Dynamic Vocabularies for Web-based Concept Detection by Trend Discovery

Damian Borth
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
d_borth@cs.uni-kl.de

Adrian Ulges
German Research Center for
Artificial Intelligence (DFKI)
D-67663 Kaiserslautern,
Germany
adrian.ulges@dfki.de

Thomas M. Breuel
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
tmb@cs.uni-kl.de

ABSTRACT

We present a novel approach towards automatic vocabulary selection for video concept detection. Our key idea is to expand concept vocabularies with *trending topics* that we mine automatically on other media like Wikipedia or Twitter. We evaluate several strategies for extending concept detection to auto-detect these topics in new videos, either by linking them to a static concept vocabulary, by a visual learning of trends on the fly, or by an expansion of the vocabulary.

Our study on 6,800 YouTube clips and the top 23 target trends (covering a timespan of 6 months) demonstrates that a direct visual classification of trends (by a "live" learning on trend videos) outperforms an inference from static vocabularies. However, further improvements can be achieved by a combination of both approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Concept Detection, Social Media, Trends, Vocabulary

1. INTRODUCTION

With the the growing proliferation of images and videos over the last years, the demand for search and retrieval tools has increased. Here, *concept detection* [12] – the automatic recognition of objects, locations or actions – is a prominent approach, which is of particular interest for web-based services like YouTube hosting huge amounts of content [15].

An open issue with concept detection is the selection of suitable *vocabularies* of target concepts: These are usually picked manually by experts [10], which narrows their applicability and suitability to deal with the enormous diversity of web-based video. Instead, we would like concept vocabularies to evolve while new topics of interest arise in media and society. By detecting such *trending topics* and building new detectors for them, we could tailor concept detection to the latest user interest, widen its scope significantly, and help to disambiguate i.e. narrow down potential concepts to the most likely ones. In this paper, we study such a trend-based evolution of concept detection vocabularies, using YouTube as an application domain (Fig. 1):

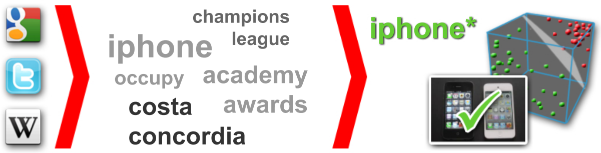


Figure 1: To align concept detection with users' information need, trending topics are discovered by mining multiple media streams to serve as a concept vocabulary extension.

- We mine Google searches, Twitter posts and Wikipedia access statistics for *trending topics*, i.e. terms that experience a spike in popularity over a certain time period. We show that these trends – like "Super Bowl" or "iphone 4" – are strongly correlated with YouTube uploads, i.e. if a trend emerges on other media, video uploads on YouTube spike correspondingly.
- Our second goal is to auto-detect trends in uploaded videos. To do so, there are two general detection strategies, namely **(1)** linking a targeted trend like "Super Bowl" with pre-trained concepts like "American Football" or "Commercial", or **(2)** by training a new "Super Bowl" detector as the trend emerges and videos tagged with it are uploaded. We compare these two strategies and present a combination of both that merges trends into the concept vocabulary.

We present experiments over a timespan of 6 months in winter 2011/12. Out of 200,000 topics we selected the 23 most prominent ones and evaluated their detection on a dataset of 6,800 YouTube clips (541 hours of video) covering the same test period. Our results show that training direct trend recognition outperforms a static vocabulary (233 detectors trained on YouTube videos). However, a combination of both strategies can improve accuracy further.

2. RELATED WORK

Research in concept-based video retrieval [12] is strongly driven by benchmarks like TRECVID [11], where various

MM'12, October 29–November 2, 2012, Nara, Japan.

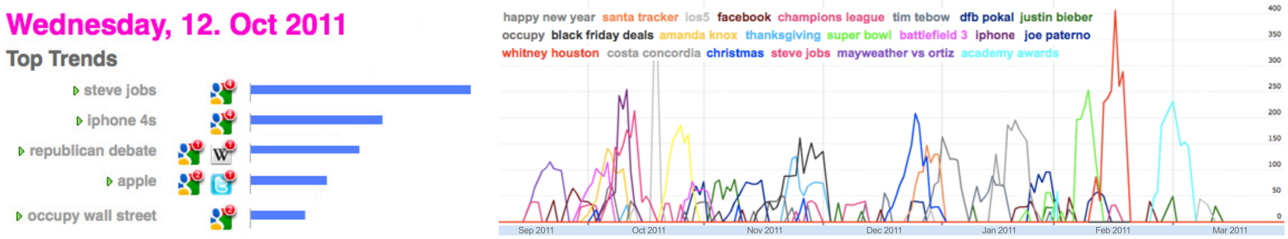


Figure 2: Left: For each day, top trends are discovered by aggregating feeds from Google, Wikipedia, and Twitter, and trend *scores* are computed. **Right**: The trend scores for the 23 most prominent trends, plotted over the observation period.

concept detection systems are evaluated on common datasets. Typically, vocabularies of such systems are expert-defined, where visual discrimination, utility for retrieval and availability of training material have been identified as important characteristics of "good" concepts [10, 11]. Though the time-consuming acquisition of training data [1] poses a limiting factors to vocabulary size, large-scale concept sets exist, like ImageNet [8] or Google's Video2Text system [5]. Our work bears similarities to the latter in a sense that we focus on web video as a domain, and that we exploit web video content with user-generated tags as a source of training and test data, which allows us to learn concept detectors "on the fly". The key difference, however, is that we link up web-based concept detection with trend discovery to develop *dynamic vocabularies* adapted to evolving user interests.

Evolving tag vocabularies have also been studied in [3], where an inductive transfer was applied upon a fixed black-box vocabulary to adapt to a users' personalized tagging behavior over time. The approach differs from ours as we train new concept detectors, and as we employ a discovery of trending topics over large user communities, which – to the best of our knowledge – has not been investigated before.

Prior work on trend discovery focuses on blogs and Twitter content [4, 7]. Some approaches employ aggregated trends provided by platforms like Twitter [7], while others perform an analysis on the raw data [2, 4]. Similar to the former, we utilize platform provided lists of trending events. However, we further process and aggregate those trending topics and group them to real-world events.

3. TREND DISCOVERY

We discover *trending topics* – terms that experience a spike in user popularity – by analyzing statistics of Google searches, posts on Twitter and Wikipedia site accesses. These are clustered to account different spellings and paraphrases, and finally combined over the different services to obtain *trend scores* describing the popularity of topics.

Step 1: Trending Topic Raw Sources We employ Google, Twitter and Wikipedia as a basis by retrieving a daily ranked list of popular terms from 10 different sources[1], namely 5 Google feeds (*Search* and *News* for USA and Germany as well as the *Trends* feed), Twitter (*daily trends* for USA and Germany) and Wikipedia (access statistics for English and German language). For each feed, we retrieve 10-20 ranked topics per day (110 topics total).

Step 2: Unification, Clustering and Aggregation Each topic is mapped to a corresponding Wikipedia URI by selecting the top English Wikipedia site for a Google Search with the topic. After this, we cluster the given URIs by thresholding the Levenshtein distances of term pairs (the threshold is set to 0.35 × the word length). This results in a grouping of terms like "super bowl time", "super bowl 2012" or "superbowl" into a consistent cluster, i.e. we unify trending topic from heterogeneous sources.

For each each day and for each of our 10 feeds, we record the rank at which a topic appears. These ranks are combined over the different feeds using Borda count, obtaining a score for each day (Fig. 2 [left] illustrates an example day, with the top 5 topics ranked by their scores, and with the feeds on which they appeared). To take the overall "life cycle" of an event into account, we measure its impact by summing up all of its daily scores over the observation period, obtaining a global *trend score*. This score serves as the basis for picking the 23 most prominent trending topics, whose scores are plotted over our observation period in Fig. 2 [right].

4. EXTENDING CONCEPT DETECTION

Given a set of trends $t_1, .., t_m$ and a new video or keyframe (described by content-based features x), our goal is to estimate $P(T = t_j|x)$. We also assume an initial static concept vocabulary $C_1, .., C_n$ to be given. For these concepts, trained detectors exist that estimate concept scores $P(C_1 = 1|x), .., P(C_n = 1|x)$.

Strategy 1: Concept-to-Trend Mapping Our first strategy is to work only with the static concept vocabulary and then map the detected concepts to target trends t_j. To do so, we estimate concept-trend similarities using the normalized Flickr distance $D(c_i, t_j)$ as proposed in [6], i.e. a concept c_i and trend t_j are considered the more similar the more often they co-occur as tags on Flickr. The distance D is mapped to a similarity $exp\{-D(c_i, t_j)/\gamma\}$, and these similarities are normalized to probabilities $P(C_i = 1|T = t_j)$ (more information on the estimation of γ will follow later).

The concept detection results $P(C_i = 1|x)$ and concept-trend-similarities $P(C_i = 1|T = t_j)$ are now combined by marginalizing over all possible concept appearances (we found this approach to work well in a similar scenario before [14]):

$$
\begin{aligned}
& P(T = t_j|x) \\
&= \sum_{c_1, c_2, .., c_n \in \{0,1\}} P(T = t_j, C_1 = c_1, .., C_n = c_n|x) \\
&\approx \sum_{c_1, c_2, .., c_n \in \{0,1\}} \Big[P(C_1 = c_1, .., C_n = c_n|x) \cdot \\
&\qquad\qquad P(T = t_j|C_1 = c_1, .., C_n = c_n) \Big].
\end{aligned}
$$

[1] using Google Insights for Search and the Twitter API

Assuming independence of the individual concepts and applying Bayes' rule, we can rewrite this as:

$$\approx \sum_{c_1,c_2,..,c_n \in \{0,1\}} \left[\prod_{i=1}^{n} P(C_i = c_i|x) \cdot \frac{P(T=t_j) \prod_{i=1}^{n} P(C_i = c_i|T=t_j)}{\prod_{i=1}^{n} P(C_i = c_i)} \right] \quad (1)$$

$$= P(T=t_j) \cdot \prod_{i=1}^{n} \left[\frac{P(C_i = 0|x) \cdot P(C_i = 0|T=t_j)}{P(C_i = 0)} + \frac{P(C_i = 1|x) \cdot P(C_i = 1|T=t_j)}{P(C_i = 1)} \right],$$

whereas the priors $P(C)$ and $P(T)$ are set to uniform distributions. This way, we estimate trends via concept detection.

Strategy 2: Training Visual Trend Detectors We expect the videos of a trend to bear similarities with certain concepts, but also to be quite specific (for example, "Super Bowl" videos show a certain stadium and certain teams). Therefore, our second strategy is to train a trend-specific detector from the web [13]: As the trend emerges, videos tagged with it are uploaded. We exploit these as positive training samples to train a "trend detector" on the fly (this training set is more focused and smaller compared to a "regular" concept training set). The resulting detector can be applied to detect the trend in other videos, estimating $P(T|x)$.

Strategy 3: Expanding the Concept Vocabulary Finally, we test a combination of the former two strategies by *expanding* the concept vocabulary: The trend detector is simply added to the vocabulary as a new concept c_{n+1}. Its normalized Flickr distance is set to $D(c_i, t_j) := 0$ (after all, the newly added concept represents the trend itself), i.e. the "trend detector" has a strongest influence on the result than other concept detection scores. Concept-to-trend (*Strategy 1*) is then applied with the extended vocabulary.

5. EXPERIMENTS

Our experiments cover an observation period of 6 months (Sep 16 '11 - Mar 15 '12) over which we analyzed 200,000 topics[2] using the procedure outlined in Section 3. We ranked all trending topics according to its *trend score* and pick the top 23 ones (see Fig. 2): Some of them are obviously very challenging to detect (like "happy new year"), others seem feasible (like "battlefield 3").

Correlation of Trends and YouTube Uploads For each trend, we downloaded 150 YouTube videos (i.e. videos being tagged with the trend name) and filtered clips outside our 6 months test period, obtaining 2,500 clips (31-147 per trend). We first confirm our hypothesis that uploads on YouTube correlate with emerging trends: averaged over all trends, 57.3% of videos were uploaded on a "trend" day or the day after (a uniform distribution over time would correspond to 8.8%). Thereby, event-based trends like "whitney houston" (referring to the death of the famous singer) display the strongest alignment between YouTube and our trend recognition, while long-lasting/periodic trends like "facebook" or "champions league" the lowest. Overall, this result indicates that YouTube uploads are closely aligned with trending topics.

Visual Trend Detection Our second goal is a visual detection of trending topics. To do so, we query YouTube for additional background material distributed randomly over the observation period, focusing on daily *most recent* video clips with no tags. From the resulting 4,300 "background clips" and from the 2,500 "trend clips", we extract 78,000 keyframes using a change detection. To learn the direct visual trend detectors (*Strategy 2*), a 60%-40% split of all clips into a training and test set was conducted. Results are reported in terms of mean average precision on the test set (2,720 videos). As a static concept vocabulary (*Strategy 1*), we choose one from our prior work [13], covering 233 concepts that range from "concert" over "demonstration" to "phone" (detectors were pre-trained on a held-out dataset of YouTube clips from before the observation period).

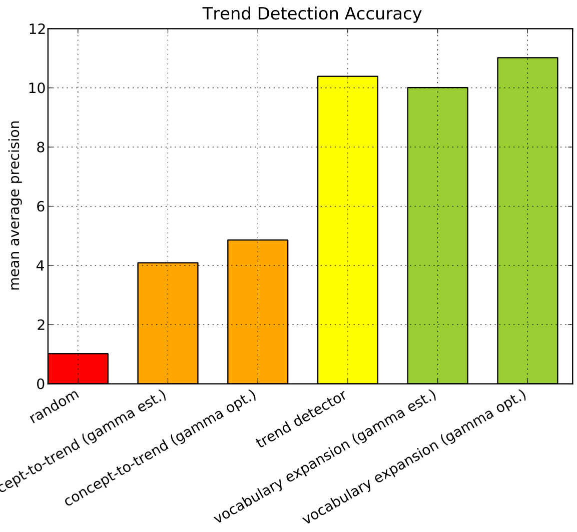


Figure 3: Quantitative results of trend recognition. A specialized trend detector (yellow) outperforms a static concept vocabulary (orange). Expanding the vocabulary with the new detector gives further improvements (green).

Both – concept detection and visual trend detection – are conducted on keyframe level, using visual words features (obtained by a regular multi-scale sampling of about 3,600 SIFT features [9], vector-quantized to 3,000 clusters using K-Means) in combination with Support Vector Machines (SVMs) using a χ^2 kernel and fitted by a grid-search cross-validation.

Quantitative results of our experiment are illustrated in Fig. 3. The vocabulary expansion strategy (VE: green bars) performs best, with an mean average precision (MAP) of 11.2% (γs optimized by grid search) and MAP 10.01% (estimated γ based the average of pairwise Flickr distances). Further, a direct training of "trend detectors" (TD: yellow) performs comparably with a MAP of 10.39%. The concept-to-trend-mapping (CTM: orange) with a of MAP 4.86% and 4.09% give the lowest accuracy. This indicates that training new detectors seems a promising approach for adapting concept detection to new emerging trends, while a static concept vocabulary can help to improve accuracy further.

A closer inspection of system performance is given in Fig. 4 for the three trends "ios5" (referring to the release of Apple's operating system), "Mayweather-vs-Ortiz" (a box fight) and "Whitney Houston" (the death of the singer). For each

[2] details: http://www.dfki.uni-kl.de/~borth/trending_topics/

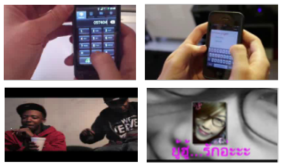
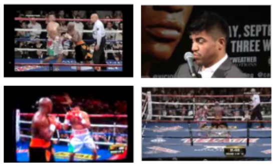
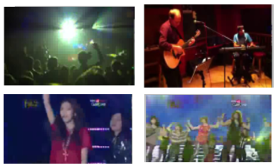
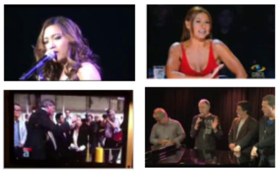
trend	**top results (*trend detector*)**	**top results (concept-to-trend mapping)**	**matched concepts (with $P(c_i\|t_j)$)**	**best performing rankers (AP)**
ios5			(1) safari: 0.56% (2) phone. 0.51% (3) cathedral: 0.49%	(1) TD: 43.3% (2) VE: 41.3% (3) phone: 37.1% (4) iphone: 27.3% (5) windows-desktop: 25.8%
Mayweather vs. Ortiz			(1) press-conf.: 0.57% (2) boxing: 0.53% (3) rugby: 0.53%	(1) VE: 26.5% (2) boxing: 23.8% (3) TD: 21.7% (4) interview: 7.8% (5) wrestling: 6.9%
Whitney Houston			(1) bill-clinton: 0.57% (2) singing: 0.55% (3) videoblog: 0.54%	(1) VE: 11.4% (2) TD: 6.9% (3) CTM: 6.2% (4) interview: 5.2% (5) obama: 5.2%

Figure 4: The 4 top-ranked videos by the direct trend detector (TD) and concept-to-query mapping (CTM) for 3 sample trends. The last column lists the best detectors by their accuracy, including trend detectors (TD), the concept-to-trend-mapping (CTM), vocabulary expansion (VE) (γ optimized by grid search), and the best individual concept detectors.

trend, the top-ranked videos for direct trend detection (TD) and concept-to-trend mappings (CMT) are displayed. We also see the the corresponding concepts and their similarities: Some can be considered outliers (e.g. “cathedral” for the trend “ios5”), while others are reasonable (like “singing” for “Whitney Houston”). The last column displays the best systems for detecting the different events. Here, we also rank individual concept detectors, which indicates that some matched concepts are suitable for recognition (like “boxing” for “Mayweather-vs-Ortiz”). In general, for most trends either the (TD) or (VE) strategy ranks at the top for all evaluated concept detectors (with some exception for poorly recognized trends).

6. DISCUSSION

We have presented an approach towards dynamically adapting concept detection on web-based video sharing portals such as YouTube, based on an automatic discovery of *trending topics*. Our experimental results have indicated that expanding concept vocabularies improves the detection accuracy of these topics. The challenge of dynamic concept vocabulary evolution opens other research questions: Most prominently, comprehensive strategies need to be developed to decide which new concepts to train. Here, user interest might be an additional criterion beside others like recognition feasibility [5] or utility [10].

7. ACKNOWLEDGMENTS

This work was partially funded by the Google Research Awards Program and supported by the PiCloud Academic Research Program.

8. REFERENCES

[1] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *Proc. ECIR*, 2008.

[2] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-world Event Identification on Twitter. In *Proc. ICWSM*, 2011.

[3] R. Datta, D. Joshi, J. Li, and J. Wang. Tagging over Time: Real-world Image Annotation by Lightweight Meta-Learning. In *Proc. Int. Conf. on Multimedia*, pages 393–402, 2007.

[4] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated Trend Discovery for Weblogs. In *Workshop on the Weblogging Ecosystem: Aggregation, analysis and dynamics*, 2004.

[5] A. Hrishikesh, G. Toderici, and J. Yagnik. Video2Text: Learning to Annotate Video Content. In *Proc. Int. Workshop on Internet Multimedia Mining*, 2009.

[6] Y.G. Jiang, C.W. Ngo, and S.F. Chang. Semantic Context Transfer across Heterogeneous Sources for Domain Adaptive Video Search. In *Proc. Int. Conf on Multimedia*, 2009.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. WWW*, 2010.

[8] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale Image Classification: Fast Feature Extraction and SVM Training. In *Proc. CVPR*, 2011.

[9] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[10] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[11] P. Over, G. Awad, M. Michael, J.Fiscus, W. Kraaij, and A. Smeaton. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2011.

[12] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Inf. Retrieval*, 4(2), 2009.

[13] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger – YouTube-based Concept Detection. In *Proc. Int. Workshop on Internet Multimedia Mining*, December 2009.

[14] Adrian Ulges, Markus Koch, and Damian Borth. Linking Visual Concept Detection with Viewer Demographics . In *Int. Conf. on Multimedia Retrieval (ICMR)*, 2012.

[15] YouTube Press Statistics. available from youtube.com/t/press_statistics (retrieved: Sep’11).