МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«Гомельский государственный технический университет имени П.О. Сухого»

КАФЕДРА «Белорусский и иностранный язык»

РЕФЕРАТ

на тему

БИЗНЕС-АНАЛИЗ И АНАЛИТИКА: ИЗ БОЛЬШИХ ДАННЫХ К БОЛЬШОМУ РЕЗУЛЬТАТУ

подготовленный для прохождения итоговой аттестации по общеобразовательной дисциплине «Основы информационных технологи»

Выполнил:
магистрант гр. МАГ 40-12 специальности 1–40 80 04 «Математическое моделирование, численные методы и комплексы программ»
Бурим Илья Павлович

Проверил:
старший преподаватель
Войтишенюк Е.В.

Гомель 2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Бизнес-анализ и аналитика (BI&A) и связанные с ними поле аналитики больших данных становится все более важным как в академических, так и в деловых сферах последние два десятилетия. Отраслевые исследования подчеркнули это значительное развитие. Например, на основе опроса более 4000 специалистов в области информационных технологий (ИТ) из 93 стран и 25 отраслей промышленности, отчет IBM Tech Trends (2011) определили бизнес-аналитику как одну из четырех основных технологических тенденций в 2010-м. В обзоре состояния бизнес-аналитика от Bloomberg Businessweek (2011), 97 процентов компаний с доходом, превышающим 100 миллионов долларов США, как было установлено, используют некоторую форму бизнес-аналитики. Отчет Глобальным институтом McKinsey (Manyika et al., 2011) предсказано что к 2018 году только Соединенные Штаты столкнутся с нехваткой от 140 000 до 190 000 человек с глубокими аналитическими навыками, а также дефицит 1,5 млн. менеджеров ориентированных на данные, с новыми идеями для анализа больших данных для принятия эффективных решений. Возможности, связанные с данными и анализом в разных организациях помогли создать значительный интерес в BI&A, который часто называют техникой, технологиями, системами, методами, методологиями и приложениями которые анализируют критически важные бизнес-данные, чтобы помочь предприятию лучше понимать свой бизнес и рынок и своевременно внести бизнес решения. В дополнение к базовой обработке данных и аналитические технологий, BI&A включает бизнес-ориентированные практики и методологии, которые могут применяться к различным сферам как электронная коммерция, рыночная разведка, электронное правительство, здравоохранение и безопасность.

АННОТАЦИЯ

Данный реферат рассказывает о исследованиям Business Intelligence Research. Реферат дает обзор этой захватывающей и успешной области, подчеркивая её многочисленные проблемы и возможности. В статье рассказаны ключевые направления, включая эволюцию BI&A, приложения и новые возможности исследований аналитики. Затем идет повествование по сферам применения BI&A завязаных на исследовании критических BI&A публикаций, исследователей и исследовательских тем, основанных на более чем десятилетии связанных научных и отраслевых публикаций BI. Образование и возможности разработки программ в BI&A.

ГЛАВА 1

BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT

1.1 BI&A 1.0

In the first place the term intelligencehas been used by researchers in artificial intelligence since the 1950s. Business intelligence became a popular term in the business and IT communities only in the 1990s. In the late 2000s, business analyticswas introduced to represent the key analytical component in BI [**?**]. More recently big dataand big data analyticshave been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data

As a data-centric approach, BI&A has its roots in the longstanding database management field. It relies heavily on various data collection, extraction, and analysis technologies (Chaudhuri et al. 2011; Turban et al. 2008; Watson and Wixom 2007). The BI&A technologies and applications currently adopted in industry can be considered as BI&A 1.0, where data are mostly structured, collected by companies through various legacy systems, and often stored in commercial relational database management systems (RDBMS). The analytical techniques commonly used in these systems, popularized in the 1990s, are grounded mainly in statistical methods developed in the 1970s and data mining techniques developed in the 1980s.

Data management and warehousing is considered the foundation of BI&A 1.0. Design of data marts and tools for extraction, transformation, and load (ETL) are essential for converting and integrating enterprise-specific data. Database query, online analytical processing (OLAP), and reporting tools based on intuitive, but simple, graphics are used to explore important data characteristics. Business performance management (BPM) using scorecards and dashboards help analyze and visualize a variety of performance metrics. In addition to these well-established business reporting functions, statistical analysis and data mining techniques are adopted for association analysis, data segmentation and clustering, classification and regression analysis, anomaly detection, and predictive modeling in various business applications. Most of these data processing and analytical technologies have already been incorporated into the leading commercial BI platforms offered by major IT

vendors including Microsoft, IBM, Oracle, and SAP (Sallam et al. 2011).

Among the 13 capabilities considered essential for BI platforms, according to the Gartner report by Sallam et al. (2011), the following eight are considered BI&A 1.0: reporting, dashboards, ad hocquery, search-based BI, OLAP, interactive visualization, scorecards, predictive modeling, and data mining. A few BI&A 1.0 areas are still under active development based on the Gartner BI Hype Cycle analysis for emerging BI technologies, which include data mining workbenchs, column-based DBMS, in-memory DBMS, and realtime decision tools (Bitterer 2011). Academic curricula in Information Systems (IS) and Computer Science (CS) often include well-structured courses such as database management systems, data mining, and multivariate statistics.

In conclusion the opportunities associated with data and analysis in different organizations have helped generate significant interest in BI&A, which is often referred to as the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions.

1.2 BI&A 2.0

This chapter presents the Internet and the Web. These technologies began to offer unique data collection and analytical research and development opportunities. In 2000s the HTTP-based Web 1.0 systems, characterized by Web search engines such as Google and Yahoo and e-commerce businesses such as Amazon and eBay, allow organizations to present their businesses online and interact with their customers directly. In addition to porting their traditional RDBMS-based product information and business contents online, detailed and IP-specific user search and interaction logs that are collected seamlessly through cookies and server logs have become a new gold mine for understanding customers' needs and identifying new business opportunities. Web intelligence, web analytics, and the user-generated content collected through Web 2.0-based social and crowd-sourcing systems (Doan et al. 2011; O'Reilly 2005) have ushered in a new and exciting era of BI&A 2.0 research in the 2000s, centered on text and web analytics for unstructured web contents.

An immense amount of company, industry, product, and customer information can be gathered from the web and organized and visualized through various text and web mining techniques. By analyzing customer clickstream data logs, web analytics tools such as Google Analytics can provide a trail of the user's online activities and reveal the user's browsing and purchasing patterns. Web site design, product placement optimization, customer transaction analysis, market structure analysis, and product

recommendations can be accomplished through web analytics. The many Web 2.0 applications developed after 2004 have also created an abundance of user-generated content from various online social media such as forums, online groups, web blogs, social networking sites, social multimedia sites (for photos and videos), and even virtual worlds and social games (O'Reilly 2005). In addition to capturing celebrity chatter, references to everyday events, and socio-political sentiments expressed in these media, Web 2.0 applications can efficiently gather a large volume of timely feedback and opinions from a diverse customer population for different types of businesses.

Many marketing researchers believe that social media analytics presents a unique opportunity for businesses to treat the market as a "conversation" between businesses and customers instead of the traditional business-to-customer, one-way "marketing" (Lusch et al. 2010). Unlike BI&A 1.0 technologies that are already integrated into commercial enterprise IT systems, future BI&A 2.0 systems will require the integration of mature and scalable techniques in text mining (e.g., information extraction, topic identification, opinion mining, question-answering), web mining, social network analysis, and spatial-temporal analysis with existing DBMS-based BI&A 1.0 systems.

Except for basic query and search capabilities, no advanced text analytics for unstructured content are currently considered in the 13 capabilities of the Gartner BI platforms. Several, however, are listed in the Gartner BI Hype Cycle, including information semantic services, natural language question answering, and content/text analytics (Bitterer 2011). New IS and CS courses in text mining and web mining have emerged to address needed technical training.

In conclusion it can be notified that in 2000s, BI&A created big jump in infrastructure natural language, information extraction, topic identification, opinion mining, question-answering. And created new directions in IT.

1.3 BI&A 3.0

This chapter describes most of the academic research on mobile BI, opening up exciting new steams of innovative applications and describes business intelligence and analytics in Web 3.0 area.

Whereas web-based BI&A 2.0 has attracted active research from academia and industry, a new research opportunity in BI&A 3.0 is emerging. As reported prominently in an October 2011 article in The Economist (2011), the number of mobile phones and tablets (about 480 million units) surpassed the number of laptops and PCs (about 380 million units) for the first time in 2011. Although the number of PCs in use surpassed 1 billion in 2008, the same article projected that the number of mobile connected devices would reach

10 billion in 2020. Mobile devices such as the iPad, iPhone, and other smart phones and their complete ecosystems of downloadable applicationss, from travel advisories to multi-player games, are transforming different facets of society, from education to healthcare and from entertainment to governments. Other sensor-based Internet-enabled devices equipped with RFID, barcodes, and radio tags (the "Internet of Things") are opening up exciting new steams of innovative applications. The ability of such mobile and Internet-enabled devices to support highly mobile, location-aware, person-centered, and context-relevant operations and transactions will continue to offer unique research challenges and opportunities throughout the 2010s. Mobile interface, visualization, and HCI (human–computer interaction) design are also promising research areas. Although the coming of the Web 3.0 (mobile and sensor-based) era seems certain, the underlying mobile analytics and location and context-aware techniques for collecting, processing, analyzing and visualizing such largescale and fluid mobile and sensor data are still unknown.

No integrated, commercial BI&A 3.0 systems are foreseen for the near future. Most of the academic research on mobile BI is still in an embryonic stage. Although not included in the current BI platform core capabilities, mobile BI has been included in the Gartner BI Hype Cycle analysis as one of the new technologies that has the potential to disrupt the BI market significantly (Bitterer 2011). The uncertainty associated with BI&A 3.0 presents another unique research direction for the IS community. Table 1 summarizes the key characteristics of BI&A 1.0, 2.0, and 3.0 in relation to the Gartner BI platforms core capabilities and hype cycle.

In conclusion it can be notified that the decade of the 2010s was an exciting one for high-impact BI&A research and development for both industry and academia. IS research and education programs need to carefully evaluate future directions, curricula, and action plans, from BI&A 1.0 to 3.0. The business community and industry have already taken important steps to adopt BI&A for their needs. The IS community faces unique challenges and opportunities in making scientific and societal impacts that are relevant and long-lasting (Chen 2011a).

1.4 BI&A Applications: From Big Data to Big Impact

This chapter describes new streams where business intelligence and analytics big data will be used. Streams like international travel, high-speed network connections, global supply-chain, and outsourcing have created a tremendous opportunity for IT advancement. It predicted by Thomas Freeman in his seminal book, The World is Flat (2005).

Several global business and IT trends have helped shape past and present BI&A research directions. In addition to ultra-fast global IT connections, the development and deployment of business-related data standards, electronic data interchange (EDI) formats, and business databases and information systems have greatly facilitated business data creation and utilization. The development of the Internet in the 1970s and the subsequent large-scale adoption of the World Wide Web since the 1990s have increased business data generation and collection speeds exponentially. Recently, the Big Data era has quietly descended on many communities, from governments and e-commerce to health organizations. With an overwhelming amount of web-based, mobile, and sensorgenerated data arriving at a terabyte and even exabyte scale (The Economist 2010a, 2010b), new science, discovery, and insights can be obtained from the highly detailed, contextualized, and rich contents of relevance to any business or organization.

In addition to being data driven, BI&A is highly applied and can leverage opportunities presented by the abundant data and domain-specific analytics needed in many critical and highimpact application areas. Several of these promising and high-impact BI&A applications are presented below, with a discussion of the data and analytics characteristics, potential impacts, and selected illustrative examples or studies: (1) ecommerce and market intelligence, (2) e-government and politics 2.0, (3) science and technology, (4) smart health and well-being, and (5) security and public safety. By carefully analyzing the application and data characteristics, researchers and practitioners can then adopt or develop the appropriate analytical techniques to derive the intended impact. IS departments thus face unique opportunities and challenges in developing integrated BI&A research and education programs for the new generation of data/analyticssavvy and business-relevant students and professionals (Chen 2011a).

In conclusion to technical system implementation, significant business or domain knowledge as well as effective communication skills are needed for the successful completion of such BI&A projects.

1.5 E-Commerce and Market Intelligence

In addition new impacts and development directions in business intelligence and analytics big data for e-commerce organizations.

The excitement surrounding BI&A and Big Data has arguably been generated primarily from the web and e-commerce communities. Significant market transformation has been accomplished by leading e-commerce vendors such Amazon and eBay through their innovative and highly scalable ecommerce

platforms and product recommender systems. Major Internet firms such as Google, Amazon, and Facebook continue to lead the development of web analytics, cloud computing, and social media platforms. The emergence of customer-generated Web 2.0 content on various forums, newsgroups, social media platforms, and crowd-sourcing systems offers another opportunity for researchers and practitioners to "listen" to the voice of the market from a vast number of business constituents that includes customers, employees, investors, and the media (Doan et al. 2011; O'Rielly 2005). Unlike traditional transaction records collected from various legacy systems of the 1980s, the data that e-commerce systems collect from the web are less structured and often contain rich customer opinion and behavioral information. For social media analytics of customer opinions, text analysis and sentiment analysis techniques are frequently adopted (Pang and Lee 2008). Various analytical techniques have also been developed for product recommender systems, such as association rule mining, database segmentation and clustering, anomaly detection, and graph mining (Adomavicius and Tuzhilin 2005). Long-tail marketing accomplished by reaching the millions of niche markets at the shallow end of the product bitstream has become possible via highly targeted searches and personalized recommendations (Anderson 2004). The Netfix Prize competition for the best collaborative filtering algorithm to predict user movie ratings helped generate significant academic and industry interest in recommender systems development and resulted in awarding the grand prize of $1 million to the Bellkor's Pragmatic Chaos team, which surpassed Netflix's own algorithm for predicting ratings by 10.06 percent. However, the publicity associated with the competition also raised major unintended customer privacy concerns.

In conclusion much BI&A-related e-commerce research and development information is appearing in academic IS and CS papers as well as in popular IT magazines.

1.6 E-Government and Politics 2.0

This chapter describes cahnges e-commerce area with coming Web 2.0 and with coming new technologies in BI&A research. The advent of Web 2.0 has generated much excitement for reinventing governments.

The 2008 U.S. House, Senate, and presidential elections provided the first signs of success for online campaigning and political participation. Dubbed «politics 2.0», politicians use the highly participatory and multimedia web platforms for successful policy discussions, campaign advertising, voter mobilization, event announcements, and online donations. As government and political processes become more transparent, participatory, online, and

multimedia-rich, there is a great opportunity for adopting BI&A research in e-government and politics 2.0 applications. Selected opinion mining, social network analysis, and social media analytics techniques can be used to support online political participation, e-democracy, political blogs and forums analysis, e-government service delivery, and process transparency and accountability (Chen 2009; Chen et al. 2007). For e-government applications, semantic information directory and ontological development (as exemplified below) can also be developed to better serve their target citizens.

Despite the significant transformational potential for BI&A in e-government research, there has been less academic research than, for example, e-commerce-related BI&A research. Egovernment research often involves researchers from political science and public policy. For example, Karpf (2009) analyzed the growth of the political blogosphere in the United States and found significant innovation of existing political institutions in adopting blogging platforms into their Web offerings. In his research, 2D blogspace mapping with composite rankings helped reveal the partisan makeup of the American political blogsphere. Yang and Callan (2009) demonstrated the value for ontology development for government services through their development of the OntoCop system, which works interactively with a user to organize and summarize online public comments from citizens.

In conclusion it can be notified that e-commerce area will be grow with BI&A science. Also it can be notified that more important for e-commerce will be the BI&A in e-government research and e-commerce-related BI&A research directions.

1.7 Science and Technology

This chapter describes using business intelligence and analytics big data for science. Describes tools and algorithms for successful opening of new achievements and researchs.

Many areas of science and technology (S&T) are reaping the benefits of high-throughput sensors and instruments, from astrophysics and oceanography, to genomics and environmental research. To facilitate information sharing and data analytics, the National Science Foundation (NSF) recently mandated that every project is required to provide a data management plan. Cyber-infrastructure, in particular, has become critical for supporting such data-sharing initiatives.

The 2012 NSF BIGDATA program solicitation is an obvious example of the U.S. government funding agency's concerted efforts to promote big data analytics. The program aims to advance the core scientific and technological

means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life.

Several S&T disciplines have already begun their journey toward big data analytics. For example, in biology, the NSF funded iPlant Collaborative is using cyberinfrastructure to support a community of researchers, educators, and students working in plant sciences. iPlant is intended to foster a new generation of biologists equipped to harness rapidly expanding computational techniques and growing data sets to address the grand challenges of plant biology. The iPlant data set is diverse and includes canonical or reference data, experimental data, simulation and model data, observational data, and other derived data. It also offers various open source data processing and analytics tools.

In astronomy, the Sloan Digital Sky Survey (SDSS) shows how computational methods and big data can support and facilitate sense making and decision making at both the macroscopic and the microscopic level in a rapidly growing and globalized research field. The SDSS is one of the most ambitious and influential surveys in the history of astronomy. Over its eight years of operation, it has obtained deep, multicolor images covering more than a quarter of the sky and created three-dimensional maps containing more than 930,000 galaxies and over 120,000 quasars. Continuing to gather data at a rate of 200 gigabytes per night, SDSS has amassed more than 140 terabytes of data. The international Large Hadron Collider (LHC) effort for high-energy physics is another example of big data, producing about 13 petabytes of data in a year (Brumfiel 2011).

In conclusion it can be notified that business intelligence and analytics big data contributed to the big jump in astronomy, physics high-energy. The relationship between BI&A and science contributed to delevopment of new data processing and analytics tools, algorithms.

ГЛАВА 2

Translate

2.1 Smart Health and Wellbeing

Much like the big data opportunities facing the e-commerce and S&T communities, the health community is facing a tsunami of health- and healthcare-related content generated from numerous patient care points of contact, sophisticated medical instruments, and web-based health communities. Two main sources of health big data are genomics-driven big data (genotyping, gene expression, sequencing data) and payer–provider big data (electronic health records, insurance records, pharmacy prescription, patient feedback and responses) [1]. The expected raw sequencing data from each person is approximately four terabytes. From the payer–provider side, a data matrix might have hundreds of thousands of patients with many records and parameters (demographics, medications, outcomes) collected over a long period of time. Extracting knowledge from health big data poses significant research and practical challenges, especially considering the HIPAA (Health Insurance Portability and Accountability Act) and IRB (Institutional Review Board) requirements for building a privacy-preserving and trustworthy health infrastructure and conducting ethical healthrelated research [2]. Health big data analytics, in general, lags behind e-commerce BI&A applications because it has rarely taken advantage of scalable analytical methods or computational platforms [1].

Over the past decade, electronic health records (EHR) have been widely adopted in hospitals and clinics worldwide. Significant clinical knowledge and a deeper understanding of patient disease patterns can be gleanded from such collections (Hanauer et al. 2009; Hanauer et al. 2011; Lin et al. 2011). Hanauer et al. (2011), for example, used large-scale, longitudinal EHR to research associations in medical diagnoses and consider temporal relations between events to better elucidate patterns of disease progression [3]. used symptom–disease–treatment (SDT) association rule mining on a comprehensive EHR of approximately 2.1 million records from a major hospital. Based on selected International Classification of Diseases (ICD-9) codes, they were able to identify clinically relevant and accurate SDT associations from patient records in seven distinct diseases, ranging from cancers to chronic and infectious diseases.

In addition to EHR, health social media sites such as Daily Strength and PatientsLikeMe provide unique research opportunities in healthcare decision support and patient empowerment (Miller 2012b), especially for chronic diseases such as diabetes, Parkinson's, Alzheimer's, and cancer. Association rule mining and clustering, health social media monitoring and analysis, health text analytics, health ontologies, patient network analysis, and adverse drug side-effect analysis are promising areas of research in health-related BI&A. Due to the importance of HIPAA regulations, privacy-preserving health data mining is also gaining attention [2].

Partially funded by the National Institutes of Health (NIH), the NSF BIGDATA program solicitation includes common interests in big data across NSF and NIH. Clinical decision making, patient-centered therapy, and knowledge bases for health, disease, genome, and environment are some of the areas in which BI&A techniques can contribute [6] [5]. Another recent, major NSF initiative related to health big data analytics is the NSF Smart Health and Wellbeing (SHB)6 program, which seeks to address fundamental technical and scientific issues that would support a much-needed transformation of healthcare from reactive and hospital-centered to preventive, proactive, evidence-based, person-centered, and focused on wellbeing rather than disease control. The SHB research topics include sensor technology, networking, information and machine learning technology, modeling cognitive processes, system and process modeling, and social and economic issues [5], most of which are relevant to healthcare BI&A.

2.2 Security and Public Safety

Since the tragic events of September 11, 2001, security research has gained much attention, especially given the increasing dependency of business and our global society on digital enablement. Researchers in computational science, information systems, social sciences, engineering, medicine, and many other fields have been called upon to help enhance our ability to fight violence, terrorism, cyber crimes, and other cyber security concerns. Critical mission areas have been identified where information technology can contribute, as suggested in the U.S. Office of Homeland Security's report «National Strategy for Homeland Security», released in 2002, including intelligence and warning, border and transportation security, domestic counter-terrorism, protecting critical infrastructure (including cyberspace), defending against catastrophic terrorism, and emergency preparedness and response. Facing the critical missions of international security and various data and technical challenges, the need to develop the science of "security informatics" was recognized, with

its main objective being the development of advanced information technologies, systems, algorithms, and databases for securityrelated applications, through an integrated technological, organizational, and policy-based approach [7].

BI&A has much to contribute to the emerging field of security informatics.

Security issues are a major concern for most organizations. According to the research firm International Data Corporation, large companies are expected to spend $32.8 billion in computer security in 2012, and small- and medium-size companies will spend more on security than on other IT purchases over the next three years [8]. In academia, several security-related disciplines such as computer security, computational criminology, and terrorism informatics are also flourishing [9].

Intelligence, security, and public safety agencies are gathering large amounts of data from multiple sources, from criminal records of terrorism incidents, and from cyber security threats to multilingual open-source intelligence. Companies of different sizes are facing the daunting task of defending against cybersecurity threats and protecting their intellectual assets and infrastructure. Processing and analyzing security-related data, however, is increasingly difficult. A significant challenge in security IT research is the information stovepipe and overload resulting from diverse data sources, multiple data formats, and large data volumes. Current research on technologies for cybersecurity, counter-terrorism, and crimefighting applications lacks a consistent framework for addressing these data challenges. Selected BI&A technologies such as criminal association rule mining and clustering, criminal network analysis, spatial-temporal analysis and visualization, multilingual text analytics, sentiment and affect analysis, and cyber attacks analysis and attribution should be considered for security informatics research. The University of Arizona's COPLINK and Dark Web research programs offer significant examples of crime data mining and terrorism informatics within the IS community [7]. The COPLINK information sharing and crime data mining system, initially developed with funding from NSF and the Department of Justice, is currently in use by more than 4,500 police agencies in the United States and by 25 NATO countries, and was acquired by IBM in 2011. The Dark Web research, funded by NSF and the Department of Defense (DOD), has generated one of the largest known academic terrorism research databases (about 20 terabytes of terrorist web sites and social media content) and generated advanced multilingual social media analytics techniques. Recognizing the challenges presented by the volume and complexity of defense-related big data, the U.S. Defense Advanced Research Project Agency (DARPA) within DOD initiated the XDATA program in 2012 to help develop computational techniques and software tools for processing and analyzing the vast amount of mission-oriented information for defense activities.

XDATA aims to address the need for scalable algorithms for processing and visualization of imperfect and incomplete data. The program engages applied mathematics, computer science, and data visualization communities to develop big data analytics and usability solutions for warfighters.7 BI&A researchers could contribute significantly in this area.

2.3 Text Analytics

A significant portion of the unstructured content collected by an organization is in textual format, from e-mail communication and corporate documents to web pages and social media content. Text analytics has its academic roots in information retrieval and computational linguistics. In information retrieval, document representation and query processing are the foundations for developing the vector-space model, Boolean retrieval model, and probabilistic retrieval model, which in turn, became the basis for the modern digital libraries, search engines, and enterprise search systems (Salton 1989). In computational linguistics, statistical natural language processing (NLP) techniques for lexical acquisition, word sense disambiguation, part-of-speech-tagging (POST), and probabilistic context-free grammars have also become important for representing text [14]. In addition to document and query representations, user models and relevance feedback are also important in enhancing search performance. Since the early 1990s, search engines have evolved into mature commercial systems, consisting of fast, distributed crawling; efficient inverted indexing; inlink-based page ranking; and search logs analytics. Many of these foundational text processing and indexing techniques have been deployed in text-based enterprise search and document management systems in BI&A 1.0.

Leveraging the power of big data (for training) and statistical NLP (for building language models), text analytics techniques have been actively pursued in several emerging areas, including information extraction, topic models, questionanswering (Q/A), and opinion mining. Information extraction is an area of research that aims to automatically extract specific kinds of structured information from documents. As a building block of information extraction, NER (named entity recognition, also known as entity extraction) is a process that identifies atomic elements in text and classifies them into predefined categories (e.g., names, places, dates). NER techniques have been successfully developed for news analysis and biomedical applications. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. New topic modeling algorithms such as LDA (latent Dirichlet allocation) and other probabilistic models have attracted recent research [11]. Question answering (Q/A) systems rely on techniques

from NLP, information retrieval, and human–computer interaction. Primarily designed to answer factual questions (i.e., who, what, when, and where kinds of questions), Q/A systems involve different techniques for question analysis, source retrieval, answer extraction, and answer presentation [12]. The recent successes of IBM's Watson and Apple's Siri have highlighted Q/A research and commercialization opportunities. Many promising Q/A system application areas have been identified, including education, health, and defense. Opinion mining refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated contents. Sentiment analysis is often used in opinion mining to identify sentiment, affect, subjectivity, and other emotional states in online text. Web 2.0 and social media content have created abundant and exciting opportunities for understanding the opinions of the general public and consumers regarding social events, political movements, company strategies, marketing campaigns, and product preferences [13].

In addition to the above research directions, text analytics also offers significant research opportunities and challenges in several more focused areas, including web stylometric analysis for authorship attribution, multilingual analysis for web documents, and large-scale text visualization. Multimedia information retrieval and mobile information retrieval are two other related areas that require support of text analytics techniques, in addition to the core multimedia and mobile technologies. Similar to big data analytics, text analytics using MapReduce, Hadoop, and cloud services will continue to foster active research directions in both academia and industry.

2.4 Web Analytics

Over the past decade, web analytics has emerged as an active field of research within BI&A. Building on the data mining and statistical analysis foundations of data analytics and on the information retrieval and NLP models in text analytics, web analytics offers unique analytical challenges and opportunities. HTTP/HTML-based hyperlinked web sites and associated web search engines and directory systems for locating web content have helped develop unique Internetbased technologies for web site crawling/spidering, web page updating, web site ranking, and search log analysis. Web log analysis based on customer transactions has subsequently turned into active research in recommender systems. However, web analytics has become even more exciting with the maturity and popularity of web services and Web 2.0 systems in the mid-2000s [15].

Based on XML and Internet protocols (HTTP, SMTP), web services offer a new way of reusing and integrating third party or legacy systems. New types of web services and their associated APIs (application programming interface) allow developers to easily integrate diverse content from different web-enabled system, for example, REST (representational state transfer) for invoking remote services, RSS (really simple syndication) for news "pushing," JSON (JavaScript object notation) for lightweight data-interchange, and AJAX (asynchronous JavaScript + XML) for data interchange and dynamic display. Such lightweight programming models support data syndication and notification and "mashups" of multimedia content (e.g., Flickr, Youtube, Google Maps) from different web sources—a process somewhat similar to ETL (extraction, transformation, and load) in BI&A 1.0. Most of the e-commerce vendors have provided mature APIs for accessing their product and customer content [16]. For example, through Amazon Web Services, developers can access product catalog, customer reviews, site ranking, historical pricing, and the Amazon Elastic Compute Cloud (EC2) for computing capacity. Similarly, Google web APIs support AJAX search, Map API, GData API (for Calendar, Gmail, etc.), Google Translate, and Google App Engine for cloud computing resources. Web services and APIs continue to provide an exciting stream of new data sources for BI&A 2.0 research.

A major emerging component in web analytics research is the development of cloud computing platforms and services, which include applications, system software, and hardware delivered as services over the Internet. Based on serviceoriented architecture (SOA), server virtualization, and utility computing, cloud computing can be offered as software as a service (SaaS), infrastructure as a service (IaaS), or platform as a service (PaaS). Only a few leading IT vendors are currently positioned to support high-end, high-throughput BI&A applications using cloud computing. For example, Amazon Elastic Compute Cloud (EC2) enables users to rent virtual computers on which to run their own computer applications. Its Simple Storage Service (S3) provides online storage web service. Google App Engine provides a platform for developing and hosting Java or Python-based web applications. Google Bigtable is used for backend data storage. Microsoft's Windows Azure platform provides cloud services such as SQL Azure and SharePoint, and allows .Net framework applications to run on the platform. The industry-led web and cloud services offer unique data collection, processing, and analytics challenges for BI&A researchers.

In academia, current web analytics related research encompasses social search and mining, reputation systems, social media analytics, and web visualization. In addition, webbased auctions, Internet monetization, social marketing, and web privacy/security are some of the promising research directions related to web analytics. Many of these emerging research areas may

rely on advances in social network analysis, text analytics, and even economics modeling research.

2.5 Smart Health and Wellbeing

Подобно большим возможностям данных, с которыми сталкивается электронная коммерция и сообщество S&T,так же и сообщество здравоохранения сталкивается с цунами медицинских и медицински связаннхы данных из многочисленных пунктов обслуживания пациентов, сложных медицинских инструментов и веб-сообществ здравоохранения. Двумя основными источниками медицинских данных являются большие геномики данныех (генотипирование, экспрессия генов, данные секвенирования) и большие данные от клиентов (электронные медицинские записи, страховка записи, рецепт аптеки, отзывы пациентов и ответы) [1]. Ожидаемые необработанные данные от каждого человека - около четырех терабайт. Со стороны получатель-поставщик, матрица данных может иметь сотни тысячи пациентов со многими записями и параметрами (демография, медикаменты, результаты), собранные в течение длительного период времени. Извлечение знаний из больших данных о здоровье ставит значительные исследовательские и практические задачи, особенно учитывая HIPAA (переносимость медицинского страхования и Закон о подотчетности) и IRB (Комиссия по институциональному обзору) требования для создания конфиденциальности и надежности инфраструктуры здравоохранения и обеспечения этического медицинского исследования [2]. Аналитика больших медицинских данных, в целом, отстает от приложений электронной коммерции BI&A потому что они редко используют преимущества масштабируемых аналитических методов или вычислительных платформ [1].

За последнее десятилетие электронные медицинские записи (EHR) широко применяется в больницах и клиниках по всему миру. Значительные клинические знания и более глубокое понимание образцы болезни пациента могут быть оценены из таких коллекций (Hanauer et al., 2009; Hanauer et al., 2011; Lin et al., 2011). Hanauer et al. (2011), например, использовали крупномасштабные продольные EHR для исследований ассоциаций в медицинских диагнозах и рассмотреть временные отношения между событиями к лучшему выявить закономерности прогрессирования заболевания [3]. используется правило ассоциации с симптомами болезни (SDT) Получение полезных данных на всесторонних EHR записях приблизительно 2.1 миллион записей из крупной больницы. На основе выбранных Кодов международной классификации болезней (ICD-9), они смогли идентифицировать клинически релевантные и точные SDT ассоциации из записей пациентов

в семи различных заболеваниях, от рака до хронических и инфекционных заболеваний.

В дополнение к EHR, сайты социальных сетей здравоохранения, такие как Daily Strength и PatientsLikeMe предоставляют уникальные возможности для исследований в поддержке принятия решений в области здравоохранения и расширении прав и возможностей пациентов [4], особенно для хронических заболеваний, таких как диабета, болезни Паркинсона, болезни Альцгеймера и рака. Разработка общих правил и кластеризация, мониторинг социальных сетей и анализ, аналитика по медицинским текстам, медицинские онтологии, анализ связей пациента и побочных эффектов наркотиков это перспективные направления исследований в области здравоохранения и бизнеса. Из-за важность правил HIPAA, сохранение конфиденциальности получение полезных данных также привлекает внимание [2].

Частично финансируемый Национальными институтами здравоохранения (NIH), программа NSF BIGDATA включает в себя общие интересы в больших данных через NSF и NIH. Клиническое решение терапии, ориентированной на пациента, и базы знаний для здоровья, болезней, геномов и окружающей среды является одним из области, в которых BI&A методы могут помочь [6] [5]. Еще одна недавняя инициатива NSF связанная со здоровьем большая аналитика данных - это NSF Smart Health and Wellbeing (SHB) 6 программ, в рамках которых фундаментальные технические и научные вопросы, которые столь необходимы для трансформации здравоохранения из реактивных и ориентированных на медицинских данных на профилактические, профилактические, основанные на фактических данных, ориентированных на человека и ориентированных скорее на благополучие, чем на контроль болезний. Темы исследований SHB включают сенсорную технологию, сети, информацию и технологии машинного обучения, моделирование когнитивных процессов, моделирование систем и процессов, и социально-экономические вопросы [5], большая часть которые имеют отношение к здравоохранению BI&A.

2.6 Security and Public Safety

После трагических событий 11 сентября 2001 года безопасность исследования получили много внимания, особенно учитывая растущей зависимости бизнеса и нашего глобального общества от цифровое включение. Исследователи вычислительной науки, информационные системы, социальные науки, инженерия, медицина, и многие другие области призваны помочь нашей способности бороться с насилием, терроризмом, киберпреступностью и другими проблемы с кибербезопасностью. Критические места ин-

формационных технологий были обозначены в докладе Управления национальной безопасности США « Национальная стратегия государственной безопасности », выпущенная в 2002 году, включая разведку и обнаружения, пограничную и транспортную безопасность, внутренняя борьба с терроризмом, защита критической инфраструктуры (включая киберпространство), защищая от катастроф терроризма, готовности к чрезвычайным ситуациям и реагирование. Встали критические цели для международной безопасности и различных данныех и технических проблем, необходимость разработки науки о «информационной безопасности» чья главная задача является развитие передовых информационных технологий, систем, алгоритмов и баз данных для приложений, связанных с безопасностью, с помощью интегрированных технологических, организационный и политический подходов [7].

BI&A может многое внести в новую область инфармационной безопасности.

Вопросы безопасности являются серьезной проблемой для большинства организаций. По данным исследовательской фирмы International Data Corporation, ожидается, что крупные компании потратят $ 32,8 млрд. компьютерной безопасности в 2012 году, а также малых и средних компании будут тратить больше средств на безопасность, чем на другие ИТ покупки в течение следующих трех лет [8]. В академических кругах некоторые связанные с безопасностью дисциплины, такие как компьютерная безопасность, вычислительная криминология и терроризм информатика также процветают [9].

Интелектуальная собственность, безопасности и общественная безопасность это большое количество данных из нескольких источников, от криминальных отчеты о случаях терроризма и угрозах кибербезопасности до многоязычной интелектуальной собственности с открытым исходным кодом. Компании разных размеров сталкиваются с непростой задачей защиты от угрозы кибербезопасности и защиты их интеллектуальных активов и инфраструктуры. Обработка и анализ связанных с безопасностью данные все труднее. Значительная проблема в области безопасности ИТ-исследования - это информационный «чёрный ящик» и большая сложность, возникающая из разных источников данных, множества форматов данных и больших объемов данных. Текущие исследования технологий для кибербезопасности, борьбы с терроризмом и борьбы с преступностью в приложениях отсутствует согласованная структура для решения этих проблем. Избранные технологии BI&A таких как криминальная ассоциация, управление и кластеризация, анализ криминальной сети, пространственно-временный анализ и визуализация, многоязычная текстовая аналитика, ха-

рактеристический анализ окраски текста и анализ кибер-атак и атрибуция должны быть рассмотренный для исследования информационной безопасности. Университет Аризоны COPLINK и Dark Web исследовательские программы предлагают примеры важныех данных о преступности и террористическую информацию от сообществе IS [7]. Система сбора информации о совместном использовании информации и преступности COPLINK, первоначально разработанная с финансированием от NSF и Министерства юстиции, в настоящее время используется более 4500 полицейских агентств в Соединенных Штатах и 25 стран НАТО, и была приобретена IBM в 2011 году. Исследование Dark Web, финансируемое NSF и Департаментом обороны (DOD), создали одну из самых больших известных академические базы данных исследований терроризма (около 20 терабайт: террористические веб-сайты и контент в социальных сетях) и передовые многоязычные методы анализа социальных медиа. Признавая проблемы, связанные с объемом и сложность больших данных, связанных с обороной, защита США Агентство перспективных исследований (DARPA) в рамках DOD инициировал программу XDATA в 2012 году, чтобы помочь разработать вычислительные методы и программные средства для обработки и анализ огромного количества ориентированной на использования информации для обороны. XDATA направлена на удовлетворения потребностей в масштабируемыех алгоритмах для обработки и визуализации несовершенных и неполных данных. Программа применяется математиками, информатиками и сообществом визуализации данных для разработки анализа больших данных и удобства использования для военнослужащих. BI&A исследователи могли внести значительный вклад в этой области.

2.7 Text Analytics

Значительная часть неструктурированного контента, собранная организацией в текстовом формате, от электронной почты и корпоративных документов на веб-страницы и социального медиаконтента. Текстовая аналитика имеет свои академические корни в поиске информации и вычислительная лингвистика. В информационном поиске представление документов и обработка запросов являются основой для разработки векторного пространства модели, модель булевых поисков и модель вероятностного поиска, которая, в свою очередь, стала основой для современной цифровой библиотеки, поисковые системы и поисковые системы предприятия [10]. В вычислительной лингвистике статистические методы обработки естественного языка (NLP) для лексического приобретения, смысловое значение слова, частота речи (POST) и вероятностные контекстно-свободные правила

также стали важными для представления текста [14]. В дополнение к представлениям документов и запросов пользователь модели и релевантность обратной связи также важны для повышая эффективность поиска. С начала 1990-х годов поисковые системы превратились в зрелые коммерческие системы, состоящие из быстрых, распределенных систем; эффективная инвертированная индексация; ранжирование основанное на ссылках; и аналитики поисковых журналов. Многие из этих основополагающих методы обработки текста и индексирования были развернуты в текстовом поиске предприятия и документах систем управления в BI&A 1.0.

Использование мощности больших данных (для обучения) и статистических NLP (для создания языковых моделей), методы текстовой аналитики активно проводились в нескольких новых областях, включая извлечение информации, тематические модели, опрос (Q/A) и интеллектуального анализа. Извлечение информации является областью исследований, целью которой является автоматическое извлечение конкретных видов структурированной информации из документов. В виде строительных блок извлечения информации, NER (метод назвающийся распознавание объектов, также известный как извлечение сущности) является процесс, который идентифицирует атомные элементы в тексте и классифицирует их в предопределенные категории (например, имена, места, даты). Технологии NER были успешно разработаны для новостного анализа и биомедицинских приложений. Модели темы - это алгоритмы для открытия основных тем, которых больше всего и в противном случае неструктурированный сбор документов. Новые алгоритмы моделирования темы, такие как LDA (latent Dirichlet allocation) и другие вероятностные модели недавних исследований [11]. Системы ответа на вопрос (Q / A) полагатся на методы из NLP, поиск информации и взаимодействие человека с компьютером. Прежде всего, чтобы ответить на фактические вопросы (то есть, кто, что, когда и где вопросы), системы Q/A включают различные методы для анализ вопросов, поиск источников, извлечение ответов и [12]. Недавние успехи IBM Watson и Apple Siri выделили Q/A возможности исследований и коммерциализации. Были определены многообещающие области применения системы Q / A, включая образование, здравоохранение и защиту. Знания относящихся к вычислительным методам для извлечения, классификации, понимания и оценки мнений, выраженных в различные онлайн-источники новостей, комментарии в социальных сетях и другое пользовательское содержимое. Анализ настроений часто используется в интеллектуальном анализе для выявления настроений, эффектов, субъективности, и других эмоциональных состояний в онлайн-тексте. Web 2.0 и социальный медиа-контент создал множество богатых и интересных возможностей для понимания об-

щего мнения общественности и потребителей в отношении социальных событий, политических движений, стратегии компании, маркетинговых кампании и предпочтения продуктов [13].

В дополнение к вышеуказанным направлениям исследование текстового анализа также предлагает значительные исследовательские возможности и проблемы в несколько более узконаправленных областях, включая веб-стилометрические анализ авторства, многоязычный анализ для веб-документов и широкомасштабную визуализацию текста. Мультимедийный поиск информации и поиск мобильной информации это две другие связанные области, которые требуют поддержки текст-аналитических методов, в дополнение к основным мультимедиа и мобильных технологий. Подобно большой аналитике данных, текст аналитики с использованием MapReduce, Hadoop и облачных сервисов продолжать улучшать активные направления исследований в обеих академических и промышленных кругах.

2.8 Web Analytics

За последнее десятилетие веб-аналитика стала активной области исследований в BI&A. Основываясь на данных и основах статистического анализа, аналитики данных и модели поиска информации и NLP в текстовой аналитике, веб-аналитика предлагает уникальные аналитические задачи и возможности. HTTP/HTML-гиперссылки веб-сайтов и ассационные поисковые системы и системы каталогов для размещение веб-контента помогли разработать уникальный интернет-сайт технологий для обхода/паутинга веб-сайта, веб-страницы обновлений, ранжирование веб-сайтов и анализ поискового журнала. Веб-журнал анализа, основанный на транзакциях с клиентами, впоследствии превратились в активные исследования в рекомендательных системах. Однако, веб-аналитика стала еще более захватывающей с зрелостью и популярностью веб-сервисов и систем Web 2.0 в середине 2000-х годов [15].

На основе XML и интернет-протоколов (HTTP, SMTP), web услуги предлагают новый способ повторного использования и интеграции сторонних или устаревшие системы. Новые типы веб-сервисов и их ассоциированные API (интерфейс прикладного программирования) позволяют разработчикам легко интегрировать разнообразный контент из разных веб-система, например, REST (репрезентативная передача состояния) для вызова удаленных сервисов, RSS (на самом деле простое синдицирование) для новостей «pushing», JSON (JavaScript объектная нотация) для облегчения обмена данными и AJAX (асинхронный JavaScript + XML) для обмена

данными и динамическим дисплеем. Такие легкие модели программирования поддерживают синдикации данных и уведомлений и «mashups» мультимедийного контента (например, Flickr, Youtube, Google Maps) из разных веб-источников - процесс, несколько схожий с ETL (извлечение, преобразование и загрузка) в BI&A 1.0. Большинство поставщиков электронной коммерции предоставили готовые API-интерфейсы для доступа к их продукту и клиентскому контенту [16]. Например, через Amazon Web Services разработчики можете получить доступ к каталогу продукции, отзывам клиентов, сайту ранжирования, истории цен и Amazon Elastic Compute Cloud (EC2) для вычислительной мощности. Аналогичным образом, веб-сайт Google API поддерживают поиск AJAX, API карт, API GData (для Calendar, Gmail и т. д.), Google Translate и приложение Google Engine для облачных вычислительных ресурсов. Веб-службы и API продолжают предоставлять захватывающий поток новых данных, которые являются источниками для исследования BI&A 2.0.

Одним из основных компонентов анализа веб-аналитики является разработка облачных вычислительных платформ и сервисов, которые включают приложения, системное программное обеспечение и аппаратное обеспечение предоставляемых как услуги через Интернет. На основе сервисно-ориентированной архитектуры (SOA), виртуализации серверов и утилит вычислений, облачные вычисления могут предлагаться как программное обеспечение как услугу (SaaS), инфраструктуру как услугу (IaaS) или платформу как услугу (PaaS). В настоящее время только несколько ведущих поставщиков ИТ-услуг позиционируется для поддержки высоконагруженных высокопроизводительных BI&A приложений с использованием облачных вычислений. Например, Amazon Elastic Compute Cloud (EC2) позволяет пользователям арендовать виртуальные компьютеры, на которых позволяет запускать собственные компьютерные приложения. Их простая служба хранения (S3) оказывает услугу предоставления онлайн-хранилище. Google App Engine предоставляет платформу для разработки и хостинга веб-приложений на основе Java или Python. Google Bigtable используется для хранения данных на бэкэнд. Microsoft's Windows Azure платформа предоставляет облачные сервисы, такие как SQL Azure и SharePoint, а также позволяет .Net framework приложениям работать на платформе. Отраслевая сеть и облачные сервисы предлагают уникальный сбор, обработку и аналитические задачи для исследователей BI&A.

В академических кругах текущие исследования, связанные с веб-аналитикой, охватывают социальный поиск и поиск полезных данных, системы репутации, социальные медиа-аналитика и веб-визуализация.

Кроме того, веб-сайт аукционы, интернет-монетизация, социальный маркетинг и конфиденциальность/безопасность в Интернете - вот некоторые из перспективных направлений исследований связанныех с веб-аналитикой. Многие из этих исследовательских областей могут полагаться на успехи в анализе социальных сетей, текстовая аналитика и даже исследование экономического моделирования.

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Miller, K. 2012a. «Big Data Analytics in Biomedical Research,» Biomedical Computation Review (available at http:// biomedicalcomputationreview.org/content/big-data-analyticsbiomedical-research; accessed April 2, 2018).

2. Gelfand, A. 2011/2012. "Privacy and Biomedical Research: Building a Trust Infrastructure—An Exploration of Data-Driven and Process-Driven Approaches to Data Privacy," Biomedical Computation Review, Winter, pp. 23-28 (available at http://biomedicalcomputationreview.org/content/privacy-andbiomedical-research-building-trust-infrastructure, accessed April 2, 2018).

3. Lin, Y., Brown, R. A., Yang, H. J., Li, S., Lu, H., and Chen, H. 2011. "Data Mining Large-Scale Electronic Health Records for Clinical Support," IEEE Intelligent Systems (26:5), pp. 87-90

4. Miller, K. 2012b. "Leveraging Social Media for Biomedical Research: How Social Media Sites Are Rapidly Doing Unique Research on Large Cohorts," Biomedical Computation Review (available at http://biomedicalcomputationreview.org/content/ leveraging-social-media-biomedical-research; accessed August 2, 2012).

5. Wactlar, H., Pavel, M., and Barkis, W. 2011. "Can Computer Science Save Healthcare?" IEEE Intelligent Systems (26:5), pp. 79-83.

6. Chen, H. 2011b. "Smart Health and Wellbeing," IEEE Intelligent Systems (26:5), pp. 78-79.

7. Chen, H. 2006. Intelligence and Security Informatics for International Security: Information Sharing and Data Mining, New York: Springer.

8. Perlroth, N., and Rusli, E. M. 2012. "Security Start-Ups Catch Fancy of Investors," New York Times, Technology Section, August 5.

9. Brantingham, P. L. 2011. "Computational Criminology," Keynote Address to the European Intelligence and Security Informatics Conference, Athens, Greece, September 12-14.

10. Salton, G. 1989. Automatic Text Processing, Reading, MA: Addison Wesley

11. Blei, D. M. 2012. "Probabilistic Topic Models," Communications of the ACM (55:4), pp. 77-84.

12. Maybury, M. T. (ed.). 2004. New Directions in Question Answering, Cambridge, MA: The MIT Press.

13. Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval (2:1-2), pp. 1-135.

14. Manning, C. D., and Schütze, H. 1999. Foundations of Statistical Natural Language Processing, Cambridge, MA: The MIT Press.

15. O'Reilly, T. 2005. "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software," September 30, (http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html).

16. Schonfeld, E. 2005. "The Great Giveaway," Business 2.0 (6:3), pp. 80-86.

ПРИЛОЖЕНИЕ

BI&A 1.0

As a data-centric approach, BI&A has its roots in the longstanding database management field. It relies heavily on various data collection, extraction, and analysis technologies (Chaudhuri et al. 2011; Turban et al. 2008; Watson and Wixom 2007). The BI&A technologies and applications currently adopted in industry can be considered as BI&A 1.0, where data are mostly structured, collected by companies through various legacy systems, and often stored in commercial relational database management systems (RDBMS). The analytical techniques commonly used in these systems, popularized in the 1990s, are grounded mainly in statistical methods developed in the 1970s and data mining techniques developed in the 1980s.

Data management and warehousing is considered the foundation of BI&A 1.0. Design of data marts and tools for extraction, transformation, and load (ETL) are essential for converting and integrating enterprise-specific data. Database query, online analytical processing (OLAP), and reporting tools based on intuitive, but simple, graphics are used to explore important data characteristics. Business performance management (BPM) using scorecards and dashboards help analyze and visualize a variety of performance metrics. In addition to these well-established business reporting functions, statistical analysis and data mining techniques are adopted for association analysis, data segmentation and clustering, classification and regression analysis, anomaly detection, and predictive modeling in various business applications. Most of these data processing and analytical technologies have already been incorporated into the leading commercial BI platforms offered by major IT vendors including Microsoft, IBM, Oracle, and SAP (Sallam et al. 2011).

Among the 13 capabilities considered essential for BI platforms, according to the Gartner report by Sallam et al. (2011), the following eight are considered BI&A 1.0: reporting, dashboards, ad hocquery, search-based BI, OLAP, interactive visualization, scorecards, predictive modeling, and data mining. A few BI&A 1.0 areas are still under active development based on the Gartner BI Hype Cycle analysis for emerging BI technologies, which include data mining workbenchs, column-based DBMS, in-memory DBMS, and realtime decision tools (Bitterer 2011). Academic curricula in Information Systems (IS) and Computer Science (CS) often include well-structured courses such as database management systems, data mining, and multivariate statistics.

BI&A 2.0

Since the early 2000s, the Internet and the Web began to offer unique data collection and analytical research and development opportunities. The HTTP-based Web 1.0 systems, characterized by Web search engines such as Google and Yahoo and e-commerce businesses such as Amazon and eBay, allow organizations to present their businesses online and interact with their customers directly. In addition to porting their traditional RDBMS-based product information and business contents online, detailed and IP-specific user search and interaction logs that are collected seamlessly through cookies and server logs have become a new gold mine for understanding customers' needs and identifying new business opportunities. Web intelligence, web analytics, and the user-generated content collected through Web 2.0-based social and crowd-sourcing systems (Doan et al. 2011; O'Reilly 2005) have ushered in a new and exciting era of BI&A 2.0 research in the 2000s, centered on text and web analytics for unstructured web contents.

An immense amount of company, industry, product, and customer information can be gathered from the web and organized and visualized through various text and web mining techniques. By analyzing customer clickstream data logs, web analytics tools such as Google Analytics can provide a trail of the user's online activities and reveal the user's browsing and purchasing patterns. Web site design, product placement optimization, customer transaction analysis, market structure analysis, and product recommendations can be accomplished through web analytics. The many Web 2.0 applications developed after 2004 have also created an abundance of user-generated content from various online social media such as forums, online groups, web blogs, social networking sites, social multimedia sites (for photos and videos), and even virtual worlds and social games (O'Reilly 2005). In addition to capturing celebrity chatter, references to everyday events, and socio-political sentiments expressed in these media, Web 2.0 applications can efficiently gather a large volume of timely feedback and opinions from a diverse customer population for different types of businesses.

Many marketing researchers believe that social media analytics presents a unique opportunity for businesses to treat the market as a "conversation" between businesses and customers instead of the traditional business-to-customer, one-way "marketing" (Lusch et al. 2010). Unlike BI&A 1.0 technologies that are already integrated into commercial enterprise IT systems, future BI&A 2.0 systems will require the integration of mature and scalable techniques in text mining (e.g., information extraction, topic identification, opinion mining, question-answering), web mining, social network analysis, and

spatial-temporal analysis with existing DBMS-based BI&A 1.0 systems.

Except for basic query and search capabilities, no advanced text analytics for unstructured content are currently considered in the 13 capabilities of the Gartner BI platforms. Several, however, are listed in the Gartner BI Hype Cycle, including information semantic services, natural language question answering, and content/text analytics (Bitterer 2011). New IS and CS courses in text mining and web mining have emerged to address needed technical training.

BI&A 3.0

Whereas web-based BI&A 2.0 has attracted active research from academia and industry, a new research opportunity in BI&A 3.0 is emerging. As reported prominently in an October 2011 article in The Economist (2011), the number of mobile phones and tablets (about 480 million units) surpassed the number of laptops and PCs (about 380 million units) for the first time in 2011. Although the number of PCs in use surpassed 1 billion in 2008, the same article projected that the number of mobile connected devices would reach 10 billion in 2020. Mobile devices such as the iPad, iPhone, and other smart phones and their complete ecosystems of downloadable applicationss, from travel advisories to multi-player games, are transforming different facets of society, from education to healthcare and from entertainment to governments. Other sensor-based Internet-enabled devices equipped with RFID, barcodes, and radio tags (the "Internet of Things") are opening up exciting new steams of innovative applications. The ability of such mobile and Internet-enabled devices to support highly mobile, location-aware, person-centered, and context-relevant operations and transactions will continue to offer unique research challenges and opportunities throughout the 2010s. Mobile interface, visualization, and HCI (human–computer interaction) design are also promising research areas. Although the coming of the Web 3.0 (mobile and sensor-based) era seems certain, the underlying mobile analytics and location and context-aware techniques for collecting, processing, analyzing and visualizing such largescale and fluid mobile and sensor data are still unknown.

No integrated, commercial BI&A 3.0 systems are foreseen for the near future. Most of the academic research on mobile BI is still in an embryonic stage. Although not included in the current BI platform core capabilities, mobile BI has been included in the Gartner BI Hype Cycle analysis as one of the new technologies that has the potential to disrupt the BI market significantly (Bitterer 2011). The uncertainty associated with BI&A 3.0 presents another unique research direction for the IS community. Table 1 summarizes the key characteristics of BI&A 1.0, 2.0, and 3.0 in relation to the Gartner BI platforms

core capabilities and hype cycle. The decade of the 2010s promises to be an exciting one for high-impact BI&A research and development for both industry and academia. The business community and industry have already taken important steps to adopt BI&A for their needs. The IS community faces unique challenges and opportunities in making scientific and societal impacts that are relevant and long-lasting (Chen 2011a). IS research and education programs need to carefully evaluate future directions, curricula, and action plans, from BI&A 1.0 to 3.0.

BI&A Applications: From Big Data to Big Impact

Several global business and IT trends have helped shape past and present BI&A research directions. International travel, high-speed network connections, global supply-chain, and outsourcing have created a tremendous opportunity for IT advancement, as predicted by Thomas Freeman in his seminal book, The World is Flat (2005). In addition to ultra-fast global IT connections, the development and deployment of business-related data standards, electronic data interchange (EDI) formats, and business databases and information systems have greatly facilitated business data creation and utilization. The development of the Internet in the 1970s and the subsequent large-scale adoption of the World Wide Web since the 1990s have increased business data generation and collection speeds exponentially. Recently, the Big Data era has quietly descended on many communities, from governments and e-commerce to health organizations. With an overwhelming amount of web-based, mobile, and sensorgenerated data arriving at a terabyte and even exabyte scale (The Economist 2010a, 2010b), new science, discovery, and insights can be obtained from the highly detailed, contextualized, and rich contents of relevance to any business or organization.

In addition to being data driven, BI&A is highly applied and can leverage opportunities presented by the abundant data and domain-specific analytics needed in many critical and highimpact application areas. Several of these promising and high-impact BI&A applications are presented below, with a discussion of the data and analytics characteristics, potential impacts, and selected illustrative examples or studies: (1) ecommerce and market intelligence, (2) e-government and politics 2.0, (3) science and technology, (4) smart health and well-being, and (5) security and public safety. By carefully analyzing the application and data characteristics, researchers and practitioners can then adopt or develop the appropriate analytical techniques to derive the intended impact. In addition to technical system implementation, significant business or domain knowledge as well as effective communication skills are needed for

the successful completion of such BI&A projects. IS departments thus face unique opportunities and challenges in developing integrated BI&A research and education programs for the new generation of data/analyticssavvy and business-relevant students and professionals (Chen 2011a).

E-Commerce and Market Intelligence

The excitement surrounding BI&A and Big Data has arguably been generated primarily from the web and e-commerce communities. Significant market transformation has been accomplished by leading e-commerce vendors such Amazon and eBay through their innovative and highly scalable ecommerce platforms and product recommender systems. Major Internet firms such as Google, Amazon, and Facebook continue to lead the development of web analytics, cloud computing, and social media platforms. The emergence of customer-generated Web 2.0 content on various forums, newsgroups, social media platforms, and crowd-sourcing systems offers another opportunity for researchers and practitioners to "listen" to the voice of the market from a vast number of business constituents that includes customers, employees, investors, and the media (Doan et al. 2011; O'Rielly 2005). Unlike traditional transaction records collected from various legacy systems of the 1980s, the data that e-commerce systems collect from the web are less structured and often contain rich customer opinion and behavioral information. For social media analytics of customer opinions, text analysis and sentiment analysis techniques are frequently adopted (Pang and Lee 2008). Various analytical techniques have also been developed for product recommender systems, such as association rule mining, database segmentation and clustering, anomaly detection, and graph mining (Adomavicius and Tuzhilin 2005). Long-tail marketing accomplished by reaching the millions of niche markets at the shallow end of the product bitstream has become possible via highly targeted searches and personalized recommendations (Anderson 2004). The Netfix Prize competition for the best collaborative filtering algorithm to predict user movie ratings helped generate significant academic and industry interest in recommender systems development and resulted in awarding the grand prize of $1 million to the Bellkor's Pragmatic Chaos team, which surpassed Netflix's own algorithm for predicting ratings by 10.06 percent. However, the publicity associated with the competition also raised major unintended customer privacy concerns.

Much BI&A-related e-commerce research and development information is appearing in academic IS and CS papers as well as in popular IT magazines.

E-Government and Politics 2.0

The advent of Web 2.0 has generated much excitement for reinventing governments. The 2008 U.S. House, Senate, and presidential elections provided the first signs of success for online campaigning and political participation. Dubbed «politics 2.0», politicians use the highly participatory and multimedia web platforms for successful policy discussions, campaign advertising, voter mobilization, event announcements, and online donations. As government and political processes become more transparent, participatory, online, and multimedia-rich, there is a great opportunity for adopting BI&A research in e-government and politics 2.0 applications. Selected opinion mining, social network analysis, and social media analytics techniques can be used to support online political participation, e-democracy, political blogs and forums analysis, e-government service delivery, and process transparency and accountability (Chen 2009; Chen et al. 2007). For e-government applications, semantic information directory and ontological development (as exemplified below) can also be developed to better serve their target citizens.

Despite the significant transformational potential for BI&A in e-government research, there has been less academic research than, for example, e-commerce-related BI&A research. Egovernment research often involves researchers from political science and public policy. For example, Karpf (2009) analyzed the growth of the political blogosphere in the United States and found significant innovation of existing political institutions in adopting blogging platforms into their Web offerings. In his research, 2D blogspace mapping with composite rankings helped reveal the partisan makeup of the American political blogsphere. Yang and Callan (2009) demonstrated the value for ontology development for government services through their development of the OntoCop system, which works interactively with a user to organize and summarize online public comments from citizens.

Science and Technology

Many areas of science and technology (S&T) are reaping the benefits of high-throughput sensors and instruments, from astrophysics and oceanography, to genomics and environmental research. To facilitate information sharing and data analytics, the National Science Foundation (NSF) recently mandated that every project is required to provide a data management plan. Cyber-infrastructure, in particular, has become critical for supporting such data-sharing initiatives.

The 2012 NSF BIGDATA program solicitation is an obvious example of

the U.S. government funding agency's concerted efforts to promote big data analytics. The program aims to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life.

Several S&T disciplines have already begun their journey toward big data analytics. For example, in biology, the NSF funded iPlant Collaborative is using cyberinfrastructure to support a community of researchers, educators, and students working in plant sciences. iPlant is intended to foster a new generation of biologists equipped to harness rapidly expanding computational techniques and growing data sets to address the grand challenges of plant biology. The iPlant data set is diverse and includes canonical or reference data, experimental data, simulation and model data, observational data, and other derived data. It also offers various open source data processing and analytics tools.

In astronomy, the Sloan Digital Sky Survey (SDSS) shows how computational methods and big data can support and facilitate sense making and decision making at both the macroscopic and the microscopic level in a rapidly growing and globalized research field. The SDSS is one of the most ambitious and influential surveys in the history of astronomy. Over its eight years of operation, it has obtained deep, multicolor images covering more than a quarter of the sky and created three-dimensional maps containing more than 930,000 galaxies and over 120,000 quasars. Continuing to gather data at a rate of 200 gigabytes per night, SDSS has amassed more than 140 terabytes of data. The international Large Hadron Collider (LHC) effort for high-energy physics is another example of big data, producing about 13 petabytes of data in a year (Brumfiel 2011).