2011 International Conference on Information and Network Technology
IPCSIT vol.4 (2011) © (2011) IACSIT Press, Singapore

Leveraging Unstructured Data into Intelligent Information – Analysis & Evaluation

Ayaz Ahmed Shariff K [1], Mohammed Ali Hussain [2] , Sambath Kumar[3]

[1]Department of CSE, Birla Institute of Technology (Mesra) Noida, India
ayazahmedsk@gmail.com

[2]Department of CSE Birla Institute of Technology (Mesra)
ali.alitimate@gmail.com

[3]Department of CSE, Royal College of Applied Science & Technology, UAE
sampathhari@gmail.com

Abstract. Unstructured data constitutes about 70% of the data collected or stored in larger organizations which are difficult to access, use or retrieved. This topic deals with this uncertainty to convert the unstructured data in actionable form. Knowing the business value and IT value of the structured data, the amount of effort and time wasted in accessing the necessary information lying in the back bench of collected data, cost spent on searching the information, it becomes highly necessary to manage the unstructured data. In this research, the aim is to retrieve the structured information out of unstructured data using feature extraction, analyzing this data syntactically, organize the analyzed data into entities, rules, associations, facts. Represent this data into structured form either in form of XML or data tables. XML language is very suitable for data storage and data exchange. Data transformation utility was developed using Microsoft Visual Studio 2005. The textual data in documents can be transformed into text file, the data in which can be imported into database. So the transformation of unstructured data can be accomplished with this utility. Feature extraction categorizes the data into entities, events and builds the relations among these entities and events. Due to complexity involved in extracting, mining and structuring the data, research is considered for textual data either in form of documents or web pages. The structured information can be used in decision support systems or serve the purpose intended for the process. We aim at developing a simple approach to extract the key information from scattered unstructured data lying across websites, database, emails etc. The goal is to have effective, improved information retrieval system with this approach. As an application of the approach, we are developing a news retrieval system incorporating the features discussed in this paper. In this paper, an application "Intelligent news retrieval system" has been proposed as model which pulls out the news (same or different) from various web pages (blogs, news websites) and processes them on the basis of popularity or page ranking and display on a single web page. This model collects news from various sources. The use of regular expressions is to recognize the required patterns of the data, anything inside header and title tags. To carry out the procedure, convert the web pages into plain text. This plain text analyzed for entities, facts, relationships, synonyms, thematic analysis, and verb phrases. Data dictionary is used to recognize English words. Extracted data is stored in database inform of tables or XML. Database models can be constructed using constructive information by inference rules or actionable intelligence. The structured information can be used for the purposes intended. The goal of the proposed model is to develop a simple, effective filtered online news reading website which highlights news based on priorities of users, number of hits in source websites, explicit and implicit ratings, likes by users.

Keywords : Unstructured data, Information retrieval, extracted data, model,

1. Introduction

A database is organized collection of data for many uses typically in digital form. Data can be text, numbers, graphs, images. The "unstructured data is any data without a well defined model or schema for accessing information, like word documents, emails etc." Then what is structured data? Structured data is data with a proper model organized into the likes of tables, tags or like objects.

Unstructured Data contains

- Text
- Audio
- Images, etc

Large companies may have presences in many places, each of which generate a large volume of data. For example, insurance companies may have data from thousands of local branches. Further, large organizations have complex data structure with or without schemas.

Unstructured data can take many forms like word documents, spread sheets, email messages, blogs, pictures, movies. Unstructured data by nature is raw data, data mining or "analysis" of the UD to arrive at the results or statistics that will be placed in the structured world equivalent to business rules.

In my opinion, they should unstructured data mining should contain the document name & title, location of source, discovered context, raw term, context, and exact position within the document, and possibly a few other key notions. The mining engine should be capable of "clustering" terms together to form an idea, a context.

Data mining is the process of semi automatically and analyzing large databases to find useful patterns. Data mining attempts to discover rules and patterns from the data. Unstructured data analysis and mining is much more than this. Unstructured Data can be scattered, complex and different structures, different schemas. The tools available for data mining techniques may or may not be very useful to extract and represent the structured information out of unstructured data.

2. Significance & Need of Unstructured Data Management

"The process of mining, exercising and analyzing the unstructured data to capture actionable form.[3]" The need arises due to some of the following facts[7]:-

- Amount of Unstructured Data in large corporations doubles every 2 months.
- Companies with unstructured data management can at least 15% more productive.
- The average knowledge worker spends on an average of 2.5 hours/day in search of documents.
- Merrill lynch estimates that more than 85% of all business information exists as unstructured data in form of emails, memos, notes from call centres, news, user groups, reports, letters, white papers, marketing material, research and web pages.
- More than 80% of information on internet is unstructured.
- More than 2 billion web pages have been created since 1995, with an additional 200 million new web pages being added every month according to market-research firm IDC.
- International Data Corporation (IDC) reports that an organization with 1000 workers loses a minimum of $6 million searching the information.

3. Theory

3.1. How Unstructured Data is Different from Structured Data?

We know unstructured data is one without a defined data model or cannot be easily usable by a computer program. With a structured document, certain information always appears in the same location on the page. For example, in an employment application the applicant's name always appear in the same box in the same place on the document. In contrast, an unstructured document has the opposite characteristics – information can appear in unexpected places on the document.

Value of Unstructured Data:

- Business Value:
- Better information
- Timely information
- Relevant Information

- Greater business impact
- More information is available to store, manage and modelled

3.2. Unstructured Data Management:

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

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

3.3. Text Mining:

"Process of extracting information from textual data (emails, documents) and utilizing for better decisions is called as text mining." Business Intelligence (BI) tools are used for this process and focus on semantics is made.

The following categories to mine the text - Syntactic and Semantic feature extraction:

- Structure Determination: names, companies, places, locations, people, verbs, objects etc.
- Event extractions like sales, elections, anniversaries, birthday events, etc.
- Extract the relationships among the identified entities and events.
- Categorizing the documents in an order or defined structure.
- Summarization of data and thematic analysis to find the theme or context in the documents.

Let's have a look into the process of information extraction. Once the elements of information is extracted like identifying "named attributes" (people, places) or other quantifiable variables like date, measurements, then relationships among these connecting elements are captured which express facts. For example, determine the roles of various entities and relationships among them, like, the person identified may be the "boss" of organization and also a member of other organization. This forms a link creation that quickly uses the facts in documents to understand connections in the larger world.

Fact extraction can enable more forms of querying like document preview and content packaging. Extracted entities and facts when displayed in search results might provide clue to particular document which can be useful to specific task.

4. Approach for the Work

For leveraging of 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.

Unstructured Data → Data extraction → Syntactic & Semantic Analysis → Data classification → Inference rules → Representation into structured format (XML or Data Relations)

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

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.

To extract the data, consider following tasks:

- Define its input: Input can be unstructured data; semi structured data, and structured data.
- Using text pattern matching also known as Regular expression: To identify small or large-scale structure e.g. records in invoices and their associated data from headers and footers.
- Target the extraction: Extraction target can be a relation of 'k' tuples, where k is number of attributes in a record or object.

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.

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.

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.

There are many tools available to extract data as follows:

- HTML-aware tools: for HTML documents that require HTML document to be represented in parsing tree. Ex: Roadrunner
- NLP techniques: RAPDIER, SRV tools build relationships between sentences, elements and phrases.

5. Applications of Leveraging the Unstructured Data into Intelligent Information

The following are some of the sample applications evaluated out of unstructured data.

- Business applications like broadcast content management, call centre automation, CRM, manufacturing quality control, etc.
- Unlock the hidden knowledge lying in the back benches of unstructured data
- Reduces costs of analyzing text by eliminating manual work
- For better decision making, business growth opportunities.
- Marketing optimization: Organizations can search public information to gain understanding of the overall market trends to position their products.
- Health care applications: Managing patient records help doctors to identify a patient's medical history.

6. Proposed Model: Intelligent News Retrieval System:

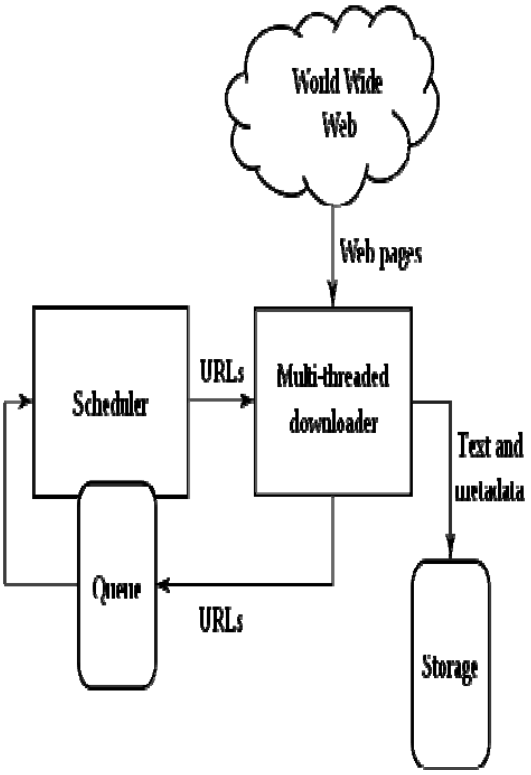

Fig 1: High level architecture of basic crawler

Figure 1 shows crawler to fetch web pages & index the document. The project aims at developing a system that would use a crawler based search method which would pull out news from major news websites, forums, portals, blogs and even Twitter. The news will be processed based on its popularity and exact upward force exerted on a news story by the internet. The news will be presented to the user based on his preferences and general behaviour that the system learns over time to produce highly relevant results.

7. Acknowledgements

We would like to thank Birla Institute of Technology (Mesra) for support of this research work. We would also like to acknowledge from Faculty, CSE of BIT for valuable guidance.

8. References

[1] Mansuri I.R. Sarawagi S. "Integrating Unstructured Data into Relational Databases" Data Engineering. ICDE '06. Proceedings of the 22nd International Conference, IIT Bombay 2006.

[2] David Alfred Ostrowski. IEEE International Conference on Semantic Computing "A Framework for the Classification of Unstructured Data." Berkeley, CA, USA 2009.

[3] Rao R. "From unstructured data to actionable form" appeared in IT professional, ieee.org computer society." Inxight, Sunnyvale, CA, USA

[4] Abraham Silberschatz, S. Sudarshan "Database Management System Concepts."

[5] http://searchbusinessanalytics.techtarget.com/feature/Managing-unstructured-data-in-the-organization

[6] Maluf D.A. Tran, P .B "Managing unstructured data with structured legacy systems" , Aerospace conference 2008 IEEE.

[7] Unstructured Data in http://en.wikipedia.org/wiki/Unstructured_data

[8] Seth Grimes. "is unstructured data merely modelled" published in Intelligent Information week journal. 2005.

[9] Robert Malone. "Structuring unstructured data" published in Forbes magazine, USA. 04-may-2007.

[10] http://www.information management.com/issues/20030201/6287-1.html

[11] Caret Chou, Kishor Gummaraju, Muralidhar. White paper "semantics Driven Consumer Insights" for Content packaged Goods (CPG) sector of Infosys technologies ltd, Bangalore, India.