UNIVERSITI PUTRA MALAYSIA

DATA FILTERING FRAMEWORK FOR PRESERVING MEANINGFUL DATA RECORDS FROM STREAMS OF UNSTRUCTURED WEATHER DATA

WA'EL JUM'AH AL-ZYADAT

FSKTM 2014 17

DATA FILTERING FRAMEWORK FOR PRESERVING MEANINGFUL DATA RECORDS FROM STREAMS OF UNSTRUCTURED WEATHER DATA

By

WA'EL JUM'AH AL-ZYADAT

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

July 2014

COPYRIGHT

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

DATA FILTERING FRAMEWORK FOR PRESERVING MEANINGFUL DATA RECORDS FROM STREAMS OF UNSTRUCTURED WEATHER DATA

By

WA'EL JUM'AH AL-ZYADAT

July2014

Chairman: Rodziah Atan, PhD

Faculty: Computer Science and Information Technology

The aim of this research is to design and implement a data filtering framework for preserving meaningful data records from streams of unstructured weather data based on data collection, data pre-process, data filtering, classification, and visualization. The data collection involved monitoring data and data structuring; data pre-process, error checking, error validation, and correction of error; data filtering involved filtering concept, sequential process, and particles filtering while classification data involved the proposed 5M method.

The environmental data creates challenges in the context of storage and data processing, which producing large volume of data collection including a massive 30% of unwanted data. It affects the correctness of data in term of usage

The rationale and importance of this research comes in forms of creating step by step operations of data filtering method using local weather data that combines data pre-processing and data filtering steps to enhance the accuracy of data classifications. This research successfully reformatted data collected from sensor-boards which are unstructured, with features that structure raw data to a standardized data format.

The analysis of the proposed framework employed three measurements approaches for the validation purposes. First, data pre-process component measurement using

correctness and precision measures; filtering components are measured by data (indexing and item) using the context of records to discover the duplicates between two different datasets, and finally the 5M classification is measured by the percentage of total and meaningful$_{\text{ sensitivity}}$ to indicated classification of relevant items which explore the meaningful data of measurements based on percentages of classifier data from relevant data. Result shows the pre-processed data collected is reduced by 69.23 % with similar accuracy as compared to the raw data. Classification hit is 84.6 % accuracy to each clustered data.

This research has been able to classify streaming weather data. The execution of data filtering framework for preserving meaningful data records from streams of unstructured weather data produced highly accurate classification clusters.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

RANGKA KERJA PENAPISAN DATAUNTUKMEMELIHARAREKODDATABERMAKNA DARIPADAALIRANDATACUACATIDAK BERSTRUKTUR

Oleh

WA'EL JUM'AH AL-ZYADAT

Julai 2014

Pengerusi: **Rodziah Atan, PhD**
Fakulti: **Sains Komputer dan Teknologi Maklumat**

Tujuan kajian ini adalah untuk mereka bentuk dan melaksanakan rangka kerja penapisan data untuk memelihara rekod data yang bermakna dari aliran data cuaca tidak berstruktur berdasarkan pengumpulan data, pra-pemprosesan data, penapisan data, klasifikasi, dan visualisasi. Pengumpulan data melibatkan pemantauan data dan penstrukturan data; data pra-pemprosesan, pemeriksaan ralat, pengesahan kesilapan, dan pembetulan kesilapan; penapisan data melibatkan konsep penapisan, proses berurutan, dan penapisan zarah manakala klasifikasi data pula melibatkan kaedah 5M yang dicadangkan.

Data persekitaran mewujudkan cabaran dalam konteks penyimpanan dan pemprosesan data, penghasilan jumlah data yang besar semasa proses pengumpulan yang juga melibatkan sejumlah besar 30% data yang tidak diingini. Ia memberi kesan kepada ketepatan data bagi kegunaan seterusnya.

Rasional dan kepentingan kajian ini ditonjolkan melalui pembangunan operasi kaedah penapisan data secara langkah demi langkah menggunakan data cuaca tempatan yang menggabungkan pra-pemprosesan dan penapisan data untuk meningkatkan ketepatan klasifikasi data. Penyelidikan ini berjaya menformat semula data yang dikumpul dari papan pengesan yang tidak berstruktur, dengan fungsian yang mampu menstruktur data mentah kepada format data seragam.

Analisis yang dijalankan bagi rangka kerja cadangan ini menggunakan tiga pendekatan pengukuran untuk tujuan pengesahan. Pertama, pengukuran komponen data pra-proses menggunakan pengukuran kesahihan dan ketepatan; kedua, komponen penapisan diukur menggunakan data (indeks dan item) menggunakan rekod konteks untuk mencari pendua antara dua set data yang berbeza, dan akhir sekali, pengelasan 5M diukur melalui jumlah peratusan dan sensitivity-bermakna bagi menunjukkan klasifikasi bagi item yang relevan. Ia dijalankan bagi meneroka maklumat bermakna yang diukur berdasarkan

kepada peratusan data pengelas yang relevan. Keputusan menunjukkan data pra-proses terkumpul dapat dikurangkan sebanyak 69.23% dengan ketepatan yang sama berbanding dengan keseluruhan data mentah. Kenaan klasifikasi pula adalah berketepatan 84.6% bagi setiap kelompok data.

Kajian ini telah dapat mengklasifikasi aliran data cuaca. Pelaksanaan rangka kerja penapisan data untuk memelihara rekod data yang bermakna dari aliran data cuaca tidak berstruktur yang dicadangkan dari kajian ini mampu untuk menghasilkan ketepatan tinggi bagi setiap klasifikasi kluster.

ACKNOWLEDGEMENTS

All praise and thanks are due to Allah, the Lord of the worlds, the Most beneficent, the Most merciful. All praise and thanks to Him who said:

وَإِن تَعُدُّواْ نِعْمَةَ ٱللَّهِ لَا تُحْصُوهَآ ۗ إِنَّ ٱللَّهَ لَغَفُورٌ رَّحِيمٌ

All praise and thanks to Allah for His guidance and blessings and for granting me knowledge, patience and perseverance to accomplish this work successfully. My deepest gratitude and thanks go to Almighty Allah, who made all this and everything possible.

First of all, I would like to express my deepest and sincere gratitude to my supervisors, Associate Prof. Dr. Rodziah Atan, Prof. Dr. Hamidah Ibrahim and Associate Prof Dr. Masrah Azrifah Azmi Murad for their immeasurable wisdom, guidance, patience, valuable suggestions and supervision. Also, special thanks to my friends, Dr. Ibrahim Almarashdeh, Dr. Mutasem Alsmadi, Dr. Tirad Almalahmeh, Dr. Anas Badreen,and Ammar Sultan for their helpful suggestions and constructive criticisms of this thesis.

A big 'thank you' goes to my parents, brothers and sisters for their love, encouragement, support and for their prayers. My thanks too to all the staff at the faculty of computer science and information technology for their cooperation and kindness throughout my study. I would also like to thank all my friends and colleagues, for their encouragement, cooperation and help. I learnt a great deal from each of you.

APPROVAL

I certify that a Thesis Examination Committee has met on (10 July 2014) to conduct the final examination of (WA'EL JUM'AH AL-ZYADAT) on his thesis entitled" DATA FILTERING FRAMEWORK FOR PRESERVING MEANINGFUL DATA RECORDS FROM STREAMS OF UNSTRUCTURED WEATHER DATA" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U. (A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Abdul Azim b Abd Ghani, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Rusli bin Hj Abdullah, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Fatimah bt Sidi, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Xiaofang Zhou, PhD
Professor
School of Information Technology & Electrical Engineering
University of Queensland
Australia
(External Examiner)

ZULKARNAIN ZAINAL, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 12 March 2015

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

Rodziah Binti Atan, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Hamidah Ibrahim, Ph.D.
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Masrah Azrifah Binti Azmi Murad, Ph.D
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declarationby graduate student

I hereby confirm that:

- This thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of the thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceeding, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/ fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ---------- Date: July-2014

Wa'el Jum'ah Al-zyadat, GS24106

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ----------
Rodziah Binti Atan, Ph.D.

Signature: ----------
Hamidah Ibrahim, Ph.D.

Signature: ----------
Masrah Azrifah Binti Azmi Murad, Ph.D

TABLE OFCONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF ABBREVAITIONS

°C	Celsius
ACD	Analog-to-Digital Converter
ANN	Artificial Neural Network
BNNL	Bayesian neural network learning
CAD	Convert Analog to Digital
CD	Class Dependent
CL	Class Independent
CoCORaHS	Community Collaborative Rain Hail & Snow Network
CORDIC	COordinate Rotation DIgital Computer
DBMS	Data Base Management System
DC	Data Centre
DSMS	Data Stream Management System
DWT	Discrete Wavelet Transform
EKF	Extended Kalman Filtering
EnKF	Ensemble Kalman Filtering
FCBF	Fast Correlation Based Filtering
FFT	Fourier Filtering Bank
FIFO	First In First Out
GPS	Global Positioning System
GUI	Graphical User Interface
IADEM	Incremental Algorithm Driven by Error Margins
JMM	Jabatan Meteorological Malaysia
KDD	Knowledge Discovery in Database
KNN	K Nearest Neighbors
LP	Linear Prediction
ML	Machine Learning
MMD	Malaysian Meteorological Department
MSE	Mean Square Error
NIDS	Network Intrusion Detection System
NN	Neural Network
NWS	National Weather Service
O/D	Origin Destination
OOP	Object Oriented Programming
OS	Operating System
PBMCD	Process Based on the Minimum Class Difference
PCA	Principle Component Analysis
PVS	Potentially Visible Set
RFID	Radio Frequency Identification
SAD	Shuffle Analog into Digital

SIFTER	Smart Information Filtration Technology for Electronic Resource
SLP	Symmetric Linear Prediction
SPO	Structure Preserving Oversampling
SQP	Sequential Quadratic Programming
SU	Symmetrical Uncertainty
SVM	Support Vector Machine
UPM	University Putra Malaysia
VDBMS	Video Database Benchmarking
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years there has been a rise in the number of distributed systems supporting applications that continuously collect, aggregate and disseminate data from information sources across a network. Those data sources, such as click stream or sensor data,are often characterized as fast rate high-volume "stream" (Babcock, Babu, Datar, Motwani, and Widom, 2002). The rapid accumulation of data from an environment has become an inseparable part of human understanding of factual information that impacts several aspects of our lives (Dunham and Margaret, 2003).

Data collection in streaming data is the process of sampling signals that measure real world physical conditions and converting the resulting sample into digital numeric value that can be manipulated by a computer such as sensors that convert physical parameter to electrical signals (Kos, Tomaž, Kosar, Mernik, and Marjan, 2012; Patnaik, Marwah, Sharma, and Ramakrishnan, 2011; Sun and Lee, 2006).The main challenge in data collection is to automatically generate unstructured data from the raw data (possibly with some associated uncertainty) that are gathered by sensor devices or measurements. The raw data are spatially and temporally correlatedand are notin a format ready for analysis (Weinberg, David, Beers, Blanton, and Eisenstein, 2007).The data pre-process is the firstcomponent to perform on the raw data to prepare it for another processing procedure (Singh and Kumar, 2013).

Data pre-processis an important component in the data quality process that involves transforming raw data into an understandable format in which the real world data are often incomplete and inconsistent.It is a preliminary processing of data in order to prepare the data for the primary processing for further analysis. The term can be applied to any first or preparatory processing stage when there are several steps required to prepare data for the user. For example, extracting data from a large set, filtering it for various reasons and combining sets of data could be a pre-process step (Schmieder, Edwards, and Robert, 2011; Famili, Shen, Wei, Richard, and Simoudis, 1997).

Data filtering is a fundamental process to avoid massive data. In this millennium, data continue to grow exponentially which sometimes becomes unmanageable. Those who ignore the potential danger of massive data will face problems and suffer the consequences of data overload and missing data which are the challenges of streaming data, especially when data are captured from the environment, which clearly effect the quality of data and the storage of it (Venugopal, Srinivasa, and Patnaik, 2009).Data filtering is a component that refines datasets into meta-data based on what a user (or set of users) needs, without including repetitive, irrelevant or even sensitive (Surapong,

Glesner, and Klingbeil, 2010; Condon, Deshpande, and Hellerstein, 2009; Chandola and Kumar, 2007; Lee and Shieh, 2006).In general, data filtering consists of two main tasks, namely: pre-data filtering and post-data filtering (classification).Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class, which is performed after the pre-data filtering.Classification enables the separation and classification of data according to the dataset requirements of various domains' objectives (Zhang, Zhu, Bond, and Jeffrey 2011; Zhang et al., 2008).

There are many applications for a wide range of uses requiring data filtering which include climatic data, ecological data, and streaming data. This study indicated to propose a framework to meet the challenges in stream of weather data filtering.

1.2 Problem Statement

The issue of data filtering remain at the forefront of IT-related development such as software, website, and information systems, mainly due to a large number of negative reviews on streaming data filtering. Aggarwal (2013) states that poor and unstructured data records produced up to 30% of damaged data,which are also happens to data collected real-time or in-streams.The issue of unstructured data especially for streaming weather data captured using sensors and other electronic deviceshas affect to data filtering stage. The whole scenario setting amplifies the challenges in data filtering (Ou, Yang, GuangZhi, and Dai, 2011; Fu Zhao and Leong, 2000).The difference in type of data captured (e.g. weather temperature, humidity, wind speed, or air pressure), different value for weather elements (e.g. numerical, alphanumerical, date, time) aggregated with direct capture from the environment in unformatted data records, is surely a mega challenge to be solved. There is also limitation to structure weather data records in specific form which are: basic data identification, creating suitable set of contour, and documentations (Brown and Jones, 2001; Uri Hanani, Bracha Shapira, and Peretz Shoval, 2001; Ayoade, 1976).

Current filtering approaches faced limitations in several aspects such as difficulties in matching the weather attributes into specific format, weak validation for filtered records (in terms of significance values), re-tracing filtered (omitted or removed) records from different sources; chaotic re-combination of filtered data records (main records accepting filtered input), filtered dataset is having one specific format (not possible to match record column in other database) by which, will increase the iteration process to identify the correct record and column (Cao, Nguyen, Krishnaswamy, Shonali, and Xiao, 2012; Esseghir, 2010; Dongsheng, Jiannong, Xicheng, and Chen, 2009).

Streaming join in of weather data filtering is a fundamental operation to merge information from different streams which also involves non-quality join (Xie, Junyi, Yang, and Jun, 2007). This is especially useful in many applications such as sensor networks in which the streams arriving from different sources may need to be related

with one another (Simon Fong, Robert, and Yain whar, 2014; Shah, Dharmarajan, Ramamritham, and Krithi, 2003). In the stream setting, input tuples arrive continuously, and result tuples need to be produced continuously as well. The receivers cannot assume that the input data is already indexed, or that the input rate can be controlled by the query plan (Dasu, Tamraparni, Krishnan, Venkatasubramanian, and Suresh, 2006; Garofalakis, Minos, Gehrke, Rastogi, and Rajeev, 2002). Standard join that use blocking operations, for example, sorting is no longer effective. Conventional methods for matching and query optimization are also inappropriate, due to finite input assumptions. Moreover, the long-running nature of stream queries call for more adaptive process strategies that can react to changes and variation of streaming weather data characteristics (Cao et al., 2012; Esseghir, 2010; Dongsheng et al., 2009).

Various methods such as decision trees (Gama, João, Kosina, and Petr, 2011), rule based methods, and neural networks are normally used to classify data. These techniques are designed to build classification models for static data, including archived/stored weather datasets, where these data can be chunked into smaller set of records (several passes) (Cugola, Gianpaolo, Margara, and Alessandro, 2012). However, chunking streaming weather data is not possible and processing the entire dataset as one stream (one pass) is necessary. Furthermore, the classification approach needs to be re-designed in the context of pre-data filtering, which is unique for streams of data. Gathering and preserving data records meaning using sensor devices is also an issue where these processes require special and subtle technique to enable the formation of structured data record and type as discussed in Algarni, et al., (2009) and Sawai, et al., (2003). All interrelated issues addressed in this study are ought to be solved and signify this study.

1.3 Research Objectives

This research aims at achieving the following:

1- To characterize an unstructured streaming data with variable format that precisely preserves the sections of data for efficient retrieval.
2- To propose a data filtering framework for streams unstructured weather data preserving the meaningful and relevancy of data records.

1.4 Scope of Research

This research will focus on streaming weather data records filtering from Malaysia, consisting of streaming weather (gathered using sensors) and static weather data (stored/archived). The streaming weather data are the sensor-based data capturing for 13 days using a sensor device called Cornus Sensory Network in the area of Serdang. The static data are weather data purchased from the Malaysian Meteorological Department (MET) covering 11 locations in Malaysia; Alor Setar, Bayan Lepas, Cameron Highlands, Chuping, Ipoh, Kota Bharu, Kuala Terengganu Airport, Kuantan, Malacca, Mersing, and Petaling Jaya for the years of 2008 and 2009. *Appendix A* shows the data in its original form.

The reasons for using these two sets of data are; to benchmark and validate the data collected by Cornus Sensor device and to verify the end results of the proposed data filtering approach for streams of weather data.

Research question to be addressed in this study are; how to structure streams of weather data collected from sensor devices to the form suitable for further analysis? And how to develop a data filter that preserved meaningfulness and relevancy?

1.5 Thesis Organization

There are six chapters in this thesis, including the introductory chapter, explaining the background, problem statements, objectives, scope of research, and thesis organization.

Chapter 2 is the literature review of the related studies in data collection, data pre-process, data filtering and classification. It reviews the current published research in this area to support the research.

Chapter 3 explains the five phases of framework methodology: data collection phase; data pre-process phase, data filtering phase, data classification phase and visualization phase proposed. This chapter also explains and discusses the data collection that is used in this research and the pre-process techniques that are applied on the two datasets (static and dynamic data).

Chapter 4 describes the design and development of the data filtering framework.

Chapter 5 discusses the results component of each of data filtering framework, data pre-process effect to the relevant data, data filtering accrued from the matching data, and 5M method to preserve the meaningful data using the proposed filtering method.

Chapter 6 discuss the conclusion and future work of the research. It also contains explanation that validates and verifies the work carried out in this research. Contributions of this work and future research directions are also included in this chapter.

REFERENCES

Abiteboul, Dallan, and Janet (1997). The Lorel query language for semistructured data. *International journal on digital libraries, 1*(1), 68-88.

Aggarwal, C. (2013). *Managing and Mining Sensor Data*. NY: Springer.

Aguilar Saborit, Trancoso, Muntes Mulero, and Larriba Pey. (2008). Dynamic adaptive data structures for monitoring data streams. *Data & Knowledge Engineering, 66*(1), 92-115. doi: DOI: 10.1016/j.datak.2007.12.006

Akioka, Feihui, L., Malkowski, and Irwin. (2008). *Ring data location prediction scheme for Non-Uniform Cache Architectures.* Paper presented at the Computer Design, 2008. ICCD 2008. IEEE International.

Al-Hamadi, and Soliman. (2004). Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. *Electric Power Systems Research, 68*(1), 47-59.

Anca Vaduva, Rguwe Kietz, and Regina. (2001). *M4: a metamodel for data preprocessing*. Paper presented at the Proceedings of the 4th ACM international workshop on Data warehousing and OLAP, Atlanta, Georgia, USA.

Ankur Jain, Edward Chang, and Yuan Fang Wang. (2004). *Adaptive stream resource management using Kalman Filters*. Paper presented at the Proceedings of the 2004 ACM SIGMOD international conference on Management of data, Paris, France.

Aref, Walid, Catlin, AnnChristine, and Xingquan. (2004). VDBMS: A testbed facility for research in video database benchmarking. *Multimedia Systems, 9*(6), 575-585. doi: 10.1007/s00530-003-0129-9

Ari Benbasat, and Paradiso. (2007). *A framework for the automated generation of power-efficient classifiers for embedded sensor nodes*. Paper presented at the Proceedings of the 5th international conference on Embedded networked sensor systems, Sydney, Australia.

Audet, and Dennis. (2004). A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization, 14*(4), 980-1010.

Ayoade. (1976). On the use of multivariate techniques in climatic classification and regionalization. *Theoretical and Applied Climatology, 24*(4), 257-267. doi: 10.1007/bf02263458

Babcock, Babu, Datar, Motwani, and Widom. (2002). *Models and issues in data stream systems.* Paper presented at the Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.

Baclace. (1991). *Personal information intake filtering.* Paper presented at the Bellcore information filtering workshop.

Baclace. (1992). Competitive agents for information filtering. *Communications of the ACM, 35*(12), 50.

Baesens, Jan, Dedene, and Guido. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research, 138*(1), 191-211. doi: http://dx.doi.org/10.1016/S0377-2217(01)00129-1

Beex, and Zeidler. (2002). *Data structure and non-linear effects in adaptive filters.* Paper presented at the Digital Signal Processing, 2002. DSP 2002. 2002 14th International

Ben Greenstein, Christopher Mar, Alex Pesterev, Shahin Farshchi, Eddie Kohler, Jack Judy, and Deborah Estrin. (2006). *Capturing high frequency phenomena using a bandwidth limited sensor network.* Paper presented at the Proceedings of the 4th international conference on Embedded networked sensor systems, Boulder, Colorado, USA.

Bifet, Albert, Holmes, and Bernhard. (2010). MOA: Massive Online Analysis. *The Journal of Machine Learning Research, 11*, 1601-1604.

Bifet, Albert, Holmes, Geoff, Seidl, and Thomas. (2011). MOA: a real-time analytics open source framework *Machine Learning and Knowledge Discovery in Databases* (pp. 617-620): Springer.

Bifet, Albert, Pfahringer, and Bernhard. (2010). MOA: Massive Online Analysis. *The Journal of Machine Learning Research, 11*, 1601-1604.

Bifet, Holmes, Kirkby, and Pfahringer. (2010). Moa: Massive online analysis. *The Journal of Machine Learning Research, 11*, 1601-1604.

Blake, and Merz. (1998). {UCI} Repository of machine learning databases.

Borlund, and Pia. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54*(10), 913-925.

Brown, and Jones. (2001). Context-aware Retrieval: Exploring a New Environment for Information Retrieval and Information Filtering. *Personal Ubiquitous Comput., 5*(4), 253-263. doi: 10.1007/s007790170004

Brunk. (1965). *An introduction to mathematical statistics*: Blaisdell.

Campos Velho, and Morais Furtado. (2011). Adaptive Particle Filter for Stable Distribution. In C. Constanda & P. Harris (Eds.), *Integral Methods in Science and Engineering* (pp. 47-57): Birkhäuser Boston.

Cao, Nguyen, Krishnaswamy, Shonali, and Xiao. (2012). An Integrated Framework for Human Activity Classification.

Carney, Don, Çetintemel, Uğur, Zdonik, and Stan. (2002). *Monitoring streams: a new class of data management applications.* Paper presented at the Proceedings of the 28th international conference on Very Large Data Bases.

Carsten Lanquillon , and lngrid Renz. (1999). *Adaptive Information Filtering: Detecting Changes in Text Streams* Paper presented at the ACM, Kansas City, Missouri, United States.

Chandola, and Kumar. (2007). Summarization compressing data into an informative representation. *Knowledge and information systems, 12*(3), 355-378.

Chen. (2010). Adaptive equalizer for communication channels: Google Patents.

Chen, Tao, Morris, Julian, and Elaine. (2007). Particle filters for dynamic data rectification and process change detection. In Z. Hong-Yue (Ed.), *Fault Detection, Supervision and Safety of Technical Processes 2006* (pp. 204-209). Oxford: Elsevier Science Ltd.

Chen, Tao, Morris, Julian, and Martin. (2008). Dynamic data rectification using particle filters. *Computers & Chemical Engineering, 32*(3), 451-462. doi: DOI: 10.1016/j.compchemeng.2007.03.012

Chen, Zhiping, and Kevin. (2006). A preprocess algorithm of filtering irrelevant information based on the minimum class difference. *Knowledge-Based Systems, 19*(6), 422-429.

Chou, and Verhaegen. (1999). *Identification of Wiener models with data pre-filtering.* Paper presented at the IEEE.

Chris Olston, Jing Jiang, and Jennifer Widom. (2003). *Adaptive filters for continuous queries over distributed data streams*. Paper presented at the Proceedings of the 2003 ACM SIGMOD international conference on Management of data, San Diego, California.

Chu, Fang, Zaniolo, and Carlo. (2004). Fast and light boosting for adaptive mining of data streams *Advances in Knowledge Discovery and Data Mining* (pp. 282-292): Springer.

Condon, Deshpande, and Hellerstein. (2009). Algorithms for distributional and adversarial pipelined filter ordering problems. *ACM Transactions on algorithms, 5*(2), 24-34.

Cugola, Gianpaolo, Margara, and Alessandro. (2012). Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR), 44*(3), 15.

Das, Resul, Turkoglu, and Ibrahim. (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications, 36*(3, Part 2), 6635-6644. doi: http://dx.doi.org/10.1016/j.eswa.2008.08.067

Dasu, Tamraparni, Krishnan, Venkatasubramanian, and Suresh. (2006). *An information-theoretic approach to detecting changes in multi-dimensional data streams.* Paper presented at the In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications.

dataid. (2006). What's A Bar Code? . from http://www.dataid.com/dcsystems.htm

Di Paolo Emilio, and Maurizio. (2013). Data acquisition systems.

Ding, S., Zhao, S., Yuan, Q., Zhang, X., Fu, R., and Bergman, L. (2008). *Boosting collaborative filtering based on statistical prediction errors.* Paper presented at the Proceedings of the 2008 ACM conference on Recommender systems.

Dongsheng, Jiannong, Xicheng, and Chen. (2009). Efficient Range Query Processing in Peer-to-Peer Systems. *Knowledge and Data Engineering, IEEE Transactions, 21*(1), 78-91. doi: 10.1109/tkde.2008.99

Dunham, and Margaret. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, N.J.: Prentice Hall/Pearson Education.

Düntsch, and Gediga. (1998). Uncertainty measures of rough set prediction. *Artificial intelligence, 106*(1), 109-137.

Efromovich, and Sam. (1999). Filtering and Asymptotics *Nonparametric Curve Estimation* (pp. 259-322): Springer New York.

El faouzi, Billot, Romain, Bouzebda, and Salim. (2010). Motorway travel time prediction based on toll data and weather effect integration. *IET intelligent transport systems, 4*(4), 338-345.

Esseghir. (2010). Effective Wrapper-Filter hybridization through GRASP Schemata. *JMLR: Workshop and Conference Proceedings 10: 45-54 The Fourth Workshop on Feature Selection in Data Mining*, 10.

Falkenberry. (2002). High frequency data filtering. *Tick Data, Technical, 04*(02), 16-21.

Famili, Shen, Wei, Richard, and Simoudis. (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis, 1*(1–4), 3-23. doi: http://dx.doi.org/10.1016/S1088-467X(98)00007-9

Fidel, and Crandall. (1997). *Users perception of the performance of a filtering system*.

Fischer, and Stevens. (1991). *Information access in complex, poorly structured*

information spaces. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology.

Fletcher, and Leyffer. (1999). A bundle filter method for nonsmooth nonlinear optimization. *University of Dundee, Report NA/195*.

Foltz, and Dumais. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM, 35*(12), 51-60.

Freund, and Yoav. (2001). An adaptive version of the boost by majority algorithm. *Machine learning, 43*(3), 293-318.

Fu Zhao, and Leong. (2000). A Data Preprocessing Framework for Supporting Probability-Learning in Dynamic Decision Modeling in Medicine. *NCBI*, 5.

Gama, João, Kosina, and Petr. (2011). *Learning decision rules from data streams.* Paper presented at the IJCAI Proceedings-International Joint Conference on Artificial Intelligence.

Garofalakis, Minos, Gehrke, Rastogi, and Rajeev. (2002). *Querying and mining data streams: you only get one look a tutorial.* Paper presented at the SIGMOD Conference.

Gewehr, Jan, Szugat, Martin, Zimmer, and Ralf. (2007). BioWeka extending the weka framework for bioinformatics. *Bioinformatics, 23*(5), 651-653.

Gilad. (1988). *The business intelligence system: A new tool for competitive advantage*: American Management Association.

Gilleron, Marty, and Tommasi. (2006). *Interactive Tuples Extraction from Semi-Structured Data.* Paper presented at the Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference.

Gould, Sainvitu, and Toint. (2006). A filter trust region method for unconstrained optimization. *SIAM Journal on Optimization, 16*(2), 341-357.

Gould, and Toint. (2007). FILTRANE, a Fortran 95 filter trust region package for solving nonlinear least-squares and nonlinear feasibility problems. *ACM Transactions on Mathematical Software (TOMS), 33*(1).

Greenstein, Mar, Pesterev, Farshchi, Kohler, Judy, and Estrin. (2006). *Capturing high frequency phenomena using a bandwidth limited sensor network.* Paper presented at the Proceedings of the 4th international conference on Embedded networked sensor systems.

Grisettiyz, Stachniss, and Burgard. (2005). *Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective*

*resampling.*Paper presented at the Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International

Groen, Frank, and Smith. (2009). *Concept for the NASA risk and reliability data collection and analysis environment.* Paper presented at the Reliability and Maintainability Symposium, 2009. RAMS 2009. Annual.

Gürbüz, Feyza, Özbakir, Lale, Yapici, and Hüseyin. (2011). Data mining and preprocessing application on component reports of an airline company in Turkey. *Expert Systems with Applications, 38*(6), 6618-6626.

Heasoo Hwang, Andrey Balmin, Hamid Pirahesh, and Berthold Reinwald. (2007). *Information discovery in loosely integrated data*. Paper presented at the Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China.

Holmes, Geoffrey, Donkin, Andrew, and Witten. (1994). *Weka: A machine learning workbench.* Paper presented at the Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand

Houtekamer, and Mitchell. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review, 129*(1), 123-137.

Hua Tang, Hui Zhang, and Alex Doboli. (2003). *Synthesis of continuous-time filters and analog to digital converters by integrated constraint transformation, floorplanning and routing*. Paper presented at the Proceedings of the 13th ACM Great Lakes symposium on VLSI, Washington, D. C., USA.

Huang, R., and Záruba, G. (2007). Location tracking in mobile ad hoc networks using particle filters. *Journal of Discrete Algorithms, 5*(3), 455-470. doi: DOI: 10.1016/j.jda.2006.12.005

Huijuan, Song, Yuanhua, Jia, Zhiqiang, and Shu. (2009). *Study on Traffic Data Pre-Processing Technology Based on RTMS.* Paper presented at the Intelligent Computation Technology and Automation, 2009. ICICTA '09. .

Hulten, Geoff, Domingos, and Pedro. (2003). VFML–a toolkit for mining high-speed time-changing data streams. *URL http://www. cs. washington. edu/dm/vfml*.

Jennings, and Higuchi. (1992). A personal news service based on a user model neural network. *IEICE Transactions on Information and Systems, 75*(2), 198-209.

Juhng Perng , Ting Lee, and Ker Wei (2007). *A new two-input single-output fuzzy controller*. Paper presented at the Proceedings of the Ninth IASTED International Conference on Control and Applications, Montreal, Quebec, Canada.

Jung, and Jason. (2010). Towards Semantic Preprocessing for Mining Sensor Streams

from Heterogeneous Environments. In N. Nguyen, M. Le & J. Świątek (Eds.), *Intelligent Information and Database Systems* (Vol. 5990, pp. 103-112): Springer Berlin Heidelberg.

Kadota, Ago, Horiuchi, and Ikeura. (2002). Very small IF resonator filters using reflection of shear horizontal wave at free edges of substrate. *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions, 49*(9), 1269-1279.

KaiYu, Tresp, V., and Shipeng. (2004). *A nonparametric hierarchical bayesian framework for information filtering*. Paper presented at the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom.

Karas, Ribeiro, Sagastizábal, and Solodov. (2009). A bundle-filter method for nonsmooth convex constrained optimization. *Mathematical Programming, 116*(1), 297-320.

Khosla, and Dillon. (1997). *Engineering intelligent hybrid multi-agent systems*: Kluwer Academic Pub.

Kos, Tomaž, Kosar, Mernik, and Marjan. (2012). Development of data acquisition systems by using a domain specific modeling language. *Computers in industry, 63*(3), 181-192.

Kotsiantis, and Kanellopoulos. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71-82.

Kotsiantis, Kanellopoulos, and Pintelas. (2006a). Data Preprocessing for Supervised Leaning. *International Journal of Computer Science and Communication, 1*, 7.

Kotsiantis, Kanellopoulos, and Pintelas. (2006b). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25-36.

Kranen, Philipp, Read, and Jesse. (2012). *Stream data mining using the MOA framework*. Paper presented at the Database Systems for Advanced Applications.

Kuncheva, and Ludmila. (2013). Change detection in streaming multivariate data using likelihood detectors. *Knowledge and Data Engineering, IEEE, 25*(5), 1175-1180.

Lee, and Shieh. (2006). Packet classification using diagonal-based tuple space search. *Computer Networks, 50*(9), 1406-1423.

Lee, O. (2000). Design of discrete coefficient FIR and IIR digital filters with prefilter-equalizer structure using linear programming. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions, 47*(6), 562-565. doi: 10.1109/82.847076

Lei, Yuong Wei, Ouhyoung, and Ming. (1997). Carving: a novel method of visibility preprocess- ing for un-restricted three-dimensional environments. *The Visual Computer, 13*(6), 283-294. doi: 10.1007/s003710050104

Lermusiaux, and Pierre. (1999). Data assimilation via error subspace statistical estimation. Part II: Middle Atlantic Bight shelfbreak front simulations and ESSE validation. *Monthly Weather Review, 127*(7), 1408-1432.

Li, Xiao Bai, Jacob, and Varghese. (2008). Adaptive data reduction for large-scale transaction data. *European Journal of Operational Research, 188*(3), 910-924.

Liebchen, Twala, and Shepperd. (2007). *Filtering, robust filtering, polishing: Techniques for addressing quality in software data.* Paper presented at the Empirical Software Engineering and Measurement, 2007. ESEM 2007.

Lingam, P. (2010). *Freezing as a correctness measure for Multiversion Timestamp Ordering protocol.* Paper presented at the Computer Engineering and Technology (ICCET), 2010.

Logan, Judith, Gorman, Paul, Middleton, and Blackford. (2001). *Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data.* Paper presented at the Proceedings of the AMIA Symposium.

Lopes, N., Correia, D., and Pereira, C. (2012). *An incremental hypersphere learning framework for protein membership prediction*. Paper presented at the Proceedings of the 7th international conference on Hybrid Artificial Intelligent Systems Salamanca, Spain.

Ma, H., Huang, J., Zhu, D., Liu, J., Su, W., Zhang, C., and Fan, J. (2013). Estimating regional winter wheat yield by assimilation of time series of HJ-1 CCD NDVI into WOFOST–ACRM model with Ensemble Kalman Filter. *Mathematical and Computer Modelling, 58*(3), 759-770.

Mahmood, Moa, Becker, and Jack. (1985). Effect of organizational maturity on end-users' satisfaction with information systems. *Journal of Management Information Systems*, 37-64.

Markov, Zdravko, Russell, and Ingrid. (2006). *An introduction to the WEKA data mining system.* Paper presented at the ACM SIGCSE Bulletin.

McMahon, Michael, Dascalu, Harris, Fredrick, and Strachan. (2011). *SENSOR-Applying Modern Software Data Management Practices to Climate Research.* Paper presented at the Proceedings of the 2011 Workshop on Sensor Network Applications.

Ming, Li, and Kotz. (2007). *Group-aware Stream Filtering.* Paper presented at the Distributed Computing Systems Workshops, 2007. ICDCSW '07. 27th

International

Ming Li, and David Kotz. (2008). *Event dissemination via group-aware stream filtering*. Paper presented at the Proceedings of the second international conference on Distributed event-based systems, Rome, Italy.

Mingo, Phil Antony, Rafidha Rehiman, Balakrishnan, and Kannan. (2013). An Autonomous Framework for Classifier Selection in Weka.

Mostafa, Mukhopadhyay, Palakal, and Lam. (1997). A multilevel approach to intelligent information filtering: model, system, and evaluation. *ACM Transactions on Information Systems (TOIS), 15*(4), 368-399.

Natarajan, and Gilbert. (1997). On direct PID controller tuning based on finite number of frequency response data. *ISA Transactions, 36*(2), 139-149. doi: Doi: 10.1016/s0019-0578(97)00014-1

Nedellec, Claire, Vetah, Mohamed Ould Abdel, Bessières, and Philippe. (2001). Sentence filtering for information extraction in genomics, a classification problem *Principles of Data Mining and Knowledge Discovery* (pp. 326-337): Springer.

Ochoa, Xavier, Duval, and Erik. (2008). Relevance Ranking Metrics for learning objects. *Learning Technologies, IEEE, 1*(1), 34-48.

Oldfield, and Kotz. (2001). *Armada: a parallel file system for computational grids.* Paper presented at the Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International

Oña, Gómez, and Mérida Casermeiro. (2011). Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation. *Transportation Research Part C: Emerging Technologies, 19*(1), 29-39. doi: http://dx.doi.org/10.1016/j.trc.2010.02.005

Ou, Yang, GuangZhi, and Dai. (2011). Color Edge Detection Based on Data Fusion Technology in Presence of Gaussian Noise. *Procedia Engineering, 15*(0), 2439-2443. doi: http://dx.doi.org/10.1016/j.proeng.2011.08.458

Paix. (2011). Climate Data and Data Related Products. 2013, from http://www.wmo.int/pages/themes/climate/climate_data_and_products.php

Palit, Saha, Sethi, and Hennig. (2012). A high speed digital data acquisition system for the Indian National Gamma Array at Tata Institute of Fundamental Research. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 680*(0), 90-96. doi: http://dx.doi.org/10.1016/j.nima.2012.03.046

Patnaik, Marwah, Sharma, and Ramakrishnan. (2011). Temporal data mining

approaches for sustainable chiller management in data centers. *ACM Transactions on Intelligent Systems and Technology, 2*(4).

Peng, Feng, and Li. (2009). A filter-variable-metric method for nonsmooth convex constrained optimization. *Applied Mathematics and Computation, 208*(1), 119-128.

Perera, Alexandre, Marco, and Santiago. (2002). A portable electronic nose based on embedded PC technology and GNU/Linux: hardware, software and applications. *Sensors Journal, IEEE, 2*(3), 235-246.

Ping, de Courten, Maximilian, and Jan. (2009). The development and evaluation of a PDA-based method for public health surveillance data collection in developing countries. *international journal of medical informatics, 78*(8), 532-542.

Powers, and David Martin. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 37-67.

Rao, Doraiswamy, Thakkar, and Colby. (2006). *A deferred cleansing method for RFID data analytics*.

Read, Jesse, Bifet, Geoff, and Bernhard. (2012). Scalable and efficient multi-label classification for evolving data streams. *Machine learning, 88*(1-2), 243-272.

Rizzo, L. (2012). *netmap: A Novel Framework for Fast Packet I/O.* Paper presented at the USENIX Annual Technical Conference.

Robert Hogg, and Allen Craig. (1994). *Introduction to Mathematical Statistics* (5th ed.). New York: Prentice Hall.

Sawai, Tsukamoto, Terada, and Nishio. (2003). *Composition of filtering functions.* Paper presented at the Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International

Schmieder, Edwards, and Robert. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics, 27*(6), 863-864.

Shah, Dharmarajan, Ramamritham, and Krithi. (2003). *An efficient and resilient approach to filtering and disseminating streaming data.* Paper presented at the Proceedings of the 29th international conference on Very large data bases.

Sheth, and Maes. (1993). *Evolving agents for personalized information filtering.* Paper presented at the Artificial Intelligence for Applications, 1993. Proceedings, Ninth Conference.

Simon Fong, Robert, and Yain whar. (2014). A Lightweight Data Pre-Processing Strategy with Fast Contradiction Analysis for Incremental Classifier Learning. *Hindawi Publishing Corporation, 30*, 1-16.

Singh, and Kumar, N. (2013). *Using Event-B for Critical Device Software Systems*: Springer Science & Business.

Srivastava, and Shweta. (2014). Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Applications, 88*(10), 26-29.

Storkey, A. (2006). Data Mining and Exploration: Preprocessing. 3-14. www.inf.ed.ac.uk/teaching/courses/dme/slides/preproc-print4up.pdf

Sun, and Lee. (2006). Case study of data centers' energy performance. *Energy and buildings, 38*(5), 522-533.

Sung Ho, Jun Ho, and Yong Man. (2006). Meaningful scene filtering for TV terminals. *Consumer Electronics, IEEE Transactions, 52*(1), 263-268.

Sungho, Yukyung, and Joohyoung. (2009). *Robust horizontal target detection with cooperative spatial filtering.* Paper presented at the Infrared, Millimeter, and Terahertz Waves, 2009. IRMMW-THz 2009. 34th International

Surapong, Glesner, and Klingbeil. (2010). *Implementation of realtime pipeline-folding 64-tap filters on FPGA.* Paper presented at the Ph. D. Research in Microelectronics and Electronics (PRIME), 2010

Szekely, and Torres. (2004). A semantic data collection model for sensor network applications.

Taylor, and John. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements* (Vol. 1).

Thompson, OShea, Hussain, and Steele. (2004). Efficient single-bit ternary digital filtering using sigma-delta modulator. *IEEE Signal Processing Letters, 11*(2), 164-166.

Thorisson, Smith, Krishnan, and Stein. (2005). The international HapMap project web site. *Genome research, 15*(11), 1592-1593.

Tran, McGregor, and Diao. (2010). Conditioning and aggregating uncertain data streams: Going beyond expectations. *Proceedings of the VLDB Endowment, 3*(1-2), 1302-1313.

Uri Hanani, Bracha Shapira, and Peretz Shoval. (2001). Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction, 11*(3), 203-259. doi: 10.1023/a:1011196000674

Vasak, M., Gulin, M., Vic, Nikolic, D., Pavlovic, T., and Peric, N. (2011). *Meteorological and weather forecast data-based prediction of electrical power delivery of a photovoltaic panel in a stochastic framework.* Paper presented at the

MIPRO, 2011 Proceedings of the 34th International Convention.

Velayutham, and Thangavel. (2011). Unsupervised quick reduct algorithm using rough set theory. *Journal of Electronic Science and Technology*, *9*(3), 193-201.

Venugopal, Srinivasa, and Patnaik. (2009). *Soft computing for data mining applications* (Vol. 190): Springer Verlag.

Villarroya, S., Viqueira, J. R. R., Cotos, J. M., and Flores, J. C. (2013). GeoDADIS: A framework for the development of geographic data acquisition and dissemination servers. *Computers & Geosciences*, *52*(0), 68-76. doi: http://dx.doi.org/10.1016/j.cageo.2012.09.013

Vinod, Lai, Premkumar, and Lau. (2004). *Optimization method for designing filter bank channelizer of a software defined radio using vertical common subexpression elimination.* Paper presented at the Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium.

Wang, Enli, Zhang, HS, F., and Wang. (2011). Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data. *Water Resources Research*, *47*(5), 15.

Warren. (1973). *A continuous-discrete data filter for pre-filtered observations.* Paper presented at the Decision and Control including the 12th Symposium on Adaptive Processes, 1973 IEEE

Watanabe, Mizuno, and Makino. (2003). An all-digital analog-to-digital converter with 12-μV/LSB using moving-average filtering. *Solid-State Circuits, IEEE Journal*, *38*(1), 120-125.

Wedin, Bogren, and Grabec. (2008). D3. 1 Data filtering methods. *1*(I), 36.

Weinberg, David, Beers, Blanton, and Eisenstein. (2007). *SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems.* Paper presented at the Bulletin of the American Astronomical Society.

Witten, Ian, Frank, and Eibe. (2005). *Data Mining: Practical machine learning tools and techniques* (Vol. 2): Morgan Kaufmann.

Wu, Liu, and Li. (2006). *The Method of Data Pre-processing in Grey Information Systems.* Paper presented at the Control, Automation, Robotics and Vision, 2006. ICARCV '06. 9th International Conference

Xie, Junyi, Yang, and Jun. (2007). A survey of join processing in data streams *Data Streams* (Vol. I, pp. 209-236): Springer.

Yang, Wilson, and Wang. (2010). Development of an automated climatic data scraping,

filtering and display system. *Computers and Electronics in Agriculture, 71*(1), 77-87. doi: DOI: 10.1016/j.compag.2009.12.006

Yiqun Liu, Canhui Wang, Min Zhang, and Shaoping Ma. (2005). *Web data cleansing for information retrieval using key resource page selection*. Paper presented at the Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan.

Yu, and Liu. (2003). *Feature selection for high-dimensional data: A fast correlation-based filter solution*, USA.

Yunhong, Grossman, and Robert. (2011). Toward efficient and simplified distributed data intensive computing. *Parallel and Distributed Systems, IEEE Transactions, 22*(6), 974-984.

Zhang, Wen, and Cheng. (2010). Concurrent and Storage-Aware Data Streaming for Data Processing Workflows in Grid Environments. *Tsinghua Science & Technology, 15*(3), 335-346.

Zhang, Wen, Yoshida, Taketoshi, Tang, and Xijin. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems, 21*(8), 879-886.

Zhang, Zhu, Bond, and Jeffrey (2011). Corrective classification: Learning from data imperfections with aggressive and diverse classifier ensembling. *Information Systems, 36*(8), 1135-1157. doi: http://dx.doi.org/10.1016/j.is.2011.05.002

Zhao, Yongheng, Zhang, and Yanxia. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research, 41*(12), 1955-1959. doi: http://dx.doi.org/10.1016/j.asr.2007.07.020

Zwillinger, D. (2011). *CRC standard mathematical tables and formulae* (Vol. 1): CRC press.