Advances in Knowledge-Based and Intelligent Information and Engineering Systems
M. Graña et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-105-2-575

Automatically generating multilingual, semantically enhanced, descriptions of digital audio and video objects on the Web

Bernardo Pereira Nunes[$♯], Alexander Mera[$], Marco A. Casanova[$], Ricardo Kawase[♯]

[$] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil
{bnunes, acaraballo, casanova}@inf.puc-rio.br

[♯] L3S Research Center, Leibniz University Hannover, Germany
{nunes, kawase}@l3s.de

Abstract. Every day, millions of new images, videos and audios are uploaded to the web. However, unlike text-based content, audio and video objects cannot be indexed by search engines. Thus, much valuable multimedia content stay unreachable for a great majority of online users. To overcome this problem we introduce a technique that automatically generates semantically enhanced descriptions of audio and video objects. The goal is to facilitate indexing and retrieval of the objects with the help of traditional search engines. Basically, the technique automatically generates static Web pages that describe the content of the digital audio and video objects, organized in such a way as to facilitate locating segments of the audio or video that correspond to the descriptions. The technique is a mash-up of Web services that also provides translation of the descriptions and semantic enhancement. We thoroughly analyzed the click-data comparing accesses to the digital content before and after the automatic generation of the descriptions. The outcomes suggest that the technique significantly improve the retrieval of items, not only in terms of visibility, but also brings down language barriers, by supporting multilingual access.

Keywords: publishing multimedia content, spoken content retrieval, spoken lecture processing

1 Introduction

Internet has arguably become the main source of information. Whereas it all begun with simple hypertext systems - with mere textual information, the technical evolution of hardware and broadband internet access shifted the concept of the Web to a multimedia system. Additionally, a few years ago, the exchange of information on the Web was predominantly one-way, where webmaster could publish their content to the general internet users. However, with the catch up of the Web 2.0, new means of interaction provide the ordinary internet users tools to contribute with their own content. Due to the combination of these two factors, together with the dissemination of electronic devices with built-in digital cameras and audio recorders, multimedia content proliferated tremendously and became an important source of information, communication and social interaction on the Web. This caused the phenomenon of online multimedia content

that we see today. Every day, millions of new images, videos and audios are uploaded to the web. This huge amount of new data has facilitated and motivated numerous researches on different fields. Our focus regards the issues involving search, retrieval and access to multimedia content, specifically what we call spoken content (audio and video objects).

Today, search engines are the gatekeepers to information. Most information accesses - if not every, starts with a keyword search. Unlike text-based content, spoken content is encoded into digital audio and video objects which still cannot be indexed by search engines. These types of objects do not contain intrinsic textual description. Thus, the retrieval of these types of information from a keyword search engine is solely based on the few existing metadata (title, description, author, etc). Not to mention the low quality and descriptiveness of these data. The use of content-independent metadata has been the accepted alternative to describe multimedia files in the last decade [3, 10].

In addition to that, multilingual accessibility is yet another barrier to discover content online. While automatic translation tools do a rather good job translating textual documents and websites, very few have been done to support the cross-language retrieval of objects.

To address the problem of multimedia access on the Web caused by the lack of indexable contextual content and language barriers, in this work we present a mash-up tool that facilitates the indexing and retrieval of spoken content through the automatic generation of transcripts, semantic annotations and translation.

The contributions of this work are twofold:

- We provide an online tool to automatic generate semantically enriched transcripts and translations of spoken content.
- An evaluation of the technique regarding real learning objects (spoken content).

With the outcomes provided by our contributions, we aim to answer the following research questions:

- To which extent can automatic generated scripts improve the retrieval of spoken content?
- To which extent automatic translations of scripts enable the use of spoken material?

The remainder of the paper is structured as follows. In Section 2 we describe some related work in the area. Section 3 introduces our publishing technique. Section 4 exposes the outcomes of using our tool. Finally, Section 5 presents the discussions, conclusions and future works.

2 Related Work

A recent work explores the automatic generation of tagging and geotagging which improves video retrieval [9]. In fact, most audio, video and image repositories [2, 16, 18] on the Web uses similar approaches to index multimedia files. Although content-independent metadata, such as title, authors [6, 4], can describe some aspects of spoken content, the content itself of such objects is still not accessible to text-based search engines. Even if the spoken content can be found, a user still has to manually locate

the segment of information that meets his need, which can be very tedious. Thus, to increase the chances of retrieving spoken content and to locate not just the file, but the time-aligned segment of a spoken content, a more elaborated form of annotation is required. However, a content-descriptive metadata which transcribes audio and video contents is a tiresome and laborious task almost never done by the creator or the publisher of the content.

Alberti et. al. [1] addresses the spoken content retrieval problem in the context of last US presidential campaign, where they developed a scalable system that makes the video content searchable. Although their approach uses content-descriptive metadata and content-independent metadata to describe spoken content, the content is not prepared to be machine-readable. Repp et. al. [15] present a machine-readable approach where they use a specific ontology to annotate content and make it available using OWL-DL for semantic search engines. Although they provide semantic information to the search engines, the same semantic information is not available to assist humans on the Web. A more appropriate approach would be to adopt RDFa [17], which would help to index and retrieve Web contents for both machine and humans [7].

Glass et. al. [5] discuss the components involved in a spoken content retrieval system for make spoken lectures findable. However, in their approach, content is still hidden to search engines, since the transcriptions files can only be accessed through a search form, which is an open challenge to the data management community [10, 11, 8]. Thus, to avoid a laborious task and to make spoken content available on the Web, we present in this paper an automatic technique to create textual descriptions for audio and video objects that transforms the spoken search problem into a traditional text search problem. In addition, the technique processes spoken content in such a way that text-based search engines can help locate time-aligned segments of te content.

Furthermore, the technique also annotates entities [12] present in the transcriptions using RDFa. The annotation process uses Dublin Core [4] to describe the content-independent metadata of an asset and the content-descriptive metadata uses a specific set of ontologies from DBPedia [1] for describing the concepts and relationships with others resources on the Web. As a consequence, once the content is linked with other resources, content provenance is known by the search engines, which contributes to improve precision page ranking.

To demonstrate the effectiveness of the technique proposed in this paper, we describe an experiment with more than a thousand minutes of spoken contents, divided into 99 video objects and an in-depth hit analysis of these objects.

3 Publishing Technique

In this section, we describe the publishing process with the help of a complete publishing example of a spoken content (see Figure 1). Given a spoken content, the first step is to transcribe the spoken content using an automatic speech recognition service (ASR) [13, 14]. The set of time-aligned text excerpts thus obtained forms what we call the script of the spoken content, by analogy with the usual meaning of the word. Since

[1] http://www.dbpedia.org

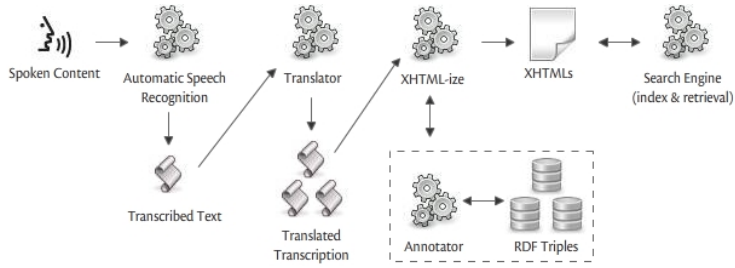

Fig. 1. Publishing technique.

the script is a textual representation of the spoken content, the spoken search problem is converted into a text search problem. Figure 2 shows an example of a script.

The second step of the publishing technique is to translate the script to other languages, if desired, to reach other user populations. Our technique uses the Google Translator API [2] to translate a script in several languages. In the next section, we discuss how this step impacts content retrieval.

The last step is decomposed into two substeps: (a) to transform a plain text script into a XHTML (eXtensible Hypertext Markup Language) file; and (b) to annotate the content using RDFa (Resource Description Framework in attributes).

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en">
...
<div>
  <a href="URL?time=04m47s">Watch this excerpt</a>
  <p>
      Against this backdrop comes Fritz Haber. The man
      who carried out the synthesis of ammonia in order
      to produce it on an industrial scale, from
      molecular hydrogen and nitrogen, abundantly
      available.
  </p>
</div>
```

Fig. 2. Example of a script.

The first substep generates static Web pages for each asset and converts each time-aligned excerpt of text from a spoken content into sections (div elements), where each section contains a hyperlink (the a element) that points to the exact segment from the spoken content that the speech occurs and the transcribed text related to the spoken content (p element) (see Figure 2). Note that each static Web Page thus generated specifies the document type and the language of that content. In the example of Figure 3, the document type is XHTML (4.01 strict) and the language of the content is in English ("xml:lang=en").

[2] http://code.google.com/apis/language/

```
...
    Against this backdrop comes <a
    about="http://dbpedia.org/resource/Fritz_Haber"
    typeof="http://dbpedia.org/ontology/Scientist"
    href="http://dbpedia.org/resource/Fritz_Haber"
    title="http://dbpedia.org/resource/Fritz_Haber">F
    ritz Haber</a>.
...
```

Fig. 3. Example of a Web page.

Although XHTML supports semantic description, the code in Figure 2 is not enriched with semantic markups. Thus, the second substep embeds into the XHTML markups a collection of attributes to enrich the semantics of the Web page content. The content transcribed in the previous step is analyzed by the Spotlight Web Services, which annotates references to DBPedia resources in the text. Hence, the Spotlight Web Services enriches the text with entity detection and name resolution. Figure 4 shows the annotation result. This substep has the potential to provide a solution for linking unstructured information sources to the Linked Open Data (LOD) cloud through DBpedia. Thus, besides contributing to enhancing search and information retrieval of the assets, the annotated text can be used for secondary tasks, such as display additional information about an asset and recommend related assets based on semantics.

Fig. 4. Result of the XHTML-ize step.

4 Experiments

We performed an extensive evaluation of our spoken content publishing technique using real Web data from the educational domain. Besides assessing the efficacy of the publishing technique, our goals included an in-depth analysis of the hits on the Web pages synthesized for the audio and video learning objects.

4.1 Experimental Setup

We evaluated the tool using a set of 99 Learning Objects (LO's), composed of video files, with 10 minutes long, containing dialogs covering elementary Chemistry topics. We conducted the experiments in two stages, for 8 months. The first stage lasted 5 months and all 99 LO's were published using content-independent metadata. The goal of the first stage was to equalize the Web page hits using content-independent metadata and use this information to create two balanced groups of LO's to assess the efficacy

of the tool. We sorted LO's by the number of hits and then, for each pair of LO's whose number of hits have the same order of magnitude, we assigned each of them to a different group.

The second stage lasted 3 months. We submitted one of the groups to the tool, but not the other. The goal of this second stage was to assess the efficacy of the tool. We hosted the static Web pages on the WordPress server and the video files on Youtube and used the statistics these services provide to evaluate the tool.

4.2 Data Analysis

Let Group A refer to the set of LO's published in the first stage, Group P refer to the set of LO's published using the tool in the second stage, and Group $\neg P$ refer to the set of LOs described only by content-independent metadata. We present in this section the results of the hit analysis for both stages of the experiment.

Total Hits Analysis. Approximately 75K hits were obtained during both stages of the experiment. However, Group A obtained just 22% from the total number of hits. Note that Group A represents the first stage with all 99 LO's. Also note that first stage collected data during 5 months, whereas the second stage lasted 3 months. Hence, 78% of the hits were performed in the second stage (see Figure 5).

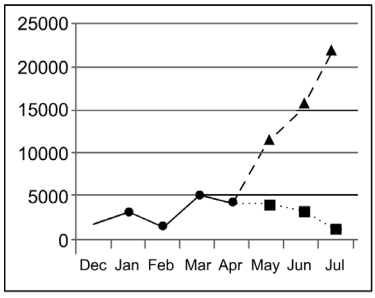

Fig. 5. Hit stats. First stage 1-5. Second stage 5-8.

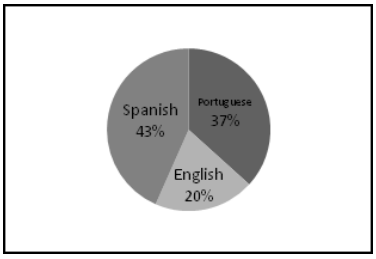

Fig. 6. Hits percentage of translated static Web pages generated by our publishing technique.

Observe that Group P obtained 66% of the total number of hits and Group $\neg P$ just 12%. Thus, Group P obtained 84% of the total number of hits in the second stage, i.e., 5.3 times more than the number of hits obtained by Group $\neg P$.

Page Hits Analysis. As described in Section 3, the translation step created a new static Web page, for each new language. In the experiment, three scripts were generated for each asset, in English, Spanish and Portuguese. Figure 6 gives, for each translated static

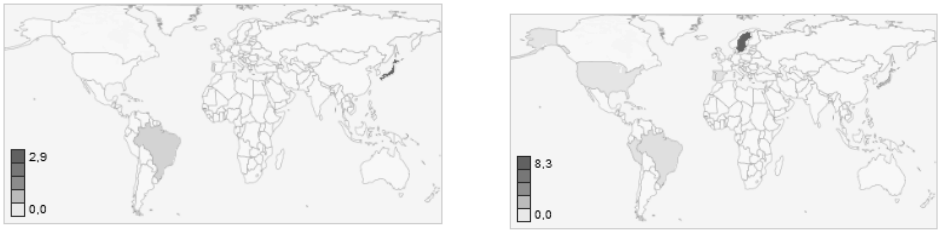

Fig. 7. Countries that have interacted with the content, on the left: Japan, Brazil and Portugal. On the right: Sweden, Japan, Brazil, Spain, Peru, USA and Portugal.

Web page, the percentage of total number of hits: 43% for pages in Spanish, 37% for pages in Portuguese and 20% for pages in English.

Although the visitors of our static Web pages are anonymous, i.e., they do not provide information about where they come from or his/her mother language, Youtube provides information about logged in users.

The information obtained from Youtube is very important because all actions (share, comment or mark as favorite) are done by users that are somehow interested in the content of an asset. Using such information, left part of Figure 7 indicates that only users from Brazil, Portugal and Japan shared, commented or marked as favorites the assets of Group A.

After applying the technique, creating Group P, users from other countries (Sweden, Japan, Brazil, Spain, Peru, United States and Portugal) could be reached, as shown in the right part of Figure 7. Note that, the native language of Brazil, Spain, Peru, United States and Portugal is indeed English, Spanish or Portuguese, the languages available through our technique. Curiously enough, this is not true of Japan and Sweden, although there is a sizable population of Brazilians living in Japan.

LO's with the largest number of hits	1	2	3	4	5	6	7	8	9	10
Group $\neg P$	1636	1615	1416	969	812	642	514	300	265	220
Group P	2466	1774	1744	1562	1499	1467	1421	1414	1386	1307

Table 1. Top 10 most accessed Learning Objects.

Least hit LO's	10	9	8	7	6	5	4	3	2	1
Group $\neg P$	32	27	27	25	24	21	21	12	11	7
Group P	333	326	287	271	250	213	195	191	108	95

Table 2. Bottom 10 most accessed Learning Objects.

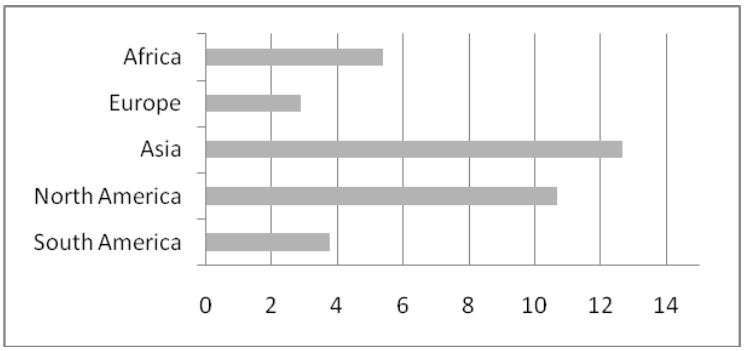

Fig. 8. The number of hits boosted in different orders of magnitude for each continent.

Assets Hits Analysis. This part of the experiment addressed the question: What is the difference between the number of hits of an LO that was published using the tool and an LO described by content-independent metadata? Table 1 shows the top 10 LO's, according to the number of hits, where the second line corresponds to LO's in Group P and the third line to LO's in Group $\neg P$. Table 1 indicates that the number of hits for LO's indexed by the tool (Group P) is indeed greater than the number of hits for LO's described only by content-independent metadata (Group $\neg P$), as expected. This is true in general for all 99 LO's. Table 2 shows the LO's with the least number of hits. It shows an even greater discrepancy.

Regional Analysis. Since Youtube provides the user's location, it was possible to tabulate the number of hits by continent (with the exception of Oceania). In the first stage analysis, Group A obtained 13,911 hits from South America, 73 hits from North America, 39 hits from Asia, 1,297 hits from Europe and 31 hits from Africa. We note that, from the hits in South America, 13,673 (approx. 99%) were from Brazil, which in a Portuguese-speaking country. The same was observed in Europe where, out of the 1,297 hits, 1,107 (approx. 85%) were from Portugal, which is also a Portuguese-speaking country.

In the second stage, the assets obtained 52,366 hits from South America, 779 hits from North America, 494 hits from Asia, 3,738 hits from Europe and 167 hits from Africa. We again highlight that, out of the number of hits from South America, most were from Brazil. However, during a short period, the number of hits from the rest of South America, mostly from Spanish-speaking countries, increased from 1% to 5.5%. (Note that the population of Brazil is almost 50% of the population of South America). The number of hits from Europe was also less concentrated: Portugal, that had 85% of the hits in the first stage of the experiment, had 67% in the second stage; the total number of hits from other European countries more than doubled from 15% to 33%. Figure 8 shows the ratio increase by continent, obtained by dividing the number of hits in the second stage of the experiment by the number of hits in the first stage.

5 Conclusion

In this paper we described a technique to automatically enhance spoken content on the Web with semantic descriptions (transcripts) and translations. This technique facilitates indexing and retrieval of the objects with the help of traditional text search engines. Based on our techniques we implemented an online tool which has been used in the evaluations demonstrated in this paper.

The tool automatically generates static Web pages that describe the spoken content, organized to facilitate locating segments of the content that correspond to the descriptions. Moreover, the tool annotates the described spoken content using RDFa and DBPedia for linking unstructured information sources to the LOD cloud and to enhance search and information retrieval of the assets. The tool also provided a way to bring down language barriers creating a multilingual resource which amplifies the range of users. The evaluation showed that the number of hits to the objects processed by the tool was significantly improved, as well the access of consumers of foreign languages.

For future work, there are several extensions that we will investigate. First, the tool may resort to semantic information to display complementary information about an asset. Second, it would be useful to enrich the semantics to work not only with DBPedia resources but also with other LOD data sources. Finally, it would be interesting to recommend related assets by taking advantage of the connected text through ontologies from the LOD and to assess how effective the recommendation was.

A complete description of the tool may be found at `http://moodle.ccead.puc-rio.br/spokenContent/`.

6 Acknowledgement

This research has been co-funded by the European Commission within the eContentplus targeted project OpenScout, grant ECP 2008 EDU 428016 (cf. `http://www.openscout.net`) and by CAPES (Process n^o 9404-11-2). Additionally, we would like to thank Gilda Helena Bernardino de Campos for providing access to the data used in this paper.

References

1. C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan. An audio indexing system for election video material. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4873–4876, Washington, DC, USA, 2009. IEEE Computer Society.
2. Baidu search engine. `http://www.baidu.com`.
3. D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416 –430, may 2008.
4. Dublin core metadata initiative. `http://www.dublincore.org`.
5. J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. Interspeech*, 2007.
6. A. Hanbury. A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19:617–627, October 2008.

7. B. Haslhofer, E. Momeni, M. Gay, and R. Simon. Augmenting europeana content with linked data resources. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 40:1–40:3, New York, NY, USA, 2010. ACM.
8. L. Jiang, Z. Wu, Q. Zheng, and J. Liu. Learning deep web crawling with diverse features. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 572–575, Washington, DC, USA, 2009. IEEE Computer Society.
9. M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 51:1–51:8, New York, NY, USA, 2011. ACM.
10. J. Madhavan, L. Afanasiev, L. Antova, and A. Halevy. Harnessing the Deep Web: Present and Future. In *4th Biennial Conference on Innovative Data Systems Research (CIDR)*, Jan. 2009.
11. J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep web crawl. *Proc. VLDB Endow.*, 1:1241–1252, August 2008.
12. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In C. Ghidini, A.-C. N. Ngomo, S. N. Lindstaedt, and T. Pellegrini, editors, *I-SEMANTICS*, ACM International Conference Proceeding Series, pages 1–8. ACM, 2011.
13. Nexiwave speech indexing. `http://www.nexiwave.com`.
14. Nuance dragon naturallyspeaking. `http://www.nuance.com`.
15. S. Repp and C. Meinel. Automatic extraction of semantic descriptions from the lecturer's speech. In *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, pages 513 –520, sept. 2009.
16. Truveo video search. `http://www.truveo.com`.
17. W3c rdfa primer. `http://www.w3.org/TR/xhtml-rdfa-primer`.
18. Youtube broadcast yourself. `http://www.youtube.com`.