Conjoint Mining of Data and Content with Applications in Business, Bio-medicine, Transport Logistics and Electrical Power Systems

Tharam Dillon, Yi-Ping Chen, Elizabeth Chang, Mukesh Mohania

Conjoint Mining of Data and Content with Applications in Business, Bio-medicine, Transport Logistics and Electrical Power Systems

Tharam S. Dillon[1,2], Yi-Ping Phoebe Chen[1], Elizabeth Chang[2], Mukesh Mohania[3]

[1]Department of Computer Science and Computer Engineering,
La Trobe University, Melbourne, Victoria 3086, Australia.
[2]School of Business, Australian Defence Force Academy,
University of New South Wales, Canberra, Australia.
[3]IBM India Research Lab.
Tharam.dillon7@gmail.com, Phoebe.Chen@latrobe.edu.au,
Elizabeth.chang@unsw.edu.au, mkmukesh@in.ibm.com

1 Introduction

Digital information within an enterprise consists of (1) *structured data* and (2) *unstructured content*. The structured data includes enterprise and business data like sales, customers, products, , accounts, inventory and enterprise assets, etc. while the content includes contracts, reports, emails, customer opinions, transcribed calls, online inquires, complements and complaints. Further, cutting edge businesses also using GPS tracking or surveillance monitors as well as sensor technologies for productivity, performance and efficiency measures, and these are provided by outsourcers etc. Similarly in the Biomedical area, resources can be structured data say in Swiss-Prot or unstructured text information in journal articles stored in content repositories such as PubMed. The structured data and the unstructured content generally reside in entirely separate repositories with the former being managed by a DBMS and the latter by a content manager frequently provided by an outsourcer or vendor [76]. This separation is undesirable since the information content of these sources is complementary. Further, each outsourcer or vendor keep the data on their own Cloud, and data are not sharable between the vendor systems, and most vendor system were not integrated with the enterprise systems, and leaves the organization to consolidate the data and information manually for data analytics. Effective knowledge and information use requires seamless access and intelligent analysis of information in its totality to allow enterprises to gain enhanced critical insights. This is becoming even more important, as the proportion of structured to unstructured information has shifted from 50-50 in the 1960s to 5-95 today [1]. Unless we can effectively utilize the unstructured content conjointly with the structured data, we will only obtain very limited and shallow knowledge discovery from an increasingly narrow slice of information. The techniques developed in our research will then be used to address significant issues in three application areas, but potential applications with significant impact are much more extensive.

2 Aims and Underlying Issues

We develop a methodology and techniques for deriving deep knowledge from structured and unstructured information conjointly. This methodology would be useful in several Business Intelligence and Advanced Analytics utilizing Data and Content (AADC), information integration applications, such as managing customer attrition, targeted marketing, fraud detection and prevention, compliance, and customer relationship management as well as a number of fields which involve information of high complexity such as Bioinformatics, Transport Logistics, Business and Electrical Power Systems (EPS). The broad aims of this research are to:

a) use structured data to disambiguate text segments and link them to records for AADC investigations
b) use ontologies to disambiguate and annotate content segments to permit AADC investigations
c) develop methods of AADC for conjoint analysis of linked structured and content elements
d) apply these techniques for investigation of business problems, biomedical problems, electrical power system problems and logistics and transportation problems.

To achieve the aims (a) to (c), we have a number of sub-aims that develop techniques for:

1. cleansing and filtering the noisy unstructured data with respect to structured data;
2. text annotation to enrich the unstructured data semantically;
3. fuzzy matching and searching over structured data based on annotated values for deriving the correlation between data and content;
4. discovering the linkages between data and content;
5. representations of the conjoint information that is suitable for AADC;
6. allowing new insights in the form of business intelligence and deep knowledge from the integrated data and content together.

The applications of the techniques developed (aim (d)) help to make better decisions and to better understand behaviors in the field of business, biomedicine, electrical power systems and logistics. In regards to aim (d), we study the following:

1. in business applications: customer experience and focused cross/up-selling and personalized marketing using the data and 'the voice of customer';
2. in biomedical area: Use of structured information from Protein Databases such as Swiss-Prot and epidemiological medical databases together with information from journal articles in PubMed to determine their relationship to specific conditions and obtain diagnostic knowledge for specific diseases;
3. In transport logistics area: movement of people, goods and services tracking for productivity, security and safety; trust, reputation and quality of service (QoS); track and trace of sub-contractor's performance (against SLA); track and trace fuel and vehicle performance, track and trace carbon emission, etc.

4. in Electric Power Systems: (i) Determination of risk events to profitability related to contractual terms for supply through bi-level contracts, in conjunction with structured data such as inflows, maintenance, failure and spot prices; (ii) Refinement of contractual terms and conditions and (iii) Remedial actions to manage and mitigate potential risk conditions.

Advanced Analytics (AADC) is an automated or semi-automated process for eliciting novel embedded knowledge, patterns and associations from data and information repositories, that is useful and understandable [2]. Thus information could be structured data and unstructured information such as freeform text. Furthermore the distribution of data both in an output class or category could be relatively flat or peaky with a multi-modal distribution.

Structured discrete data works well with (i) Association Rule Mining Techniques [3, 4], (ii) Decision Tree Methods [5, 6], and (iii) Rule Search Methods [7]. Effective methods for continuous structured data include Neural Net Based Rule extraction methods [8, 9]. Association Rules are of the form X => Y (s, c) where X and Y are sets of items and s, c are support and confidence respectively [10]. Association Rule Mining has since been extended to include efficient Apriori-like mining methods [10], query mining [11], constraint based rule mining [12, 13], mining correlations and causal structures[14] and interesting associations [15] and mining associations from semi-structured data [16-18]. An XML document possesses a hierarchical document structure, and this is frequently modelled using a labelled ordered tree.

Our research group [17] initiated an XML-enabled association rule framework. It extends the notion of associated items to XML fragments or subtrees so that one finds associations among ordered trees rather than simple items as in classical association rules. Work has been proposed for mining XML documents [16-19], including a collection of semi-structured objects [17]. To help with mining XML documents and other structures, frequent sub tree mining algorithms are being developed [20-22] which focus on extraction of different types of tree patterns for different applications. Unstructured information has no schema to describe its structure, unlike structured data. We will confine our attention to textual information, and particularly to text mining which has concentrated on two issues, namely (1) categorization of textual documents, and (2) information extraction from a document to elicit a collection of facts or named entities from text documents. The most common approach to text categorization involves three phases:

1. text pre-processing;
2. encoding key information about the document using a feature vector, which is used with the knowledge discovery technique in (3) below;
3. a classification technique or a cluster technique to categorize the document.

Encoding represents the document by a vector of features such as word or phrase or clause, weighted by an importance factor. Classification techniques used include nearest neighbor classification, decision trees, support vector machines, naive Bayes Classifier and clustering methods include SOM, K-means or statistical measures such as regression. Information Extraction essentially involves extracting factual information or named entities from textual data of a certain type. Powerful entity extraction methods include support vector machines, hidden Markov Models [23], Random

Fields Methods [24] and Maximum Entropy Models. An application in bioinformatics [25], found rather poor results showing there remain challenging issues for some domains.

3 Significance of Work

Currently when deriving business intelligence by knowledge discovery from structured data sources one can often answer the questions related to patterns of what is happening. To answer questions related to why it is happening it will be necessary to derive conjoint mining of content (unstructured/) together with structured data. Thus, for example, a bank examining its database for patterns of customers who have decided to cancel their credit cards could through Data Mining techniques determine some of the features of the customer who is cancelling their credit card. e.g. cobranded cards, branches they belong to, their addresses, length of holding card etc. They would not be able to obtain information from the database of the number and nature of complaints, requests the customer may have made which would be contained in unstructured content repositories holding emails, transcribed phone call information, etc. In order to carry out this conjoint mining it is necessary to link segments of text from the content repository with individual records of interest in the structured data base. This is the problem of semantic integration between structured data and content (unstructured). Considerable work has been done for semantic integration between different sources of structured data. Some work has been done on semantic integration of structured data sources and XML data (semi structured) for the purpose of querying [26, 27]. Very limited work has been done on semantic integration of unstructured data and structured data sources [28, 29], and all of these concentrate on semantic integration for querying these diverse data sources (whether they involve XML data or Unstructured Content). Almost no work to the best of the author's knowledge has been carried out on semantic integration of unstructured content and structured data in a form suitable for deep knowledge discovery through AADC. The information integration approaches above also assume that the schema of the different data sources is available but this is not always possible as enterprises frequently outsource supporting processes to different vendors. For example, Customer Contact Centers are generally outsourced to third parties, who maintain the unstructured information in isolation from the enterprise structured data creating information silos where the schema is not always known. Therefore, we need to develop efficient techniques for correlating the data (structured) and (unstructured) content conjointly, and this task has many technical challenges.

The first issue is that the unstructured data is typically noisy in nature. For example, unstructured data received from different contact channels, like calls, SMS, emails, is very noisy. The problems of cleaning this are somewhat different to that of cleaning structured data. We need to clean this data before we can correlate it with the structured data. The domain of noisy text correction is comparatively new, and we use new techniques for cleansing drawing on the field of automatic spelling corrections [30] and use of structured data and various ontologies to provide reference information to clarify terms in the text and fill in missing values and terms. The second

issue is we need to discover the semantic knowledge/features from the text data. Typical information extraction tools or annotators take plain text as input and identify named entities and simple relations (e.g. works-for) and other text mining annotations based on a data dictionary or a gazetteer approach. For example, given a dictionary of product names, one can identify the product name in the text document. However, the current annotation systems, like UIMA, cannot annotate the documents based on concept, ontology and hierarchy. Therefore, we will extend the UIMA type techniques for annotating the documents based on domain ontology/concept. The third issue is to carryout semantic integration of content and structured data. This requires discovery of the entities in the text data based on the semantic knowledge, which can be matched with the structured entities for data correlation. One of the challenges is that no explicit identifiers of the entity, such as a unique transaction number, may be available in the document. Additionally, the document is noisy, so that a term in the text does not exactly match the corresponding attribute of the entity in the structured data. For instance a customer may mention a different transaction amount in her email or spell her name differently from that in the database. This naturally affects recall. It also affects precision when the noisy and partial information in a document leads to an incorrect entity being identified. We propose the use of a new fuzzy matching algorithm that uses an information theoretic basis. This fuzzy matching algorithm helps overcome the problem of the lack of precise information that permits exact matching. The fourth issue is to develop an intermediate representation suitable for Data Mining. This intermediate representation should be capable of capturing the embedded structure in content as well as the values. It should be in a form suitable to enable the use of conjoint Data Mining techniques. We propose an XML based approach for doing this as it enables one to represent domain information in a more meaningful and specialized way. The fifth issue is the development of algorithms to carry out Data mining conjointly from the content and structured data.

There are also frequently complex and important relationships between information that is stored in the form of structured data and content. To date the ability to find patterns, knowledge and relationships in unstructured text mining is aimed at document classification or entity extraction or simple associations between entities. The ability to find patterns or knowledge relationships between entities that might exist within unstructured textual documents is severely limited. For instance, in the biomedical area one may wish to find the chains involved in metabolic pathways. Furthermore existing techniques for document classification or entity extraction flatten the information structure using feature vectors, losing any structure other than the immediately preceding or following words that might be implicitly embedded or emergent within the document. These implicit or emergent structures may represent how the text is arranged under headings (which are semantically meaningful) or reflect chains of argumentation. Thus when grading an essay, the grader does not only look for the presence or absence of certain words or phrases but also the logical validity of the organization and the structure and clarity of the presentation; otherwise inserting the required words in a semi random fashion could give a good grade. More sophisticated AADC techniques should be capable of detecting such implicit or emergent structures. The approach adopted here, which converts the unstructured text into

a semi structured intermediate form such as XML or RDF, will allow better representation of implicit structures. The extension of XML Mining and RDF Mining techniques [26] previously developed by the present authors allows one to investigate the presence of patterns and associations between tree structured items (sub trees within the embedded structures), graph structured items (sub graphs) and sequences within the conjoint data and content instead of just the values of attributes or terms. Correlation of data and content conjointly enables the discovery of interesting relationships and analytics that involve predicates and groupings and their arrangements in combinations. To obtain valuable insights, it is important to find useful associations among concepts which could be dimensions from content and from structured data. The applications in the fields of business, biomedicine and electric power systems and logistics allow one to discover innovative new insights that would be difficult to obtain without the use of conjoint AADC from content and data. The problems tackled in this proposal are very hard and complex problems that are becoming critical with the large proportion of content as well as data that is currently being generated. We develop novel and ground breaking techniques to address each of these issues, which are highly innovative that have the potential to produce a paradigm shift in business intelligence and deep knowledge discovery.

4 Approach And Methodology

Data and content are stored in repositories that are isolated from each other. Hence, the problem of semantic integration of data and content must be addressed in a form suitable for AADC. One also needs representations and techniques for AADC conjointly from data and content. To resolve these issues, we use two base ontologies: (i) a static concept relation ontology and (ii) an event, transformation ontology which captures the key types of transformations and events allowed. These provide a conceptual framework that enforces an agreement on the organization of information, without losing any of the flexibility of allowing people to express and view parts in their own familiar expression language. An ontology, which is a shared conceptualization of some domain [31,73], captures and represents the key concepts, relationships and constraints which permits coherent understanding of the meaning of shared information.

The base ontologies will ensure a common ground for understanding content. One way to restrict the scope of disambiguation of particular content within the base ontology is by creating sub-ontology or specialized ontology (also known as *Ontology Commitment* [32], *Ontology Version* [33], *Materialized Ontology View* [34,74] appropriate for the category of information being considered. Next, we consider the approaches used for the sub aims.

4.1 Unstructured Data Cleansing

The unstructured content is generally noisy, the extent of which is different for each problem e.g. for biomedicine, the unstructured content in journal papers is cleaner. Content like transcribed calls or emails in customer contact centers is very noisy.

We need to clean this information. Processing SMS and email requires different data cleansing, e.g., removing spam messages, disclaimers, promotional material, and previous historical exchanges by the customer can be removed using heuristics for the domain. Even the body of the message is very noisy, using incomplete product name, spelling mistakes, added binary characters, etc. We will use two different approaches to this, (1) which borrows techniques from automatic spelling and grammar checkers [30] and (2) text cleansing methods which use structured data, and various ontologies (eg. Word Net) to provide reference data to clarify terms in the text and fill in missing values and terms, and deal with term variation arising from synonyms and acronyms. Spelling error correction is related to exact and approximate pattern matching respectively. Spell checking techniques involve non-word error detection and spell correction involves isolated-word or context-dependent error correction. The task involves three steps: (i) morphological analysis to identify a word-stem from a full word-form; (ii) isolating the misspelled words using techniques such as dictionary lookup and n-gram analysis; and (iii) offering a list of suggested correct spellings using one or more of six techniques, such as minimum edit distance, similarity key techniques, rule-based techniques, *n*-gram-based techniques, probabilistic techniques, and neural networks [30]. When doing this, we compare it to structured terms in the database or synonyms provided from WordNet.

4.2 Unstructured Data Annotation

In a document, we distinguish between:

1. Entity Extraction;
2. Identification of a relationship between two entities;
3. A network of relationships between entities;
4. Associations between several entities;
5. Associations between groups of entities such as ones arranged in subtexts and/or subsections.

Most work is on Named Entity Extraction (NER) which finds terms (words or phrases) for a specific named entity. Measures used to judge the efficacy of the algorithms are precision P (classification accuracy), recall R (coverage) and the F-score (harmonic mean of precision and recall). NER methods include probabilistic methods such as Hidden Markov [23, 35] or Conditional random fields [24], rule based methods [36] or Lexicon methods. Problems in NER are due to the same word or phrase referring to different entities, or many synonyms and/or abbreviations referring to an entity. To resolve these, we use an ontology for the domain of interest which maps these terms to concepts and relationships, recasting it as named concept recognition (NCR). Unlike a lexicon, which defines various items, an ontology has definitions for concepts and their relationships. There are several text to XML Annotators such as UIMA [37], GLOSS [38], and XI [39]. We will use the UIMA (Unstructured Information Management Architecture) [37] back-end that uses both statistical and rule-based annotators for text. Typically such annotators use a data dictionary or a gazette for information extraction from text of named entities (persons, organizations) and

simple relations between terms (works-for). To enhance their capability, we will extend the UIMA techniques for annotating the documents based on domain ontology/concept. This permits disambiguation of terms through concept relationships in the same segment of text. We use Annotators to extract relevant tokens from a document and map them to a small subset of the attributes for determining matches in structured data. E.g. by using NER in an annotator one can extract names from a document, and match them against the customer and product name attributes of the transaction table. One could also extract chunked text such as noun clauses by using a part of speech tagger for matching. This allows us to determine a score for an entity in a document. The highest scoring entity or best matching one can be found without computing explicit scores for all entities. Performing fuzzy match on each extracted token results in a ranked list of possible entities. Entities and relationships determined can be used to define the XML profiles for those documents. The Extraction Methods can be used to determine the values for each tag for an instance document. We use a combination approaches namely (1) ontology-based [40, 41] ones that extract terms from text and map them to concepts in an ontology to give semantics and (2) ontology-driven [42-44] ones that make active use of an ontology to guide or constrain the analysis.

4.3 Fuzzy Matching and Data and Content Correlation

Discovering the business insights conjointly requires correlation between data and content.. How can we link information from a text document (*TD*) with structured data in a database (*DB*), i.e. find the best matching entities from the *DB* for the given *TD*. We filter the annotated *TD* to retain only the relevant terms while the *DB* is considered as a set of entity instances and their associated related information. These *DB* entities are represented for matching as a collection of entity microschematas. We consider a single-type entity identification problem. We define the microschemata, as a rooted tree with the base table as the root and the related tables as the non-root nodes. If any tables have a foreign key relationship in the schema, their nodes are linked by edges. Each row in the base table is identified as an entity, having its own attribute values $e.A_j$. Here, an entity is an instance (a row in the base table) rather than a class level abstraction. The entity row is connected to the appropriate rows in the related table through foreign keys. We also have a collection *{di}* of *TDs* (the content) that have references to the entities. Each *TD* d has a set of terms *{ti}*. The *TD* consists of sentences. One or more sentences taken together will be referred to as a segment s. Let e be the central entity for this *TD*, then each term ti may correspond to some attribute $e.A_j$ of entity *e*. For instance, a *TD* about a transaction entity refers to the customer name, shop name, date attributes of a specific transaction. Given the terms in a *TD*, our goal is to identify the central entity (e.g. the Customer A/C#) from the structured table. No explicit identifiers of the entity, e.g. a customer number, may be available in the *TD*. Also noise in the *TD* means that t_i does not exactly match the corresponding attribute of *e*. This may lead to (i) not identifying the entity associated with the piece of content – poor recall or (ii) incorrectly identifying a wrong entity with the piece of content-poor precision. We want to link a given segment of the given document with an entity in the *DB*. There may also be information related to mul-

tiple entities in the given segment and we will need to identify these. First we define the microschemata for the structured *DB*, which is a rooted tree with the base table as the root and the related tables as the non-root nodes. The *TD* is filtered to retain noun phrases using a part of speech parser and annotated using the annotation techniques referred to above. If necessary "semantic integration within text document" techniques [45] can be used to identify terms which refer to the same concept. One next annotates the term using the annotation techniques (say with UIMA) or alternatively using database look ups to identify the column it occurs in. The key idea for matching a term is to determine the information content contained in a term in predicting the entity it refers to and we use an information theoretic formulation for this purpose. From Information theory [46], for a finite probability distribution p_i (i = 1,…, m), the entropy is given by

$$H\left(p_1, p_2, \ldots, p_m\right) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

This measure of information content can be considered to be the uncertainty of the occurrence of a term corresponding to the particular entity. Let us assume that the term is contained in the contexts of $n(t)$ distinct entities and there are N entities in total. Hence the probability of occurrence of a given entity e if t_i is present is $p(t_i) = n(t_i)/N$. Hence the associated information content

$$I\left(e, t_i\right) = -\left(\frac{n\left(t_i\right)}{N}\right) * \log_2\left(\frac{n\left(t_i\right)}{N}\right)$$

Given that t_i occurs $f(t_i)$ times in a particular segment d, the information content associated with t_i occurring in the segment d linked to entity e is

$$I\left(e, d\right) = \sum_{\forall ti \in d} f\left(t_i\right) * \left\{-\left(\frac{n\left(t_i\right)}{N}\right) * \log_2\left(\frac{n\left(t_i\right)}{N}\right)\right\}$$

A larger value for I indicates a greater predictive capability. A matching cache that contains a collection of pairs (e,t) (i.e term t is contained in entity e) is populated using two queries, one which returns the set of entities containing the term t and another which returns the set of terms contained in the contexts of the entity e. This avoids repetition of the queries for the same term in a new segment. This cache can be used with the expression above to produce a ranked list of entities that match a segment of text. An alternative to the above approach is to use the Zhou and Dillon Symmetrical Tau [6] to produce the ranked list. The unstructured text is annotated and represented as an XML document which is matched and merged with the structured data into XML documents using mapping between the concepts. These will be matched using a combination of semantic concept matching, online dictionaries, thesauri, schemas, and structure/related information, and extensions using an ontology. Then both documents will be adjusted to use common concept labels. We then find the common knowledge segments using our U3 mining algorithm [47], as in [48]. The additional information from unstructured content in XML corresponds to the unmatched knowledge segments and is used to augment the XML repository.

4.4 Techniques for Conjoint Data and Content Mining

To obtain valuable insight, it is important to find useful associations among concepts, from content (unstructured) and from structured data conjointly. It is useful to pre-identify valuable relationships. E.g. identify the top five products that experienced a sharp increase in complaints, common features of the customers and the nature of their complaints, products receiving the most inquiries and the profile of the inquirer and their reasons for the inquiry. Answering such questions give the enterprise timely insights about the customers' concerns. We note that structured data and content (unstructured documents) each have a very different representation. It is important for Integrated Data Mining to get a common intermediate form and we have chosen XML, as it has been successfully used for exchanging data between heterogeneous data sources, can capture the essentials of unstructured textual information, and it allows for mining of values and structures. We have the information held in a structured database with entities, i.e. $E = \{e_1,..., e_n\}$ which we represent in its XML form as:

$\{Ex = (e_{x1},..., e_{xn})$, and an unstructured document repository $U = \{u_1,..., u_m\}$.

The corresponding XML representation of the unstructured repository is

Ux = {ux1,..., uxn); An extended XML representation transaction is defined which consists of EX = {<exi, ux11> ... <exn, uxn>}

where the tuple $<e_{xi}, u_{xi}>$ consists of a concatenation of the XML representations of the two categories of information (E_x and U_x).

We used this approach for data records to obtain inter-transactional association rules [4]. As all data is conjointly represented in an XML document, the problem now becomes one of using the powerful XML mining algorithms to tackle mining of collections of these augmented XML documents. Our recent work has demonstrated the feasibility of conjoint mining of structured databases and XML repositories [49]. An XML-enabled framework for mining of association rules in XML repositories was first presented in [17] where the rules extracted are more powerful than traditional ones in expressing association relationships at both a structural and semantic level. To extract such rules the most difficult task is to find all the frequent sub trees from an XML database. This is known as the frequent sub tree mining (FSM) problem and is defined as: Given a tree database T_{db} a minimum support threshold (σ) find all subtrees that occur at least σ times in T_{db} [17]. Being able to mine all different subtree types using different support definitions is particularly important when we work on an XML representation of textual information, since these concepts can be repeated within many fragments of text and there exist different relationships among the concepts in the text, given the flexibility in its representation and expressiveness. The present authors have developed amongst the most powerful algorithms for mining XML document repositories and tree structured data.

The current work builds on this considerable body of expertise. The performance bottlenecks in FSM algorithms are candidate generation and counting, and this is often affected by the ability to effectively represent the document structure. Our work in the FSM field is characterized by a Tree Model Guided (TMG) [47, 50] candidate

generation approach. This non-redundant systematic enumeration model uses the underlying tree structure of the data to generate only valid candidates which conform to the underlying tree structure of the data. We proposed the so called *Dictionary, Embedding List* (EL) [60] and the *Recursive List* (RL) structures [47] whose purpose is to capture the structural aspects of a document and allow for efficient access to necessary information. The RL is a more compact representation of the EL that reduces the memory and serves as a global lookup list and also encodes the embedding relationships of the subtrees to be mined. To enable efficient counting we use the *Vertical Occurrence List* (VOL) [50, 51] which stores a representation of a subtree encoding together with coordinates. Using the *TMG* framework with the above representation structures we have presented FSM algorithms for mining of following subtrees (under any support definitions): ordered induced [52] and embedded [53,77], ordered [54] and unordered [51] distance-constrained embedded, unordered induced [55] and embedded [47]. We have also extended *TMG* for sequence mining [4]. An important aspect of this process is Trust[72,75]

5 Applications to the Electric Power Industry

The deregulated market allows power consumers to purchase power from different generation companies (gencos) who price the power based on the system Marginal Price [56]. To maintain a competitive position, gencos form bilateral contracts with their clients (particularly large ones). These provide the clients with a guarantee of their required energy at a defined cost over a long period (say 5 years). These contracts are in textual form with possibly different terms for each client and are stored in a content repository. The bilateral contracts are legally binding and the genco has to ensure that it can meet the required demand from all the different clients with contracts and other customers who purchase power from the genco as needed. It may also have contracts for supply to it from other gencos. Thus, the genco has to deliver the expected Ensured Energy (EE) to meet its obligations. Otherwise there are many quite severe penalties which are normally staged according to proportion of power not delivered. Uncertainty, in a hydro thermal system being able to meet this EE, is caused by scheduled maintenance, unplanned outages arising from equipment breakdown, power from hydro plants being uncertain due to uncertain inflows, uncertainty in non-contract demand. Historical information related to these factors is stored in structured data bases together with historical spot prices for electrical energy. These will be used to produce forecasts for several of these factors and develop schedules for others such as hydro scheduling [57] and scheduled maintenance which are stored in structured DBs. This has led to approaches using the structured data to assess the risk especially the loss of load probability and expected unserved energy in the case of no complex contractual terms [58].

However what is needed is a risk assessment and management approach including textual contract terms. The profitability of a company will be impacted by any inability to meet the ensured energy. In this event the genco would try either (1) to purchase the extra energy at spot prices which are generally much higher than from its own supplies (and may exceed the contract price) or (2) to not fulfil the required demand

under some contracts, or (3) try to put in place different contracts with other suppliers for the duration of any energy shortfalls, or (4) establish potential new contracts with new clients, or (5) renegotiate existing terms in existing contracts. To make such decisions, requires the genco to extract business intelligence conjointly from (i) the content repository containing information on the different contracts with its clients and suppliers (perhaps with notes on the feasibility of term variations) and (ii) the structured information in the different databases on the different factors. By linking and analyzing this information one can find associations which could result in risky situations and also determine potential remedial actions. Examples could be maintenance patterns, inflow levels (say in a dry year and failure rates which result in energy deficits leading to non-fulfilment of contractual obligations). This would alert the genco early of the need to purchase power from other gencos using supply contracts and refrain from certain client contracts.

6 Applications to Business Problems

We propose a new approach for consumer expectation-based market segmentation through conjoint mining of content and data. Consumer expectation has long been considered as an important satisfaction determinant that represents market demand and shapes the consumer behaviors [59, 60]. Hence, customer segmentation based on their expectations is of great significance for firms to predict dynamics of targeted markets.

However, consumer expectations data are often unobservable and prohibitive to collect using traditional market research methods such as interviews, self-reported surveys, experiments, etc. Conjoint mining provides a new means by which marketers are able to understand the 'minds' of millions of consumers on a daily basis without having to physically interact with consumers (e.g. shadowing, field experiment.) or explicitly soliciting opinions (e.g. interviews, surveys) from them. Consequently, data reliability and model validity bears more substantial rigor than previous approaches. Using conjoint mining techniques, consumer expectation can be inferred from customer satisfaction data gleaned from online customer reviews which include both structured and unstructured data, e.g. the reviews on Epinions.com hold numerical ratings towards each product attribute (e.g. 'battery life' for digital cameras, 'memory capacity' for MP3 players) as well as free text comments from consumers on their opinions and experiences.

Conjoint mining allows the use of structured data to reveal latent customer expectations based on unobservable concomitant variables such as consumer preferences, taste, values and the use of content to mine consumer opinions in free text, to (1) extract product features from the reviews, and (2) obtain consumer affects and sentiments towards these. We develop algorithms which augment opinion mining methods used in the interactive data analysis from [61] and the overall processes of opinion tracking from [62]. To discover heterogeneous expectations towards different brands, comparison-based algorithms [63] will be leveraged and redeveloped. The Web usage mining [64], sentiment-based algorithms [65] and opinion holder identification algo-

rithms [66] will be integrated in order to make markets segments 'actionable' for managers and produce a custom score function to classify the sentiment [66].

Product extraction will augment the method in [67] and borrow some ideas from entity-based search engines [68] to mark items from free texts. Both product and feature extraction will allow business analysts to annotate text with an ontology and allow the unstructured customer opinions/ complaints to be linked with structured customer data in internal DBs or Customer Relationship Management systems.

7 Application in Transport Logistics industry

It is important to understand that today's transport logistics providers spent 50% of their time on managing the physical mess and 50% on managing the related information mess [78]. Here, intelligent transportation has enabled vehicle to driver, vehicle to vehicle and vehicle to infrastructure communications and emerging intelligent infrastructure that provides embedded un-manned situation awareness 24/7 that enables greater mobility, security and safety.

It is also important to realise that over the years, the Transport Logistics sector has generated and accumulated much more valuable economic information than Facebook. This informs us on Big data impacts on global financial movement and Financial forecast including financial forecasts. Logistics professionals around the world know that they are no longer just transport and logistics operators, they are required to be "Data Experts" or at least to have Data Experts in their organization. Our ARC (Australia Research Council) Logistics Industry Partners in New South Wales and Queensland have been pushing their data to the Cloud since 2009 with vendor support. However, this Big Data has not been fully utilised, due to the lack of availability of co-joint data and content mining technology.

Further, many manufactured items, goods or assets today utilizing the Internet of Things are already Internet enabled, they have capability to talk to Internet, talk to each other, talk to logistics providers and talk to logistics infrastructure. This has sped up the automated people, goods and asset movement in logistics, transportation, warehouse and distribution [78] sector.

Intelligent Tracking powered by co-joint data and content mining is the core technology that is needed in transport logistics industry today. Tracking movement of people, goods and services in the entire logistics network, tracking quality of services, service providers performance, through entire life cycle of supply chain and asset management, track and trace of data and information shared over the logistics alliances, coalition partners and joint forces, situation awareness and ambient intelligent, for productivity, security and safety. Intelligent tracking powered by conjoint predictive analytics with real time data and in real time environment is a major challenge for all modern transport logistics providers. We have been working with our industry partner to adopt conjoint predictive analytics and co-joint mining for monitoring, visualisation, sharing, control and management of physical mess (goods and assets) and information mess (data) as well as business processes for their Business Intelligence in-

cluding maximisation of human, transport and infrastructure performance and minimisation of the costs and security risk.

The co-joint data and content mining on Big Data in transport logistics sector including the combined RFID and wireless sensors data on the goods and assets handling, warehousing and transportation, GPS, GPRS and position location system for transport vehicle and shipment tracking, Surveillance Systems for Operator Performance and situation awareness, provenance of Goods and Asset tracking. The co-joint data and content mining are also needed for Inter- and intra-logistics partners transactions data monitoring, customers based tracking of trades data, smart phone, bluetooth, and black-box (on heavy vehicles and ships vessels) communication and even logistics social networks to support auto and semi-automated physical flow and information flow which enables business intelligent.

We use transport logistics ontology to help manage the Big data by defining the meaning of data through adding context that gives information on the data. Our works include Ontologies, RDF annotations and contexts. We carry out mining and visualization of big data both relational data (warehouse data or 3PL data) and complex data includes tree structured data (Geo-data), XML documents (procedures and workflows), unstructured textual data (smart phone notification and web data), image data (positions and locations), multimedia data (surveillance data), graphical data (Asset tracking data).

One of our biggest challenges in the co-joint data mining has been the assurance that the Big data are from trusted sources, the data services for Big data are trusted such as Clouds, and the Quality of Data, especially in the automated environment utilising Internet of Things and Cyber-Physical Systems. If the wrong decision is made based on the poor data set, it could result in major financial losses, high casualties and possible terrorist attack through the use of transport.

8 Applications to biomedical applications

Existing biomedical information is distributed across a large number of information resources and is heterogeneous in its content, format and structure. This hinders effective information retrieval. Targeted searches are very difficult with current search engines as they look for the specific string of letters within the text rather than its meaning. Use of highly expressive knowledge models such as ontologies enables the machines to view the text as meaningful expressions. This increases the semantics and forms the basis of a more efficient approach to finding the right information. An ontology can be used for creating metadata by semantic annotation of text through three steps: tokenization (splitting the sentences into tokens), matching the tokens against the ontology terms and matching the tokens against the ontology relationships until the best fit is found. New web pages created can be annotated automatically during their creation process. This semantic annotation allows machines to access web content, understand it, retrieve and process the data automatically rather than only display

it. In our research work , we have developed a number of ontologies, such as Protein Ontology [69], Human Disease Ontology [70] and Mental Health Ontology [71]. The ontologies can be used to annotate target information in content sources and enable intelligent retrieval of specific information, analysis of it and linking with the existing pool of knowledge. E.g., protein and bibliographical reference data available via Swiss-Prot can be linked with the related publications from PubMed and with the epidemiological data in medical databases. Conjoint content and data mining of the linked content/data provides quality knowledge that can help build effective prevention and intervention strategies. Thus the presence of the protein, PSA, at a given level or given form (free or complex) at a certain age or ethnicity or lifestyle, might be indicative of a certain probability of the existence of cancer. Conjoint mining of the two structured databases and the textual information in PubMed will help with the discovery of such knowledge. There may be situations which coincide with some ambiguity or inconsistency. This will help researchers identify what requires further investigation.

9 Conclusions

The paper presents a methodology for conjoint mining of structured and unstructured information. The potential impact in industry of use of BI and AADC can be inferred from a 2003 IDC study of 40 US and European companies that use predictive analytics KDD who achieved a median Return of Investment of 145%, achieved higher investment levels and yielded higher overall returns over five years. These improvements occurred in just effectively utilizing the 5% of information available as structured data. This effect would be considerably amplified if one could in an integrated fashion exploit the remaining 95% of content as well as the 5% of structured data.

This research, by developing an integrated approach for BI and AADC of data and content, will provide a competitive edge in handling such information. This integrated knowledge discovery techniques could improve policy formation and effectiveness by Government and non-Government in such areas as compliance by companies, reduction of aberrant behavior in areas such as health benefits allocation, pension entitlements etc. It will provide an intellectually rigorous approach to underpin the trend in electronic document handling.

References

1. M. L. Brodie, "Computer science 2.0: A new world of data management," in *Proc. 33rd VLDB Conf.*, 2007, p. 1161.
2. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in Advances in knowledge discovery and data mining: American Assoc. for Artificial Intel., 1996, pp. 1-34.
3. L. Feng, T. S. Dillon, and J. Liu, "Inter-transactional association rules for multi-dimensional contexts for prediction and their application to studying meteorological data," Data Knowl. Eng., 37 (1), 85-115, April 2001.

4. H. Tan, T. S. Dillon, F. Hadzic, and E. Chang, "Sequest: Mining frequent subsequences using DMA strips," in Proc. 7th Intl. Conf. on Data Mining and Inf. Engineering, Prague, Czech Republic, 2006, pp. 315-328.
5. X. J. Zhou and T. S. Dillon, "Theoretical and practical considerations of uncertainty and complexity in automated knowledge acquisition,"IEEE Trans. Knowl. Data Eng., 7 (5), 699-712, 1995.
6. X.-J. M. Zhou and T. S. Dillon, "A statistical-heuristic feature selection criterion for decision tree induction," IEEE Transactions on Pattern Analysis and Machine Intelligence, 13 (8), 834-841, 1991.
7. F. Hadzic and T. Dillon, "Using competitive learning between symbolic rules as a knowledge learning method," in Proc. IFIP 20th world computer congress Milan, Italy: Springer, 2008, pp. 351-360.
8. S. Sestito and T. S. Dillon, "Knowledge acquisition of conjunctive rules using multilayered neural networks," Int. J. Intell. Syst., 8 (7), 779-806, 1993.
9. S. Sestito and T. S. Dillon, Automated knowledge acquisition. Sydney: Prentice Hall, 1994.
10. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Intl. Conf. Very Large Data Bases (VLDB), Chile, 1994, pp. 487-499.
11. D. Tsur, J. D. Ullman, et al., "Query flocks: A generalization of association-rule mining," ACM Intl. Conf. on Management of Data, Seattle, USA, 1998.
12. R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in Proc. 3rd Intl. Conf. on Knowledge Discovery and Data Mining, USA, 1997, pp. 67-73.
13. L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang, "Optimization of constrained frequent set queries with 2-variable constraints," in ACM SIGMOD Intl. Conf. on Management of Data, USA, 1999, pp. 157-168.
14. C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," Data Min. Knowl. Disc., 4 (2), 163-192, 2000.
15. S. Ramaswamy, S. Mahajan, and A. Silberschatz, "On the discovery of interesting patterns in association rules," in Proc. 24rd Intl. Conf. on Very Large Data Bases (VLDB), 1998, pp. 368-379.
16. K. Wang and H. Liu, "Discovering structural associations of semistructured data," IEEE Trans. Knowl. Data Eng., 12 (3), 353-371,2000.
17. L. Feng, T. S. Dillon, H. Weigand, and E. Chang, "An XML-enabled association rule framework," in Proc. 14th Intl. Conf. on Database and Expert Systems Apps. (DEXA), Prague, Czech Republic, 2003, pp. 88 - 97.
18. M. J. Zaki and C. C. Aggarwal, "XRules: An effective structural classifier for XML data," in Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Washington D.C., USA, 2003, pp. 316-325.
19. L. H. Yang, M. L. Lee, and W. Hsu, "Efficient mining of XML query patterns for caching," in Proc. 29th Intl. Conf. on Very Large Data Bases (VLDB), Berlin, Germany, 2003, pp. 69-80.
20. T. Asai, H. Arimura, T. Uno, and S.-i. Nakano, "Discovering frequent substructures in large unordered trees," in Proc. 6th Intl. Conf. on Discovery Science (DS), Sapporo, Japan, 2003, pp. 47-61.
21. M. J. Zaki, "Efficiently mining frequent trees in a forest: Algorithms and applications," *IEEE Trans. Knowl. Data Eng.*, 17 (8), 1021-1035, August 2005.
22. A. Termier, M.-C. Rousset, and M. Sebag, "Treefinder: A first step towards XML data mining," in Proc. 2nd IEEE Intl. Conf. on DataMining (ICDM), Maebashi City, Japan, 2002, pp. 450-458.

23. D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," Mach. Learn., 34 (1), 1999.
24. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 18th Intl. Conf. on Machine Learning (ICML), 2001, pp. 282-289.
25. J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in Proc. Workshop on Natural Language Processing in Biomedicine and its Apps., 2004, pp. 70–75.
26. S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, "Semantic integration of heterogeneous information sources," Data Knowl. Eng., 36 (3), 215-249, 2001.
27. P. McBrien and A. Poulovassilis, "A semantic approach to integrating XML and structured data sources," in Proc. 13th Intl. Conf. onAdvanced Information Syst. Engineering (CAiSE), Switzerland, 2001, pp. 330-345.
28. V. T. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania, "Efficiently linking text documents with relevant structured information,"32nd Intl. Conf. on Very Large Data Bases (VLDB), Seoul, Korea, 2006, pp. 667-678.
29. M. A. Bhide, A. Gupta, et al., "LIPTUS: Associating structured and unstructured information in a banking environment," in Proc. ACMSIGMOD Intl. Conf. on Management of Data, Beijing, China, 2007, pp. 915-924.
30. K. Kukich, "Techniques for automatically correcting words in text," ACM Comp. Surv., 24 (4), 377-439, 1992.
31. T. R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis., 5 (2), 1993.
32. M. Jarrar and R. Meersman, "Formal ontology engineering in the DOGMA approach," in Confederated Intl. Conf. CoopIS, DOA, andODBASE, California, USA, 2002, pp. 1238-1254.
33. M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov, "Ontology versioning and change detection on the web," in Proc. 13th Intl. Conf.on Knowledge Eng. and Knowledge Management (EKAW), Spain, 2002, pp. 247-259.
34. C. Wouters, T. S. Dillon, J. W. Rahayu, E. Chang, and R. Meersman, "A practical approach to the derivation of a materialized ontology view," in Web information systems, D. Taniar and W. Rahayu, eds. 2004.
35. G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: A machine learning approach," Bioinformatics, 20 (7), 1178-1190, 2004.
36. L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," Bioinformatics, 18 (8), 1124-1132, 2002.
37. D. Ferrucci and A. Lally, "UIMA: An architectural approach to unstructured information processing in the corporate research environ- ment," Nat. Lang. Eng., 10 (3-4), 327-348, 2004.
38. R. Kaye, "The gloss system for trans. from plain text to XML," Proc. MathUI 2006 http://www.activemath.org/~paul/MathUI06/.
39. B. Marchal, "XI: Open-source conver. of legacy text files to XML," in http://www.ananas.org/xi/index.html, [Ac.: 20/11/‘08].
40. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle, "Ontology-based extraction and structuring of information from data- rich unstructured documents," in Proc. 7th Intl. Conf. on Inform. & Knowl. Mgmt. (CIKM),USA, 1998, pp. 52-59.
41. H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall et. al., "Automatic ontology-based knowledge extraction from web documents," IEEE Intel. Syst., 18 (1), 14-21, Jan./Feb. 2003.

42. S. Handschuh, S. Staab, and A. Maedche, "Cream: Creating relational metadata with a component-based, ontology-driven annotation framework," 1st Intl. Conf. on Knowledge Capture, Canada, 2001, pp. 76-83.
43. M. Vargas-Vera, E. Motta et.al "Mnm: Ontology driven semi-automatic and automatic support for semantic markup," in Proc. 13th Intl. Conf. on Knowl. Eng. and Know. Mgmt. , Spain, 2002, pp. 213-221.
44. R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett, "Protein structures and information extraction from biological texts: The pasta system," Bioinformatics, 19 (1), 135-143, 2003.
45. Li, X., P. Morie, et al.,"Semantic integration in text: from ambiguous names to identifiable entities." AI Mag. 26(1), 2005.
46. F. M. Reza, An introduction to information theory. New York: Dover Publications, 1994.
47. F. Hadzic, H. Tan, and T. S. Dillon, "U3 – mining unordered embedded subtrees using TMG candidate generation," in Proc. IEEE/WIC/ACM Intl. Conf. on Web Intelligence, Sydney, Australia, 2008.
48. F. Hadzic, T. S. Dillon, and E. Chang, "Tree mining application to matching of heterogeneous knowledge representations," in Proc. IEEE Intl. Conf. on Granular Computing (GRC), California, USA, 2007, pp. 351-351.
49. Q. H. Pan, F. Hadzic, and T. S. Dillon, "Conjoint data mining of structured and semi-structured data," in Proc. 4th Intl. Conf. on the Se- mantics, Knowledge and Grid (SKG), Beijing, China, 2008, pp. 87-94.
50. H. Tan, F. Hadzic, T. S. Dillon, L. Feng, and E. Chang, "Tree model guided candidate generation for mining frequent subtrees from XML," ACM Trans. Knowl. Discov. Data, 2 (2), July 2008.
51. F. Hadzic, H. Tan, and T. S. Dillon, "Mining unordered distance-constrained embedded subtrees," in Proc. 11th Intl. Conf. on Discovery Science (DS), Budapest, Hungary, 2008.
52. H. Tan, T. S. Dillon, F. Hadzic, E. Chang, and L. Feng, "IMB3-Miner: Mining induced/embedded subtrees by constraining the level of embedding," in Proc. Of PAKDD, Singapore, 2006, pp. 450-461.
53. H. Tan, F. Hadzic, L. Feng, and E. Chang, "MB3-Miner: Mining embedded subtrees using tree model guided candidate generation," 1st Intl. W'shop on Mining Complex Data in conj. with ICDM'05, USA, 2005.
54. H. Tan, T. S. Dillon, F. Hadzic, and E. Chang, "Razor: Mining distance-constrained embedded subtrees," in Proc. Workshop on Ontol- ogy Mining and Knowledge Discovery from Semistructured documents (MSD) in conjunction with 2006 Intl. Conf. on Data Mining, Hong Kong, 2006, pp. 8-13.
55. F. Hadzic, H. Tan, and T. S. Dillon, "UNI3 - efficient algorithm for mining unordered induced subtrees using TMG candidate genera- tion.," IEEE Sym. on Comp. Intel. and Data Mining (CIDM), USA, 2007, pp. 568-575.
56. B. R. Szkuta, L. A. Sanabria, and T. S. Dillon, "Electricity price short-term forecasting using artificial neural networks," IEEE Trans- actions on Power Systems (PES), 14 (3), 851-857, Aug. 1999 1999.
57. D. Sjelvgren, S. Andersson, T. Andersson, U. Nyberg, and T. S. Dillon, "Optimal operations planning in a large hydro-thermal power system," IEEE Trans. Power App. Syst., PAS-102 (11), 3644-3651, 1983.
58. T. S. Dillon, R. W. Martin, and D. Sjelvgren, "Stochastic optimization and modelling of large hydro-thermal systems for long term regu- lation," Intl Journal of Electrical Power and Energy Systems, 2 (1), 2-20, 1980.

59. R. L. Oliver, "A cogni. model of the antecedents and conseq. of satisfaction decisions," J. Marketing Rsch. , 17, 1980.
60. [60] R. T. Rust, J. J. Inman, J. Jia, and A. Zahorik, "What you don't know about customer-perceived quality: The role of customer expecta- tion distributions," Marketing Science, 18 (1), 77–92, 1999.
61. L.-W. Ku and H.-H. Chen, "Mining opinions from the web: Beyond relevance retrieval " J. Amer. Soc. Inf. Sci. Technol., 58 (12), 1532- 2882, 2007.
62. N. Glance, M. Hurst, K. Nigam, M. Siegler, et.al., "Deriving marketing intelligence from online discussion," Proc. 11th ACM SIGKDD Intl. Conf. on Knowl. Discov. in Data Mining (KDD), USA, 2005, pp. 419-428.
63. N. Jindal and B. Liu, "Identifying comparative sentences in text documents.," in Proc. 29th ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval, Seattle, USA, 2006, pp. 244-251.
64. A. G. Büchner and M. D. Mulvenna, "Discovering internet marketing intelligence through online analytical web usage mining," ACM SIGMOD Rec., 27 (4), 54-61, 1998.
65. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.," in Proc. 40th Ann. Meeting on Assoc. for Comp. Lingui.USA, 2001, pp. 417-422.
66. S.-M. Kim and E. Hovy, "Determ. the sentiment of opinions," 20th Intl. Conf. on Comp. Lingustics. Switzerland. 2004.
67. S. Drenner, M. Harper, D. Frankowski, et.al., "Insert movie refer. here: A system to bridge conversation and item-oriented web sites," SIGCHI Conf. on Human Fact. in Comp. Syst. (CHI), Canada, 2006, pp. 951-954.
68. T. Cheng, X. Yan, and K. C.-C. Chang, "Entityrank: Searching entities directly and holistically," in Proc. 33rd Intl. Conf. on Very Large Data Bases, Vienna, Austria, 2007, pp. 387-398.
69. AS Sidhu, TS Dillon, E Chang, BS Sidhu Protein ontology: vocabulary for protein data 2005. ICITA 2005. Third .Int. Conf. Information Technology and Application
70. M. Hadzic and E. Chang, "Medical ontologies to support human disease research and control," J. Web Grid Serv., 1 (2), 2005.
71. M. Hadzic, M. Chen, and T. S. Dillon, "Towards the mental health ontology," in Proc. IEEE Intl. Conf. on Bioinformatics and Biomedi- cine (BIBM), USA, 2008, pp. 284-288.
72. M Alhamad, T. Dillon, E. Chang, "SLA-based trust model for cloud computing" x 2010 13th International Conference Network-Based Information Systems (NBiS),
73. K Aberer, T Catarci, P Cudré-Mauroux, T Dillon, S Grimm, MS Hacid, et. al. Emergent semantics systems...Semantics of a Networked World. Semantics for Grid Databases, 14-43.
74. C Wouters, T Dillon, W Rahayu, E Chang, R Meersman Ontologies on the MOVE Database Systems for Advanced Applications, 812-823.
75. E Chang, TS Dillon, FK Hussain Trust and reputation relationships in service-oriented environments 2005. ICITA 2005. Third .Int. Conf. Information Technology and Applications
76. C. Wu, E. Chang"Searching services on the web: A public web services discovery approach", 2007. SITIS'07.Conf Signal-Image Technologies and Internet-Based System.
77. H. Tan, T. S. Dillon, F. Hadzic, E. Chang, L Feng MB3-Miner: efficiently mining eMBedded subTREEs using Tree Model Guided candidate generation. Department of Mathematics and Computing Science Saint Marys University.
78. E. Chang 2014 "Transport Logistics, the Grand Challenges", Australian Defence Force Academy