Boosting Retrieval of Digital Spoken Content

Bernardo Pereira Nunes[1,2], Alexander Mera[1],
Marco A. Casanova[1], and Ricardo Kawase[2]

[1] Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil
{bnunes,acaraballo,casanova}@inf.puc-rio.br
[2] L3S Research Center, Leibniz University Hannover, Germany
{nunes,kawase}@l3s.de

Abstract. Every day, the Internet expands as millions of new multimedia objects are uploaded in the form of audio, video and images. While traditional text-based content is indexed by search engines, this indexing cannot be applied to audio and video objects, resulting in a plethora of multimedia content that is inaccessible to a majority of online users. To address this issue, we introduce a technique of automatic, semantically enhanced, description generation for multimedia content. The objective is to facilitate indexing and retrieval of the objects with the help of traditional search engines. Essentially, the technique generates static Web pages automatically, which describe the content of the digital audio and video objects. These descriptions are then organized in such a way as to facilitate locating corresponding audio and video segments. The technique employs a combination of Web services and concurrently provides description translation and semantic enhancement. Thorough analysis of the click-data, comparing accesses to the digital content before and after automatic description generation, suggests a significant increase in the number of retrieval items. This outcome, however is not limited to the terms of visibility, but in supporting multilingual access, additionally decreases the number of language barriers.

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

1 Introduction

The Internet has veritably become the predominant source of information. What began as mere textual information within simple hypertext systems has evolved mutually with technical hardware and broadband Internet access and shifted the simple Web concept to one of a complex multimedia system. Where previously, the exchange of information on the Web was chiefly one-way, i.e., webmasters published content to general users, the catch up of Web 2.0 has opened up new means of interaction in which ordinary Internet users were given the tools to contribute with their own content. As a result of these two factors, together with the dissemination of electronic devices with built-in digital cameras and audio recorders, multimedia content has proliferated tremendously and become an important source of information, communication and social interaction on the Web. The product is the emanation of a multimedia phenomenon that marks the online content we see today. Millions of new images, videos and audio are uploaded

M. Graña et al. (Eds.): KES 2012, LNAI 7828, pp. 153–162, 2013.

to the web on a daily basis, spurring and motivating an expanse of new research in various fields. Our focus lies with the issues involving search, retrieval and access to multimedia content, specifically that of audio and video objects, which we refer to as "spoken content".

Today, search engines are the gatekeepers to information. Almost all information accesses begin with a keyword search. Spoken content, unlike its text-based rival, still cannot be indexed by search engines as it is encoded into digital audio and video objects which do not contain intrinsic textual description. Retrieval of information of this type from a keyword search engine is therefore based solely on the few existing metadata (title, description, author, etc), which is of low quality and descriptiveness. As an alternative, content-independent metadata has been acceptably employed to describe multimedia files over the last decade [3,9].

Multilingual accessibility, however has posed another problem to online content discovery. While automatic translation tools do a rather good job translating textual documents and websites, very few have been proven to support the cross-language re-trieval of objects. To confront the issue 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 which automatically generates semantically enriched transcripts and translations of spoken content.
- An evaluation of the technique as pertains to real learning objects (spoken content).

In accordance with the outcomes provided by our contributions, we aim to answer the following research questions:

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

The remainder of the paper is structured as follows. Section 2 describes relavant information from previous work. Section 3 introduces our publishing technique. Section 4 exposes the results of implementing our tool. Finally, Section 5 presents discussions, conclusions and future works.

2 Related Work

A recent study explored the improvement of video retrieval via automatic generation of tagging and geotagging [8]. This work concurs with similar approaches which most audio, video and image repositories [2,15,17] on the Web utilize to index multimedia les. However, while content-independent metadata, such as title or author, can describe some aspects of spoken content, the actual content of these objects remains inaccessible to text-based search engines. This results in a tedious search on behalf of the user, because even if the spoken content can be found, the user must still manually locate

the specific segment he is seeking. In order to increase the likelihood of retrieving the time-aligned segment of spoken content rather than just the file, a more elaborate form of annotation is required. This content-descriptive form of metadata, which transcribes audio and video content, is, however, a wary task rarely executed by the publisher and/or creator of the content.

Alberti et al. [1] addresses the spoken content retrieval problem in the context of last US presidential campaign, where a scalable system that makes the video content searchable was developed. Although their approach adhibits content-descriptive metadata and content-independent metadata to describe spoken content, the content is not prepared to be machine-readable. To address this issue and present a machine-readable approach, Repp et al. [14] apply a specific ontology to annotate content and make it attainable using OWL-DL for semantic search engines. This approach provides semantic information to search engines, however that same semantic information is not available to assist human Web users. A more proactive approach would be the adoption of RDFa [16], which would assist both machines and humans to index and retrieve Web content [6].

Glass et al. [5] discuss making spoken lectures findable by implementing components of a spoken content retrieval system. Yet this approach poses a challenge to the data management community [9,10,7] because transcription files can only be accessed through a search form and are therefore still hidden from search engines. To pragmatically combat these issues we present a technique of automatic video and audio text description generation, which transforms the spoken search problem into a traditional text search problem, and in so doing makes spoken content available to users on the Web. Additionally, our technique processes spoken content in a manner which permits text-based search engines to assist in locating time-aligned segments of the content.

Furthermore, the technique also annotates entities [11] present in the transcriptions using RDFa. The annotation process employs Dublin Core [4] to explicate the content-independent metadata of an asset, while the content-descriptive metadata applies a specic set of ontologies from DBpedia[1] to illustrate the concepts and relationships with other Web resources. Consequently, content provenance is known by search engines once that content has been linked with other resources, thus improving precision page ranking.

The effectiveness of the proposed technique is described by an experiment with over a thousand minutes of spoken content, divided into 99 video objects and an in-depth hit analysis of these objects.

3 Publishing Technique

This section describes the publishing process with the aid of a complete spoken content publishing example (see Figure 1). Using an automatic speech recognition service (ASR), the first step is to transcribe a given spoken content [12,13]. The set of time-aligned text excerpts thus obtained forms what we call the script of the spoken content, by analogy of the usual meaning of the word. This approach thus converts the spoken search problem to a text search problem since the script is a textual representation of the spoken content. Figure 2 shows an example of a script.

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

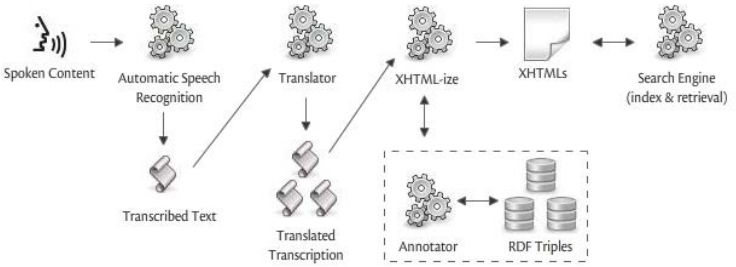

Fig. 1. Publishing technique

Step two seeks to reach additional user populations by translating, if desired, the script into other languages. Our technique translates these scripts into various languages using the Google Translator API[2]. In the following section, we discuss how this step impacts content retrieval.

The final 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 involves a conversion process, in which static Web pages for each asset are generated and every time-aligned excerpt of text from a spoken content is recast into sections (div elements). Each of these sections contains a hyperlink (the a element) that indicates the exact segment of spoken content where the speech occurs and the transcribed text related to that content (p element) (see Figure 2). Moreover, the language and document type of the content is specified by each static Web page generated by the aforementioned method. 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").

The code in Figure 2 is not enriched with semantic markups, although XHTML supports semantic description. Accordingly, the embedding of a collection of attributes into XHTML markups to enrich the semantics of the Web page content comprises substep

[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

two. Spotlight Web Services analyzes the content transcribed in the previous step, by annotating references to DBpedia resources in the text. Thus, the text is enriched with entity detection and name resolution by using. Figure 4 shows the annotation result. The potential exists in this substep to provide a solution for linking the Linked Open Data (LOD) cloud to unstructured information sources via DBpedia. Thus, the annotated text can be used for secondary tasks, such as recommending related assets based on semantics and displaying additional information about those assets in congruence with its primary use to enhance search and retrieval.

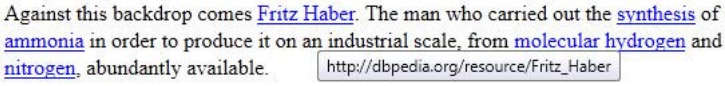

Fig. 4. Result of the XHTML-ize step

4 Experiments

Over the course of our study, real Web data from the educational domain was used to perform an extensive evaluation of our spoken content publishing technique. Our objectives included both a thorough analysis of Web page hits synthesized for the audio and video learning objects, as well as an assessment of the efficacy of the publishing technique.

4.1 Experimental Setup

The tool was evaluated by means of 99 Learning Objects (LO's) comprised of 10 minute video files each containing dialogs on diverse elementary chemistry topics. The experiment was carried out over a period of eight months in two separate stages. Stage one lasted five months and during this time all 99 LO's were published using content-independent metadata. The objective of the first stage was to equalize the Web page hits via content-independent metadata. The information gathered was then used to create two balanced groups of LO's to evaluate the efficacy of the tool. Each LO was then sorted according to the number of hits and further assigned to different groups in pairs who shared the same order of magnitude.

The second stage was a selective process in which one of the groups was submitted to the tool, while the other was not. This stage had a duration of three months. The objective of this second stage was to evaluate the efficacy of the tool. Static Web pages were hosted on the Wordpress server, while the video files were hosted on Youtube. The tool was then assessed using the statistics these services provide.

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 LOs published using the tool in the second stage, and Group $\neg P$ refer to the set of LOs described only by content-independent metadata. This section examines the results of the hit analysis during both experiment stages.

Total Hits Analysis. Approximately 75K hits were obtained during both stages of the experiment. Group A, which represents the first stage with all 99 LO's, obtained just 22% of the total number of hits. It is important to note that in stage one, data was collected for five months, whereas stage two had a three-month duration, thus 78% of the hits were performed in stage two (see Figure 5).

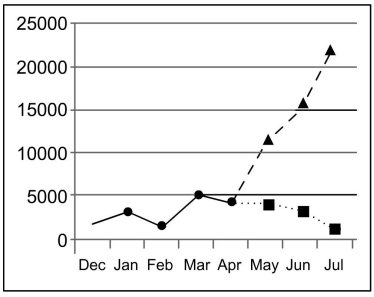

Fig. 5. Hit stats. First stage 1-5. Second stage 5-8

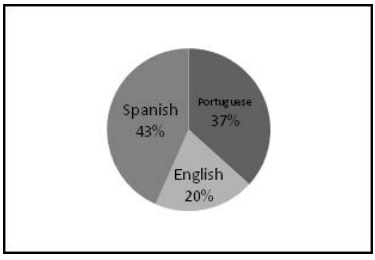

Fig. 6. Hits percentage of translated static Web pages generated by our publishing technique

Observe that Group P captured 66% of the total number of hits whereas Group $\neg P$ just 12%. Hence, Group P captured 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, a new static Web page was created for each new language in the translation step. Throughout the experiment, three language scripts were generated for each asset; English, Spanish and Portuguese. Figure 6 provides the percentage of total number of hits for each translated static Web page: 43% for pages in Spanish, 37% for pages in Portuguese and 20% for pages in English.

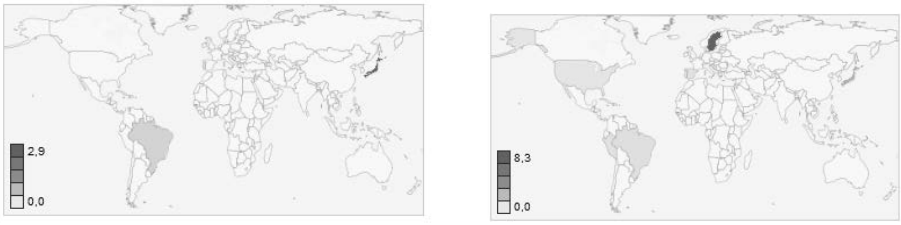

Fig. 7. Countries that have interacted with the content, on the left: Japan, Brazil and Portugal. On the right: Sweden, Japan, Brazil, Spain, Peru, USA and Portugal.

Although static Web page visitors are anonymous, that is, no information pertaining to his/her location or mother language is provided, this information is available for logged in users on Youtube.

The information attained from Youtube is highly relevant to our study because all actions (share, comment or mark as favorite) are executed by users that share in interest in the content of an asset. According to this information and as depicted in the left image of Figure 7, only users from Brazil, Portugal and Japan shared, commented or marked as favorites the assets of Group A.

The right image of Figure 7 reveals that after the technique had been applied creating Group P, users from other countries (Sweden, Japan, Brazil, Spain, Peru, United States and Portugal) could be reached. Note that, the native language of Brazil, Spain, Peru, United States and Portugal is indeed English, Spanish or Portuguese, the languages implemented during experimentation with our technique. This does not extend, however to the populations of Japan and Sweden, although there is a sizable population of Brazilians living in Japan.

Table 1. Bottom 10 most accessed Learning Objects

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

Assets Hits Analysis. At this stage of the experiment, the variance between the number of hits of an LO that was published using the tool and an LO described by content-independent metadata was addressed. Table 1 depicts 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 illustrates, as expected, that the number of hits for LO's indexed by the tool (Group P) is notably greater than the number of hits for LO's described only by content-independent metadata (Group $\neg P$). This is generally true for all 99 LO's. Table 2 depicts the LO's with the lowest number of hits, and shows an even greater discrepancy.

Table 2. 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

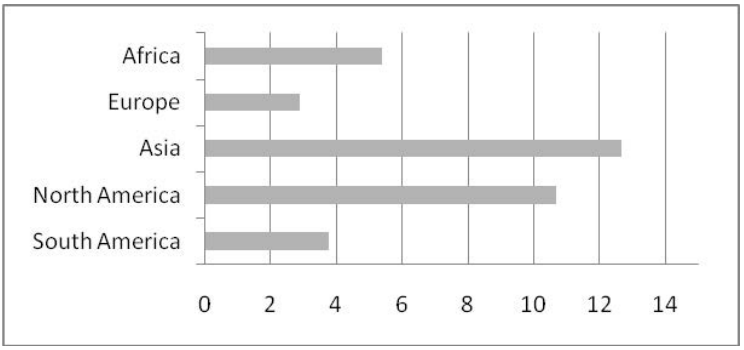

Fig. 8. The number of hits boosted in different orders of magnitude for each continent

Regional Analysis. Using the information gathered from Youtube pertaining to user location, it was possible to tabulate the number of hits by continent (with the exception of Oceania). In the stage one of 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, 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 captured 52,366 hits from South America, 3,738 hits from Europe, 779 hits from North America, 494 hits from Asia, and 167 hits from Africa. We again highlight that the vast majority of hits from South America came from Brazil, and note that the population of Brazil comprises almost 50% of that of South America. It is important to note, however that during a brief period the number of hits from other South American countries increased from 1% to 5.5%. This increase largely took place in Spanish speaking countries. Similarly, the number of hits obtained from European users was also less concentrated: Portugal, which captured 85% of the hits in the first stage of the experiment, had 67% in the second stage, whereas the total number of hits from other European countries more than doubled from 15% to 33%. Figure 8 depicts 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 presented a technique which automatically enhances spoken content on the Web using semantic descriptions (transcripts) and translations. This technique facilitates indexing and retrieval of the objects with the aid of traditional text search engines. Our techniques provided us the basis for an online tool which was used to complete the evaluations demonstrated in this paper.

The tool functions by automatically generating static Web pages which describe the spoken content, which are organized to facilitate locating segments of the content corresponding to the descriptions. The tool further annotates the described spoken content using RDFa and DBpedia to link unstructured information sources to the LOD cloud and enhances search and information retrieval of the assets. The tool also provides a means of amplifying user range by breaking down language barriers through the creation of a multilingual resource. Evaluation proves that the number of hits to the objects processed by the tool was significantly improved, as well the access of consumers of foreign languages.

Future work will investigate several extensions to the tool. First, the tool may resort to semantic information to display complementary information about an asset. Second, the semantics will be enriched to function not only with DBpedia resources, but also with other LOD data sources. Finally, we will recommend related assets by taking advantage of the connected text through ontologies from the LOD and to assess its effectiveness.

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

Acknowledgement. This work has been partially supported by CAPES (Process n^o 9404-11-2). Additional thanks to Chelsea Candra Schmid for her cooperation.

References

1. Alberti, C., Bacchiani, M., Bezman, A., Chelba, C., Drofa, A., Liao, H., Moreno, P., Power, T., Sahuguet, A., Shugrina, M., Siohan, O.: An audio indexing system for election video material. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 4873–4876. IEEE Computer Society, Washington, DC (2009)
2. Baidu search engine, `http://www.baidu.com`
3. Brezeale, D., Cook, D.: Automatic video classification: A survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38(3), 416–430 (2008)
4. Dublin core metadata initiative, `http://www.dublincore.org`
5. Glass, J., Hazen, T.J., Cyphers, S., Malioutov, I., Huynh, D., Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project. In: Proc. Interspeech (2007)
6. Haslhofer, B., Momeni, E., Gay, M., Simon, R.: Augmenting europeana content with linked data resources. In: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS 2010, pp. 40:1–40:3. ACM, New York (2010)

7. Jiang, L., Wu, Z., Zheng, Q., Liu, J.: 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, WI-IAT 2009, vol. 01, pp. 572–575. IEEE Computer Society, Washington, DC (2009)
8. Larson, M., Soleymani, M., Serdyukov, P., Rudinac, S., Wartena, C., Murdock, V., Friedland, G., Ordelman, R., Jones, G.J.F.: Automatic tagging and geotagging in video collections and communities. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR 2011, pp. 51:1–51:8. ACM, New York (2011)
9. Madhavan, J., Afanasiev, L., Antova, L., Halevy, A.: Harnessing the Deep Web: Present and Future. In: 4th Biennial Conference on Innovative Data Systems Research (CIDR) (January 2009)
10. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep web crawl. Proc. VLDB Endow. 1, 1241–1252 (2008)
11. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Ghidini, C., Ngomo, A.-C.N., Lindstaedt, S.N., Pellegrini, T. (eds.) I-SEMANTICS. ACM International Conference Proceeding Series, pp. 1–8. ACM (2011)
12. Nexiwave – speech indexing, http://www.nexiwave.com
13. Nuance – dragon naturallyspeaking, http://www.nuance.com
14. Repp, S., Meinel, C.: Automatic extraction of semantic descriptions from the lecturer's speech. In: IEEE International Conference on Semantic Computing, ICSC 2009, pp. 513–520 (September 2009)
15. Truveo video search, http://www.truveo.com
16. W3c – rdfa primer, http://www.w3.org/TR/xhtml-rdfa-primer
17. Youtube – broadcast yourself, http://www.youtube.com