Information Processing and Management 49 (2013) 1281–1300

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Investigating the document structure as a source of evidence for multimedia fragment retrieval

Mouna Torjmen-Khemakhem [a], Karen Pinel-Sauvagnat [b,*], Mohand Boughanem [b]

[a] ReDCAD, DGIMA, National School of Engineering of Sfax, Tunisia
[b] SIG, IRIT, University of Toulouse, France

ARTICLE INFO

Article history:
Received 2 May 2012
Received in revised form 28 May 2013
Accepted 3 June 2013
Available online 18 July 2013

Keywords:
XML document
Multimedia fragment
Context-based image retrieval
Hierarchical structure

ABSTRACT

Multimedia objects can be retrieved using their context that can be for instance the text surrounding them in documents. This text may be either near or far from the searched objects. Our goal in this paper is to study the impact, in term of effectiveness, of text position relatively to searched objects. The multimedia objects we consider are described in structured documents such as XML ones. The document structure is therefore exploited to provide this text position in documents. Although structural information has been shown to be an effective source of evidence in textual information retrieval, only a few works investigated its interest in multimedia retrieval. More precisely, the task we are interested in this paper is to retrieve multimedia fragments (i.e. XML elements having at least one multimedia object). Our general approach is built on two steps: we first retrieve XML *elements* containing multimedia objects, and we then explore the surrounding information to retrieve relevant *multimedia fragments*. In both cases, we study the impact of the surrounding information using the documents structure.

Our work is carried out on images, but it can be extended to any other media, since the physical content of multimedia objects is not used. We conducted several experiments in the context of the Multimedia track of the INEX evaluation campaign. Results showed that structural evidences are of high interest to tune the importance of textual context for multimedia retrieval. Moreover, the proposed approach outperforms state of the art approaches.

1. Introduction

Multimedia Information Retrieval (MIR) aims at retrieving multimedia contents such as images, videos or audio objects, in response to a user information need. Two classes of approaches were developed in literature. *Content-based approaches* exploit the physical content of multimedia objects such as the color and the texture in image retrieval, or the pitch and the timbre in audio retrieval. The second class of approaches, called *context-based*, extract information around multimedia objects, which is then used to represent the objects. In this case, the physical content of multimedia objects is not exploited at all, and objects can be retrieved independently of the media type. Contextual information can be for example the text surrounding the multimedia object or the associated document title (one can for instance cite approaches of Gong, Hou, & Cheang (2006) or Noah, Azilawati, Sembok, & Meriam (2008) for image retrieval Müller, Kurth, Damm, Fremerey, & Clausen (2007) for audio retrieval or Volkmer & Natsev (2006) for video retrieval). Other contextual information such as

* Corresponding author. Tel.: +33 5 61 55 63 22.
E-mail addresses: torjmen.mouna@redcad.org (M. Torjmen-Khemakhem), Karen.Sauvagnat@irit.fr (K. Pinel-Sauvagnat), Mohand.Boughanem@irit.fr (M. Boughanem).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.06.001

hyperlinks or semantic resources is also considered in Dunlop and Rijsbergen (1993) and Popescu, Grefenstette, and Moëllic (2008).

The basic assumption of approaches exploiting the text surrounding the multimedia object is that this text is included to describe the multimedia objects. Therefore it may contribute to evaluate the relevance of these objects with respect to a query. Our aim in this paper is to study the impact of text proximity in the relevance of search objects. We exploit structural information as a contextual source for multimedia retrieval. Indeed, although structural information is now extensively used in documents, only a few studies exploited the document structure to tune the importance of the different textual parts surrounding the multimedia objects.

XML documents are natural candidates for our study. Indeed, XML (*eXtended Markup Language*) is the most common language used to structure documents. This encoding standard can be used either to annotate and describe multimedia objects (as for MPEG, SVG, or SMIL formats), or to hierarchically organize documents content (text and images, videos, etc.). In the first case, all documents share the same standard structure defined by the format specification whereas in the second one, structure is heterogeneous across the different collections of documents. In this paper, we focus on this latter type of structured documents, where textual content can be easily understood by human readers.

In the particular context of XML multimedia retrieval and as defined in Westerveld and Zwol (2006) and Tsikrika and Westerveld (2008), two types of results can be returned to users queries: *multimedia elements*, i.e. the multimedia objects themselves (images for example) or *multimedia fragments*, which are composed of multimedia objects and associated text. They can be considered as document parts containing at least one multimedia object.

The main issue in multimedia element retrieval is the evaluation of the relevance of multimedia objects using contextual information composed of structure and associated text. In multimedia fragment retrieval, in addition to the object relevance, the challenge is to identify and select the most relevant multimedia fragments to be returned to the user. The resulting fragments should have an appropriate granularity, they can be composed either of the multimedia object itself, or of both text and multimedia objects.

In this paper, we focus on multimedia fragment retrieval. The approach we propose is based on two steps, first we retrieve multimedia elements and then we explore the surrounding information to retrieve relevant multimedia fragments. For both steps, we will study the impact of text proximity for relevance evaluation thanks to the underlying structural information.

Although our multimedia retrieval approach is applicable to any media type as it is only based on the multimedia object context and not on its content, we chose to illustrate and evaluate it on images for two reasons: first, the image is the the most used and easiest media (other than text) to integrate into digital documents, and secondly, to the best of our knowledge, existing collections to evaluate the use of document structure in multimedia retrieval only contain images (e.g. INEX Multimedia[1] and CLEFImage[2]).

The rest of the paper is structured as follows. We first discuss related work in Section 2 and describe our approach in Section 3. Section 4 presents evaluation and results, and our approach is compared to the state-of-the-art approaches in Section 5. Results and future works are discussed in Section 6.

2. Related work

We review in this section existing approaches for context-based multimedia retrieval, more precisely for context-based image retrieval, where queries are expressed using keywords (text) and the images annotated (indexed) by keywords provided manually or built automatically. We then focus in the second part of the section on approaches using also structural context to index images.

2.1. Using textual context

A first way to index images is to manually or automatically annotate them by concepts provided by the user and/or derived from semantic resources (Akbas & Yarman-Vural, 2007; Fan & Li, 2006; Hliaoutakis, Varelas, Voutsakis, Petrakis, & Milios, 2006; Piotrowski, 2009; Popescu et al., 2008).

Other approaches state that there is a strong correlation between an image and its surrounding text in the document. Therefore images search is often carried out using the textual content of the image name (and sometimes its extension) or using the associated text of the image, extracted from the document. In web collections for instance (HTML pages), the associated text of the image is generally extracted from `src` and `alt` tags (Shen, Ooi, & Tan, 2000), title of the web page, other particular tags (Noah et al., 2008), or also from text close to the image (Chen, Liu, Zhang, Li, & Zhang, 2001; Guglielmo & Rowe, 1996; LaCascia, Sethi, & Sclaroff, 1998; Gong et al., 2006; Srihari, Zhang, Rao, Baird, & Chen, 2000). Many images search engines on the Web (as Google[3] and Lycos[4]) use such methods.

The context of images can also be enlarged to other documents, thanks for example to (hyper) links (Chakrabarti et al., 1998; Chakrabarti, Punera, & Subramanyam, 2002; Dunlop, 1991; Dunlop & Rijsbergen, 1993; Haveliwala, Gionis, Klein, &

[1] INEX: Initiative for the Evaluation of XML Retrieval, multimedia track.
[2] CLEFImage: Cross-Language Evaluation Forum, Image Track.
[3] http://www.google.com.
[4] http://www.lycos.fr.