Abstract With the commoditization of digital devices, personal information and media sharing is becoming a key application on the pervasive Web. In such a context, data annotation rather than data production is the main bottleneck. Metadata scarcity represents a major obstacle preventing efficient information processing in large and heterogeneous communities. However, social communities also open the door to new possibilities for addressing local metadata scarcity by taking advantage of global collections of resources. We propose to tackle the lack of metadata in large-scale distributed systems through a collaborative process leveraging on both content and metadata. We develop a community-based and self-organizing system called PicShark in which information entropy—in terms of missing metadata—is gradually alleviated through decentralized instance and schema matching. Our approach focuses on semi-structured metadata and confines computationally expensive operations to the edge of the network, while keeping distributed operations as simple as possible to ensure scalability. PicShark builds on structured Peer-to-Peer networks for distributed look-up operations, but extends the application of self-organization principles to the propagation of metadata and the creation of schema mappings. We demonstrate the practical applicability of our method in an image sharing scenario and provide experimental evidences illustrating the validity of our approach.

Keywords Metadata scarcity · Metadata heterogeneity · Metadata entropy · Peer-to-Peer collaboration · Peer data management

1 Introduction

Until recently, the creation of digital artifacts—such as electronic documents, images, or videos—was constrained by the limited availability of devices capable of capturing and handling information in binary form. Today, the situation has radically changed with the commoditization of digital devices. Typewriters have now totally disappeared from the office space, whereas email has become one of the main communication channels. Mobile phones can handle information written as bidimensional bar-codes, while personal computers casually store and process gigabytes of personal images. In this new context, we argue that the lack of metadata, rather than the lack of data, has become the main bottleneck.

The problem became apparent a few years ago when end-users suddenly had to resort to third-party tools to find relevant pieces of information on their own computer. At that time, several projects proposed to index information based on metadata to enhance the search process. Microsoft’s Stuff I’ve Seen [14], for instance, relies on time-stamp metadata like Last Time Modified or Last Time Opened to display search results, while Google Desktop[^1] indexes documents based

The lack of metadata resurfaces today as a new problem in distributed settings. More and more platforms allow end-users to share their digital content in large communities: Flickr, YouTube, and MySpace are well-known examples of that trend. In distributed environments, however, automatically generated metadata such as *Last Time Opened*, *Filename*, or *Size* often cannot be exploited in a meaningful way by arbitrary users searching for a specific file. In a distributed setting, users typically have never encountered the file they are searching for and are thus unaware of its technical details. Higher-level, more meaningful metadata like *Description*, *Event*, or *Location* are much more relevant in distributed environments, but still often require human attention, one of the scarcest resources in our digital society. As a result, the majority of digital content on current collaborative platforms simply cannot be retrieved by third-parties because of the lack of adequate metadata.

In the following, we tackle the problem of metadata scarcity in large-scale collaborative environments. We focus on semi-structured metadata formats and propose a radically new approach to foster global search capabilities from incomplete, local, and heterogeneous metadata. Our approach is based on a new metric for metadata scarcity and on Peer-to-Peer (P2P) interactions. The main contributions of our work are:

- the formalization of the problem of sharing semi-structured metadata in distributed settings, explicitly taking into account metadata incompleteness and metadata heterogeneity
- the definition of a new metric—called *metadata entropy*—to capture the degree of incompleteness or uncertainty related to semi-structured metadata
- the description of a bottom-up and recursive process based on instance and schema matching to infer metadata in collaborative P2P contexts
- the presentation of a system architecture supporting metadata inference in distributed environments
- the experimental evaluation of our metadata inference process on a large set of several hundreds of annotated images.

We start with a general description of the problems related to the sharing of semi-structured metadata in Sect. 2 and formalize our problem in Sect. 3. Our metadata sharing approach is presented in detail in Sect. 4. We describe the architecture of our prototype and the results of our experimental evaluation in Sect. 5. Finally, we give a survey of related work in Sect. 6 before presenting our conclusions.

2 Sharing semi-structured metadata

2.1 On semi-structured metadata

While the use of unstructured metadata drew considerable attention on the Web in recent years—e.g., through keyword annotation of images or HTML pages—the focus recently shifted back to more structured metadata formats. Unstructured metadata such as tags are ambiguous by nature and lack precise semantics, making it very difficult to support structured searches à la SQL. Structured representations such as relational tables are much easier to process automatically, as they constrain the representation of data through complex data structures and schemas.

In the following, we focus on recent formats that let end-users freely define and extend their own schemas according to their needs. We qualify those formats as semi-structured formats since they tend to blur the separation between the data and schemas and to impose looser constraints than the relational model to the data. Such formats are today sprouting from various contexts and encompass a large variety of data models. The Extensible Markup Language (XML) [6], for example, relies on hierarchies of elements to organize data or metadata. Ontological metadata tie metadata to formal descriptions where classes of resources (and properties) are defined and interrelated. This class of metadata standards is currently drawing a lot of attention with the advent of the Semantic Web and its associated languages (e.g., RDF/S [24], Adobe’s XMP2 or OWL [25]). Semi-structured formats are gaining momentum. They are flexible enough to allow easy definition and extension of schemas, while sufficiently structured to support automated processing and complex searches (e.g., through languages such as XQuery [5] or SPARQL [27]).

2.2 On the difficulty of sharing semi-structured metadata

Our goal is to enable global search capabilities for shared resources based on semi-structured metadata in large-scale, heterogeneous and distributed settings. Although semi-structured metadata formats are getting increasingly popular, support for meaningfully sharing semi-structured metadata outside of their original context or community of interest is often lacking. Semi-structured metadata are intrinsically difficult to share, since their values only make sense in a given context—as opposed to keyword metadata or textual tags, which supposedly convey predefined, global semantics. Hence, large-scale collaborative applications typically disregard semi-structured metadata or treat them as simple unstructured keywords ignoring their intrinsic structure.