Information Processing and Management 51 (2015) 148–162

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

IntoNews: Online news retrieval using closed captions

Roi Blanco, Gianmarco De Francisci Morales, Fabrizio Silvestri *

Yahoo Labs, Barcelona, Spain

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 18 July 2014
Accepted 29 July 2014
Available online 11 September 2014

Keywords:
Second screen
News retrieval
Continuous retrieval
IntoNow
IntoNews

ABSTRACT

We present IntoNews, a system to match online news articles with spoken news from a television newscasts represented by closed captions. We formalize the news matching problem as two independent tasks: closed captions segmentation and news retrieval. The system segments closed captions by using a windowing scheme: sliding or tumbling window. Next, it uses each segment to build a query by extracting representative terms. The query is used to retrieve previously indexed news articles from a search engine. To detect when a new article should be surfaced, the system compares the set of retrieved articles with the previously retrieved one. The intuition is that if the difference between these sets is large enough, it is likely that the topic of the newscast currently on air has changed and a new article should be displayed to the user. In order to evaluate IntoNews, we build a test collection using data coming from a second screen application and a major online news aggregator. The dataset is manually segmented and annotated by expert assessors, and used as our ground truth. It is freely available for download through the Webscope program.[1] Our evaluation is based on a set of novel time-relevance metrics that take into account three different aspects of the problem at hand: precision, timeliness and coverage. We compare our algorithms against the best method previously proposed in literature for this problem. Experiments show the trade-offs involved among precision, timeliness and coverage of the airing news. Our best method is four times more accurate than the baseline.

1. Introduction

Television has been the most important communication medium of the last century. However, in the last few years the Web has started to take over this role thanks to a wider offer of content and the possibility of interaction. Recently, a new breed of applications for mobiles and tablets has started appearing on the market, offering the so-called "*second screen*" experience. The goal of these applications is to enhance the TV-watching experience by providing additional content related to the program airing at the moment, thus bridging the TV and Web worlds. By allowing the audience to interact with the program on TV, second screen applications ultimately aim at increasing user engagement. These applications are the natural evolution of a widely recognized trend: between 75% and 85% of TV viewers use another device at the same time.[2]

Second screen[3] from Yahoo! is an example of a second screen application, and the focus of the current work. When second screen launched in 2011, users immediately acclaimed this application as a fun way of watching TV programs. The user

* Corresponding author.
E-mail addresses: roi@yahoo-inc.com (R. Blanco), gdfm@yahoo-inc.com (G. De Francisci Morales), silvestr@yahoo-inc.com (F. Silvestri).

[1] http://webscope.sandbox.yahoo.com.
[2] http://www.guardian.co.uk/technology/appsblog/2012/oct/29/social-tv-second-screen-research.
[3] http://www.intonow.com.

http://dx.doi.org/10.1016/j.ipm.2014.07.010
0306-4573/

experience for people watching TV program is greatly improved. For instance, while watching a football game on TV it can show statistics about the teams playing, or show the title of the song performed by a contestant in a talent show. Other services include forums, episode synopsis, real-time meme generator (CapIt), polls and much more. second screen aims at enhancing the experience of watching TV transforming it into a "*large scale*" social activity. The additional content provided by second screen is a mix of editorially curated and automatically selected one.

From a research perspective, one of the most interesting and challenging use cases for these applications is related to news programs (newscasts). When a user is watching a newscast, they might want to delve deeper into the news airing at the moment. This work presents IntoNews, a system that finds an online news article that matches the piece of news discussed in the newscast currently airing on TV, and displays it to the user in real-time.

The main problem underlying IntoNews is matching different data sources that speak about the same piece of news. On one side we have the text from online news articles. On the other, we obtain the content of the newscast currently airing from the streams of *Closed Captions* (cc) broadcasted along with it by television networks.

The challenges in making IntoNews effective are multiple. The news article we surface to the user must match exactly the news currently airing. The problem is even more challenging given that the matching article has to be selected among the thousands published online every day. The language used on TV and in news articles has different characteristics. Furthermore, the cc tend to be noisy, lack proper capitalization and contain many typos and misspellings. Finally, and most importantly, news articles must be surfaced as soon as possible to be valuable to the user.

We propose a solution based on techniques from the realm of information retrieval (IR). Fig. 1 shows the conceptual schema of the components of our system and how they interact with each others. We decompose the main news matching task into two sub-tasks: find a good segmentation of the stream of cc, and retrieve relevant news for the segment as soon as possible. We model a newscast as a series of contiguous segments, each matching a single cohesive topic. The *segmentation* problem consists in finding the boundaries of these news segments in the stream of cc. The *retrieval* problem consists in formulating a query given a segment, and issuing the query to an underlying IR engine.

While the user is watching the newscast, the system continuously processes the cc and tries to detect a segment boundary ("new story"). If a new segment is detected, it examines the incoming text until it has enough information to retrieve the news article from the IR engine. When enough information has been accumulated, the system submits the query to the IR engine, retrieves the results, and shows them to the user.

There are differences between this problem and those faced by traditional IR systems. Users of a typical IR system issue queries to retrieve a set of top-k most relevant items from a collection. We can identify three distinct phases in a typical IR process: (i) the user formulating the query and issuing it, (ii) the IR system processing the query and retrieving the top-k items, and (iii) the user checking a subset of the resulting items to satisfy their information need.

IntoNews differs in phases (i) and (iii). First, it does not require the user to formulate a query, rather the system "*implicitly*" formulates one for the user by using the content of the newscast airing on TV. In fact, formulating a query by observing only a continuous stream of text without any indication on topic boundaries, query keywords, important concepts or entities is a challenging task, which is fundamentally different from typical IR tasks.

Second, the user sees a small number of results that are continuously changing as new cc arrive. Usually, IR quality assessment only evaluates the amount of relevant documents ranked at the top of the result list. However, IntoNews has to account

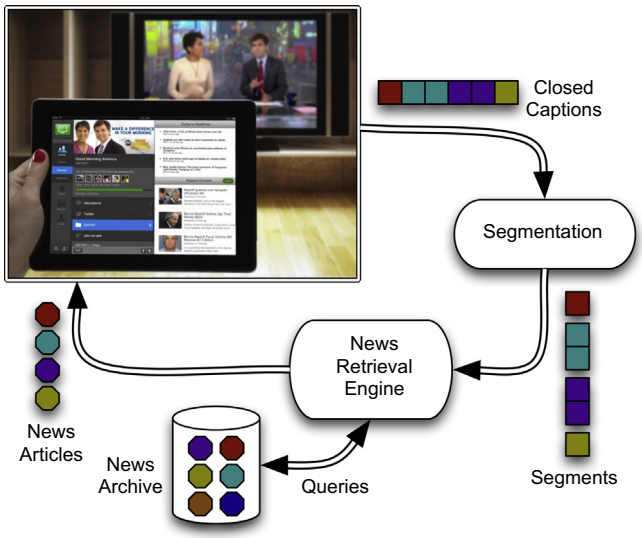

Fig. 1. Conceptual schema of IntoNews.