Generating Advertising Keywords from Video Content

Michael J. Welch
UCLA Computer Science Dept
4732 Boelter Hall
Los Angeles, CA 90095
mjwelch@cs.ucla.edu

Junghoo Cho
UCLA Computer Science Dept
4732 Boelter Hall
Los Angeles, CA 90095
cho@cs.ucla.edu

Walter Chang
Advanced Technology Labs
Adobe Systems Inc
San Jose, CA 95110
wachang@adobe.com

ABSTRACT

With the proliferation of online distribution methods for videos, content owners require easier and more effective methods for monetization through advertising. Matching advertisements with related content has a significant impact on the effectiveness of the ads, but current methods for selecting relevant advertising keywords for videos are limited by reliance on manually supplied metadata. In this paper we study the feasibility of using text available from video content to obtain high quality keywords suitable for matching advertisements. In particular, we tap into three sources of text for ad keyword generation: production scripts, closed captioning tracks, and speech-to-text transcripts. We address several challenges associated with using such data. To overcome the high error rates prevalent in automatic speech recognition and the lack of an explicit structure to provide hints about which keywords are most relevant, we use statistical and generative methods to identify dominant terms in the source text. To overcome the sparsity of the data and resulting vocabulary mismatches between source text and the advertiser's chosen keywords, these terms are then expanded into a set of related keywords using related term mining methods. Our evaluations present a comprehensive analysis of the relative performance for these methods across a range of videos, including professionally produced films and popular videos from YouTube.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Advertising, keyword selection, related term mining

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

1. INTRODUCTION

The rapidly growing user base and movement towards online video distribution necessitates new methods for content owners to monetize their videos and for advertisers to effectively market their products. Traditionally, television networks have monetized their content by selling time slots to advertisers, who in turn rely on estimated audiences and target demographics to determine which programs they should advertise during. Advertisements online have the potential to be more directly relevant to the video content or the interest of viewers, since ads can be selected from a large pool of advertisements individually for each viewing. The effectiveness of online ads are also easier to quantify by measuring clicks from the viewers.

Current methods for selecting the advertising keywords for a video often rely on user supplied metadata, such as the video title, summary, comments, anchor text from adjacent pages, and so on. This text is often sparse compared to the much richer video content, and many professionally produced videos are only available online for a short period of time. It is difficult to adequately identify all of the keywords manually, leading to missed opportunities for ad placement.

In this paper we study the effectiveness of generating advertising keywords using the *content* of the video. Note that we use the term *keyword* to refer to text of arbitrary length, which may be individual words or multi-word phrases. We focus on text sources such as production scripts, closed captioning tracks, and automatically generated speech-to-text transcripts. Text sources tend to be more reliable than image-based analysis in practice today, and require significantly less domain-specific knowledge or offline training.

Even with the text content for a video, several challenges remain. Identifying relevant keywords from text is non-trivial, and made more difficult when only error-filled speech transcripts are available. Methods for identifying advertising keywords on Web pages often rely on external links and explicit structural markup or formatting [6, 12], which the text from a video lacks. Unlike documents, which generally convey information through a single medium (text), the intended user experience for a video is communicated through both visual and auditory components. Dialog is often sparse and may fail to capture this complete experience, and the relevant keywords for advertisers may not necessarily directly appear in the text for a video, particularly when only dialog-based data is available.

We address these issues in two stages. In Section 2 we describe statistical and generative models to determine a set of dominant keywords within a text source (script, closed cap-

tioning track, or speech transcript). The vocabulary of the extracted keywords does not always coordinate well with the keywords advertisers have in mind. That is, while we may have a set of relevant keywords for the video, they may not overlap with the terms advertisers intend to bid on. To address this *vocabulary impedance problem*[9], we extract related keywords from multiple data sources to increase the likelihood of matching an advertiser's keywords. In both steps, keywords are identified and ranked without consulting an inventory of ads or advertiser supplied keywords. We evaluate each of the text inputs as sources for advertising keywords across a wide range of videos, including professionally produced films and amateur videos on YouTube.

2. PROCESSING SOURCE TEXT

In the first stage of processing, we address the complexities of video-based text sources, such as scripts, and describe methods for selecting keywords using statistical analysis and topic modeling. We consider three sources of text data for a video: (1) complete movie scripts, which contain descriptions of the scenes, actions, and dialog, along with corresponding metadata (e.g. name of character speaking the dialog), typically formatted in a human readable layout, (2) closed captioning tracks (CC), which contain the text of the spoken dialog and timecodes indicating when that dialog is spoken, and (3) speech transcripts (STT), which consist of a series of words, each with an associated timecode and duration. While scripts and CC tracks are manually generated, and thus highly accurate, speech transcripts are created through an automatic process which converts audio data to text, and frequently contain errors and omissions.

2.1 Script Processing

Understanding the semantics of a text element in a script is helpful when processing it. For example, character names appear frequently in a script prior to each of their lines of dialog, though we generally find them to be a poor choice for advertising keywords. We add semantics to each text segment of a script using a finite state machine based parser derived from conventional screenplay writing rules.

Movie scripts (and thus keywords extracted from their text) do not contain any associated timecodes. To add timecode information to script keywords, we generate the speech transcript and use the Levenshtein Word Edit Distance [7] algorithm to find the best word alignment between script dialog and the STT transcript. Note that the parsing and alignment steps described in this section are specific to scripts. CC and STT input contain only text of the spoken dialog, and the corresponding timecodes are already present.

2.2 Statistical Generation of Keyword Terms

In the final step for a source text (script, CC, or STT), the timecoded text elements are used to build an N-gram tree that is pruned by N-gram term frequency to discover the most dominant terms, based in large part on the work of Chim and Deng [3]. In our experiments, we use N-grams of length $N = 4$, and evaluate the top $M = 20$ keywords.

2.3 Generative Models For Noisy Data

The statistical N-gram method works well when keywords and phrases are repeated multiple times. While this is often the case for longer or well-formed text input, short or noisy text often results in the majority of (non-stopword) N-grams only being mentioned once. With this type of input, statistical models are unable to decipher which keywords are most important. To better handle short or noisy text input, we use a keyword selection method based on generative topic modeling. In this model, we assume that a video comprises a small number of hidden topics, which can be represented as keyword probabilities, and that a video's text is generated from some distribution over those topics. The highly probable keywords in those topics are likely to be most representative of the video content. We use Latent Dirichlet Allocation (LDA) [1] to learn the topics and corresponding topic-keyword probability distribution from the input text. We then combine these topics to form a ranked keyword list.

2.3.1 Generating Topics

To discover the underlying topics in a video, we segment the input text into sentences and perform topic modeling with LDA. In our experiments, we set the number of topics $K = 5$ with the LDA parameters $\alpha = 0.3$ and $\beta = 0.1$. The resulting topic-term distribution ϕ is a KxV matrix, where K is the number of topics, V is the size of the input vocabulary, and $\phi[i][j]$ is the probability of keyword j in topic i. We form an ordered list of keywords k_i for each topic, sorted by their probability in $\vec{\phi[i]}$. This results in K ranked lists of keywords, one per topic, which must then be merged into a single list to select the top M. While simply selecting the top $\frac{M}{K}$ keywords from each topic is one option, we describe a more general solution for merging multiple ranked lists when we discuss our approach to finding related keywords in Section 3.3.

2.4 Filtering the Keywords

We apply two filters, when possible, to remove frequently occurring words which are often not useful in the context of matching advertisements. From all input sources, keywords matching a list of English profanity are removed. We also find that main character names are often amongst the top ranked keywords, but generally do not retrieve relevant advertisements. When given a complete script, we remove character names from the keywords using a dictionary automatically constructed during the parsing and tagging stage. For closed captioning and speech transcripts, however, these names are unknown and thus may still appear in the top keywords. This is more common for closed captioning than speech transcripts, however, as spoken character names are less likely to be correctly transcribed by the STT engine.

3. DISCOVERING RELATED TERMS

The keywords selected by processing the source text can provide a useful set of terms to represent the content of a video. These keywords are limited, however, to the vocabulary used by the original script authors. Closed captioning and speech transcripts are limited further to only the spoken dialog. An advertiser may have a particular set of semantically related keywords in mind which do not necessarily overlap with any of the selected keywords. These vocabulary mismatches result in missed opportunities to connect advertisers with relevant content. In this section we investigate two simple techniques for identifying related terms to help bridge the gap between the vocabularies used in videos and keywords chosen by advertisers. We explore term mining approaches based on (1) co-occurring terms using the Web, and (2) the Wikipedia graph.

Search results	Wikipedia	Combined
digital camera	photography	digital camera
lens	pornography	photography
canon	visual arts	canon
nikon	photograph	nikon
zoom	digital camera	pornography
film camera	photojournalism	lens
digital slr	photographic film	digital photography
megapixels	aperture	photograph
digital photography	canon	aperture
compact	photographic lens	shutter speed

Table 1: Example related terms for keyword "camera"

3.1 Mining with Web Search

Buckley et al. [2] noted that related terms will typically co-occur non-randomly in documents relevant to a query. To find candidate related keywords for term(s) T, we first submit T as a query to a Web search engine. For each of the top 50 search results, we identify a set of relevant keywords and construct a vector space model M from the results. Based on the popular TF-IDF [10] term weighting, we compute the corpus frequency (CF) and inverse-document-frequency (IDF) weight for each term in M, and rank the keywords according to their CF*IDF score.

3.2 Mining with Wikipedia

Graphical models for term expansion have been studied using random walks and multiple semantic links [4]. Within the text of a Wikipedia article, numerous *inter-wiki* links point to other Wikipedia pages, which allows us to model Wikipedia as a directed graph $G = \{V, E\}$. We use the link structure of the graph to both identify and rank candidate related terms. We require the relatedness between two article nodes a and b to be a symmetric relationship: a is related to b if and only if b is related to a. To identify candidate related terms for term T, we first locate the Wikipedia page with T as the title. Given the node t for T, we identify any nodes in the graph which form a direct cycle with t as candidate related terms. That is, keyword k is related to t if (t, k) and (k, t) are both in E. We then approximate the relative importance of terms by computing PageRank [8] over the Wikipedia graph. Candidate terms are assigned a score equal to their PageRank value, and ranked accordingly.

3.3 Combining Ranked Lists

The CF*IDF ranking metric for search result keywords has no inherent range, whereas PageRank assigns a value to each node such that the score of all pages sums to one. To combine these two sources, we normalize scores by assigning a score to each term within a list based on its reciprocal rank. For an ordered list of terms l, we assign a score to the term at rank i as $s_l(t_i) = (1 + \log i)^{-1}$, with any term not existing in the list assigned a score of 0. We may then combine the terms from any n ranked keyword lists into a single list, with a final score for each term t as $S(t) = \sum_{j=1}^{n} \alpha_j s_j(t)$, where the weight placed on list j is defined as α_j, such that $\sum_{j=1}^{n} \alpha_j = 1$. In our experiments, we placed equal weight ($\alpha = 0.5$) on both the search result and Wikipedia sources. Table 1 shows an example of the suggested related terms generated by the methods described above.

4. EVALUATION

We conducted a user survey to evaluate the keywords chosen from the source text and related term mining across a range of videos including 12 films, 3 clips from news and educational content, and 5 amateur clips from YouTube. Users were shown a 3-4 minute video (or film trailer) and a set of keywords. We show 5 of the top 20 keywords for each method and text source, and 1 of the top 10 related terms for each of those keywords, chosen and ordered at random. Users made binary assessments on the relevance of each keyword. Over 23 people participated in the survey (personally identifiable information was not required), with a minimum of 9 and average of 13 users evaluating each video.

Source	Statistical	Generative	S-Related	G-Related
Script	0.389	0.353	**0.254**	0.215
CC	**0.443**	0.397	**0.260**	0.221
STT	0.291	0.307	0.208	0.186

Table 2: Precision of Source and Related Keywords

4.1 Evaluation Metrics

We evaluate the generated keywords using two metrics (additional metrics are discussed in the full version of this paper [11]). The average relevancy of the keywords displayed to users we call the *precision*. The second metric we define is *popularity*, which serves as an indicator of how pertinent the keywords are to advertisers. We define precision and popularity as:

$$\text{Precision}(S) = \frac{1}{i} \sum_{i} \frac{|K_i(S) \cap R_i|}{|K_i(S)|}$$

$$\text{Popularity}(S) = \frac{1}{|R(S)|} \sum_{k \in R(S)} A_k^*$$

R_i is the set of keywords judged relevant in evaluation i and $K_i(S)$ are the keywords displayed to the user for evaluation i which come from source S. $R(S)$ are the keywords from source S judged relevant by at least one user, and A_k^* is the number of advertisers bidding for keyword k. Since we do not have an inventory of ads available to exactly know A_k^*, we estimate it with a Web search engine using the number of ads returned for query k. Although most commercial search engines limit $0 \leq A_k^* \leq 8$, we are primarily concerned with relative performance across text sources. Note that, because the popularity of a keyword is meaningless if it is not relevant to the content, we compute popularity for the set of keywords identified as relevant by at least one user.

4.2 Precision

Table 2 shows the precision of the keyword selection methods and their identified related terms. Cells in bold indicate a significant difference in performance ($p < 0.05$) between the two methods. For example, in Table 2, the precision of the statistical method on closed captioning tracks was higher than the generative method with $p = 0.037$.

As we expected, for "well formed" text such as scripts, the statistical method generally achieves higher precision. The generative method shows slightly better performance on the noisier speech transcripts, though the difference is not large enough to be statistically significant. The precision of related terms is lower than the corresponding terms identified directly from the source text. Interestingly, we also see that closed captioning data outperforms full scripts. This may indicate that viewers more closely associate dialog with the main points or themes of a video than the additional props, scenery, and actions described in a complete script.

We take a closer look at the performance for speech transcripts across three different video types in Table 3. Here we

Video Type	Statistical	Generative
Studio Films	0.268	0.252
News/Educational	0.442	0.473
User Generated	0.268	**0.368**

Table 3: Precision for Speech Transcripts

Video Type	WER	Statistical	Generative
Studio Films	0.857	0.723	0.690
News/Educational	0.406	0.731	0.961

Table 4: Relative Precision and Word Error Rate

see that for the longer, professionally produced films, the statistical method achieves marginally higher precision even on speech transcripts. The generative method performs significantly better on the shorter news and user generated clips, which supports our earlier intuition that statistical methods alone would likely have insufficient data to find the best keywords in such cases. We also note that news and educational content, on which the speech-to-text engine is expected to be most accurate, achieves the highest precision.

4.3 Relative Precision

Hauptmann's work indicates that STT word error rates (WER) under 0.4 result in retrieval performance comparable (approx. 80% relative retrieval precision) to a perfect transcript [5]. We compute the average WER for films and news/educational videos (using the STT engine's "default" language models), and compare the relative precision of STT with respect to CC in Table 4. User generated videos are not included because no "correct" transcripts are available for the content. As expected, the average WER for news and educational videos is substantially lower, though still around 0.4. For this type of content, the relative precision of STT is 96% of the closed captioning. For the higher word error rate of films we can still achieve over 70% average relative precision. These results further support use of the statistical selection methods on longer text inputs and the generative methods on shorter text, and suggest that speech transcripts alone may be sufficient to find meaningful advertising keywords for videos where the background noise is reasonably contained and the STT language models are appropriately trained, such as news and educational content.

4.4 Popularity

Popularity estimates the utility of keywords for advertising by measuring the average number of ads returned when each relevant keyword is issued as a search query, shown in Table 5. Popularity is notably higher for related terms in most cases, suggesting they would be more beneficial for advertising. While closed captioning was generally considered the most precise source of keywords, we also see it produces the least meaningful keywords for advertisers. This may be a result of character names appearing in the closed captioning keywords, which we noted earlier are filtered out from script input text and are less likely to retrieve relevant ads.

We also look closer at the popularity of keywords from speech transcripts in Table 6. In all cases, the keywords

Source	Statistical	S-Related	Generative	G-Related
Script	3.59	3.96	3.00	**4.18**
CC	2.11	**3.81**	2.00	**3.77**
STT	2.54	**4.39**	2.56	**4.30**

Table 5: Popularity of Keywords

Video Type	Statistical	S-Related	Generative	G-Related
Studio Films	2.97	**4.35**	2.67	**4.39**
News/Educational	1.69	**4.11**	2.21	**3.50**
User Generated	1.89	**4.83**	2.63	**4.75**

Table 6: Popularity for Speech Transcripts

identified through related term mining have higher popularity than the keywords from the source text by a statistically significant margin. It also again shows that news and educational content contains less popular keywords for advertisers.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have explored the suitability of a range of text sources for generating advertising keywords for video content. We have demonstrated that statistical N-gram keyword selection methods are effective when a sufficient amount of text data is available, while methods based on generative topic modeling perform better when the data is short or error prone, as is often the case with automatic speech recognition and user generated clips on sites such as YouTube. We have also shown that expanding the keywords from the source text with related term mining can substantially improve the likelihood of matching relevant and more marketable advertiser keywords. We used simple methods for identifying related terms to demonstrate improvements for advertising, though related works in term expansion (e.g. [4]) may provide even more relevant related keywords.

Although not studied in this short paper, clearly a tradeoff between precision and popularity can be played using a combination of source and related keywords. Readers are encouraged to view the full version of this paper [11] for a discussion of precision-popularity tradeoffs, as well as additional details and evaluations omitted for this short version.

6. ACKNOWLEDGEMENTS

This work is partially supported by NSF grants, IIS-0534784 and IIS-0347993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institutions.

7. REFERENCES

[1] D. Blei, A. Ng, M. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.
[2] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *TREC*, 1994.
[3] H. Chim and X. Deng. A new suffix tree similarity measure for document clustering. In *WWW*, 2007.
[4] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *CIKM*, 2005.
[5] A. Hauptmann. Lessons for the future from a decade of informedia video analysis research. In *CIVR*, 2005.
[6] D. Kelleher and S. Luz. Automatic hypertext keyphrase detection. In *IJCAI*, 2005.
[7] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, 1966.
[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Library Technologies Project*, 1998.
[9] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. Silva de Moura. Impedance coupling in content-targeted advertising. In *SIGIR*, 2005.
[10] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, 1988.
[11] M. J. Welch, J. Cho, and W. Chang. Generating advertising keywords from video content. *UCLA Computer Science Technical Report #100025*, 2010.
[12] W. T. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *WWW*, 2006.