# Data Warehousing and Knowledge Discovery: A Chronological View of Research Challenges

Tho Manh Nguyen[1], A Min Tjoa [1], and Juan Trujillo [2]

[1] Institute of Software Technology and Interactive Systems,
Vienna University of Technology,
Favoritenstr. 9-11/188, A-1040 Vienna, Austria
{tho, amin}@ifs.tuwien.ac.at
[2] Dept. of Language and Information Systems, University of Alicante,
Apto. Correos 99. E-03080. Spain
jtrujillo@dlsi.ua.es

Data Warehousing and Knowledge Discovery has been widely accepted as a key technology for enterprises to improve their abilities in data analysis, decision support, and the automatic extraction of knowledge from data. Historically, the phrase knowledge discovery in databases was coined at the first KDD (Knowledge Discovery and Data Mining) workshop in 1989 to emphasize that knowledge is the end-product of a data-driven discovery process. Since then, much research has been accomplished in this field. This paper which is written as an epilogue of the DAWAK 2005 proceedings by the programme committee chairpersons together with Nguyen Manh Tho, should reflect the past development of DAWAK-results and other significant research outcomes in the area and above all should deliver a rough sketch of the current development and possible future work.

In the early 1990s, most businesses realized that there was an urgent need for more sophisticated tools for analyzing their business data, customer profiles and product information. Data mining and data warehousing technology mainly originated from these needs.

Data mining can be defined as the automated extraction of hidden predictive information from large amount of data. Data Mining is considered as an important step in the Knowledge Discovery -process that produces a particular enumeration of patterns (or models) over the data [9]. The most commonly used techniques in data mining and knowledge discovery in the late 1980s and early 1990s are artificial neural networks, decision trees, genetic algorithms, nearest neighbourhood, and rule induction [12,13].

William H. Inmon, who is widely accepted as the mental-father of data warehousing has been working on data warehousing concepts since 1983, and used for the first time this term in 1992 [10]. In 1993, Ted Codd coins the term OLAP (On-Line Analytical Processing) and defined the famous 12 OLAP rules [11]. Based on their definitions, a data warehouse is actually a comprehensive system that includes all the processes, tools and technologies necessary for extracting data from many operational, legacy, possibly heterogeneous data sources and managing them in a separate storage (warehouse) to provide end-user decision-support access. OLAP tools are well-suited for complex data analysis, such as multidimensional data analysis, and to assist in

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2005, LNCS 3589, pp. 530 – 535, 2005.

decision support activities. The multidimensional (MD) data model has been proved to be the most suitable for OLAP applications.

The two sides of the coin for decision making are principally formed by Data Warehousing and Knowledge Discovery. Knowledge discovery in data warehouses focuses upon the extraction of interesting and previously unknown knowledge [3]. Researchers and application developers have designed knowledge discovery systems for a large number of application domains including finance, health, telecommunications and marketing. Consequently, many areas of research in data warehousing and knowledge discovery mushroomed in the late 1990s. Initially, when the concepts of OLAP and multi-dimensional databases have not yet the desired level of maturity, most researchers focus on the topics of data warehouse design, view selection and maintenance, multiple query optimization using views, on-line view maintenance, OLAP operators and environment, fragmentation of multidimensional database [1, 23, 24]. In the data mining community, interests are focused on data & web mining, pattern recognition and time series databases [1, 14, 15, 16]. More attention is also devoted to scientific data exploration dealing with such research objects as mining and discovery on biological data, spatial, text and multimedia data. Distributed and parallel mining techniques become more and more in use and some large-scale parallel and distributed knowledge discovery systems start to appear [17]. With the widespread of web applications web usage analysis and user profiling builds a special focus of investigations which huge relevance for e-commerce [17].

In the early 2000 the research community gradually solve the most basic issues in data warehouse design, materialized view maintenance and selection, OLAP query design and evolution. However, with the development of the increasing number of commercial products in OLAP and data warehousing, the data warehouse research activities force to concentrate towards more advanced techniques such as integrating active rules, update filtering, parallel processing, summarizability problem, data expiry, data indexing [2,3]. These activities include also the increasing consideration of security issues in OLAP and generally in the data warehousing environment [2, 25] - as well as the advanced interest on the heterogeneous and distributed data warehousing environment. [3,26]. The concept of object-orientation and the emerging XML technology cause significant implications on the design and development of data warehouse applications. [3, 26].

In the data mining and knowledge discovery research fields, the early 2000's witness new mining algorithms and techniques which are proposed and applied in a variety of applications such as text mining, outlier detection in scientific data, mining of temporal patterns, optimizing inventory in E-commerce, telephony and ISP applications [18]. Special attention is devoted to Web mining, interactive knowledge exploration, matchmaking and visualization [2, 18]. Mining of Web-log data, and multimedia data are still most important research topics while mining in bioinformatics has become an emerging field. [3,19].

In 2002 and 2003, with the advances of modern monitoring technologies (i.e. sensors, RFID, transmission of huge amount of digital satellite monitoring data) and with the demand of high speed business changes, the integration of a special type of data source namely of continuous data streams is becoming more and more essential. Data streams occur in applications such as sensor networks, networking flow analysis, web clicks stream analysis, telecommunication fraud detection, e-business and stock mar-

ket online analysis. One of the main characterics of data streams is the impracticality of complete storage – hence we are usually restricted to the storage of samples or aggregations. It is demanding to conduct advanced analysis over fast and huge data streams to capture the trends, patterns, and exceptions. As a further consequence, the concept of near-real time data warehouses was initially announced in [4] while concurrently we still observe further intense investigations on parallel and distributed warehousing. [4, 27]. Ontology Structures, which are a foundation of the Semantic Web [7], has been applied among others for the integration of heterogeneous Data Quality Improving techniques [5]. With the increasing embedding of Web data as one of the main data sources, in DWH-research the Web-Warehousing concept is thus investigated in-depth together with its accompanying concepts and related technologies such as XML OLAP Cube, XML Warehouse, Warehouse design based on XML Schema [5, 28]. OLAP researchers proceed in the intense investigation of advanced improvements in Cube presentation, management, and performance (which obviously is not restricted on traditional data, but also includes XML data and other non-standard data sources such as spatial data [5, 28]).

Evidently the data mining community recognizes the important role of streams and time series analysis applications. A plethora of algorithms and techniques were proposed to mine high-speed data streams, to analyse click streams, and to correlate synchronised and asynchronised online streams [20, 21]. Furthermore, with the huge amount of web data, research conducted on web search tools and web classification applications builds a main focus[4, 20, 21]. Data mining in Bioinformatics, multimedia and complex data still receives a lot of interesting and research efforts [20, 21]. Data mining techniques are also applied to improve database-engine issues by using techniques which have been successfully deployed in other areas of applied computer science and systems theory – a prominent example herefore is the use of Rough Set Theory, as an alternative of fuzzy sets [5].

The complexity of existing data warehousing and enterprise systems has reached a new quality in the recent years. However, there is still a lack of comprehensive documentation and dissemination of requirement engineering methods. Therefore, conceptual modelling still plays an essential role in integrating higher level of abstractions for the description of processes all components of the data warehouse architecture [6]. Spatial data warehouses reach some maturity with the design framework of Geographical Dimensional Schema. OLAP-techniques are now enhanced with innovations in Range Aggregation and approximate queries answering [6, 29]. Data streams analysis and time series mining techniques receive an enduring boosted interest from the researchers [6, 22]. Pattern discovery and event sequence mining has emerged as a new field of interest while data semantics became an increasingly important issue. The topic of data visualization and exploration gets increased interests while traditional mining techniques such as association rules, clustering remain steadily prevailing [6, 22]. In parentheses it should be remarked that not only academic researchers, but increasingly the industrial community is concentrating in these activities [6, 22].

Due to the fact of the tremendous amount of existing data warehouse systems, more current efforts are taken to integrate and transform the different heterogeneous data warehousing systems. An important innovation can be observed in extending existing relevant and successful CASE tools used in software development (most notably UML-tools) for data warehouse design and development. With the expanding

spread of open source tools we can observe a trend to solve the Business Intelligence issues in the open source community.

Artificial intelligence techniques such as machine learning, neural network, case-based reasoning have inspired a continuous fast growing attention within the data mining community. An expanding interest can be observed in the integration of text processing and the mining unstructured data using data semantics approaches in combination with traditional techniques i.e. clustering, association rules, pattern recognition...

Although research in data warehousing and knowledge discovery has generated successful and remarkable results, new applications still continuously generate new challenges to the research community. With the exponential growing amount of information to be included in the decision making process, the data to be considered becomes more and more complex in both structure and semantics. Consequently, the process of retrieval and knowledge discovery from this huge amount of heterogeneous complex data builds the litmus-test for the research in the area. Current emerging real world applications such as real-time data warehousing, analysis of spatial and spatio-temporal data, OLAP mining, mobile OLAP and more recent applications in natural sciences (especially bioinformatics) requires novel representation and manipulation techniques for non-standard data and tailored efficient algorithms for the computation of dedicated aggregate queries and application-specific index structures.

Vendors like Oracle, IBM and Microsoft already tightly integrate OLAP and data mining in their DBMS commercial tools. Therefore, we are observing a trend to close the gap between data warehousing and data mining. It is even imaginable that in the medium-term future the separation of OLAP-data warehouses (with its semantic redundant storage requirement) from its OLTP-database-sources could be bridge by innovative view-generation techniques as originally proposed in the three-level architecture.

## References

[1] Mukesh K. Mohania, A. Min Tjoa (Eds.): *Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99*, Florence, Italy, August 30 - September 1, 1999, Proceedings. LNCS 1676 Springer 1999, ISBN 3-540-66458-0

[2] Yahiko Kambayashi, Mukesh K. Mohania, A. Min Tjoa (Eds.): *Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000*, London, UK, September 4-6, 2000, Proceedings. LNCS 1874 Springer 2000, ISBN 3-540-67980-4

[3] Yahiko Kambayashi, Werner Winiwarter, Masatoshi Arikawa (Eds.): *Data Warehousing and Knowledge Discovery, Third International Conference, DaWaK 2001*, Munich, Germany, September 5-7, 2001, Proceedings. LNCS 2114 Springer 2001, ISBN 3-540-42553-5

[4] Yahiko Kambayashi, Werner Winiwarter, Masatoshi Arikawa (Eds.): *Data Warehousing and Knowledge Discovery, 4th International Conference, DaWaK 2002*, Aix-en-Provence, France, September 4-6, 2002, Proceedings. LNCS 2454 Springer 2002, ISBN 3-540-44123-9

[5] Yahiko Kambayashi, Mukesh K. Mohania, Wolfram Wöß (Eds.): *Data Warehousing and Knowledge Discovery, 5th International Conference, DaWaK 2003*, Prague, Czech Republic, September 3-5,2003, Proceedings. LNCS 2737 Springer 2003, ISBN 3-540-40807-X

[6] Yahiko Kambayashi, Mukesh K. Mohania, Wolfram Wöß (Eds.): *Data Warehousing and Knowledge Discovery, 6th International Conference, DaWaK 2004*, Zaragoza, Spain, September 1-3, 2004, Proceedings. LNCS 3181 Springer 2004, ISBN 3-540-22937-X

[7] Tim Berners-Lee: *Semantic Web Road map*, September 1998. http://www.w3.org/DesignIssues/Semantic.html

[8] Mukesh K. Mohania, Yahiko Kambayashi, A. Min Tjoa, Roland Wagner, Ladjel Bellatreche: *Trends in Database Research, Database and Expert Systems Applications, 12th International Conference, DEXA 2001* Munich, Germany, September 3-5, 2001

[9] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth: *From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence*, 0738-4602-1996

[10] W. H. Inmon:*Building the Data Warehouse 1st edition*, 1992.

[11] E.F. Codd & Associates : *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*, white paper ,1993

[12] Usama M. Fayyad, Ramasamy Uthurusamy (Eds.): *Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop*, Seattle, Washington, July 1994. AAAI Press, Technical Report WS-94-03, ISBN 0-929280-73-3

[13] Usama M. Fayyad, Ramasamy Uthurusamy (Eds.): *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, August 20-21, 1995. AAAI Press, ISBN 0-929280-82-2

[14] Evangelos Simoudis, Jiawei Han, Usama M. Fayyad (Eds.): *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996, ISBN 1-57735-004-9

[15] David Heckerman, Heikki Mannila, Daryl Pregibon (Eds.): *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, Newport Beach, California, USA, August 14-17, 1997. AAAI Press, 1997, ISBN 1-57735-027-8

[16] Rakesh Agrawal, Paul E. Stolorz, Gregory Piatetsky-Shapiro (Eds.): *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, August 27-31, 1998, New York City, New York, USA. AAAI Press, 1998, ISBN 1-57735-070-7

[17] Usama Fayyad, Surajit Chaudhuri, David Madigan (Eds.): *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 15-18, 1999, San Diego, CA, USA. ACM, 1999

[18] Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo, Ismail Parsa (Eds.): *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, August 20-23, 2000, Boston, MA, USA. ACM, 2000

[19] Doheon Lee, Mario Schkolnick, Foster Provost, Ramakrishnan Srikant (Eds.): *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, August 26-29, 2001, San Francisco, CA, USA. ACM, 2001

[20] Osmar R. Zaïane, Randy Goebel, David Hand,Daniel Keim, Raymond Ng (Eds.): *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 23-26, 2002, Edmonton, Alberta, Canada. ACM 2002, ISBN 1-58113-567-X

[21] Lise Getoor, Ted E. Senator, Pedro Domingos, Christos Faloutsos (Eds.): *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 24 - 27, 2003. ACM 2003, ISBN 1-58113-737-0

[22] Won Kim, Ron Kohavi, Johannes Gehrke, William DuMouchel (Eds.): *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, August 22-25, 2004. ACM 2004, ISBN 1-58113-888-1

[23] Il-Yeol Song, Toby J. Teorey (Eds.) : *Proceedings of the 1st ACM international workshop on Data warehousing and OLAP, DOLAP 1998*, Washington, D.C., United States November 02 - 07, 1998

[24] Il-Yeol Song, Toby J. Teorey (Eds.) : *Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP, DOLAP 1999*, Kansas City, Missouri, United States November 02 - 06, 1999

[25] Rokia Missaoui, Il-Yeol Song (Eds.): *Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP, DOLAP 2000*, McLean, Virginia, United States November 06 - 11, 2000

[26] Joachim Hammer (Eds.): *Proceedings of the 4th ACM international workshop on Data warehousing and OLAP, 2001*, Atlanta, Georgia, USA November 09 - 09, 2001

[27] Dimitri Theodoratos, Il-Yeol Song (Eds.): *Proceedings of the 5th ACM international workshop on Data warehousing and OLAP,DOLAP 2002*, November 8, 2002, McLean, VA, Proceedings. ACM 2002, ISBN 1-58113-5904

[28] Stefano Rizzi, Il-Yeol Song (Eds.): *Proceedings of the 6th ACM international workshop on Data warehousing and OLAP,DOLAP 2003*, New Orleans, Louisiana, USA November 07 - 07, 2003

[29] Il-Yeol Song, Karen C. Davis (Eds.): *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP,DOLAP 2004*, Washington, DC, USA, November 12-13, 2004, Proceedings. ACM 2004, ISBN 1-58113-977-2