Survey on Data-intensive Applications, Tools and Techniques for Mining Unstructured Data

Santhosh Voruganti
Assistant.Professor in IT Department,
CBIT, Hyderabad

ABSTRACT

Due to the swift growth of WWW there has been large volume of information is produced and shared by various administrations in nearly every business, industry and other fields. Due to this high explosion it’s really a big challenge to store, manage and access knowledge. Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly. Often many times faster than structured databases .Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, pdfs ,videos, photos, audio files, presentations, web pages and many other kinds of business documents. A huge amount of information spread across the web poses a major challenge in identifying relevant information. Existing tools lack analysis and visualization capabilities and traditional result displays long list of documents instead of providing concrete answers. This paper discusses various methods,tools and techniques for mining unstructured data that enables better data analysis and visualization.

Keywords

Unstructured data, structured data, data mining, text mining, machine learning, DGE model.

1. INTRODUCTION

Historically Natural Language Processing (NLP) focuses on unstructured data (speech and text) understanding while Data Mining (DM) mainly focuses on massive, structured or semi-structured datasets. The general research directions of these two fields also have followed different philosophies and principles. For example, NLP aims at deep understanding of individual words, phrases and sentences (“micro-level”), whereas DM aims to conduct a high-level understanding, discovery and synthesis of the most salient information from a large set of documents when working on text data (“macro-level”). But they share the same goal of distilling knowledge from data. In the past five years, these two areas have had intensive interactions and thus mutually enhanced each other through many successful text mining tasks. This positive progress mainly benefits from some innovative intermediate representations such as “heterogeneous information networks”.

The major obstacle in extracting information from such unstructured document is its unorganized, ambiguous and collapsed content. This requires a higher level of pre-processing rather than the typical preprocessing techniques adopted for structured and semi structured documents. Techniques such as rule based feature extraction using association rule mining, term frequency, inverted document frequency, dimensionality cube formation, dimensionality reduction etc., are popular. However, applying these techniques directly to unstructured documents will provide low precision and recall. Though the recent trends in ontology based feature identification methodologies show promising results and improvements over existing solution, they concentrate only on the concept and relationship identifications using document terms. Information loss or shortage of effectiveness of information extraction cannot be acknowledged in critical applications like medical records, tender documents etc. since they involve risk to life and are also cost intensive.

There are a number of applications where unstructured data needs to be integrated with structured databases on an ongoing basis so that at the time of extraction a large database is available Consider the example of publications portals like Cite seer and Google Scholar. When integrating publication data from personal homepages, it does not make sense to perform the extraction task in isolation. Structured databases from, say ACM digital library or DBLP are readily available and should be exploited for better integration. Another example is resume databases in HR departments of large organizations. Resumes are often stored with several structured fields like experience, education and references. When a new resume arrives in unstructured format, for example, via email text, we might wish to integrate it into the existing database and the extraction task can be made easier by consulting the database. Another interesting example is from personal information management (PIM) systems where the goal is to organize personal data like documents, emails, projects and people in a structured inter-linked format. The success of such systems will depend on being able to automatically extract structure from the existing predominantly file-based unstructured sources. Thus, for example we should be able to automatically extract from a PowerPoint file, the author of a talk and link the person to the presenter of a talk announced in an email. Again, this is a scenario where we will already have plenty of existing structured data into which we have to integrate a new unstructured file.

Labeled unstructured data is the classical source of information for extraction model. Traditionally, three kinds of information have been derived from unstructured records. First, the labeled entities hold the same kind of pattern information as the database of entities as discussed above. Second, the context in which an entity appears is a valuable clue that is present solely in the labeled unstructured data. Context is typically captured via few words before or after the labeled entity. For example, the word “In” often appears before a journal name. Finally, the likely ordering of entity labels within a sequence is also useful. For example, author names usually appear before or after a title.

Unstructured data exists in two main categories: bitmap objects and textual objects. Bitmap objects are non-language based (e.g. image, audio, or video files) whereas textual objects are "based on written or printed language" **a**nd predominantly include text documents . Text mining is the discovery of previously unknown information or concepts from text files by automatically extracting information from several written resources using computer software. In text mining, the files mined are text files which can be in one of two forms. Unstructured text is usually in the form of summaries and user reviews whereas structured text consists of text that is organized usually within spreadsheets. This evaluation focused specifically on mining unstructured text files.

2. RELATED WORK

Data Mining is Knowledge detection and resolution process of databases. It has obtained previously unknown, secret, meaningful and useful patterns being automatically established from large scaled databases,. So, data mining knowledge discovery in databases is looking for patterns in data. Likewise, text mining looking for patterns in text.

Text mining is the process of analyzing text to extract information that is useful for particular purposes. Text is unstructured, amorphous, and complicated to deal with. Nevertheless, text is the most common vehicle for the formal exchange of information.

Data Mining techniques and their tools are designed to exert structured data from databases. Text mining functionality] is similar to data mining, but text mining can work with unstructured data such as PDF files or semistructured data sets such as emails, XML and HTML files and etc. So, text mining is a superior way for companies in business fields. Since the most of information in these places is saved as text or in text files.

The management of unstructured data is acknowledged as one of the most critical unsolved problems in data management and business intelligence fields in current times. The major reason for this unresolved problem is primarily because of the actuality that the methods, systems and related tools that have established themselves so successfully converting structured information into business intelligence, simply are ineffective when we try to implement the same on unstructured information. New methods and approaches are very much necessary. It is a known realism that huge amount of information is shared by the organizations across the world over the web. It is, however, significant to observe that this information explosion across the globe has resulted in opening a lot of new avenues to create tools for data management and business intelligence primarily focusing on unstructured data. In this paper, we explore the challenges being faced by information system developers during mining of unstructured data in the context of semantic web and web mining. Opportunities in the wake of these challenges are discussed towards the end of the paper.

Machine learning is a powerful and essential tool for accomplishing many tasks and problems associated with big data, but current researches and developments are still faced with a lot of great research challenges for big data processing. In order to realize the full potential of big data, we need to address several major research challenges and open issues, including the following (but not limited to): l How to explore and exploit the useful information hidden in big data by using machine learning techniques should draw further attention, as large quantities of useful data are getting lost since new data is largely untagged and unstructured data. l In most existing machine learning applications, the researchers just apply single learning algorithm or technique to deal with practical problems, but it is important to realize that each approach has strengths and weaknesses. Thus the idea of hybrid learning should be further considered at present big data background. l The characteristics of big data make the data visualization an enormously challenging task. The recent visualization techniques l like dimension reduction can only give an abstract view of the data. Therefore, how to use machine learning techniques to give true geometric representations for big data also needs to be investigated.

Big data is large volume, heterogeneous, distributed data. Big data applications where data collection has grown continuously, it is expensive to manage, capture or extract and process data using existing software tools. For example Weather Forecasting, Electricity Demand Supply, social media and so on. With increasing size of data in data warehouse it is expensive to perform data analysis. Data cube commonly abstracting and summarizing databases. It is way of structuring data in different n dimensions for analysis over some measure of interest. For data processing Big data processing framework relay on cluster computers and parallel execution framework provided by Map-Reduce. Extending cube computation techniques to this paradigm. MR-Cube is framework (based on map reduce) used for cube materialization and mining over massive datasets using holistic measure.

Unstructured data processing is therefore a very important emerging class of applications. There are a number of unstructured data processing applications that are already in use today. These applications include text searches (exact and approximate searches) ,content-based searches of image, video, and audio files , and data fusion. Although some of these applications are used in relatively niche domains (e.g., geo-spatial data fusion is used in urban planning and forestry), the core *methods* used in these applications are expected to become common place across a wider range of applications in the future. For example, Content-Based Image Retrieval (CBIR)], which is very processing and I/O intensive, is now used in the field of medicine for querying biomedical digital libraries.

However, CBIR has been identified by Intel as an important emerging application even for the home user, say, for content-based querying of digital photo collections stored on her personal computer, or on a photo repository, such as, Flickr. This growing demand for unstructured data management has already started creating a market for hardware appliances that are specifically designed for searching and processing unstructured data.

2.1 Entail Of Security In Big Data

For marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective. If a security breach occurs to big data, it would result in even more serious legal repercussions and reputational damage than at present. In this new era, many companies are using the technology to store and analyze petabytes of data about their company, business and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful.

The challenge of detecting and preventing advanced threats and malicious intruders must be solved using big data style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources.

Not only security but also data privacy challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset, therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security.

3. METHODS

3.1 Unstructured Data Management

Generally, unstructured data can be represented with multiple interpretations. For example, multiple semantic constraints can be extracted from a document and multiple features can be extracted from image content (color, texture,etc.). The dynamic nature of these data demands the search for resources to ensure maintenance and evolution of the data schemas and representation.

To manage unstructured data, information from various sources has to be extracted, organized, characterized ,analyze the data, data mining, classification of data, text mining and modeling of the processed data.

- Extract Information
- Feature extraction
- Organized the facts
- Text mining
- Modeling and defined the structure of processed data

For managing unstructured data in web pages for database using XML: It's hard to find a tool that deals the unstructured data which can be stored, retrieve data extracted into structured database. The following steps to be carried out to get the output into actionable form from unstructured data.

3.2 Unstructured Data

Unstructured data to be analyzed is considered as input either a web page or a document.

3.2.1 Data Extraction:

Data extraction is a process of retrieving and capturing the data from one medium to another medium. Medium can be web pages, documents, database, and stack of information. Web pages are typically considered unstructured data though web pages are defined by HTML, which has rich structure. This is because web pages also contains lot of static text, links and references to external, images, XML files, animations and databases. Therefore extract and categorized information out of data. A wrapper access HTML document and exports it into structured format XML or data relations. Maintaining the Integrity of the Specifications.

3.2.2 Target the extraction:

Extraction target can be a relation of 'k' tuples, where k is number of attributes in a record or object

3.2.3 Syntactic & Semantic Analysis:

For syntactic analysis, structure is determined by generating a parse tree by classifying sentence into subjects, verb phrase (verb, object). Similarly semantic analysis finds synonyms.

3.2.4 Data classification:

Data classification is to categorize data based on required models like object oriented model or ER model. There are many algorithms to classify in data mining like 'K-nearest neighbour (KNN)' algorithm. Some more algorithms include Bayesian algorithm and concept vector based (CVB) algorithm to classify words in documents. 'Page rank algorithm' uses search ranking technique based on hyperlinks on the web.

3.2.5 Inference rules and Representation into structured format:

Inference rules can be employed to draw conclusions of the classified data by preserving the semantic property. XML is used to store and transport the data. The classified data is stored in the form of data tables or XML is used to store the data based on the requirement of the desired action planned from the unstructured data.

3.3 The Need For A Data Generation And Exploitation Model

We now argue that to manage unstructured data effectively,a clear data generation and exploitation model (or DGE model for short) will have to emerge. Unfortunately, no such model has been identified by our community. We then speculate on such a model.

A DGE model explains the interaction between the data, the system, and the users. It explains how the data is generated inside the system, who the users are, what their information needs are, how they express the needs, and how they interact with the system to satisfy these needs.

A DGE model for unstructured data should use a combination of Information Extraction, Information Integration, and Human Intervention to generate structured data from the originally unstructured data. The model should allow a broad range of data exploitation modes (e.g., keyword search, structured querying, browsing, visualization, monitoring), as well as seamless transition from one mode to another, in an iterative fashion through interaction with the user.

To really find relationships and patterns between different types of data, agencies are turning to predictive analytics. These tools help organizations uncover patterns from large amounts of both structured and unstructured data patterns that not only find current trends, but can help predict future occurrences. There is a reason why the most effective analytics tools today are based on Hadoop .It's the perfect complement to big data, because it is designed to process, store and analyze petabytes and exabytes of distributed, unstructured and structured data. Apache Hive is a Datawarehouse built on top of Hadoop that allows you to query and manage large sets in scattered storage space using a HiveQL. Hive translates queries into a series of Map Reduce jobs.

4. TOOLS AND TECHNIQUES FOR MINING UNSTRUCTURED DATA

Different software tools to help them organize and manage unstructured data. These can include the following:

4.1 Big Data Tools*:* Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly..

4.2 Business Intelligence Software: Also known as BI, this is a broad category of analytics, data mining, dashboards and reporting tools that help companies make sense of their structured and unstructured data for the purpose of making better business decisions.

4.3 Data Integration Tools*:* These tools combine data from disparate sources so that they can be viewed or analyzed from a single application. They sometimes include the capability to unify structured and unstructured data.

4.4 Document management systems: Also called "enterprise content management systems," a DMS can track, store and share unstructured data that is saved in the form of document files. Information.

4.5 Management solutions*:* This type of software tracks structured and unstructured enterprise data throughout its lifecycle.

4.6 Search and indexing tools: These tools retrieve information from unstructured data files such as documents, Web pages and photos.

The different techniques used to search analyse and deliver unstructured data are

- Content management system
- Relational Database
- Data Mining
- Text Analytics. Federal search or enterprise
- search data base
- Non-relational database
- Real time data visualization tools
- E-discovery application

The new technologies for unstructured data are

- Log monitoring and reporting tools
- In-memory databases
- NOSQL databases
- Hadoop
- MPP data warehouses

5. CONCLUSION AND FUTURE SCOPE

The identification, collection and analysis unstructured data from web-based sources presents both an opportunity and a challenge to individuals, organizations and institutions that use the internet as a source for primary and secondary data to support decision making.

In the unstructured data world, we argue that it is highly desirable to have a similar example system, one that can and unify the work, and hopefully enable rapid progress.

The next important issue in a real-life setting concerns the assumptions about the data. In general one may claim that data mining deals with all sorts of structured tabular data (e.g., non-numeric, highly unbalanced, unclean data) as well as with non-structured data (e.g., text documents, images, multimedia), and does not make assumptions about the distribution of the data.

6. REFERENCES

[1] Data Mining with Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

[2] 10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH QIANG YANG Department of Computer Science Hong Kong University of Science and Technology Clearwater Bay, Kowloon, Hong Kong, China XINDONG WU Department of Computer Science University of Vermont 33 Colchester Avenue, Burlington, Vermont 05405, USA xwu@cs.uvm.edu International Journal of Information Technology & Decision Making Vol. 5, No. 4 (2006) 597–604 c_World Scientific Publishing Company

[3] Evaluating the Effectiveness of Keyword Search William Webber Computer Science and Software Engineering The University of Melbourne Victoria 3010, Australia wew@csse.unimelb.edu.au

[4] Integrating Predictive Analytics and Social Media Yafeng Lu, Robert Kr¨uger, Student Member, IEEE, Dennis Thom, Feng Wang,Steffen Koch, Member, IEEE, Thomas Ertl, Member, IEEE, and Ross Maciejewski, Member.

[5] Handling Unstructured Data for Semantic Web – A Natural Language Processing Approach Hemant Kumud Scholars Journal of Engineering and Technology (SJET) ISSN 2321-435X (Online) Sch. J. Eng. Tech., 2014; 2(2A):193-196

[6] A Recent Survey on Unstructured Data to Structured Data in Distributed Data Mining Padmapriya et al, Int.J. Computer Technology & Applications ,Vol 5 (2),338-344 ISSN:2229-6093

[7] Managing Unstructured Data With Structured Legacy Systems David A. Maluf david.a.maluf@nasa.gov Peter B. Tran peter.b.tran@nasa.gov NASA Ames Research Center Intelligent Systems Division Mail Stop 269-4 Moffett Field, CA 94035

[8] Approaches for Managing and Analyzing Unstructured Data N. Veeranjaneyulu, M. Nirupama Bhat, A. Raghunath School of Computing, Vignan's University, Guntur, India International Journal on Computer Science and Engineering (IJCSE)

[9] F. S. Gharehchopogh Hacettepe University Department Computer Engineering Ankara, Turkey Z. A. Khalifelu IAU Branch of Shabestar Department of Computer Engineering Shabestar, Iran Analysis and Evaluation of Unstructured Data:Text Mining versus Natural Language Processing

[10] The Case for a Structured Approach to Managing Unstructured Data AnHai Doan, Jeffrey F. Naughton, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, Ba-Quy Vuong University of Wisconsin-Madison

[11] The Acumen and Acuity of Data Mining and Text Mining Anjali Jivani, Jait Purohit, Kaushal Patel Comp

Science & Engineering MSU Baroda, India Volume 4, Issue 11, November 2014 ISSN: 2277 128X Available online at: www.ijarcsse.com

[12] A Benchmark Suite for Unstructured Data Processing Clinton Wills Smullen, IV Shahrukh Rohinton Tarapore Sudhanva Gurumurthi Department of Computer Science University of Virginia Charlottesville VA 22904 {cws3k,shahrukh,gurumurthi}@cs.virginia.edu

[13] Storing of Unstructured data into Mongo DB using Consistent Hashing Algorithm Saranraj Sankarapandi PG Scholar, IIIT-Srirangam, Tiruchirapalli, Tamilnadu, India. Dr. M. Sai Baba Associate Director, RMG, Indira Gandhi Centre for Atomic Research, Kalpakkam, Tamilnadu, India. S. Jayanthi Assistant Professor, Department of Computer Science & Engineering,Anna University, BIT Campus Tiruchirapalli, Tamilnadu, India. E.Soundararajan Scientific Officer/E,SIRD ,Indira Gandhi Centre for Atomic Research, Kalpakkam, Tamilnadu, India. International Journal of Emerging Technologies in Engineering Research (IJETER) Volume 3, Issue 3, December (2015)