset:** MCG-WEBV [1] is employed in this paper which is available at <http://mcg.ict.ac.cn/mcg-webv.htm>. The 1.0 version of the dataset is consisted of 80031 videos from 15 categories of YouTube. We choose the tag feature from the dataset to conduct our tag transformer experiment.

### 3.2 Web Video Categorization

We study the performance of web video categorization with different representations to verify the effectiveness of tag transformer. By comparing the performance of web video

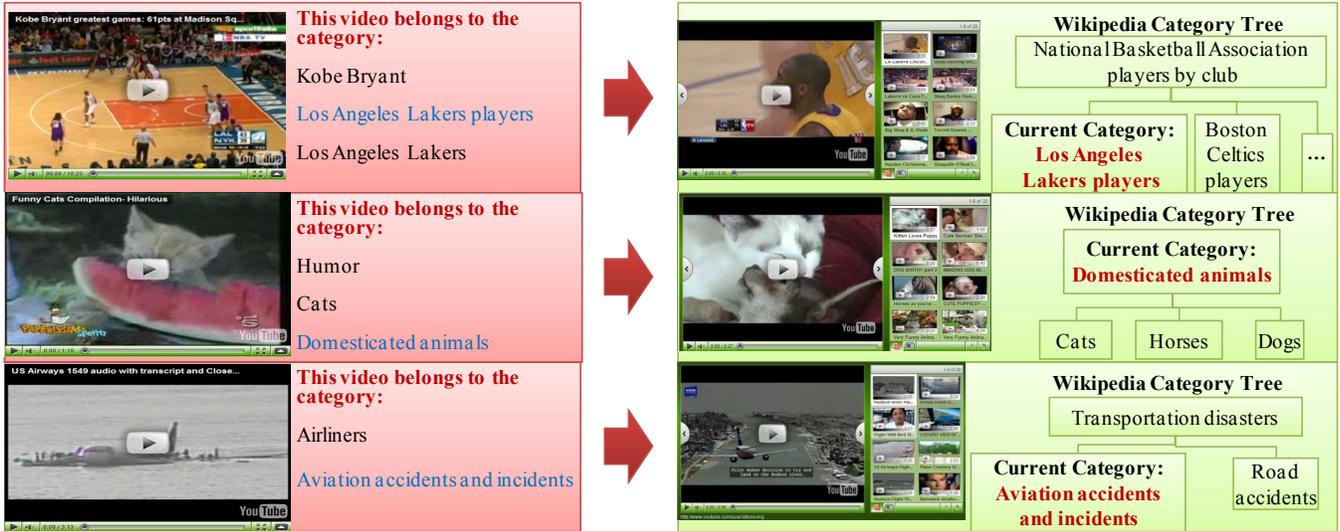


Figure 2. Three example scenes from the structural web video recommendation demo

categorization based on original tags, tag transformer without tag ranking, tag transformer, we demonstrate that tag transformer provides a more effective representation than the original tags.

We adopt the web video categorization method similar in [8]. We randomly split the 80031 videos in 1.0 version of MCG-WEBV into 3:6:1, with 3/10 videos as training data, 6/10 as testing data and 1/10 for cross validation. The original YouTube category labels are treated as the ground truth. Then SVM classifiers are trained on the training set based on original tags space, tag transformer without tag ranking and tag transformer space separately, then predicted the testing data.

As it is well known that original tags are very noisy for the web applications, the overall performance is relatively low (MAP is 0.521). After transformer those tags to Wikipedia categories, we get a more precise, enriched and structural description of the video. It has better performance compared to original tag. The MAP of tag transformer is 0.576. Furthermore, the MAP of tag transformer without tag ranking is 0.548. This drop shows that tag ranking is an effective technique in removing those noise tags.

### 3.3 The Scalability of Tag Transformer

As mentioned in [1], there are 111,913 unique tags in 1.0 version of MCG-WEBV dataset and 6,392 of them appear more than 30 times are regarded as frequent tags or popular tags. We adopt Wikipedia to validate each tag to check whether this tag is recruited in Wikipedia. The result is that: all 6,392 frequent tags are included in Wikipedia; compare to that of 68.3% for total 111,913 tags. We find that those tags which are not recruited in Wikipedia are mainly spelling errors or the too specific nouns (e.g. “Titue” and “CSDK”). The result indicates that Wikipedia can cover most tags on the web and is an promising choice to tackle the scalability challenge of original tags.

## 4. APPLICATION

Based on tag transformer result and structural Wikipedia category system, we propose a novel web video organization and related video recommendation approach. A video is related to different videos from different points of view. Two videos are related maybe they talk about the same person or different objects in the same filed just like “Donald Duck” and “Mickey Mouse”. But current video recommendation studies [4] mainly focus on

whether videos are related and rank them by the visual or textual relevance, ignoring in which aspect they are related.

According to the category tree  $Tree = \{C_1, C_2, \dots, C_N\}$  of each tag  $T$  of the video  $V$ , we assign the video to the category  $C_i$  if probability  $\Pr(C_i | T, V)$  is higher than the threshold (we adopt 0.6 in this paper). By this way, we index all the videos by the structural Wikipedia category system. This organization introduces novel opportunity for the web video recommendation not only structural displays the related videos but also tells the user in which aspect they are related.

When the user is viewing video  $V$ , he will get Wikipedia categories of this video belong to. Then he can choose a particular category  $C_i$  to see related videos about this subject, the related video list is consisted of videos  $(V_1 \dots V_M)$  in category  $C_i$  and sort by the following criterion:

$$Weight(V_i | V, C_i) = \Pr(C_i | T, V_i) + Sim(V, V_i) \quad (2)$$

Where  $\Pr(C_i | T, V_i)$  is the probability of video  $V_i$  to the category  $C_i$ ;  $Sim(V, V_i)$  is the classical *Cosine* similarity of video  $V_i$  to the original video  $V$  based on their tag feature. By this way, those videos with high probability in category  $C_i$  and close related to the original video  $V$  will be promoted to top of the recommended video list. The user can also navigate through the Wikipedia category system to browse other related videos nested in the father categories, neighbor categories and so on. We upload the demo system to: <http://mm2010.byethost4.com> and post 3 scenes in Figure 2 to show some unique characteristics:

**Clear Structure:** The structural organization of videos allows the users to find more relevance videos about the same topic. For example, the video in the first scene belongs to the category “Kobe Bryant”, “Los Angeles Lakers Players” and “Los Angeles Lakers”. Users can browse the category “Kobe Bryant” to see videos about this basketball player, or click category “Los Angeles Lakers Players” to see videos about Kobe’s teammates in Los Angeles Lakers. If he is interested in players in other NBA teams, just navigate through the father category “National Basketball Association players by club” to browse its brother categories.

**Multi Dimension Relevance:** Current web video recommendation methods focus to select those close related videos based on the textual or visual features. Different from those one dimension relevance recommendation methods, our system provides a multi dimension relevance to find related videos from different views. For example, the video in the second scene is tagged with “Funny Cats ...” You can find those closed related videos in the category “Humor” and “Cats”. You can also navigate into the son category of “Domesticated animals” to see videos about dogs and horses.

**Comprehensive Classification of Topic:** The precise Wikipedia category system allows our method to comprehensive classify the topics. The video in third scene is talk about a plane accident in New York and assigned to the category “Aviation accidents and incidents”. This category is the son category of “Transportation disasters” and is the brother of category “Road accidents”, “Maritime incidents” and “Railway accidents”. This mechanism detailed and comprehensively represent this topic to the users.

We design two experiments to evaluate our system. The goals of the experiments are to: 1) Verify if our approach can recommend those closed related videos to the user. 2) Verify whether our web video recommendation approach can provide better user experience to the user.

#### 4.1 Evaluation of Closed Related Videos

In order to validate whether our web video recommendation method can provide those closed related videos to the user, we take the original Youtube video relationship (i.e. closed related videos are listed for each video in YouTube) as the ground truth. Because the video relationship is not complete in 1.0 version of MCG-WEBV, we chose the subsequent version which is consisted of monthly most viewed videos and their related videos from 15 categories of YouTube.

With the 1457 most viewed videos of July, 2009, there exist 36364 related videos. For each most viewed video, we check whether its related videos in YouTube are nested in our recommended related video list. The result is that 90.7% of YouTube related videos are nested in the top 10 of our recommended related video list, and 96.5% for the top 20. The result suggests that our approach can provide those YouTube suggested related videos to the user.

#### 4.2 Online User Study

In order to measure the user experience of our method, we performed an online user study. The study attempts to verify whether our structural web video recommendation approach can provide better user experience than the original YouTube suggestion of related videos.

Until to May 4, 2010, there are 172 web users participated in the user study. Participants are required to browse 10 structural video recommendation scenes and rating each scene from 4 aspects (interesting, rich, correct and useful). For each aspect, the participants give a score ranging from 1 to 5 (original YouTube suggestion of related videos is assigned with 3 for each aspect). The higher the score, the better the assessment of our method. The average result of 10 scenes is reported in Table 1, more details of the user study result can be founded at our online demo:

**Table 1. Online User Study Result**

Aspects	Interesting	Rich	Correct	Useful
Score	4.72	4.64	4.06	4.68

Based on the above user study result, we can find that those web users consider our web video recommendation approach is an interesting, richness, correct and useful way to browse web videos. After interactive with some of them, we conclude that the superiority of our method is due to the following reasons: many web users visit YouTube or similar online video sharing websites to browse videos to pass some spare time. Comparing to the original YouTube suggestion of related videos, our method not only recommend those closed related videos but also provide more richness and structural related videos to the end users, which is in accordance with their browsing habits.

### 5. CONCLUSION

In this paper, we propose an approach to transform the tags to Wikipedia categories which supplies a precise, enriched and structural description of the video. Experimental results on the web video categorization demonstrate the superiority of tag transformer than the original tags. Inspired by the structural nature of tag transformer, we proposal a structural web video recommendation method which brings the users wonderful experience for web video browsing. It is worth noting that although we illustrated only one application of tag transformer, we believe that this new representation of videos will introduces novel opportunity to many web video related applications.

### 6. ACKNOWLEDGEMENTS

This work was supported by the National Nature Science Foundation of China (60902090), National Basic Research Program of China (973Program, 2007CB311100), Co-building Program of Beijing Municipal Education Commission, Beijing New Star Project on Science & Technology (2007B071).

### 7. REFERENCES

- [1] J. Cao, Y.D. Zhang, Y.C. Song, Z.N. Chen, X. Zhang, and J.T. Li. MCG-WEBV: A Benchmark Dataset for Web Video Analysis, Technical Report, ICT-MCG-09-001, Institute of Computing Technology, May, 2009.
- [2] Dong Liu, Xian-Sheng Hua, Lin-jun Yang, Meng Wang, Hong-Jiang Zhang. Tag Ranking. WWW 2009: 351–360.
- [3] Jim Giles. Internet Encyclopaedias Go Head to Head. Nature, 438(7070):900–901, December 2005.
- [4] Tao Mei, Bo Yang, Xian-Sheng Hua, Lin-jun Yang, Shi-Qiang Yang, Shipeng Li: VideoReach: An Online Video Recommendation System. SIGIR 2007: 767-768.
- [5] Simon E. Overell, Börkur Sigurbjörnsson, Roelof van Zwol: Classifying Tags Using Open Content Resources. WSDM 2009: 64-73.
- [6] Morgan Ames, Mor Naaman: Why We Tag: Motivations for Annotation in Mobile and Online Media. CHI 2007: 971-980.
- [7] Kilian Q. Weinberger, Malcolm Slaney, Roelof van Zwol: Resolving Tag Ambiguity. ACM Multimedia 2008: 111-120.
- [8] Yicheng Song, Yongdong Zhang, Xu Zhang, Juan Cao, Jintao Li: Google challenge: incremental-learning for web video categorization on robust semantic feature space. ACM Multimedia 2009: 1113-1114
- [9] Rada Mihalcea: Using Wikipedia for Automatic Word Sense Disambiguation. HLT-NAACL 2007: 196-203.