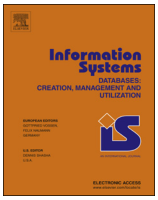
Contents lists available at ScienceDirect

Information Systems

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



Information system for image classification based on frequency curve proximity

L. Sánchez [a], Javier Alfonso-Cendón [a,*], Tiago Oliveira [b], Joaquín B. Ordieres-Meré [c], Manuel Castejón Limas [a], Paulo Novais [b]

[a] *University of León, Leon, Spain*
[b] *University of Minho, Braga, Portugal*
[c] *Polytechnic University of Madrid, Madrid, Spain*

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form
29 August 2016
Accepted 31 August 2016

Keywords:
Information system
Similarity search
Frequent itemset mining
Metadata
Image classification

ABSTRACT

With the size digital collections are currently reaching, retrieving the best match of a document from large collections by comparing hundreds of tags is a task that involves considerable algorithm complexity, even more so if the number of tags in the collection is not fixed. For these cases, similarity search appears to be the best retrieval method, but there is a lack of techniques suited for these conditions. This work presents a combination of machine learning algorithms put together to find the most similar object of a given one in a set of pre-processed objects based only on their metadata tags. The algorithm represents objects as character frequency curves and is capable of finding relationships between objects without an apparent association. It can also be parallelized using MapReduce strategies to perform the search. This method can be applied to a wide variety of documents with metadata tags. The case-study used in this work to demonstrate the similarity search technique is that of a collection of image objects in JavaScript Object Notation (JSON) containing metadata tags.

1. Introduction

Due to the current diversity and availability of image capturing devices, such as digital cameras, digital scanners and smartphones, and the use of the Internet to disseminate content, the size of digital image collections is continuously increasing. The predictions from a technical report from the International Data Corporation [1] point towards a growth of digital content from 130 exabytes to 40,000 exabytes, between 2005 and 2020. This implies a heavy investment in Information Technology hardware, software, services, telecommunications, and staff, in short, of all the components that make up the infrastructure of the digital universe. Most of this information is produced by average consumers in their interaction with social media, by sending camera phone images and videos between devices and around the Internet, and so on [2]. While the information holding potential analytical value is growing at an unbelievable rate, only a small fraction of this information has been explored. The effective management of these collections has become a necessity for both companies and the general public.

Classical database management systems (DBMSs) are designed to handle data objects that have a pre-established structure. Normally, this structure is acquired by treating every feature of a data object as an independent dimension, and then building representations in the

[*] Correspondence to: Dpto. Ingenierías Mecánica, Informática y Aeroespacial, Escuela de Ingenierías Industrial e Informática, Universidad de León, 24071 León, Spain.

E-mail addresses: lidia.sanchez@unileon.es (L. Sánchez), javier.alfonso@unileon.es (J. Alfonso-Cendón), toliveira@di.uminho.pt (T. Oliveira), j.ordieres@upm.es (J.B. Ordieres-Meré), manuel.castejon@unileon.es (M.C. Limas), pjon@di.uminho.pt (P. Novais).

http://dx.doi.org/10.1016/j.is.2016.08.001
0306-4379/© 2016 Elsevier Ltd. All rights reserved.

form of records. These records are then stored according to a certain database model which can be relational, object-oriented, object-relational, hierarchical, etc. However, these models require that the data objects have a fixed, and typically reduced, number of features because queries are usually performed by exact matching, partial matching, and joining applied to some of the features. Yet, there are applications that demand the use of data with a simplified structure, and, thus, less organized and precise [3]. The problem with this type of data is that it is nearly impossible to order it and it is not meaningful to perform equality comparisons on it. For these cases, proximity, or similarity, is a more suitable search criterion. Similarity search is a central component to content-based retrieval in multimedia database systems. It is a general term that includes a wide range of techniques whose main goal is normally one of the following [4,5]: (1) to find objects whose feature values fall within a range of distance, using a defined metric, from a query object (range queries); (2) to find a certain number of objects whose features are the closest to an object query (nearest neighbor queries); and (3) to find pairs of objects within the same set which are similar to each other.

As such, efficient search and retrieval mechanisms are a basic need in systems that deal with these collections in a wide variety of domain applications. Photography, fashion, crime prevention, architecture, publishing, journalism and academic research itself are only a few examples of domains where image search systems are necessary. However, going through large collections of documents is a hazardous task and involves the use of expensive computational resources. There is a clear need for an object search method that is quick, lightweight, and easy to apply to large item collections.

Metadata is normally referred to as *data about data*. It provides additional information that supplements the content of images. As such, it has become a powerful mechanism to search through the content of image libraries and other digital media such as audio and video [6]. Using metadata is considered advantageous because it is still impractical, namely in the field of digital photography, to organize and query images based on millions of image pixels. Considering this, it is preferable to use metadata properties describing what the picture represents and details (where, when and how) of its capture.

The premise of this work is that the structural and descriptive metadata of an image can provide useful cues, independent of the captured scene content, for image retrieval and matching. To test this hypothesis, one developed an algorithm that constructs characteristic curves of image objects by analyzing all the metadata tags in a document. Using these curves, the algorithm can perform fast searches in the document database and retrieve a list of images sorted by proximity to a given one. The advantage of the algorithm lies in being possible to group similar objects in order to determine if different objects have the same origin. This kind of relationship may be a great advantage in order to know more about the history of an image, to know if it has been modified or tampered with. The setting used to test the algorithm includes a collection of JavaScript Object Notation (JSON) objects containing the metadata tags of images in multiple formats. JSON is an emerging data transfer format and it is used as an access method in many NoSQL database [7], which are an example of the systems that house the simplified and less structured data mentioned earlier. NoSQL provides horizontal scaling and, thus, in particular conditions enables a faster retrieval. The computation can be divided in concurrent tasks across distributed machines. To achieve this, these systems have to relax some of the characteristics of traditional DBMSs, one of which is data structure. At the same time, this is also a desirable feature for certain data types, such as images, which come in multiple formats, each one of them with different tags. The algorithm was implemented using Go [8], a programming language developed by Google that provides more facilities for the implementation of concurrency and parallelism in order to get the most out of multicore and networked machines.

The paper is organized as follows. Section two provides related work in the fields of similarity search, itemset mining, and image metadata. Section three is considered a materials and methods section which has a description of the technique and search strategy, of how the frequency curves for the documents are constructed, and of how to perform a search query using the developed algorithm. In this section, there are also results that demonstrate their effectiveness. Section three features a discussion where the main strengths and limitations of the approach are highlighted. Finally, in section five conclusions are drawn about the main contributions of the work.

2. Related work

This section provides information on the three main topics of this work: similarity search, frequent itemset mining (FIM), and image metadata. Given the vastness of the work developed in similarity search, only the aspects and approaches that bear a resemblance or can offer a good counterpoint to the approach followed herein will be mentioned. Central to this work is also discovering which features from a given set in an object collection are the most important for conducting similarity search queries, thus the inclusion of FIM in the topics of interest. The section ends with a description of what metadata is, its purposes and its issues.

2.1. Similarity search

Similarity search has established itself as one of the fundamental paradigms in modern applications. This is an important task when trying to find patterns in applications, involving the exploration of data such as images, videos, time series, text documents, and so forth.

In essence, it consists in a problem of finding, within a set of objects, those which are more similar to a given query object. Normally, data collections are treated as metric objects, which brings significant advantages because many data classes and information-seeking strategies conform to the metric view. There are four fundamental aspects of similarity search: the distance measure,

Please cite this article as: L. Sánchez, et al., Information system for image classification based on frequency curve proximity, Information Systems (2016), http://dx.doi.org/10.1016/j.is.2016.08.001