An Online System with End-User Services: Mining Novelty Concepts from TV Broadcast Subtitles

Mika Rautiainen
CSE, University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
mika.rautiainen@ee.oulu.fi

Jouni Sarvanko
CSE, University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
jouni.sarvanko@ee.oulu.fi

Arto Heikkinen
CWC, University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
arto.heikkinen@ee.oulu.fi

Mika Ylianttila
CIE, University of Oulu
P.O.BOX 1001
FIN-90014 UNIVERSITY OF OULU, Finland
mika.ylianttila@cie.fi

Vassilis Kostakos
CSE, University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
vassilis@ee.oulu.fi

ABSTRACT

Better tools for content-based access of video are needed to improve access to time-continuous video data. Particularly information about linear TV broadcast programs has been available in a form limited to program guides that provide short manually described overviews of the program content. Recent development in digitalization of TV broadcasting and emergence of web-based services for catch-up and on-demand viewing bring out new possibilities to access data. In this paper we introduce our data mining system and accompanying services for summarizing Finnish DVB broadcast streams from seven national channels. We describe how data mining of novelty concepts can be extracted from DVB subtitles to augment web-based "Catch-Up TV Guide" and "Novelty Cloud" TV services. Furthermore, our system allows accessing media fragments as Picture Quotes via generated word lists and provides content-based recommendations to find new programs that have content similar to the user selected programs. Our index consists of over 180 000 programs that are used to recommend relevant programs. The service has been under development and available online since 2010. It has registered over 5000 user sessions.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information System**]: Video; H.3.1 [**Content Analysis and Indexing**]: [Indexing methods, Linguistic processing]; H.2.8 [**Database Applications**]: [Data mining]; I.2.7 [**Natural Language Processing**]: [Text analysis]

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

ACM 978-1-4503-2174-7/13/08 ...$15.00.

General Terms

Algorithms, Experimentation, Measurement

Keywords

video analysis, novelty concept detection, broadcast data mining, online TV

1. INTRODUCTION

Video broadcasts are often supplemented with subtitles, which provide very rich but underutilized data source for semantic video access. A popular way of summarizing text and web document collections is extracting key words or tags and generating word or tag clouds, which are typically based on frequency or popularity of a set of user assigned tags or document key words. Several methods exist for key word extraction [7][6][11]. Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training [5]. Another definition was given by Soboroff & Harman for NIST TREC Novelty track [9]: "the task was to highlight sentences containing relevant and new information in a short, topical document stream. This is analogous to highlighting key parts of a document for another person to read, and this kind of output can be useful as input to a summarization system." We introduce our novelty detection system as following: highlight new and relevant information from broadcast TV stream to summarize topical content for novel online services. This paper describes our data mining demo set-up with web-based end-user applications. In section 2 we give technical specification and in section 3 we present two end-user applications that have utilized the methodology.

2. TECHNICAL SPECIFICATION

Overview of our system architecture can be seen in Figure 1. It shows the main components of TV data mining system: DVB stream recording, novelty word detection using data mining and machine learning techniques, metadata source for mined data and end-user services that utilize the results

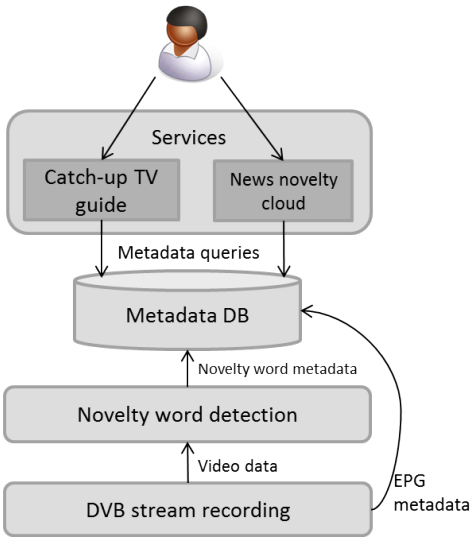

Figure 1: Content-based Broadcast TV Analysis System

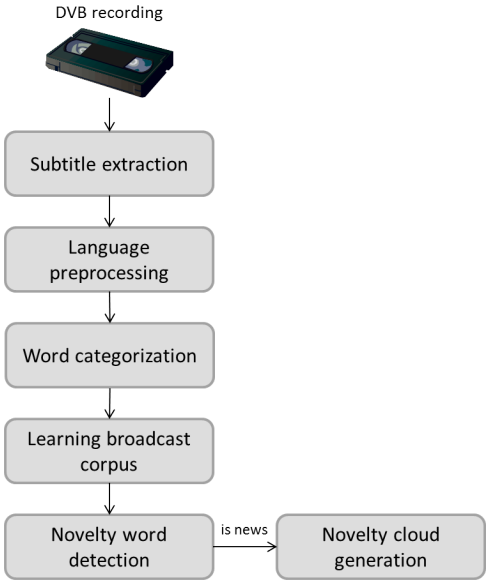

Figure 2: Content-based Broadcast TV Analysis System

of data mining to make broadcast information accessible on the web.

Our system records TV programs continuously from seven national TV channels while skipping chat shows. We use MythTV [1] to schedule recordings into a FIFO cache of MPEG-2 TS files. Our data mining process starts with the extraction of subtitle data from the DVB recordings. We look for all possible sources for subtitle data: Teletext, DVB subtitle bitmaps and subtitles embedded in video frames using OCR. Novelty word detection extracts unique and descriptive text concept word lists and stores this extracted metadata in a database along with the EPG metadata of the programs. The "Catch-up TV Guide" and "Novelty Cloud" end-user services use extracted metadata to summarize broadcast news as well as other content.

Figure 1 shows our data mining process in detail. Novelty word detection uses machine learning and statistical data mining techniques to extract descriptive novelty concepts automatically from TV program subtitles. Since subtitles may be extracted directly from video frames, they need to go through language preprocessing to reduce errors and prepare the text for further analysis. Next, the words are divided into categories before they are incorporated into trained broadcast corpus and used in novelty word detection. Novelty word detection extracts unique and descriptive words for a single program in the Catch-up TV Guide service and a set of descriptive concepts to summarize a group of programs in the Novelty Cloud service.

2.1 Video capturing and subtitle extraction

We capture TV programs from Finnish DVB-C broadcast channels and look for DVB subtitles or teletext if available. We also detect embedded subtitles from video frames if both DVB subtitles and teletext subtitles are unavailable. Depending on the source of bitmap subtitles, we apply image morphology and histogram analysis to preprocess bitmaps for optical character recognition (OCR). For OCR we use gocr and tesseract software. For video frame embedded subtitles, we post-process detected text data using error correction based on multi-sample word recognition and difference image verification. This allows us to clean srt encoded subtitles from sampling errors. In general, we estimate that we are able to obtain well above 90% of all subtitles in the selected TV channels. The channels use all aforementioned subtitling techniques with preference in teletext and DVB subtitle formats. Embedded subtitles are common in commercial channels.

2.2 Language preprocessing

Before analysis we filter out malformed words and fix OCR errors using heuristic rules. An important clean-up task is to remove stop words, i.e. frequent words that did not convey useful information, e.g. conjunctions and adverbs. We collected these words into a black list, which we used to separate the uninformative words from the words that contain unique and relevant concepts to highlight novel information from program content data.

Also, we implemented specific logic to find non-inflected word stems and lemma. We used two Finnish morphology libraries for Malaga: Voikko [2] to obtain part of speech of the words and Sukija [10] to get their lemma. Snowball [8] was used to find the stem of words.

2.3 Word categorization

We categorized words into generic words, names and abbreviations. Generic words are all the recognized words that are not names or abbreviations. Names and abbreviations were either recognized by Voikko [2] or detected using heuristic rules. Common abbreviations such as 'jne.' ('etc.' in English) were filtered out due to their uninformative nature.

2.4 Learning broadcast corpus

We constructed a normative broadcast word corpus which we use as a baseline for statistical novelty detection. This broadcast corpus is trained from the entire collection of indexed broadcast TV program subtitles and updated daily.

Table 1: Variable notations for two corpora

	Corpus 1	Corpus 2
Frequency of a word	O_1	O_2
Total amount of words	N_1	N_2

It stores the frequencies of all the recognized words and contains over 274 000 generic words, 53 000 unique names and 5 000 abbreviations. The word frequencies in a program were normalized using respective ordinal numbers to even out differences between longer and shorter programs. Normalization was achieved by sorting the list into descending order by word frequency and by dividing the ordinal number with the number of words on the list.

2.5 Novelty word extraction

Novelty word extraction carries out statistical comparison between a group of TV programs (hereafter subcorpus) and the broadcast baseline corpus. We used three time ranges (day, week, month) and 10 program genres (news and current affairs, cooking, nature, lifestyle, documentaries, science, travel, comedy, crime stories and all programs) to construct subcorpora. Each subcorpus contained a set of themed programs from which we wanted to detect novelty words.

We repurposed equations 1 and 2 from [7] to detect novelty words, i.e. words that are non-normative in a typical TV broadcast. Table 1 explains the used notations. These equations were used to compute novelty values for generic, name and abbreviation word lists. We also modified Eq. 2 to detect low frequency novelty words that are most unique for a single program. We did this by calculating the logarithm of the subcorpus and leaving out the weight balance multiplier O_i. The modified equation was used on generic words and as an outcome we obtained novel and unique information entities to describe program contents. These words were stored into a list of special novelty words.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \quad (1)$$

$$LL = 2 * ((O_1 * log(O_1/E_1)) + (O_2 * log(O_2/E_2))) \quad (2)$$

We use the extracted novelty concept lists to add descriptive metadata for each individual program in a sub-corpus. In the demonstration we show how novelty concept metadata is used to create semantic access to the parts of TV programs in the Catch-up TV Guide service and how content-based similar program recommendations are carried out using novelty concept metadata.

2.6 Novelty cloud generation

Novelty Cloud service provides a word cloud visualization that is generated from the extracted novelty concept words. In order to generate novelty cloud, word lists from novelty word extraction are combined to visualize novel concepts from generic, special and name word lists. To account for recurrence of novelty concepts over a series of programs in a sub-corpus, more frequent novelty words are given more weight in the cloud by increasing the font size for the most

Figure 3: Catch-up TV Guide program view for a program about Ötzi the Iceman.

frequent novelty words. Due to statistical data mining techniques against a trained corpus, the resulting word cloud produces different visualization of textual content than a traditional word cloud, which is typically based on simple popularity or frequency statistics. With novelty cloud we propose a visualization technique for broadcast video that is able to highlight unique and novel content entities over a time range. In the demonstration we show how our Novelty Cloud service summarizes weekly and monthly news topics from Finnish broadcast news. The method can also be used in other program genres.

3. DEMONSTRATION SETUP

The demonstration consists of two applications that show how the detected novelty words can be used to summarize unique and relevant information about TV content and access relevant parts of a program metadata.

Catch-up TV Guide [3] summarizes recently aired TV programs and facilitates content-based browsing of program content and finding interesting programs for on-demand catch-up TV viewing. Catch-up TV Guide links program summaries to broadcaster's catch-up TV services to enable viewing of video stream when interesting content is found. Default view shows programs from most recent days, with newest program on top. Figure 3 shows a program summary box in the Catch-up TV Guide. It displays a TV documentary about Ötzi the Iceman. The box shows a list of extracted novelty words for this program. Starting from the top, in order from left to right, the words in the novelty word cloud are: Borreliosis, human, iceman, iceman institute, Stone Age, contamination, copper axe, hand, lactose intolerance, world, stomach, arrowhead, cavity organ, made of/containing flint, melt edge.

Extracted novelty words bring forward quickly skimmable highlights from the TV program content, using both text and video frames from the program content. The textual summaries are presented in the form of a tag cloud, where the most novel words are highlighted using a larger font size. Each word in the cloud acts as a hyperlink that opens a program content browser from the point in the program where the word was first mentioned. Program content browser shows dynamically extracted picture quotes with subtitle excerpts to help people make relevance judgment on the program. The program content browser allows browsing through all the parts of the program where the selected novelty word appears. With the proposed design users do not need to resort to video timeline navigation at catch-up TV

streams to find out if program content is relevant and interesting for their needs. Instead they can assess the relevancy of the information before committing to viewing the stream. When relevant content has been found, The Catch-up TV Guide enables easy access to the actual program content by providing links to the program's web stream on the broadcaster's site, if available. The aims of Catch-up TV Guide service are to give easy access to new and interesting content inside TV programs and facilitate finding new and interesting TV content for on-demand viewing. In addition to program summaries, we demonstrate how extracted novelty concepts are used in recommending programs similar to selected program from the dynamically updating index of 180 000 TV programs. The service is first in Finland that is able to dynamically promote similar content across different TV channels using content-based techniques.

Another prototype end-user service is Novelty Cloud [4]. It is a service that summarizes novel news topics from TV news broadcasts for a week or a month. It combines generic, special and name novelty word lists and generates a large novelty word cloud from them. Novelty Cloud service displays novelty word summaries from a group of programs over a period of time instead of displaying individual programs. It portrays an overview of news events and allows users to navigate overviews monthly or weekly from the recent events to the beginning of 2012. Figure 4 shows a Novelty Cloud for the week 4 of 2013. The most novel words in the word cloud are Katainen (Finnish prime minister), Turku (a major Finnish city), Million euro loss (2012 Helsinki European Athletics Championships), Algeria, Border guard, Mali, church visit (Obama inauguration), trade union, Ice Hockey Federation, Islamic extremist etc. Users may click any word on the Novelty Cloud to retrieve TV programs that have mentioned the topic during the time period. Furthermore, users may see the original news broadcast at the broadcaster's site if the stream is still available. With the proposed design, Novelty Cloud service allows users to skim through broadcasts news topics for the past weeks and months in a quick and effortless manner. We have collected user logs of over 5000 sessions. The statistics of 1267 clicks show that selecting a novelty word in our Catch-up TV Guide service collected 51% of the clicks in the program summary. Key frame images had 29% of the clicks whereas program titles had 20%. Since 2013 Novelty Cloud service has collected 410 sessions with average click rate of 1.72 words per session.

4. CONCLUSIONS

We demonstrate our broadcast TV data mining system with two end-user applications that utilize rich text content in time-continuous video. Novelty concept extraction produces semantic content descriptions that facilitates finding new and relevant information from dynamically updating TV program index in a content-rich manner. We demonstrate the applicability of our data mining system with new end-user services: Catch-up TV Guide for browsing recently aired programs and Novelty Cloud for quick overviews of broadcast news topics. Novelty word summaries were the most popular way to examine program content in our Catch-up TV Guide service.

Figure 4: Novelty cloud from week 4 of the year 2013.

5. ACKNOWLEDGMENTS

We would like to thank Academy of Finland and Finnish Funding Agency for Technology and Innovation for supporting this work.

6. REFERENCES

[1] Mythtv, open source dvr. `http://www.mythtv.org/`.

[2] Voikko - free linguistic software for finnish. `http://voikko.sourceforge.net/`.

[3] Kuukkelitv.fi. KuukkeliTV - Mediaseinä. http://www.kuukkelitv.fi/mediaseina.

[4] Kuukkelitv.fi. KuukkeliTV - Uutispilvi. http://www.kuukkelitv.fi/uutispilvi.

[5] M. Markou and S. Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481 – 2497, 2003.

[6] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[7] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, CompareCorpora '00, pages 1–6, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[8] Snowball. Snowball. `http://snowball.tartarus.org/`.

[9] I. Soboroff and D. Harman. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[10] Sukija. Sukija. http://sourceforge.net/projects/sukija/.

[11] K. Zhang, H. Xu, J. Tang, and J. Li. Keyword extraction using support vector machine. In J. Yu, M. Kitsuregawa, and H. Leong, editors, *Advances in Web-Age Information Management*, volume 4016 of *Lecture Notes in Computer Science*, pages 85–96. Springer Berlin Heidelberg, 2006.