

The World-Wide Web is a sky survey: The orbit of Comet Holmes

Dustin Lang^{1,2,3}, David W. Hogg^{4,5}, and others

ABSTRACT

We performed a commercial image search on the Web with the search term “Comet Holmes”. Many of the images obtained by this search were calibrated successfully by the automated *Astrometry.net* system. The image centers and sizes form a set of data points to which we fit a test-particle orbit in the Solar System, parameterizing, fitting, and marginalizing out the distribution of outliers. The marginalized posterior probability distribution function for the orbital parameters includes at high probability the published orbital parameters, and provides for each image an estimated date for the exposure and probability of being an outlier. Comparing the estimated dates to the reported dates in the image EXIF headers (where available), we find that most cameras have accurately set clocks. This project demonstrates that discoveries are possible with data of extreme heterogeneity and unknown provenance. Implications for next-generation data sharing are discussed.

Subject headings: celestial mechanics — comets: individual (Holmes) — ephemerides — methods: statistical — surveys — time

1. Introduction

The Web bristles with billions of images, on web pages, in public photo-sharing sites, on social networks, and in private email and file-sharing conversations. A tiny fraction but

¹Department of Computer Science, University of Toronto, 6 King’s College Road, Toronto, Ontario, M5S 3G4, Canada

²Princeton University Observatory, Princeton, NJ, 08544, USA

³to whom correspondence should be addressed: dstn@astro.princeton.edu

⁴Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY, 10003, USA

⁵Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117, Heidelberg, Germany

enormous number of these images are *astronomical* images—images of the night sky in which astronomical sources are visible. This is true even if we exclude from consideration scientific collections such as those of professional observatories and surveys and only count the images of hobbyists, amateurs, and sight-seers. In principle these images, taken together, contain an enormous amount of information about the astronomical sky. Of course they have no scientifically responsible provenance, have never been “calibrated” in any sense of that word, and were (mainly) taken for purposes that are not at all scientific. But having been generated from CCD-like measurements of the intensity field, they cannot help but contain important scientific information. The Web is, therefore, an enormous and virtually unexploited sky survey.

It is difficult to estimate the total number of astronomical images on the web, and even harder to estimate the total data throughput (*étendu* or equivalent measure of scientific information content). [DSTN: do some relevant searches on *Flickr* and also give some statistics for the “astrometry” group there.]

The technical obstacles to making use of a data collection of this heterogeneity and unreliability are immense: If anything has been learned from our interaction with electronic communication, it is that publisher-supplied or provider-supplied meta-data about Web content are consistently missing, misleading, in error, and obscure. Indeed, when it comes to the astronomical properties of imaging discovered on the Web, most providers don’t even know what we want in terms of “meta-data”; we want calibration parameters relating to image date, astrometric coordinate system, photometric sensitivity, and point-spread function.

Two important changes are occurring in astronomy that are opening up the possibility that we might exploit data collections as radically confusing as that of the entire Web. The first is that tools are beginning to appear that can perform completely hands-free data analysis tasks. The best example so far is the *Astrometry.net* system, which can take astronomical imaging of completely unknown provenance, and calibrate it astrometrically using the data in the image pixels alone (Lang et al 2009).

The second change is that there has been an enormous increase in the amount and diversity of publicly available professional data—that is, calibrated, trustworthy, science-oriented data in observatory, sky-survey, and individual-investigator collections. These collections are so large and diverse that automated data analysis tools that can trivially interact with extremely heterogeneous data are necessary in many scientific domains. That is, much of the technology required for exploitation of the Web-as-sky-survey is required for *any* mature, data-intensive scientific investigation.

We have been exploring some of these ideas with the *Astrometry.net* project. Not only

has the system calibrated thousands of images taken by amateurs and hobbyists, we have interfaced the system to the *Flickr* photo sharing site (Stumm et al, forthcoming). *Flickr* users who add an image from their online collection to a *pool* (group of related images) called “astrometry” find the image calibrated automatically by an unmanned bot that downloads the image, calibrates it with *Astrometry.net*, and then posts to the image’s page on *Flickr* astrometric calibration meta-data, or what have been called “astro-tags”. We make use of the *Flickr* Application Programming Interface, something many image and data-sharing sites employ. The success of this suggests that automated maintenance of a heterogeneous crowd-sourced sky survey might be possible in the future.

In this paper, we explore some of the ideas around the Web-as-sky-survey concept, by performing a scientific investigation of Comet Holmes using Web-discovered images alone. Although we do learn things about Comet Holmes, our main interest is in developing and testing new technologies for observational astrophysics.

2. Data and calibration

Yahoo image search, wget, simplexy, astrometry.net, statistics.

Show table of URLs (short form in journal, long form online). Show example images (with permissions). Show footprint map.

3. Orbit determination

We turn the image footprints into data as follows.

Our attitude towards image time EXIF info is the following.

We take the approach of generative modeling; that is, we evaluate the probability of the data given the model. The images are treated as independent (caveats?) so the probability density for the full data set is given by a product of individual-image probabilities.

$$\mathcal{L} = \prod_i p(\mathbf{I}_i | \boldsymbol{\omega}, p_{\text{bad}}) \quad , \quad (1)$$

where i is an index specifying the image, \mathbf{I}_i are the (modeled) properties of image i (discussed below), $\boldsymbol{\omega}$ is an ordered list of orbital parameters, and p_{bad} is the probability that the image is not relevant to the question at all. Each individual-image probability is approximated by

$$p(\mathbf{I}_i | \boldsymbol{\omega}, p_{\text{bad}}) = \frac{1}{Z_i} \int_{t_{\text{min}}}^{t_{\text{max}}} dt \left[\frac{[1 - p_{\text{bad}}]}{\Omega_i} \Theta(\mathbf{I}_i, \boldsymbol{\omega}, t) + \frac{p_{\text{bad}}}{4\pi} \right] \quad , \quad (2)$$

where Z_i is a normalization constant, the integral is over a prior time interval $[t_{\min}, t_{\max}]$, Ω_i is the solid angle of image i in steradians, and Θ is a function that returns unity when the comet is inside the solid-angular footprint of the image, and zero when it isn't.

This likelihood is approximate; we have made the following simplifying assumptions:

- Independence. How to do better?
- We know little about the position of the Comet inside each image. How to do better: detect.
- We have no time information at all, yet we do.

Orbit generation is performed with more simplifying assumptions:

- All observations are taken from the Earth–Moon barycenter.
- The Earth and the Comet are both on Kepler (1609) orbits.

4. Discussion

Grant numbers.