Expert Systems With Applications 69 (2017) 201–213

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Combining heterogeneous sources in an interactive multimedia content retrieval model

Julián Moreno-Schneider[a,*], Paloma Martínez[b], José L. Martínez-Fernández[c]

[a] Deutches Forschungzentrum für Künstliches Intelligenz - DFKI, Alt-Moabit, 91c, 10559, Berlin, Germany
[b] Computer Science Department, Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911, Leganés, Madrid, Spain
[c] MeaningCloud LLC, USA

ARTICLE INFO

Article history:
Received 7 April 2016
Revised 20 October 2016
Accepted 20 October 2016
Available online 21 October 2016

Keywords:
Multimodal information retrieval
User adaptation
Retrieval engines
Rule-based expert systems

ABSTRACT

Interactive multimodal information retrieval systems (IMIR) increase the capabilities of traditional search systems, by adding the ability to retrieve information of different types (modes) and from different sources. This article describes a formal model for interactive multimodal information retrieval. This model includes formal and widespread definitions of each component of an IMIR system. A use case that focuses on information retrieval regarding sports validates the model, by developing a prototype that implements a subset of the features of the model. Adaptive techniques applied to the retrieval functionality of IMIR systems have been defined by analysing past interactions using decision trees, neural networks, and clustering techniques. This model includes a strategy for selecting sources and combining the results obtained from every source. After modifying the strategy of the prototype for selecting sources, the system is re-evaluated using classification techniques. This evaluation compares the normalised discounted cumulative gain (NDCG) measure obtained using two different approaches: the multimodal system using a baseline strategy based on predefined rules as a source selection strategy, and the same multimodal system with the functionality adapted by past user interactions. In the adapted system, a final value of 81,54% was obtained for the NDCG.

1. Introduction

Present day society is characterised by a constant technological revolution, where the generation and consumption of information is attaining huge levels. The amount content on the internet, the main container of information, is increasing exponentially. There are plenty of services that offer multimedia content. Among well-known examples, Google (www.google.com) specialises in text content, YouTube (www.youtube.com) provides searches for videos, SoundCloud (soundcloud.com) facilitates music sharing, and Flicker (www.flickr.com) allows users to publish and search for photos. When dealing with multimedia, such systems are mainly based on textual metadata.

When dealing with audio, images, or videos, commercial systems are mainly based on the characterisation of resources in terms of textual metadata, which is later compared against user query expressions. Some examples of metadata for documents are *'author'*, *'date of creation'*, *'title'*, and *'language'*.

Thus, retrieval methods must evolve to become dependent on the device used to query (PC, smartphone, tablet, etc.), what is being queried, and who is querying. Furthermore, advances in the devices available to users are leading to a change in the formats applied in the definition of queries. Google has introduced voice query, where users can interact with the search engine by using a microphone to formulate a query, and previously they have included queries using images to search for other similar images.

The nature of internet access is also changing. The use of smartphones has exceeded the use of traditional computers for browsing the internet, but other devices are also becoming popular, such as smartwatches (9%), smart televisions (34%), games consoles (37%), smart wristbands (7%), and tablets (47%).[1]

The main problem presented by this growing presence of multimedia content is that users need to access larger and larger quantities of information in different formats and sources, and

* Corresponding author.
E-mail addresses: julian.moreno_schneider@dfki.de (J. Moreno-Schneider), pmf@inf.uc3m.es (P. Martínez), jmartinez@meaningcloud.com (J.L. Martínez-Fernández).

[1] Data extracted from http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/ accessed at 16/07/2016.

http://dx.doi.org/10.1016/j.eswa.2016.10.049
0957-4174/

they wish to do so in a faster and easier manner, without having to query several sources.

If we consider a scenario where a journalist (say a sports editor for television) has to prepare news regarding F1, the journalist must cover information from all F1 races, travelling to all F1 Grand Prix locations for live broadcasts. They must document and archive all audiovisual material captured in a race. At the same time, they must develop additional pieces of information related to the last race. Retrieving these pieces of information is a difficult task, which can be simplified by using a multimodal retrieval system.

This retrieval must be simple, quick, and transparent to the user. Web search engines are the most well-known retrieval systems, but these do not allow the mixing of formats in queries. That is, a query composed of a combination of text and image cannot be performed.

When dealing with a technology that can query several retrieval engines, two problems arise for each engine that is considered. Namely, when this engine should be queried, and how its results are processed. Most techniques rely on the mode of the query to select an engine, and a simple mixture to present the final list as a combination. Therefore, they do not really adapt to different environments or queries. Furthermore, new techniques may have to be developed if we would like to work with several multimodal engines.

The main goal of this study is to adapt the functionality of an *interactive multimodal information retrieval (IM-IR)* system based on past user behaviour. In particular, the aim is to exploit past interactions of an IMIR system through the use of classification algorithms, in order to avoid the need for expert-defined rules. We attempt to employ semi-supervised machine learning-type decision trees and neural networks. To accomplish this goal, we must fulfil two preliminary tasks. First, we define a multimodal information retrieval model that queries multiple heterogeneous sources, emphasising which sources are queried and how the results are combined. Second, we implement a working IMIR prototype based on this model. This implementation will later be adapted to take into account past user interactions.

The remainder of this article is organised as follows. Section 2 reviews work related to multimodal information retrieval and expert systems. Section 3 describes the defined formal model. The implementation of a basic prototype based on this model is described in Section 4. The functionality adaptation techniques of the IMIR prototype based on past user interactions are presented in Section 5, and evaluated in Section 6. Finally, conclusions and directions for future research are presented in Section 7.

2. Related work

In this section, the main components of an IMIR system are described, considering the perspectives of the repositories that are queried, the manner in which the user may formulate queries, the underlying information retrieval (IR) models, the combination of retrieval engines required to answer user queries, and finally how the results are merged to obtain a list to be displayed to the user.

2.1. Multimedia information

Information collections are divided according to the modes of the objects that compose them. A monomodal collection contains items from a single mode, such as the Wikipedia dataset used in Hong and Si (2012). By contrast, a multimodal collection contains objects from different modes, such as in the work of Yilmaz, Gulen, Yazici, and Kitsuregawa (2012), which manages video and text; or the work of Camargo and González (2016), which employs two data sets of images, Flickr4Concepts and MIRFlickr (Huiskes & Lew, 2008).

Furthermore, there is a special case in this division: monomodal collections containing multimedia objects accompanied by metadata, such as the ImageCLEF 2011 Medical Retrieval Task dataset (Kalpathy-Cramer et al., 2011), which encompasses images and metadata. Finally, it is interesting to mention two completely multimodal collections. The work of Jou, Li, Ellis, Morozoff-Abegauz, and Chang (2013) employs a multimodal collection composed of 18,000 h of broadcast news, 3.58 million of new articles, and 430 million Twitter messages; and the TREC Federated Web Search (FedWeb) Track 2013 Forum (Demeester, Trieschnigg, Nguyen, & Hiemstra, 2013) offers a multimodal collection composed of results obtained from 157 real web search engines, divided into 24 categories (ranging from news, academic articles, and images to jokes and lyrics). The collection contains both the search result snippets (1,973,591) and the pages (1,894,463) that the search results link to (that is, the HTML of the corresponding web pages).

2.2. Representation of information needs: query

In most cases, queries arise in the textual mode, such as on commercial internet search engines (Yahoo, Bing, Google, etc.). For further details, see the work of Sushmita (2012). Some studies have used multimedia elements as queries, such as image queries (Wong, Cheung, & Po, 2005), voice queries (Hauptmann, Jin, & Ng, 2002), and short videos (Yang, Cai, Hanjalic, Hua, & Li, 2012). Some researchers have studied multimodal query representation using specific languages, such as rich unified content description (Daras et al., 2011). These types of languages are interesting, because they offer the capability of representing every multimedia element in a query. In our work, we include a formal representation in the model definition.

2.3. Retrieval techniques

Considering that retrieval techniques are not the focus of this study, only a brief introduction is provided, in order to provide some context to the reader. For a complete review of IR techniques, see Baeza-Yates and Ribeiro-Neto (2011) and Manning, Raghavan, and Schütze (2008). Retrieval through the matching of a text query and document content (keywords) is the most commonly employed method, such as in Hong and Si (2012) or Görg et al. (2010). One study that investigates image retrieval based on low-level features is Romberg, Lienhart, and Hörster (2012). There exist other studies that have used metadata to perform the retrieval of multimedia elements, such as Hauptmann et al. (2002) for retrieving videos or Lana-Serrano, Villena-Román, and Cristóbal (2011) and Benavent, García-Serrano, Granados, Benavent, and de Ves (2013) for retrieving images.

Available internet search engines (such as Yahoo or Bing) have also been used as *retrieval engines* in some studies, such as Sushmita (2012). Another interesting work is Torres (2005), which defines the visual object information retrieval (VOIR) prototype, combining two layers (conceptual and feature-based) to perform retrieval.

Multimedia retrieval based on annotations, relevance feedback, and concepts is similar to a metadata based search. Documents are retrieved based on the similarity of the document and query annotations. Some methods employing this approach are the Mediamill system (Worring, Snoek, de Rooij, Nguyen, & Smeulders, 2007) and the ESCRIRE project or EsCosServer architecture (Medina-Ramírez, 2007).

Other types of multimodal retrieval systems create combined or centralised indexes, containing all modes of documents, such as Marchand-Maillet et al. (2011).