Comput Syst Sci & Eng (2020) 5: 321–334
© 2020 Tech Science Press

International Journal of
Computer Systems
Science & Engineering

Conjoint Knowledge Discovery Utilizing 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§] **and Vish Ramakonar**[4]

[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*
[4] *Alsys MSC Sdn Bhd, Kuala Lumpur, Malaysia*

In Digital Enterprises Structured Data and Semi/Unstructured Content are normally stored in two different repositories, with the first often being stored in relational Databases and the second in a content manager which is frequently at an external outsourcer. This storage of complementary information in two different silos has led to the information being processed and data mined separately which is undesirable. Effective knowledge and information use requires seamless access and intelligent analysis of information in its totality to allow enterprises to gain enhanced insights. In this paper, we develop techniques to carry out correlation of the information across different sources and then carryout out knowledge discovery across these complementary sources in a conjoint manner. The techniques developed in our research will then be used to address significant issues in four application areas namely Business, Logistics, Bioinformatics and Electric Power Systems but potential applications with significant impact are much more extensive.

Keywords: **Please Provide Keywords**

1. INTRODUCTION

Digital information within an enterprise consists of (1) *structured data* and (2) semi/*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, on-line 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

*Tharam.dillon7@gmail.com
†Phoebe.Chen@latrobe.edu.au
‡Elizabeth.chang@unsw.edu.au
§mkmukesh@in.ibm.com

the latter by a content manager frequently provided by an outsourcer or vendor [75]. 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 insights. Enterprises are fast realizing the need to bridge this separation, and are demanding integrated retrieval, management and intelligent analysis of both the structured and semi/unstructured content to gain critical insights into customer and market trends, optimizing the decision-making process, and business intelligence. 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 [9]. Estimates from IDC state that in 2018 there was 33 zettabytes (33 trillion gigabytes) of digital data in the world [33]. This is estimated to grow to 175 zettabytes in 2025. IDC found that 90% of data in the digital universe is unstructured data [34]. According to various sources, the growth rate for un-structured data can be up to 80% per annum while structured data is around 40%. We should note that it is often useful to cast un-structured information as semi-structured to understand and utilize it for further processing.

Unless we can effectively utilize the semi/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 four 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 semi/unstructured information conjointly. This methodology would be useful in several Business Intelligence and Advanced Analytics utilizing Data and Content (AADC), information integration applications in Business, 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, 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 semi/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 Relationship Management (CRM) systems utilizing information from different information sources that integrate the various business processes of an organization such as (a) customer support, complaints and issues, (b) marketing and management of existing customers, (c) sales targeting of new and existing customers.
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 [18]. Thus information could be structured data and semi/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 [19,63], (ii) Decision Tree Methods [79,80], and (iii) Rule Search Methods [26]. Effective methods for continuous structured data include Neural Net Based Rule extraction methods [53,54]. Association Rules are of the form $X \Rightarrow Y$ (s, c) where X and Y are sets of items and s, c are support and confidence respectively [2]. Association Rule Mining has since been extended to include efficient Apriori-like mining methods [2], query mining [68], constraint based rule mining [44,58], mining correlations and causal structures[56] and interesting associations [50] and mining associations from semi-structured data [20,71,77].

The advent of the web has led to the development of XML as a semi-structured data representation. An XML document possesses a hierarchical document structure, where an XML element may contain further embedded elements, and each element can be attached with a number of attributes. It is therefore frequently modelled using a labelled ordered tree. Our research group [20] 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. Basically, the problem of mining XML Documents can be recast as mining ordered trees. Work has been proposed for mining XML documents [20,71,76,77], including a collection of semi-structured objects [20]. To help with mining XML documents and other structures, frequent sub tree mining algorithms are being developed [5,67,78] 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 [8], Random Fields Methods [43] and Maximum Entropy Models. An application in bioinformatics [38], found rather poor results showing there remain challenging issues for some domains.

3. INNOVATION AND 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 (semi/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 semi/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 [6,47]. Very limited work has been done on semantic integration of unstructured data and structured data sources [7,11], 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 semi/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 (semi/unstructured) content conjointly, and this task has many technical challenges.

The first issue is that the semi/unstructured data is typically noisy in nature. For example, semi/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 [42] 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 [6] 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 [1,24], 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* [35], *Ontology Version* [40], *Materialized Ontology View* [72,73] appropriate for the category of information being considered. Next, we consider the approaches used for the sub aims.

4.1 Semi/Unstructured Data Cleansing

The semi/unstructured content is generally noisy, the extent of which is different for each problem e.g. for biomedicine, the semi/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 [42] 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 [42]. 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 [38,81] or Conditional random fields [43], rule based methods [66] 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 [21], GLOSS [37], and XI [46]. We will use the UIMA (Unstructured Information Management Architecture) [21] 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 [3,17] ones that extract terms from text and map them to concepts in an ontology to give semantics and (2) ontology-driven [22,32,70] 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 $e.Aj$. 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 $e.Aj$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 ti 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 multiple 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 [51], for a finite probability distribution $pi(i = 1, \ldots, m)$, the entropy is given by

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

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

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

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

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

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 [79] 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 [28], as in [31]. 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 (semi/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 (semi/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. $E = \{e1, \ldots, e_n\}$ which we represent in its XML form as:

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

The corresponding XML representation of the unstructured repository is $U_x = \{u_{x1}, \ldots, u_{xn}\}$; An extended XML representation transaction is defined which consists of $EX = \{< exi, ux11 > \ldots < exn, uxn >\}$

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

We used this approach for data records to obtain inter-transactional association rules [19]. 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 [20] 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} [20]. 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. Within this FSM framework the two most commonly mined types of subtrees are induced and embedded. An induced subtree preserves the parent-child relationships of each node in the original tree. In addition to this, an embedded subtree allows a parent in the subtree to be an ancestor in the original tree and hence ancestor-descendant relationships are preserved over several levels.

In the current frequent subtree mining literature two different support definitions can be found. The *transaction-based* [78] support is a conventional support derived from the support definition used in association mining for structured data. The second support definition is called *weighted support* in [78] and is referred to as *occurrence-match support* in [60].

We will extend this to *hybrid support* which is a combination of the transaction-based support and occurrence-match support. Hybrid support of a subtree t is denoted by '$x|y$', where 'x' denotes the number of transactions that support subtree t, and y denotes the least number of times that t has occurred in those x transactions. This will be more useful for our purposes.

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) [28,61,65] 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.

There are two aspects of tree mining that are addressed in our previous work namely (i) suitable representation and (ii) the algorithmic aspect. The representation aspect has to (a) model and represent the actual subtree in memory or secondary storage and (b) ensure that complex computations and data manipulations can be performed efficiently and effectively. To address (a) a space efficient depth-first string encoding was proposed. To address (b) several data-structures were proposed namely the so called *Dictionary, Embedding List* (EL) [28] and the *Recursive List* (RL) structures [28] 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) [27,61] which stores a representation of a subtree encoding together with coordinates. The *RMP coordinate list* allows storage of only the right-most-path coordinate of enumerated subtrees instead of the full coordinate which on average is shorter in size. 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 [64] and embedded [60,65], ordered [62] and unordered [27] distance-constrained embedded, unordered induced [29] and embedded [28]. We have also extended *TMG* for sequence mining [63]. An important aspect of this process is Trust [4,13].

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 [59]. 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 [15].

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 Customer Relationship Management (CRM). Customer Relationship Management (CRM) systems integrate the various business processes of an organization such as (a) customer support, complaints and issues, (b) marketing and management of existing customers, (c) sales

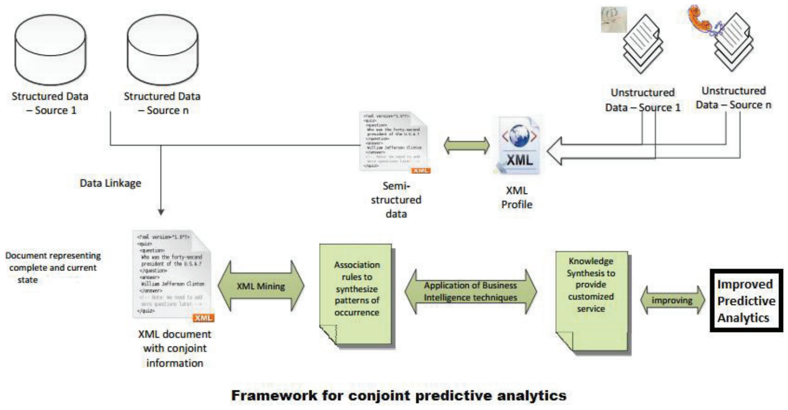
Structured Data – Source 1
Structured Data – Source n
Unstructured Data – Source 1
Unstructured Data – Source n
Data Linkage
Semi-structured data
XML
XML Profile
Document representing complete and current state
XML document with conjoint information
XML Mining
Association rules to synthesize patterns of occurrence
Application of Business Intelligence techniques
Knowledge Synthesis to provide customized service
improving
Improved Predictive Analytics
Framework for conjoint predictive analytics

Figure 1 Framework for conjoint predictive analytics.

targeting of new and existing customers and by using the customers' information assist in providing answers to the customer queries. Such CRM systems, apart from reducing operational costs, enable an organization to manage their customer base and utilize their information in an efficient way to synthesize knowledge from it and provide them with improved services.

Due to the necessity of budget tightening; organizations want to improve on the analysis achieved by CRM systems and introduce a new innovation in customer management known as Customer Management Excellence (CME). CME pushes the barriers of Customer Management to new limits beyond current CRM with the aim of providing improved and customized service to customers, thereby improving their satisfaction and loyalty. This is achieved by introducing intelligence into the customer management process.

However as mentioned earlier, the amount of semi/unstructured information being generated has increased as compared to the structured data. For example, in today's world a customer's relationship with an organization (in terms of complaints, preferences, interaction experiences and follow-ups) is mostly obtained from sources such as records, emails, and transcribed calls which will be semi-/unstructured in format. Such information presents an incomplete picture of a customer and should be linked dynamically to the corresponding fields in the structured databases to have a complete and current state of the customer that will assist an organization in answering the queries appropriately or perform intelligent knowledge discovery and synthesis to introduce intelligence into the CRM systems. In this paper, we propose such a framework for creating a Customer Relationship Management Ecosystem that will enable organizations to conjointly consider the relevant information from different information sources (semi/unstructured) and utilize it to synthesize knowledge from it.

Customer Information which is gathered and stored by organizations can broadly be classified into two types namely Structured, and b) Semi/unstructured. Structured information refers to data stored in databases with a defined schema (for example the customer details stored in the database), whereas semi structured information refers to information that normally has defined tags but the information stored within those tags could be free form text (for example information communicated by the customer to the organization either by email or phone). Unstructured information after annotation can be treated as semi structured information with relevant tags.

While on the one hand, an increase in the amount of data is beneficial as it increases the scope on which organizations perform various analyses for the customer, on the other hand in introduces the challenge of considering the relevant data at the same time during the analysis. This is because even though such semi-structured information may be related to the records defined in the structured format, they are stored in different data repositories from the structured data, and they exist in silos. Due to this separation, the literature approaches these different sources of information separately and essentially data mine the structured data for which many techniques exist. However, although the results obtained from such analysis may be valid; they may not be complete and do not capture the current state of the customer, for example the customer may have emailed the organization about his products preferences (which maybe in semi/unstructured format stored in a different repository as compared to data in the structured format) and this has to be considered during targeted marketing planning. To achieve this, these two sources of relevant information which complement each other have to be considered together to not only achieve ***valid*** but also ***current*** information about the customer. Limited work has gone into the development of techniques for conjointly mining both data sources.

Conjoint mining allows the use of structured data and Semi/unstructured information to not only reveal what events are occurring but also why they are occurring.

We also 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 [48,52]. Hence, customer segmentation based on their expectations is of great significance for firms to predict dynamics of targeted markets.

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 [41] and opinion tracking from

[23]. To discover heterogeneous expectations towards different brands, comparison-based algorithms [36] will be leveraged and redeveloped. The Web usage mining [10], sentiment-based algorithms [69] and opinion holder identification algorithms [66] will be integrated in order to make markets segments 'actionable' for managers and produce a custom score function to classify the sentiment [39].

Product extraction will augment the method in [16] and borrow some ideas from entity-based search engines [14] to mark items from free texts. Both product and feature extraction will allow business analysts to annotate text with an ontology and allow the semi/unstructured customer opinions/ complaints to be linked with structured customer data.

7. APPLICATION IN TRANSPORT LOGISTICS INDUSTRY

It is important to understand that today's transport logistics providers spend 50% of their time on managing the physical mess and 50% on managing the related information mess [12]. Here, intelligent transportation has enabled vehicle to driver, vehicle to vehicle and vehicle to infrastructure communications and emerging intelligent infrastructure that provides embedded unmanned 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 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 conjoint 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 [12] sector.

Intelligent Tracking powered by conjoint data and content mining is the core technology that is needed in the transport logistics industry today. Tracking movement of people, goods and services in the entire logistics network, tracking quality of services, service providers performance, through the 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 may be necessary. Intelligent tracking powered by conjoint predictive analytics with real time data and in the 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 conjoint mining for monitoring, visualisation, sharing, control and management of the physical mess (goods and assets) and the information mess (data) as well as business processes. This Business Intelligence leads to maximisation of human, transport and infrastructure performance and minimisation of the costs and security risk. In Logistics data on goods movement and storage and customer profiles are in structured form. Customer interactions are often in semi/unstructured form. In addition customer manifests, Import/Export documentation is often in semi/unstructured form.

The conjoint data and content mining is necessary 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 conjoint data and content mining are also needed for Inter- and intra-logistics partners transactions data monitoring, customers based tracking of trades data, smart phone, blue-tooth, 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 intelligence.

We use a 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 which includes tree structured data (Geo-data), XML documents (procedures and workflows), semi/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 conjoint 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. [74]. 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

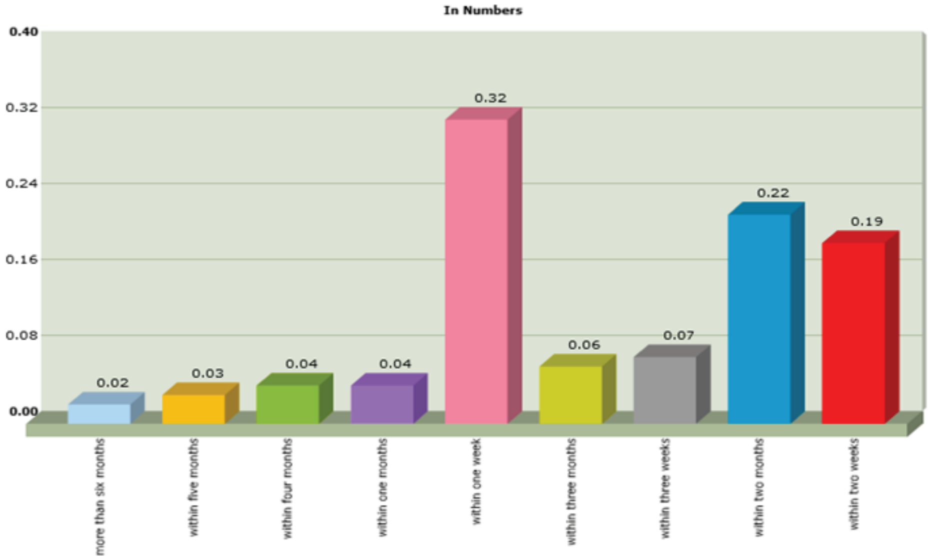

Figure 2 Number of defects in a category.

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 [55], Human Disease Ontology [25] and Mental Health Ontology [30]. 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. CASE STUDY

In order to illustrate the use of the technique of conjoint mining developed in this paper on a practical example, the conjoint data mining technique was applied to data collected from a CRM system in a Property Development company. The work was carried out in conjunction with ALYSYS SDN. BHD as part of a Brian Gain project funded by the MOSTI in Malaysia. The particular case study in this domain required a system to capture and track customer complaints related to property defects. The system tracked the complaint (in the form of a case and Ticket) till the defect was rectified. With such an architecture, although the CRM system provided us with the basic functionality to keep track of the customers' request for repairs but it did not allow us to take their feedback into consideration and/or to determine the reasons why the delay was occurring, what type of property is the most affected etc as this information is usually stored separately in a different format. In other words, the CRM system could tell us 'what' happened but not 'why' and 'how' it happened. For example, as shown in Figure 2 the CRM system can tell us the number of defects belonging to a particular category. However, it is not able to provide the reasons why these defects are occurring or the associations with other factors such as contractors and type of property. Furthermore by using the case/ticket functionality, we could extract reports on how long it is taking for an issue to be resolved. However, we cannot analyze the factors affecting the duration to rectify the defect such as the type of defect, property type or suburb, reasons for the overdue case/ticket duration etc. To address the problem, we utilized the conjoint consideration of structured and semi-structured information from different data sources to synthesize the reasons behind the observations. After applying conjoint data mining, useful patterns and associations were found that could not be found with traditional methods. Some examples of the useful patterns/associations found by the system in the property domain were:

(a) Based on a particular property type (for example apartment, commercial outlet, home etc) in a particular suburb we were able to identify and categorize the different types of defect subcategories that occur. For example as shown in Figure 3 , it can be seen that the probability of having a plumbing pipe defect in an apartment in a particular suburb/area is 24%.

(b) Furthermore in that suburb, we were able to determine the probability that a particular contractor will take a particular time period to fix a particular defect (Figure 4); e.g. the probability that the contractor will take within 1 week to fix an electrical problem is 32% whereas there is a 19% chance that he will take up to 2 weeks to fix a defect.

Based on the obtained analysis, strategies can be developed to better address the customer complaints. For example, we can

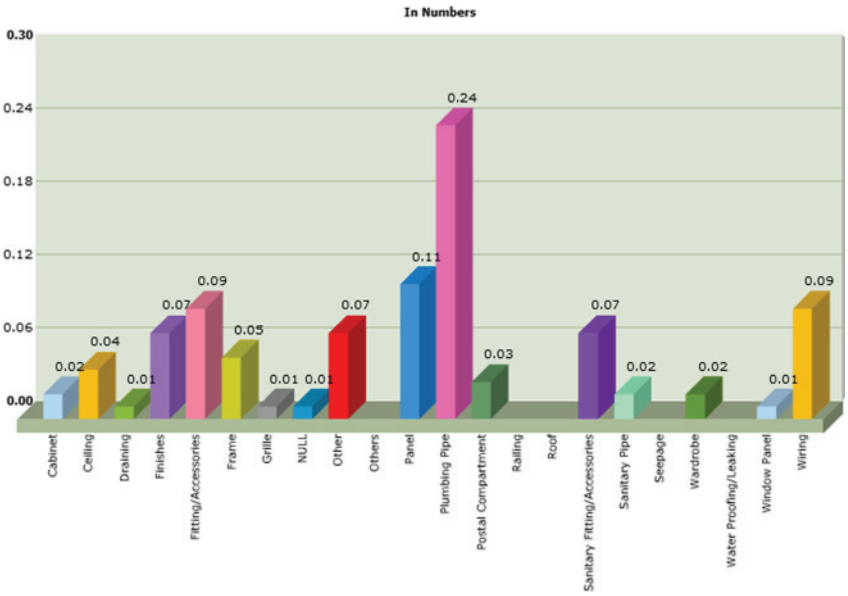

Figure 3 Different types of Subcategories of defects in a particular suburb.

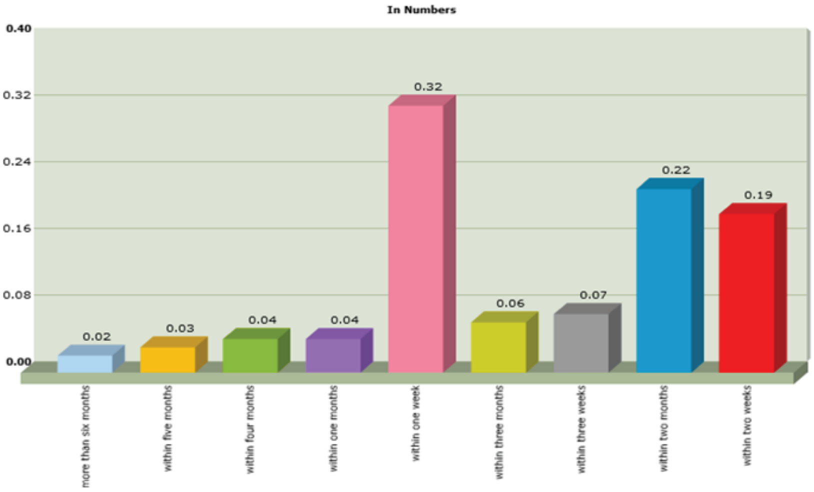

Figure 4 Probability of time taken by contractor to fix defect.

develop recommendations and strategies that will enable a real estate company to:

(i) make predictions on resource planning based on how long a particular defect category takes to fix and then schedule resources accordingly so that the contractors time is used more efficiently.

(ii) make predictions on which defects tend to occur together – for example, a plumbing pipe issue could also be associated with a leaky ceiling in 58% of cases. This may help in doing preventative maintenance at the same time rather doing it separately.

(iii) identify the most commonly occurring sub-category of defects in a building type and then identify a list of potential contractors in particular suburbs who can deal with them in a reasonable amount of time.

(iv) obtain insights for identifying the reasons why it is taking a long time by a contractor to resolve a particular case. Such analysis can be utilized to provide an adaptive service to the customers the next time a similar case occurs.

10. CONCLUSIONS

The paper presents a methodology for conjoint mining of structured and semi/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.

11. ACKNOWLEDGEMENT

The work presented in this paper is financially supported by the Australian Research Council ARC Discovery Grant DP130104770 and Brain Gain project DIP02/2010 by MOSTI (Ministry of Science Technology and Innovation, Malaysia)

REFERENCES

1. K. Aberer, T. Catarci, P. Cudré-Mauroux, T. Dillon, S. Grimm, M. S. Hacid, et. al. Emergent semantics systems. Semantics of a Networked World. Semantics for Grid Databases, 14–43
2. 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.
3. 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.
4. M Alhamad, T. Dillon, E. Chang, “SLA-based trust model for cloud computing” x 2010 13th International Conference Network-Based Information Systems (NBiS).
5. 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.
6. S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, “Semantic integration of heterogeneous information sources,” Data Knowl. Eng., 36 (3), 215–249, 2001.
7. 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.
8. D. M. Bikel, R. Schwartz, and R. M. Weischedel, “An algorithm that learns what’s in a name,” Mach. Learn., 34 (1), 1999
9. M. L. Brodie, “Computer science 2.0: A new world of data management,” in *Proc. 33rd VLDB Conf.*, 2007, p. 1161.
10. G. Büchner and M. D. Mulvenna, “Discovering internet marketing intelligence through online analytical web usage mining,” ACM SIGMOD Rec., 27 (4), 54–61, 1998.
11. 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.
12. E. Chang “Transport Logistics, the Grand Challenges”, 2014 Australian Defence Force Academy
13. E. Chang, T. S. Dillon, F. K. Hussain Trust and reputation relationships in service-oriented environments 2005. ICITA 2005. Third .Int. Conf. Information Technology and Applications.
14. 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.
15. T. S. Dillon, R. W. Martin, and D. Sjelvgren, “Stochastic optimization and modelling of large hydro-thermal systems for long term regulation,” Intl Journal of Electrical Power and Energy Systems, 2 (1), 2–20, 1980
16. 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.
17. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle, “Ontology-based extraction and structuring of information from datarich unstructured documents,” in Proc. 7th Intl. Conf. on Inform. & Knowl. Mgmt. (CIKM),USA, 1998, pp. 52–59.
18. 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.
19. 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.
20. 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
21. D. Ferrucci and A. Lally, “UIMA: An architectural approach to unstructured information processing in the corporate research environment,” Nat. Lang. Eng., 10 (3–4), 327–348, 2004.
22. 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.
23. 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.
24. T. R. Gruber, “A translation approach to portable ontology specifications,” Knowl. Acquis., 5 (2), 1993.
25. M. Hadzic and E. Chang, “Medical ontologies to support human disease research and control,” J. Web Grid Serv., 1 (2), 2005.
26. 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.
27. 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
28. 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.
29. F. Hadzic, H. Tan, and T. S. Dillon, “UNI3 – efficient algorithm for mining unordered induced subtrees using TMG candidate generation,” IEEE Sym. on Comp. Intel. and Data Mining (CIDM), USA, 2007, pp. 568–575
30. M. Hadzic, M. Chen, and T. S. Dillon, “Towards the mental health ontology,” in Proc. IEEE Intl. Conf. on Bioinformatics and Biomedicine (BIBM), USA, 2008, pp. 284–288.
31. 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.
32. 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.
33. http://idcdocserv.com/
34. http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm. (2012)
35. M. Jarrar and R. Meersman, “Formal ontology engineering in the

DOGMA approach," in Confederated Intl. Conf. CoopIS, DOA, and ODBASE, California, USA, 2002, pp. 1238–1254.

36. 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.
37. R. Kaye, "The gloss system for trans. from plain text to XML," Proc. MathUI 2006 http://www.activemath.org/∼paul/MathUI06/
38. 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
39. S.-M. Kim and E. Hovy, "Determ. the sentiment of opinions," 20th Intl. Conf. on Comp. Lingustics. Switzerland. 2004.
40. 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.
41. 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.
42. K. Kukich, "Techniques for automatically correcting words in text," ACM Comp. Surv., 24 (4), 377–439, 1992.
43. 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.
44. 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.
45. Li, X., P. Morie, et al., "Semantic integration in text: from ambiguous names to identifiable entities." AI Mag. 26(1), 2005.
46. B. Marchal, "XI: Open-source conver. of legacy text files to XML," in http://www.ananas.org/xi/index.html, [Ac.: 20/11/'08].
47. P. McBrien and A. Poulovassilis, "A semantic approach to integrating XML and structured data sources," in Proc. 13th Intl. Conf. ondvanced Information Syst. Engineering (CAiSE), Switzerland, 2001, pp. 330–345.
48. R. L. Oliver, "A cogni. model of the antecedents and conseq. of satisfaction decisions," J. Marketing Rsch., 17, 1980.
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 Semantics, Knowledge and Grid (SKG), Beijing, China, 2008, pp. 87–94.
50. 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.
51. F. M. Reza, An introduction to information theory. New York: Dover Publications, 1994.
52. R. T. Rust, J. J. Inman, J. Jia, and A. Zahorik, "What you don't know about customer-perceived quality: The role of customer expectation distributions," Marketing Science, 18 (1), 77–92, 1999.
53. S. Sestito and T. S. Dillon, "Knowledge acquisition of conjunctive rules using multi-layered neural networks," Int. J. Intell. Syst., 8 (7), 779–806, 1993.
54. S. Sestito and T. S. Dillon, Automated knowledge acquisition. Sydney: Prentice Hall, 1994.
55. A. S. Sidhu, T. S. Dillon, E. Chang, B. S. Sidhu Protein ontology: vocabulary for protein data 2005. ICITA 2005. Third Int. Conf. Information Technology and Application.
56. C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable techniques for mining causal structures," Data Min. Knowl. Disc., 4 (2), 163–192, 2000.
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. 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.
59. B. R. Szkuta, L. A. Sanabria, and T. S. Dillon, "Electricity price short-term forecasting using artificial neural networks," IEEE Transactions on Power Systems (PES), 14 (3), 851–857, Aug. 1999 1999.
60. 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
61. H. Tan, F. Hadzic, T. S. Dillon, L. Feng, and E. Chang, "Tree model guided candidate generation for mining frequent sub-trees from XML," ACM Trans. Knowl. Discov. Data, 2 (2), July 2008.
62. H. Tan, T. S. Dillon, F. Hadzic, and E. Chang, "Razor: Mining distance-constrained embedded subtrees," in Proc. Workshop on Ontology Mining and Knowledge Discovery from Semistructured documents (MSD) in conjunction with 2006 Intl. Conf. on Data Mining, Hong Kong, 2006, pp. 8–13.
63. 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
64. 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.
65. 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.
66. L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text," Bioinformatics, 18 (8), 1124–1132, 2002.
67. 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.
68. 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.
69. 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.
70. 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.
71. K. Wang and H. Liu, "Discovering structural associations of semistructured data," IEEE Trans. Knowl. Data Eng., 12 (3), 353–371, 2000.
72. C. Wouters, T. Dillon, W. Rahayu, E. Chang, R. Meersman Ontologies on the MOVE Database Systems for Advanced Applications, 812–823.
73. 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.
74. 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, Edited by D. Taniar and J. W. Rahayu, Idea Group Publishing, USA, 2004*, D. Taniar and W. Rahayu, Eds. USA: Idea Group Publishing, 2004.
75. 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.

76. 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.
77. 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.
78. M. J. Zaki, "Efficiently mining frequent trees in a forest: Algorithms and applications," *IEEE Trans. Knowl. Data Eng.*, 17 (8), 1021–1035, August 2005.
79. 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.
80. 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.
81. 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.