Big Data New Frontiers: Mining, Search and Management of Massive Repositories of Solar Image Data and Solar Events

Juan M. Banda, Michael A. Schuh, Rafal A. Angryk, Karthik Ganesan Pillai, and Patrick McInerney

Montana State University, Bozeman, MT 59717 USA
{juan.banda,michael.schuh,angryk,k.ganesanpillai,
patrik.mcinerney}@cs.montana.edu

Abstract. This work presents one of the many emerging research domains where big data analysis has become an immediate need to process the massive amounts of data being generated each day: solar physics. While building a content-based image retrieval system for NASA's Solar Dynamics Observatory mission, we have discovered research problems that can be addressed by the use of big data processing techniques and in some cases require the development of novel techniques. With over one terabyte of solar data being generated each day, and ever more missions on the horizon that expect to generate petabytes of data each year, solar physics presents many exciting opportunities. This paper presents the current status of our work with solar image data and events, our shift towards using big data methodologies, and future directions for big data processing in solar physics.

1 Introduction

With the launch of NASAs Solar Dynamics Observatory (SDO) on February 11, 2010, researchers in solar physics entered the era of Big Data. SDO is the first mission of NASA's Living With a Star (LWS) program, a long term project dedicated to studying aspects of the Sun that significantly affect human life, with the goal of eventually developing a scientific understanding sufficient for prediction. Space weather (originating from the Sun) is currently considered to be one of the most serious threats to our communication systems, power grids, and space and air travel [1]. Solar storms can interfere with radio communications and satellites (GPS, etc.), and induce geomagnetic currents in our power and communication grids, oil and gas pipelines, undersea communication lines, telephone networks, and railways. A 2008 U.S. government report prepared for the Federal Emergency Management Agency put the yearly financial impact of a massive solar storm event at more than US $1 trillion (http://bit.ly/14GjFUJ).

In the following subsections we will show how the Big Data four V-dimensions of: Volume, Velocity, Variety, and Veracity directly apply to solar data. We highlight several key points on how these dimensions need to be addressed by adapting and expanding our work using and developing big data methodologies.

B. Catania et al. (eds.), *New Trends in Databases and Information Systems*,
Advances in Intelligent Systems and Computing 241,
DOI: 10.1007/978-3-319-01863-8_17,

1.1 Volume and Velocity

The instruments onboard the Earth-orbiting SDO spacecraft currently generate about 70,000 high resolution images (4096x4096 pixels each) per day (Fig. 2a) (VELOCITY), sending back to Earth about 0.55PB of raster data every year (VOLUME). NSF is already in process of building a new ground-based instrument in Hawaii, called the Advanced Technology Solar Telescope (ATST) which is expected to capture about 1 million images per day (3-5PB of data per year).

Currently, the volumes of near-continuous SDO raster data processed all over the world are generating significant amounts of image and object data, and posing significant data mirroring issues related to the distributed character of these massive data repositories. Moreover, many automated computer vision software modules work continuously on this massive data stream to facilitate space weather monitoring. With ATST the amount of data to be processed will be too extreme to be processed in real-time and considerable sampling will need to take place if the current algorithms are not scaled to the task.

1.2 Variety and Veracity

There is a multitude of diverse data about the Sun coming from different instruments and software modules. Ongoing efforts exist to integrate the data under the Virtual Solar Observatory umbrella (http://bit.ly/18cpyk6). This situation leads to significant data integration challenges, which are of crucial importance for long-term, solar cycle-oriented (each approx. 11 years) research investigations. The VARIETY of solar data can best be described by two examples: 1) Some of the oldest solar data repositories come from space missions in the 1990s, such as Yohkoh Data Archive Center (http://bit.ly/1279tsv), which contains data from a telescope launched by Japan in 1991, and SOHO (http://1.usa.gov/17v2FaD), a joint project between the European Space Agency (ESA) and NASA originated in 1995. 2) There is a wide variety of data compacting and meta-data reporting services such as Helioviewer (http://bit.ly/13imn3i), and the Heliophysics Events Knowledgebase (http://bit.ly/14GkWeA), which provide spatiotemporal data about solar events in vector formats.

Almost all of these resources come from government-funded instruments and/or have data repositories maintained by large companies (e.g. Lockheed Martin) or governmental institutions (e.g. SAO, ESA, NASA). This guarantees high data quality, with certain data standards prototyped over 20 years ago, and assures data VERACITY.

2 Current State of Solar Physics Data Mining

In this section we will cover some of the most important areas of research that the Data Mining Lab at Montana State University (MSU) has been working on over the last several years while closely collaborating with the MSU Solar Physics department, and the Harvard-Smithsonian Center for Astrophysics. Our collaboration started with the objective of building a content-based image retrieval

system (CBIR) for the SDO mission and has developed into new and interesting areas of interdisciplinary research between the two fields. We will outline our three main contributions to the field and mention some of the initial challenges.

2.1 SDO Data Pipeline Details

In Figure 1 we present a high-level overview of the SDO pipeline and the main components that relate to our research purposes, for a more detailed and in-depth discussion of SDO and its data flow, please see [13]. SDO is currently on a geosynchronous orbit with a continuous dual-band data downlink to the ground station in New Mexico. The station's Data Distribution System is able to hold a rolling 30-day storage window before data goes to Stanford University and the JSOC/NetDRMS for distribution of HMI and AIA image data for science teams to process the data via Lockheed Martin Solar and Astrophysics Laboratory (LMSAL) and Smithsonian Astrophysical Observatory (SAO). While most of the Feature Finding Team Modules process the image data from LMSAL, our Feature Extraction Module for the SDO CBIR system codes runs at SAO. Only a handful of modules run at near-real time latency to provide space weather data for NOAA, while our module runs at a 6 minute cadency. Other science modules run at different cadencies and report to the Heliophysics Events Knowledgebase (HEK) at different intervals. Our SDO CBIR system gathers data from HEK and SAO (image parameter files, headers and thumbnails) on a daily basis. This data gets processed by several data preparation and nearest neighbor table generation/update scripts for it to be visible to users on our web-based front-end. While there are plenty of places where the whole system could be improved for better big data analysis, we are currently focusing on optimizing our nn-table update and data preparation scripts using Hadoop-based algorithms. Other potential research areas will be discussed on the following pages.

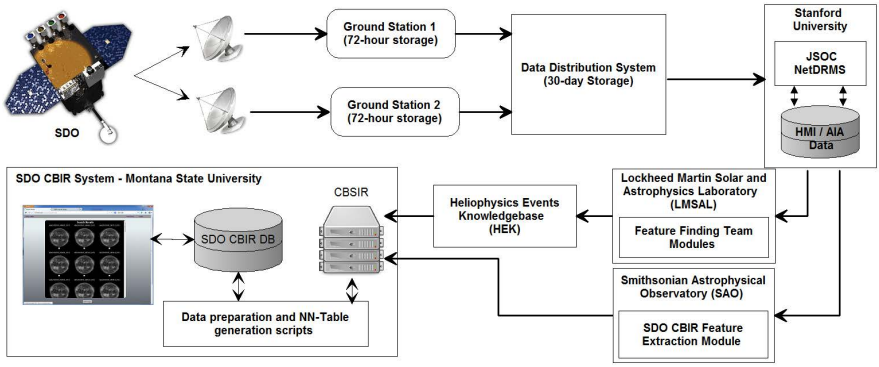

Fig. 1. SDO Pipeline outiline

2.2 SDO Content-Based Image Retrieval System

In our task to create a real-world CBIR system for solar data we have faced many interesting challenges from the unique aspects of solar data that relate directly to new and important research questions in computer science and data mining. We have addressed everything from image parameter selection, evaluation, clustering [2,3,4], dissimilarity measures evaluation for retrieval [5], dimensionality reduction analysis for retrieval [6], and evaluation of high-dimensional indexing techniques [7]. We also found very interesting relationships between our solar images and medical x-ray images [8], allowing us to further our research horizons and look into medical image retrieval and CBIR systems [9]. The first version of our system is available at http://bit.ly/17v3fVG and currently features over six months of solar data. The system is currently being enhanced with region-based querying facilities and other big data-related enhancements which will be discussed in section 3.2.

2.3 Gathering of Labeled Solar Events from Multiple Sources

The Heliophysics Event Knowledgebase (HEK) is an all-encompassing, cross-mission meta-data repository of solar event reports and information. This meta-data can be acquired at the official web interface http://bit.ly/ZWdRKH, but after finding several limitations for large-scale event retrieval, we decided to develop our own software application named QHEK (for Query HEK). Figure 2b shows an example of six types of solar events reported publicly to the HEK from fellow FFT modules. We color-code and overlay the events on the appropriate images (time and wavelength) and show the bounding boxes, and when available the detailed event boundary outlines. A preliminary version of this large-scale dataset is publicly available at http://bit.ly/15TFTps and contains over 24,000 event labels from six months of data [10].

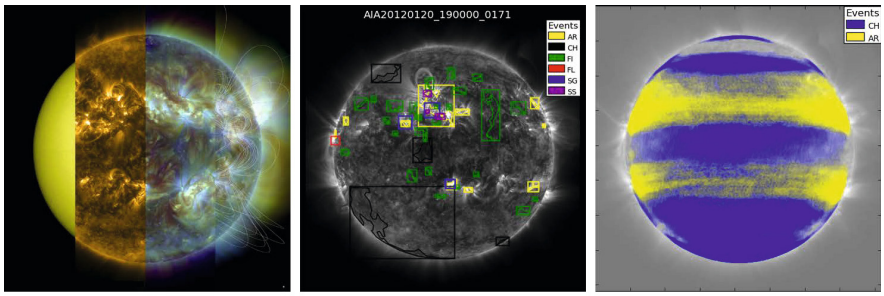

Fig. 2. Examples of SDO solar image data and meta-data. a)courtesy of NASA/SDO

2.4 Visualization of Large Scale Solar Data

To combat serious cases of information overload from the data, meta-data, and results, we have also had to develop extensive visualization tools tailored to our specific data domain and research applications. While not all of our work is directly visual, such as high-dimensional indexing techniques and spatio-temporal frequent pattern mining [7,11,12], almost all of it is related to some sort of visualizable end result. For example, with the help of visualization we can quickly analyze hundreds of solar events at once and validate a module's reporting effectiveness against known solar science, such as the confirmed distinct bands of active regions and coronal holes shown in Fig.2c. We can also use visualization to more easily assess the strengths and weaknesses of our own classification algorithms and labeling methods, whereby the human eye can keenly pick up on similar miss classified regions or poorly generated data labels.

3 Transitions into Big Data Analysis for Solar Physics

The following subsections will give some insights into our work of transitioning from traditional large-scale image retrieval and data mining approaches to big data methodologies and technologies. We also point out several of the research challenges, practical applications of current big data technologies, and the development of new big data analysis algorithms.

3.1 State of the Art in Large Scale Image Retrieval

Large scale image retrieval has been an active topic of research since the late 2000's with the likes of Google Image Search and systems like QBIC. These systems have since become closed, and in their infancy handled mostly meta-data based image search and basic color histogram matching, making them not well suited for current large scale image retrieval needs – a in depth review on CBIR could be found here [2]. With interesting works dealing with more than 10,000 image categories [16] and high-dimensional signature compression for large scale retrieval [17], it is not until 2012 where in the Neural Information Processing Systems (NIPS) conference we find the first Workshop on large scale Visual Recognition and Retrieval. Here researchers presented several algorithms that work on large scale image datasets, but almost none of them mention the use of big data technologies such as Hadoop, HBase, or HSearch. The first real mention of using big data technologies for image retrieval is in [15], where a highly speculative system using Hadoop and Lucene is proposed. We have yet to find literature with a functional system using said technologies. While most of the image retrieval algorithms have been parallelized and tested in distributed environments, either GPU or using OpenMP, they have yet to be ported to Hadoop-based environments. For the future version of our SDO CBIR system we are working on developing a Hadoop-based algorithm for nearest-neighbor index generation.

3.2 Towards Big Data Revision of the SDO CBIR System

Our first step is to verify the feasibility of migrating our traditional SDO CBIR system to a more flexible search-engine based technology using Lucene. With this initial step underway, if successful, we may then migrate to Apache's HBase hadoop-based technology for scalability with the larger amounts of data we will accumulate. We are also looking into incorporating HSearch on our HBase data repository to serve data queries for the front-end of our system. We are deploying scripts to update our CBIR system similarity indexes using MapReduce to calculate and re-calculate our similarity tables on a weekly, and eventually daily, basis in order to provide the most up-to-date results for solar scientists when important events happen (e.g. big solar flares).

The biggest research potential of our current work is the combination of image retrieval, information retrieval, and big data methodologies to create a big data content-based image retrieval system, something that will greatly benefit other areas that are starting to deal with high volumes of image data, but are currently stuck with traditional approaches. We are excited about future collaborations with image processing and retrieval researchers in expanding existing algorithms and methodologies into big data environments.

3.3 Event Labeling Module Validation

With the massive amounts of label data coming from multiple science modules, there is a need for big data technologies capable of aggregating and validating data. The current best existing computer vision tools for labeling solar images are single-object detectors, each heavily reliant on the known visual characteristics of their specific phenomenon for accurate labeling [13]. Object recognition and classification based on more general visual parameters is still limited, although some success has been seen in filament detection [14].

The development of these specific modules is expensive in terms of time, effort and domain knowledge, therefore a general purpose computer vision tool is much better suited for extension to include new phenomena, or to classify subtypes of known phenomena by visual character. For these reasons we seek to develop a tool capable of using the image texture parameters to label and classify events in solar imagery. In this environment we endeavor to construct and test a multi-label event classification scheme for solar images. A major advantage of multi-label classification is that it is known that the occurrences of solar phenomena are not independent. For example, active regions are areas of high solar activity, while coronal holes are areas of low solar activity, so they should never occur in the same location (again, see Fig.2c).

3.4 Spatio-temporal Solar Event Reporting and Mining

Spatio-temporal analysis of solar physics data is a major emerging area and the volume of data this will generate must be addressed using big data analysis methodologies. In our initial stages we are working on establishing an all-encompassing infrastructure in order to store all reported events. The current

reporting involves a single spatial label per temporal event, an event that could range from minutes to days. We are proposing to create tracking datasets with each spatial label converted to a temporal step of our solar data, exploding our dataset from thousands of records to millions. In the SDO data context, we are looking at over 70,000 images with multiple spatio-temporal labels per day.

In our second step, we are investigating the migration of data into HBase, with highly-scalable search capabilities using HSearch. This will be taking advantages of Hadoop/MapReduce environments to process and analyze the data with their clustering and mining algorithms, as well as having a front-end to serve the data for other research institutions. New algorithms will need to be developed to fit the context of spatio-temporal data analysis for big data sources.

4 Looking into the Future

As the volume of solar data keeps growing each day, the transition from using traditional data mining, machine learning, and information retrieval techniques into more scalable big data methodologies and tools is imminent. While we have outlined some of the steps we are currently taking to address these issues, we are also looking for new collaborations with big data experts to further benefit the field of solar physics. As we have shown, there are plenty of new areas of research that can be benefitted from the massive solar datasets and the new tools and algorithms expected to be developed for this domain can be greatly beneficial for other big data research areas.

References

1. Hapgood, M.A.: Towards a scientific understanding of the risk from extreme space weather. Advances in Space Research 47(12), 2059–2072 (2011)
2. Banda, J.M., Angryk, R.: Selection of Image Parameters as the First Step Towards creating a CBIR System for the Solar Dynamics Observatory. In: Proc. of Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA), pp. 528–534 (2010)
3. Banda, J.M., Angryk, R.: An Experimental Evaluation of Popular Image Parameters for Monochromatic Solar Image Categorization. In: Proc. of the 23rd Florida Artificial Intelligence Research Society Conf., pp. 380–385 (2010)
4. Banda, J.M., Angryk, R.: On the effectiveness of fuzzy clustering as a data discretization technique for Large-scale classification of solar images. In: Proc. IEEE International Conference on Fuzzy Systems, pp. 2019–2024 (2009)
5. Banda, J.M., Angryk, R.: Usage of dissimilarity measures and multidimensional scaling for large scale solar data analysis. In: Proc 2010 Conf. on Intelligent Data Understanding (CIDU), pp. 189–203 (2010)
6. Banda, J.M., Angryk, R., Martens, P.C.H.: On Dimensionality Reduction for Indexing and Retrieval of Large-Scale Solar Image Data. Solar Phys. 283, 113–141 (2012)
7. Schuh, M.A., Wylie, T., Banda, J.M., Angryk, R.A.: A comprehensive study of iDistance partitioning strategies for kNN queries and high-dimensional data indexing. In: Gottlob, G., Grasso, G., Olteanu, D., Schallhart, C. (eds.) BNCOD 2013. LNCS, vol. 7968, pp. 238–252. Springer, Heidelberg (2013)

8. Banda, J.M., Angryk, R., Martens, P.: On the surprisingly accurate transfer of image parameters between medical and solar images. In: Proceedings of the International Conference on Image Processing (ICIP), pp. 3730–3733 (2011)
9. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications: clinical benefits and future directions. International journal of medical informatics 73, 1–23 (2004)
10. Schuh, M.A., Angryk, R.A., Pillai, K.-G., Banda, J.M., Martens, P.C.H.: A large-scale solar image dataset with labeled event regions. To appear in. In: Proc. of the International Conference on Image Processing, ICIP (2013)
11. Pillai, K.-G., Angryk, R.A., Banda, J.M., Schuh, M.A., Wylie, T.: Spatio-temporal co-occurrence pattern mining in data sets with evolving regions. In: ICDM Workshops 2012, pp. 805–812 (2012)
12. Pillai, K.G., Sturlaugson, L., Banda, J.M., Angryk, R.A.: Extending high-dimensional indexing techniques pyramid and iMinMax(θ): Lessons learned. In: Gottlob, G., Grasso, G., Olteanu, D., Schallhart, C. (eds.) BNCOD 2013. LNCS, vol. 7968, pp. 253–267. Springer, Heidelberg (2013)
13. Martens, P.C.H., Attrill, G.D.R., Davey, A.R., Engell, A., Farid, S., et al.: Computer vision for the solar dynamics observatory (SDO). Solar Physics (2011)
14. Schuh, M.A., Banda, J.M., Bernasconi, P.N., Angryk, R.A., Martens, P.C.H.: A comparative evaluation of automated solar filament detection. Solar Physics (under review, 2013)
15. Gu, C., Gao, Y.: A Content-Based Image Retrieval System Based on Hadoop and Lucene. In: Cloud and Green Computing (CGC), November 1-3, pp. 684–687 (2012)
16. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
17. Sánchez, J., Perronnin, F.: High-dimensional signature compression for large-scale image classification. In: Proc. of CVPR (2011)