Content Repurposing Platform for Massive and Diverse Unstructured Data Analytics

Shoji Kodama
Nobuo Nukaga
Ryoichi Ueda
Shinya Iguchi

OVERVIEW: Data is rapidly increasing in quantity and becoming more diverse, and there are growing demands from many different fields to analyze this data and utilize it in ways other than its primary purpose. Whereas most previous data analysis software was designed to handle structured data, recent years have seen a growing need for the handling of text, images, and other unstructured data created by people. The problems with unstructured data in the past have been that its size and quantity have made processing slow and that it was difficult to analyze because, unlike structured data, it could not be processed mechanically in its original format. Therefore, Hitachi has been working on research and development of IT infrastructure technologies, such as storage systems, technologies for massive data processing, and media processing technologies for searching text, voice, image, and other media data. For the future, Hitachi plans to undertake research into content repurposing platforms that support the consistent management and utilization of unstructured data by merging these technologies and making further improvements.

INTRODUCTION

UNDERSTANDING the view that data is the oil of the 21st century, the market for data analysis software exceeded 10 billion dollars in 2011 and is expected to continue growing rapidly[1].

Most data that has been subjected to analysis in the past has been of the structured type that either has a defined format or consists of numeric values such as sales totals or inventory levels, which are easy to process computationally on a computer. However, structured data makes up only 20% of the total data held by corporations, with text, images, voice, and other unstructured data produced by people accounting for the remainder[2].

Recent years have seen growing demand from a range of sectors including healthcare, finance, corporate information, government institutions, and video surveillance for analysis and repurposing of unstructured data that in the past was limited to collection, storage, and viewing, where “repurposing” means using data in ways other than its primary purpose such as in academic research or marketing. The problem with unstructured data, however, has been that it has not been able to be put to full use in the past because of the difficulty of subjecting it to computational processing by computers.

Given these circumstances, Hitachi has been looking at massive data processing and information

Table 1. Trends in Unstructured Data Repurposing Needs

Demand for analyzing text, images, voice, and other types of unstructured data for use in business is growing in a range of different sectors.

Sector	Needs	Use cases	Unstructured data
Healthcare	Use of ICT in healthcare	Educational and academic use, analysis and application of medical data such as diagnostic reason management and DPC analysis[3]	Scanned data, medical images, voice notes
Corporate IT system	Disclosure of electronic evidence (e-discovery)	Investigation of a person’s past activities and motivations based on e-mail and other evidential data	E-mail, documents, logs
Video surveillance	Use of surveillance images in marketing	Identify people’s gaze points from video captured by a vending machine to improve product positioning.[4]	Surveillance images
Government	Data.gov	Over 400,000 items of government data have been put into standardized formats and posted on the web.[5]	Questionnaire results, list of bankrupt banks
Finance	Use of social media to predict stock prices	Research into relationship between a company’s stock price and its popularity on social media.[6]	User-generated content

IT: information technology   ICT: information and communication technology   DPC: diagnosis procedure combination extraction as technologies capable of supporting the repurposing of unstructured data.

This article looks at the trends and technical challenges associated with the repurposing needs of various different types of unstructured data collected by corporations, and summarizes past research initiatives along with the content repurposing platforms being devised by Hitachi.

TRENDS IN DEMAND AND TECHNICAL CHALLENGES

Trends in Unstructured Data Repurposing Needs

Currently, repurposing needs such as search and analysis of large quantities of unstructured data are growing in a range of sectors. Table 1 lists the needs of each sector.

As the healthcare industry has adopted ICT (information and communication technology) incrementally, it faces problems due to different systems being used for different departments and applications, as well as poor interconnectivity between systems from different vendors. Although progress has been made in recent years on integrating the data from different systems by standardizing data formats such as the DICOM (Digital Imaging and Communications in Medicine) standard, hospitals still contain large amounts of unstructured data including scanned data and voice notes. There is a demand for this data to be used in other ways, such as for diagnostic reason management or academic purposes.

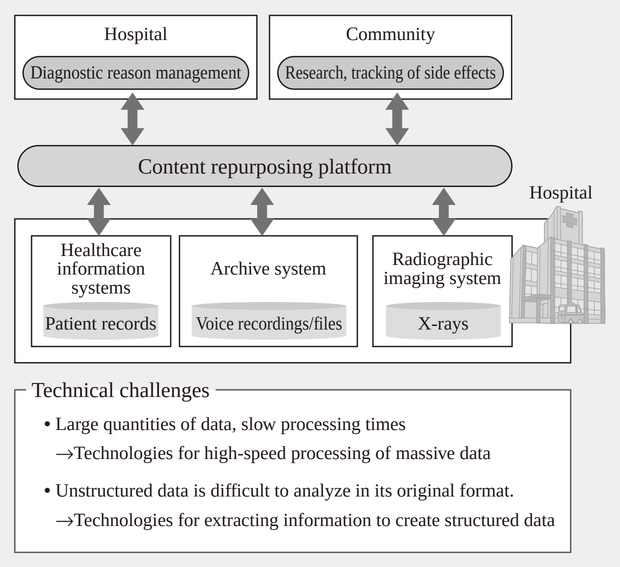
Hospital
Diagnostic reason management
Community
Research, tracking of side effects
Content repurposing platform
Hospital
Healthcare information systems
Patient records
Archive system
Voice recordings/files
Radiographic imaging system
X-rays
Technical challenges
• Large quantities of data, slow processing times
→Technologies for high-speed processing of massive data
• Unstructured data is difficult to analyze in its original format.
→Technologies for extracting information to create structured data

Fig. 1—Healthcare Examples of Data Repurposing and Technical Challenges.
Technologies for high-speed processing of massive data and technologies for extracting information to create structured data are needed to implement content repurposing platforms able to analyze and utilize data stored across a range of different systems.

In the corporate information sector, the ability to use e-discovery (electronic evidence disclosure) systems has become important. These systems search for and consolidate litigation-related electronic evidence in unstructured data such as e-mail and files that are spread around the company and analyze it to decide whether it needs to be disclosed.

Elsewhere, unstructured data repurposing needs are also growing in the video surveillance, government, and finance sectors. Meeting these needs requires consolidated management of data accumulated on a variety of different systems and common platforms that can handle repurposing across sectors (see Fig. 1).

Issues when Repurposing Data

Compared to structured data, the typical example of which is a database, the problems with repurposing unstructured data are that factors such as its size and quantity make processing slow, and that it is difficult to use for statistical processing or other analysis because its format is not machine-understandable like numerical data.

For example, platform technologies for massive data processing are required for tasks such as searching for similar images in large amounts of image data, using multiple computers to execute similarity calculations in parallel, or techniques for optimally positioning data on an HDD (hard disk drive)(7). Also, tasks such as similar case searches of the scanned images of medical records on a paper medium and statistical and other analytical processing require technologies able to extract information in a form that can be processed on a computer and convert it to structured data. Examples of this information include named entities and their attributes, or numerical data and its meaning, taken from test results, for example, and obtained by searching scanned images.

Accordingly, technologies for massive data processing and information extraction are likely to be important for implementing content repurposing platforms that support the repurposing of unstructured data.

EXISTING TECHNOLOGIES AND HITACHI'S ACTIVITIES

The following section describes Hitachi activities in the fields of massive data processing technology and information extraction technology for data repurposing.

Massive Data Processing Technology

The history of massive data processing, such as atmospheric phenomena simulations, has been driven by progress in supercomputers. Whereas supercomputers were built using special-purpose hardware until the 1990s, systems based on general-purpose CPUs (central processing units) and OSs (operating systems) predominated from 2000, and rapid progress has been made on technologies for building large-scale cluster systems by connecting together large numbers of conventional PC servers to process high volumes of data in realtime. The driving forces behind this technical innovation have been major net companies like Google Inc.*1, a provider of Internet search services, Amazon.com, Inc.*2, an online retail site operator, and Facebook, Inc.*3 which runs an SNS (social networking service) site.

In particular, Google attracted attention in 2004 with its announcement of the MapReduce(8) framework for parallel distributed processing of massive data on large-scale clusters.

MapReduce consists of “map” and “reduce” phases. In the map phase, the input data is split up into smaller units, and allocated to the individual machines in the cluster, and then processed in parallel. The map phase processing is implemented in a way that is independent of the processing on other machines. In the reduce phase, the outputs from the individual machines are collated to generate the final output. If the data required to produce the final output can be split into partial data sets that are not interdependent, the reduce phase also can be executed in parallel.

MapReduce was implemented as the Hadoop(9)*4 open source software as part of the Apache*4 Project and continues to be enhanced.

As Hadoop can be used to build massive data processing systems easily and cheaply, it is being applied in a wide range of fields.

Meanwhile, there is interest in stream data processing in which a stream of data is processed as it is generated, unlike the past stored data processing approach in which data is collated in an RDB (relational database) from which it can be read and processed as required. Stream data processing implements real-time processing of large quantities of data by executing the assigned processing as soon as data is received and outputting the result immediately.

Hitachi has been implementing advanced data analysis algorithms such as machine learning or frequent pattern recognition on MapReduce since 2008 and has embarked on research into KaaS (knowledge as a service) for extracting valuable information from large quantities of data. This work is being applied to maintenance and other fields(10). Hitachi also released a stream data processing platform in 2009 that allows processing to be specified simply using CQL (continuous query language), an extension of SQL (structured query language).

Information Extraction Technology

The creation of practical applications that process large quantities of text is becoming widespread in the world of the Internet. Google has released a trial version of Google Squared,(11)*1 which displays search results in tabular form. Unlike a conventional keyword search, this displays associated names and attributes under the search keywords in tabular form. A search for “cat” for example, displays names such as “American Shorthair” or “Persian” next to the images, text, or other results. These images or text are displayed as links to resources on the Internet. Another feature of the service is an interface that allows users to modify the displayed names or other attributes. Meanwhile, IBM*5 built the Watson*5 question and answer system, which competed on a popular US quiz program where it won the largest share of the prize money(12). Watson rapidly produces answers to complex questions on a wide range of subjects written in natural language.

A common feature of these technologies is that they collect information from different sources from which they obtain and apply practical knowledge. Whereas the primary objective of search engines in the past was to search web pages at high speed, it is anticipated that information extraction technologies for obtaining useful information from large volumes of text will be a key requirement in the future. However, technologies for extracting information from video, audio, and other multimedia data continue to pose many challenges.

Hitachi has worked on technologies for extracting information from text such as synonym extraction(13) and extraction of bibliographic information(14). Synonym extraction is a technique for automatically

*1 Google and Google Squared are trademarks of Google Inc.

*2 Amazon and the Amazon logo are trademarks of Amazon in the U.S. and/or other countries.

*3 Facebook and the Facebook logo are registered trademarks of Facebook, Inc.

*4 Apache Hadoop and Hadoop are trademarks of the Apache Software Foundation.

*5 IBM and Watson are trademarks or registered trademarks of International Business Machines Corporation in the USA and other countries.

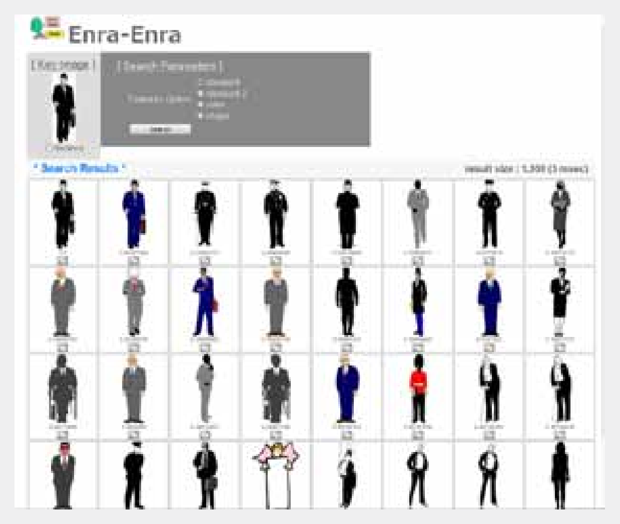

Fig. 2—The Similar Image Search Platform.
The similar image search platform extracts similarity information from images for use in searching. This is achieved by calculating features for each image and using these to determine the level of similarity among large numbers of images.

producing a thesaurus, which in the past would have been compiled manually. The accuracy of information extraction was improved by using existing thesaurus entries as "supervised data"(13). Bibliographic information extraction is a technique for improving the ease with which PDF (portable document format) documents can be searched by automatically extracting the title, author, and other metadata from the document(14). Hitachi is also developing methods for searching media data including high-speed similar image search(7) and voice search(15).

For the future, Hitachi intends to continue using these technologies in the research and development of information extraction from multimodal information with the aim of implementing solutions to be able to repurpose various different types of data (see Fig. 2).

CONTENT REPURPOSING PLATFORMS

To facilitate the repurposing of unstructured data, there is a need for content repurposing platforms that can be applied across specific sectors and which are able to process large quantities of data in a variety of formats at high speed in order to extract information in a structured format suitable for analysis by computer.

These platforms also need an environment in which it is possible to perform unified management of unstructured data in a range of formats spread across different organizations or communities in order to facilitate its utilization for a variety of purposes in conjunction with existing systems.

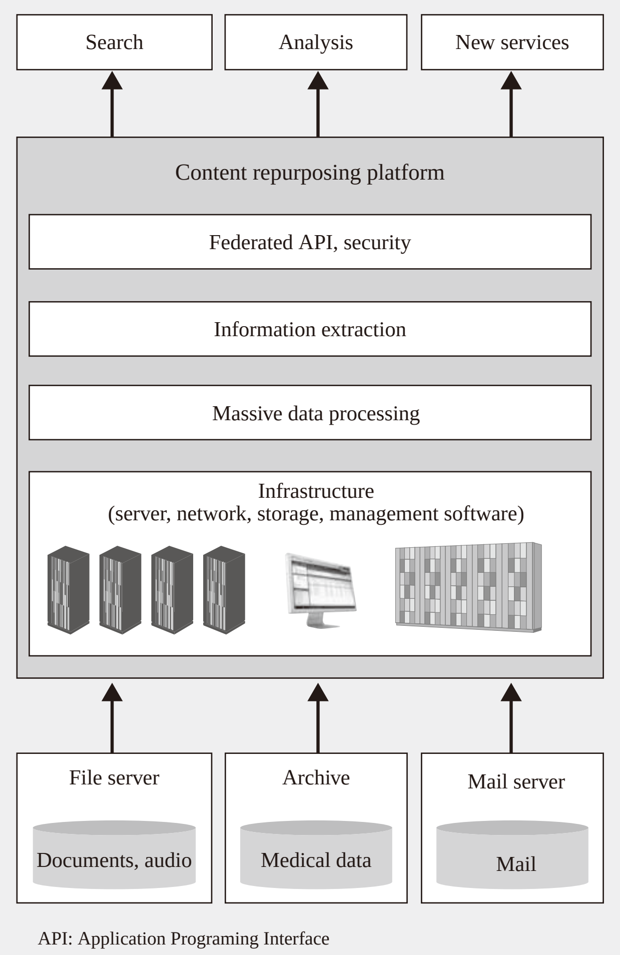

Fig. 3—Content Repurposing Platform.
The platform supports the unified management and application of unstructured data spread across multiple systems.

Hitachi intends to enhance further the technologies it has built up over many years, including storage systems and other platform technologies and technologies for massive data processing and media processing, and to proceed with research aimed at implementing these content repurposing platforms (see Fig. 3).

CONCLUSIONS

This article has looked at the trends and technical challenges associated with the repurposing needs of various different types of unstructured data collected by corporations, and summarized past research initiatives along with content repurposing platforms being devised by Hitachi.

With quantities of a wide range of different types of unstructured data, from multimedia to sensor information, increasing along with demands for applying this data, advances in technologies that can take the place of people in analyzing and handling this information will lead to the creation of systems with human-like recognition capabilities and encourage the realization of environments in which people and information systems can communicate in a more natural way. As a result, Hitachi believes that people being able to interact with information systems in a similar way to their human colleagues will allow the building of more creative and empathetic relationships with information systems and accelerate the realization of a "society of knowledge creation" overflowing with human values and benevolence. Hitachi intends to continue working on research and development aimed at achieving this goal.

REFERENCES

(1) Gartner, "Gartner Forecasts Global Business Intelligence Market to Grow 9.7 Percent in 2011," http://www.gartner.com/it/page.jsp?id=1553215

(2) Datacentrix, "5th SARMAF Seminar Non Proprietary (Open Source) vs Proprietary Software," July 2009.

(3) R. Watanabe et al., "Technology Development for New Trend of Health Information Management," Hitachi Hyoron **93**, pp. 292–297 (Mar. 2011) in Japanese.

(4) "Trial of Gaze Point Detection (Technology) for Marketing in Cigarette Vending Machines," http://www.hitachi.co.jp/Div/jkk/research/jt/ in Japanese.

(5) Data.gov, http://data.gov

(6) Facecount, "Academic Study Reveals Correlations of Stock Prices with Consumer Brand Fan," http://pressroom.blogs.pace.edu/2011/10/20/news-release-academic-study-reveals-correlations-of-stock-prices-with-consumer-brand-fan-counts/

(7) D. Matsubara et al., "High-Speed Similarity-Based Image Retrieval with Data-Alignment Optimization Using Self-Organization Algorithm," ISM2009.

(8) J. Dean et al., "MapReduce: Simplified Data Processing on Large Clusters," OSDI 2004.

(9) Apache Hadoop Project, http://hadoop.apache.org/

(10) R. Ueda et al., "KaaS Knowledge Service Platform Facilitating Innovation in Social Infrastructure," Hitachi Review **59**, pp. 224–228 (Dec. 2010).

(11) Google Squared, http://dces.essex.ac.uk/staff/udo/ecir2010/slides/ECIR_Industry_Day_2010_Crow.pdf

(12) "IBM's Watson Question Answering System Challenges Quiz Program!," http://www-06.ibm.com/ibm/jp/lead/ideasfromibm/watson/ in Japanese.

(13) Y. Morimoto et al., "Supervised Synonym Identification System Based on Contextual and Representational Similarity," Annual Conference of The Association for Natural Language Processing (Mar. 2010) in Japanese.

(14) M. Fujio et al., "Document Meta Extraction System Based on Layout Analysis," Conference of the Information Processing Society of Japan (Mar. 2010) in Japanese.

(15) N. Kanda et al., "Open-vocabulary Keyword Detection from Super-large Scale Speech Database," MMSP, 2008.

ABOUT THE AUTHORS

Shoji Kodama

Joined Hitachi, Ltd. in 1998, and now works at the Software Platform Research Department, Information Platform Research Center, Yokohama Research Laboratory. He is currently engaged in the research and development of contents re-purposing platform. Mr. Kodama is a member of the Information Processing Society of Japan (IPSJ).

Nobuo Nukaga

Joined Hitachi, Ltd. in 1994, and now works at the Intelligent Media Systems Research Department, Information Systems Research Center, Central Research Laboratory. He is currently engaged in the research and development of a speech processing system. Mr. Nukaga is a member of The Japanese Society for Artificial Intelligence (JSAI), Information Processing Society of Japan (IPSJ), The Institute of Electronics, Information and Communication Engineers (IEICE).

Ryoichi Ueda

Joined Hitachi, Ltd. in 1994, and now works at the Software Platform Research Department, Information Platform Research Center, Yokohama Research Laboratory. He is currently engaged in the research and development of massive data processing systems.

Shinya Iguchi

Joined Hitachi, Ltd. in 1998, and now works at the Software Platform Research Department, Information Platform Research Center, Yokohama Research Laboratory. He is currently engaged in the research and development of contents re-purposing platform. Mr. Iguchi is a member of the Information Processing Society of Japan (IPSJ).