

Harnessing the Wealth in your Data: Integrating Predictive Modeling and Text Analytics with xPatterns™



Table of Contents

• Executive Summary	3
• Background on Traditional BI, Predictive Modeling & Text Mining	4
○ The “Capture, Predict, Report, Act” BI and Predictive Modeling Process Flow	
○ Bringing Unstructured Data into the cycle – today’s Text Analytics Solutions	
• Introduction to xPatterns™	7
○ Data Capture	
○ Hierarchy Free Ontology Discovery, Maintenance and Refinement (Capture)	
○ Deriving Cohorts for Analysis (Predict)	
○ Unstructured Semantic Queries (Act)	
• xPatterns Value Proposition: Unlocking Unstructured Data to enrich Data Capture, Prediction and Drive Action	14
○ Example Use Case – Non-iPhone Telco Customer Service Call	
• Conclusion	16

The xPatterns platform allows real-time unstructured data to be directly integrated into current Business Intelligence (BI) and Predictive Modeling solutions in the enterprise.

Executive Summary

xPatterns enables enterprises to act upon the dynamic nature and richness of unstructured text data in a way that makes it actionable while reducing solution time-to-market and reducing overall solution maintenance.

The xPatterns platform allows real-time unstructured data to be directly integrated into current Business Intelligence (BI) and Predictive Modeling solutions in the enterprise. Moreover, xPatterns can leverage current investments in “linguistics-based” text analytics technology. The benefit realized is a solution that allows adaptive learning of semantic ontologies, while retaining the flexibility and transparency that comes with tools that enable BI and predictive modeling with structured data, and linguistics-based text analytics.

Uniquely, the xPatterns solution adapts and refines its models through usage by BI analysts, marketers and end-consumers in the field. This results in performance that continually improves and reacts to dynamically changing trends over time.

In this whitepaper we:

- Describe the "Capture, Predict, Report, Act" cycle of enterprise business intelligence.
- Introduce conventional means of leveraging unstructured data in the enterprise.
- Outline the role of the xPatterns platform in deriving rich, broad views of the customer that respond to changing enterprise, industry and customer information dynamically, down to the individual level.
- Describe how xPatterns reduces time to market and boosts ROI by applying novel machine learning techniques that make new unstructured business information and context actionable for BI.
- Illustrate the effectiveness of the technology by highlighting its simple integration into enterprise systems for traditional reporting and predictive modeling based on structured data.



capture



predict



report



act

Background on Traditional BI, Predictive Modeling & Text Mining

The “Capture, Predict, Report, Act” BI and Predictive Modeling Process Flow

"Capture, Predict, Report, Act" is the four-step cycle of bringing data through a path resulting in true business impact. The different phases of the process can be described as follows;

- **Capture** - Business intelligence, predictive modeling, and text mining are driven by data insight. The goal of the capture phase is the physical or logical aggregation of data, internal and external to the enterprise, that provides the broadest view of the customer. Data sources include:
 - *Descriptive data*: usually a mix of customer registration data, that may be self-declared information and externally-sourced (geo)demographics, business categorization and customer product/plan/subscription information.
 - *Behavioral data*: this can be as simple as transaction records, such as who bought what and when, or details of how customers use a product or service. For example, call data records for a telco, television viewing patterns for a cable company, time spent online and time-of-day usage information for an internet provider, details of how customers spend with their credit card and how they pay off their bill for a credit card company.
 - *Interaction data*: the details of how customers interact with the company through its various channels. This might include details of Web site visits by registered customers (mapped from “click-level” data to business-meaningful “events” that happen during visits), third-party web analytics data (generated by Omniture, Quantcast, or similar service), categorizations/tags/labeling of customer service contacts - including reasons for contact and outcomes via online chat, email or telephone.
 - *Attitudinal data*: people’s needs, preferences, opinions and desires. Such information is most often collected in the course of surveys conducted for market research or to assess customer satisfaction, and used in aggregate. The industry is now pushing for more actionable data as it refers to the individual.
- **Predict** - Given a very broad set of data captured, analysis allows the generation of a number of attributes that are indicative of certain propensities of customers or trends in the business. Applying

“Capture, Predict, Report, Act” takes us from collected data, through advanced analysis, reporting, and interrogation of the data by multiple business functions in the company, to the successful deployment of analytical results to improve business processes.

predictive modeling techniques to the data captured, associates derived temporal, behavioral, interaction or attitudinal attributes with ultimate customer action. This makes it possible to make predictions of customer outcomes in current or future cases. A prediction is the “raw” output of a model. It may be the propensity for a customer to behave in a particular way. For example, the probability that a wireless customer will churn within the next 30 days, based on their customer support call complaining about text overbilling, and the impending end of the 2-year contract. An example of this may be estimating the most appropriate set of features to describe to a mother adding a line for her child to her calling plan.

- **Report** - Value is realized in the cycle when the captured, analyzed and predicted views of the customer data are communicated to the organization. Rich reporting solutions allow a variety of enterprise functions to leverage the data into financial, marketing, customer care, and business planning functions. Monitoring reporting systems and dashboards through real-time and periodic (day-to-day, weekly, monthly and quarterly) reviews help the company to identify business triggers through trending and suddenly changing data. Examples include groups of customers that need additional support or customer interaction, emergent issues with a particular product or device, and the need for a new product or service that responds to a particular behaviorally or attitudinally expressed need.
- **Act** - Based on knowledge derived from the reporting of the data in the capture and predict phases of the process, appropriate actions can be formulated to optimize aspects of the business. For example, observing a particular set of behaviors may activate a prediction of high churn probability for a customer, and analysis of descriptive and behavioral data associates the customer to a particular group with which a new product offering is proving successful in retention. The “act” phase leverages a combination of data attributes, predictive models, business rules and logic to determine the most-appropriate actions at the appropriate points/timing/channels/context in the business process. Feedback from the collection of action data allows richer prediction, detailed reporting and refined action back through the cycle.

“Capture, Predict, Report, Act” takes us from collected data and advanced analysis through reporting and interrogation of the data by the company to successful deployment. This cycle enables continuous improvement: new data captured at the point of interaction (for example, during customer interactions) enhances the analytical data view, enabling more accurate predictions, richer reporting and driving better decisions with a greater proportion of positive outcomes and, hence, higher returns.

The volume of unstructured data within the enterprise (recently estimated as comprising about 85% of all enterprise data¹) is growing exponentially.

Bringing Unstructured Data into the cycle - today's Text Analytics Solutions

The volume of unstructured data within the enterprise (recently estimated as comprising about 85% of all enterprise data¹) is growing exponentially. In tandem, the era of Web 2.0 is yielding large volumes of highly dynamic, real-time, and unstructured feeds of data containing knowledge essential to the enterprise. This magnifies the need for approaches that integrate unstructured data analytics into the Capture, Predict, Report, Act cycle.

The domain of "linguistics-based" text analytics is where unstructured data is leveraged in enterprise analytics today. Such solutions are designed to first capture information from text data describing interactions and leverage rules around entities, actions, and attitudes as they are expressed in written text. The solutions then "annotate" unstructured data fields with labels that can be leveraged in structured data analytics; predictive modeling and reporting.

Text Mining generally involves the following 4 major steps:

- 1. Preparing text for analysis** - this involves document conversion, segmentation and identifying language.
- 2. Extracting concepts**
 - *Applying linguistic resources* - this is where linguistic understanding is encoded in the solution, leveraging what are termed "templates" or "rules" - that are specialized to specific application areas, such as CRM, market intelligence, gene ontology, genomics, Medical Subject Headings or MeSH®, IT, opinions, and security intelligence. These templates generally encode semantic data dictionaries and existing semantic ontologies where they exist.
 - *Term extraction* - identifying and preserving n word terms or identifying combinations of terms matching a templated concept.
 - *Type assignment* - assigning a *type* to a specific term (e.g.: person, automobile, movie, etc.)
 - *Creation of equivalence classes* - identifying similar or related concepts expressed differently in the content.
 - *Indexing* - processing data to facilitate performant lookup and retrieval.
- 3. Uncovering opinions, relationships, facts, events, & categorizing them**

NLP-based text analytics enables analysts to identify and segregate positive and negative concepts in text responses. In addition to simple

¹ Merrill Lynch, 2009

Underlying the xPatterns product is novel technology that allows the automated creation and dynamic maintenance of semantic ontologies.

positive/negative statements, text analytics solutions provide insight into positive or negative attitudes by leveraging configurable contextual cues, such as sentence structure. In this way, sentiments like those in the sentences below would be grouped correctly, despite the fact that one opinion is positive, one is negative, and one is mixed.

The hotel manager was very courteous.
The hotel manager was really rude.
The hotel staff was courteous but the room was too small.

With the appropriate rule creation, NLP-based text analytics can also uncover connections between facts and events to support initiatives as diverse as market intelligence, fraud detection, and life sciences research. NLP-based text analytics would determine that the following three phrases all have the same meaning

Company A was acquired by Company B
Company B acquired Company A
Company B's acquisition of Company A is complete

And, if a text document should read “Company B failed to acquire company A,” would correctly identify that the transaction did not occur.

4. Deploying text analytics results in predictive models

Deployment of text analytics results, in predictive models, is the step that links text analytics to decision making. Essentially the annotations of the text documents can themselves become attributes of the document, the document source, or the entity the document is referring to. For example, the frequency of expressions of negative sentiment in a customer support call transcript may be assigned as additional predictive modeling attributes applicable to the customer engaged in that call. Furthermore, the aggregate of calls for a given CSR may be related to modeling or reporting attributes associated with the performance and quality of that CSR.

Introduction to xPatterns

A fundamental challenge exists in today's NLP-based text analytics solutions, specifically when it comes to the integration of these technologies into enterprise BI and Predictive Modeling. Many aspects of the advanced text mining capabilities available that we described above, and particularly items 2 and 3, rely on the existence or creation of ontologies and semantic rules - exemplified by the suite of specialized templates provided as a part of many commercial text analytics solutions. Such templates are required to allow the following aspects of understanding a segment of text:



Linkage Analysis enables analysts to identify and segregate positive and negative concepts in text responses.

- To resolve ambiguity – needed to determine the right meaning
- To determine synonym relationships – to determine common themes across documents where those themes are described in different terms
- To identify and categorize entities, actions and events

Creating deep ontologies and semantic rules for a given semantic domain (eg. CRM, genomics, security intelligence) is a manual, and at times necessary, process. It creates inertia to deploying text analytics solutions in domains for which semantic ontologies do not currently exist or where the volume and dynamics of the data involved in describing the domain is such that creating and maintaining a deployment-specific ontology is untenable.

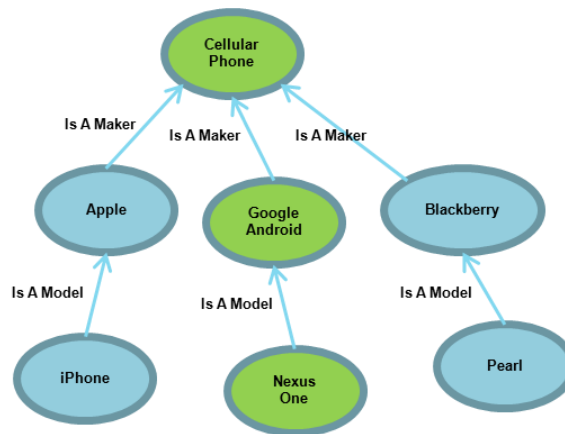


Figure 1 - Example of a simple ontology for smartphone platform manufacturers and devices

Until the Semantic Web² is a reality, there will be very few rich ontologies available. Underlying the xPatterns product is novel technology that allows the automated creation and dynamic maintenance of semantic ontologies. The technology, represents "IsAssociatedWith" relationships for domains, derived simply from reading and reviewing large bodies of unstructured text information about the domain that can be streamed to the system in real-time. The technology can be leveraged to determine indirect semantic relationships between queried concepts, and to facilitate understanding of the relevance of a specific document to a specific concept. Moreover, using *the xPatterns Persona*, it is possible to maintain unstructured semantic profiles of consumers based on implicit and explicit preference data and enable privacy-controlled querying for relevant content (an action, interaction, offer or item to display) for a consumer.

² Berners-Lee et. al, Scientific American Magazine, 2001



Data Capture

The xPatterns Persona makes it possible to maintain unstructured semantic profiles of consumers based on implicit and explicit preference data and enable privacy-controlled querying for relevant content.

xPatterns' document analysis techniques including HTML tree parsing can be employed to determine most significant items of text in a poorly constructed HTML page.

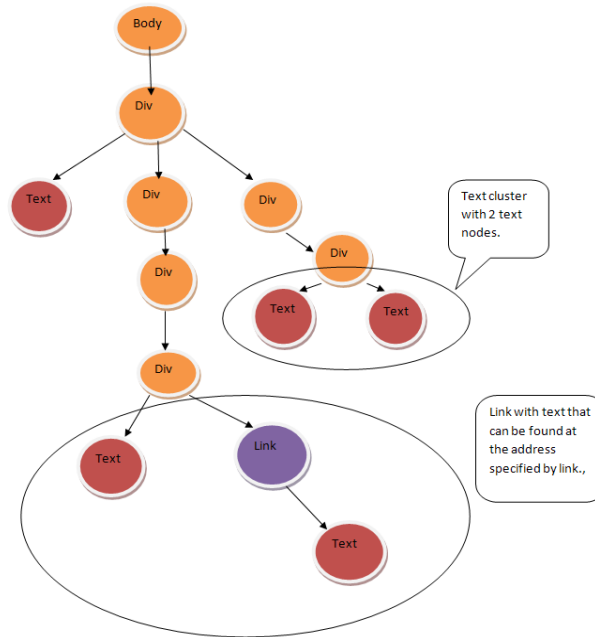


Figure 2-xPatterns HTML Document Analysis

Furthermore, it is possible to augment the data capture and concept extraction process using annotations of documents by conventional text analytics tools from companies like IBM, SPSS, Attensity, and Clarabridge. These technologies embed pre-existing custom data dictionaries and ontologies into the processing of documents for modeling within xPatterns.

Hierarchy Free Ontology Discovery, Maintenance and Refinement (Capture)

The data flow may take any of the following paths depending on the intended use of the unstructured data:

1. *Ontology Discovery* - which we term creating a "domain expert" model for a particular domain. We take a "relevance discovery" approach to extracting a model of concept relevance by mining a corpus of text content in the category of interest. The corpus is mined from the internet or elsewhere (for example using the Wikipedia Commons data set) which we reasonably believe might span the majority of content and context in the category, and remain updated and accessible. This allows

Ontology Discovery allow us to develop a reproducible process for building domain-expert relevance indices and keeping them up to date, with minimal manual intervention.

us to develop a reproducible process for building hierarchy-free ontologies and keeping them up to date with minimal manual intervention.

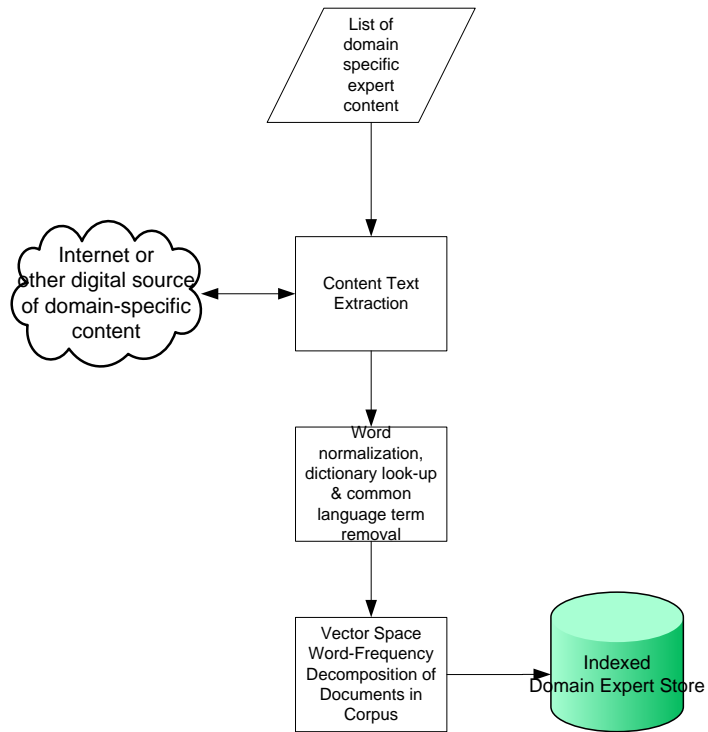


Figure 3- Ontology discovery, capture and processing of domain-specific unstructured text corpora

2. *Content Modeling/Representation* - content to be indexed for relevance-based querying/action rather than ontology discovery passes through the same basic processing flow as in Figure 4, and can be managed by the Text Mining customer via a SOAP web services interface.

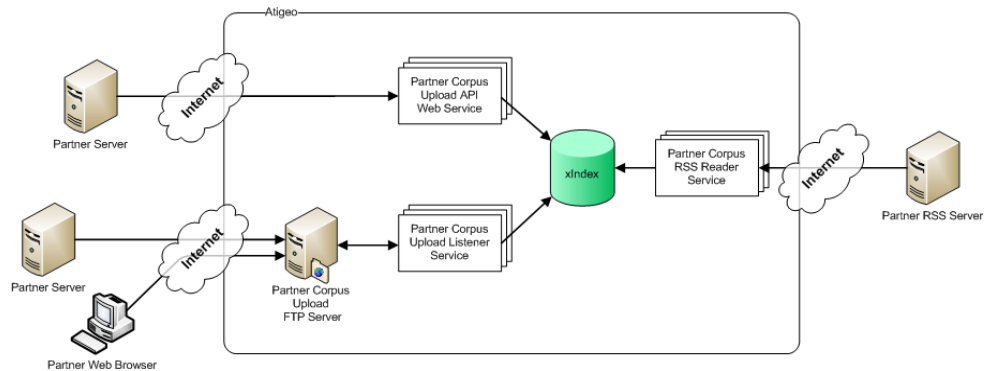


Figure 4 – Content modeling / representation

Having captured and cleaned the data, it is decomposed into a tokenization of "concepts"; single-word or multi-word terms, or broader semantic

The frequency of discovered concepts in the domain, relative to concepts in the corpus as a whole, are expressed in a vector-space representation of a document.

concepts (e.g. negatives, date and time statements, location statements) using conventional text analytics tools. The frequency of discovered concepts in the domain, relative to concepts in the corpus as a whole, are expressed in a vector-space representation of a document. The relationship between concepts and the documents/content in which they are expressed is used to specify the weights of a neural network model that;

- for content, presents as an output the projection of a given concept, or combination of concepts in a set of content - the relevance of a given content item jointly to those concepts (content corpus in part a of Figure 5)
- represents "domain expert" relationships between concepts and other concepts, mediated by the discovered content specifying the category or domain (domain expert corpus shown in parts a and b of Figure 5)

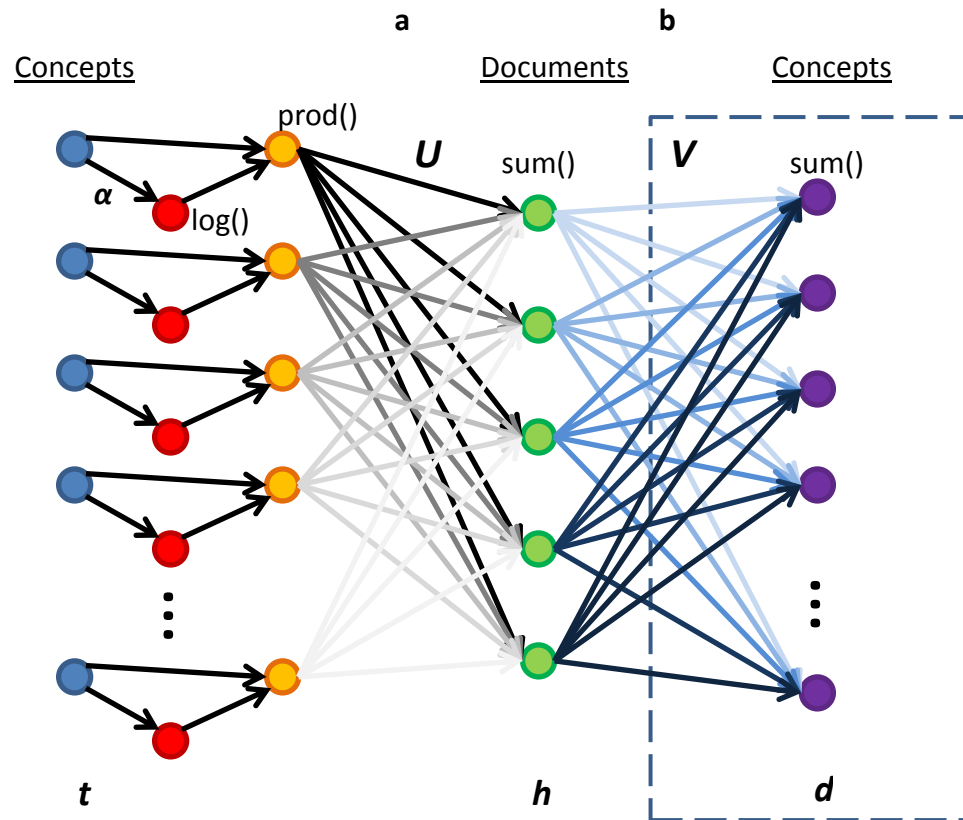


Figure 5 - xPatterns Representation of Content Corpus and Domain Expert Corpus Neural Networks

User interaction refines the semantic relationships of concepts to one another and to content in real time within the system, which translates into more relevant query results.

For example, querying a domain expert in sports in the U.S. for “giants” might



return the following top concepts:

- “giants”->“San Francisco Giants”
- “giants”->“New York Giants”
- “giants”>“Standing on the Shoulders of Giants”

Using xPatterns technology allows captured customer data attributes to be leveraged in situations where there is no direct match with a query made against the users, but there is a match based on "IsAssociatedWith" semantic relationships, as mediated by xPatterns.

Repeated user interaction with the San Francisco and New York concepts increases their association with “giants” and decreases association of “Standing on the Shoulders of...”

Deriving Cohorts for Analysis (Predict/Report)

Using xPatterns technology allows captured customer data attributes to be leveraged in situations where there is no direct match with a query made against the users, but there is a match based on "IsAssociatedWith" semantic relationships as mediated by xPatterns. This allows enrichment of the predictive modeling attributes that can be leveraged for a given user base. For example, we might define a cohort of users by those to whom "Classic Rock Music" is at least 60% relevant and use that as an input attribute to a classification model for user up sell opportunities.

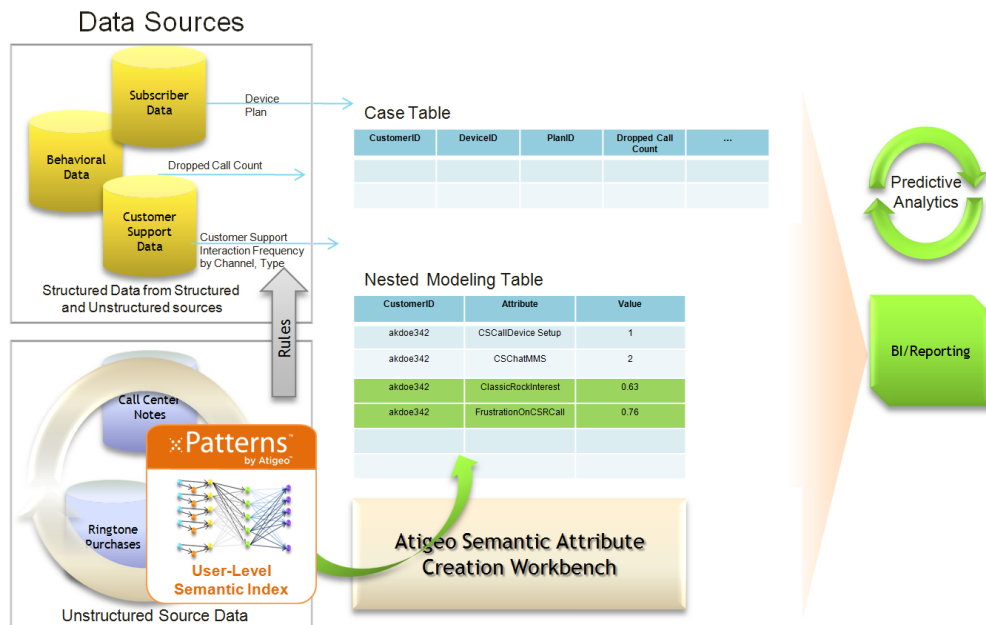


Figure 6 - Integration of xPatterns into conventional structured BI and predictive modeling.

As noted in Figure 6, “xPatterns’ Semantic Attribute Creation Workbench” provides a straightforward approach to interrogating unstructured customer data and integrating those findings directly into structured data. The combined results populate reporting and predictive modeling solutions and allow a BI analyst to create structured data attributes that are associated with a semantic query. For example, creating an attribute indicating a customer's

Actions such as offers, content to be displayed, items to suggest, tones of voice or phrases to use with a customer, can be selected based on semantic relevance.

interest in Classic Rock Music can be implemented by creating a semantic query for the terms "classic rock music" perhaps against the data source of ringtone titles and descriptions that the user has purchased. While none of the ringtones may be tagged as "classic rock", xPatterns' music domain expert will understand the association of Eric Clapton and Cream with classic rock, and the lack of affinity of Britney Spears to the same, allowing the relevance of classic rock to each individual to be assessed based on their purchased ring tones. The relevance score is converted into the "Attribute Value" corresponding to a customer attribute named "ClassicRockInterest," and then can be analyzed and reported on across the customer base and leveraged as input attributes to predictive models. As new ringtone purchases are made by the user, more is revealed about their musical taste, and the attribute value will be automatically updated.

Unstructured Semantic Queries - Determining Relevant Action (Act)

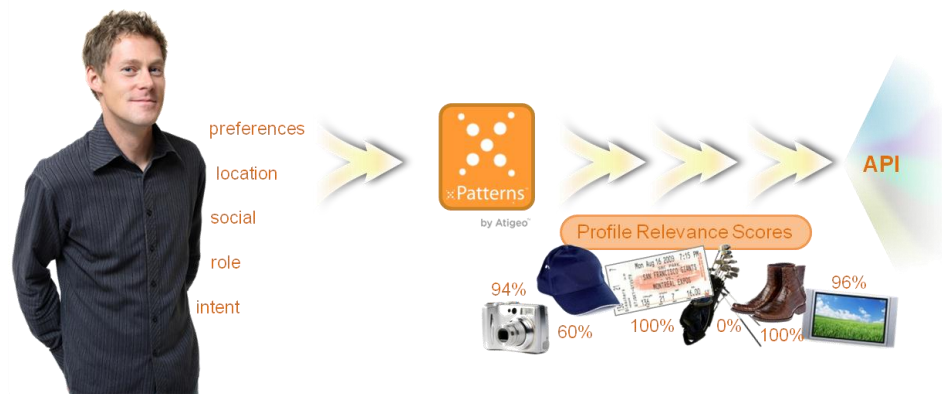


Figure 7 - Customer profile-based relevant action.

Actions such as offers, content to be displayed, items to recommend, tones of voice or phrases to use with a customer, can be selected based on semantic relevance, as illustrated in Figure 7. For example, assume a customer has been flagged as having a strong propensity for churn and there are a number of different retention offers that are available to propose to the customer. There may be rules governing offers that should or should not be made to the customer e.g., "don't recommend a smartphone to a user who has just opted out of a data plan." Using xPatterns technology allows us to assess the more subtle affinity of a customer to a particular potential offer or to expand the types of offers that are possible. For example, offering concert tickets to a customer who has bought a particular ringtone, or pre-populating their device with a number of ringtones they might like.

xPatterns Value Proposition: Unlocking Unstructured Data to Enrich



Using xPatterns technology allows us to assess the more subtle affinity of a customer to a particular potential offer or to expand the types of offer that are possible.

Data Capture, Prediction, Reporting and Drive Action

- Provides a mechanism for leveraging unstructured enterprise data into present business intelligence data flow and processes.
- Allows access to external real-time unstructured data sources, including social media, for business intelligence, CRM and marketing.
- Compliments existing business intelligence, reporting analytics, and data mining solutions by incorporating unstructured data while leveraging existing data warehouse investments.
- Improves productivity over conventional text analytics platforms by reducing the time required to create and maintain ontologies for text analytics.
- Allows businesses to associate findings in structured and unstructured data to provide new dimensions for reporting and attributes to drive predictive modeling and business decision processes.

Example Use Case –Telco Customer Retention

In this example, xPatterns participates in the Capture, Predict, Report, Act cycle, to both retain and cross-sell the customer of a Communication Service Provider (CSP). The Telco has a number of structured and unstructured data attributes across disparate data silos, pertaining to the customer such as; plan and device information, network behavior information and call data records allowing the discovery of dropped calls, and topic identification in messaging, together with IT system information like purchased ringtones and user searches made on the CSP web portal.

Common approaches today intersect behavioral and billing information on customers with survey data on a subset of the customer base, to identify and then learn to classify in the broader customer base, different segments of customers. For example, groups of customers with similar behavioral and billing attributes that link them. Such segments allow business rules to be created determining broadly how to treat those users in retention and cross-sell and up-sell situations. The challenge is that the interactions are then guided at an aggregate level of user commonality, rather than at the level of a direct, personal understanding of the user and can miss the mark.

In the scenario of Figure 8, current device plan and network behavior data are used to determine an offer to the customer, which does not leverage additional unstructured, semantically informative attributes of the user that are available to the CSP.

xPatterns unlocks previously unleveraged data elements of the user, by focusing on unstructured sources which require a level of semantic understanding.

xPatterns unlocks previously unleveraged data elements of the user, by focusing on unstructured data elements which require a level of semantic understanding.

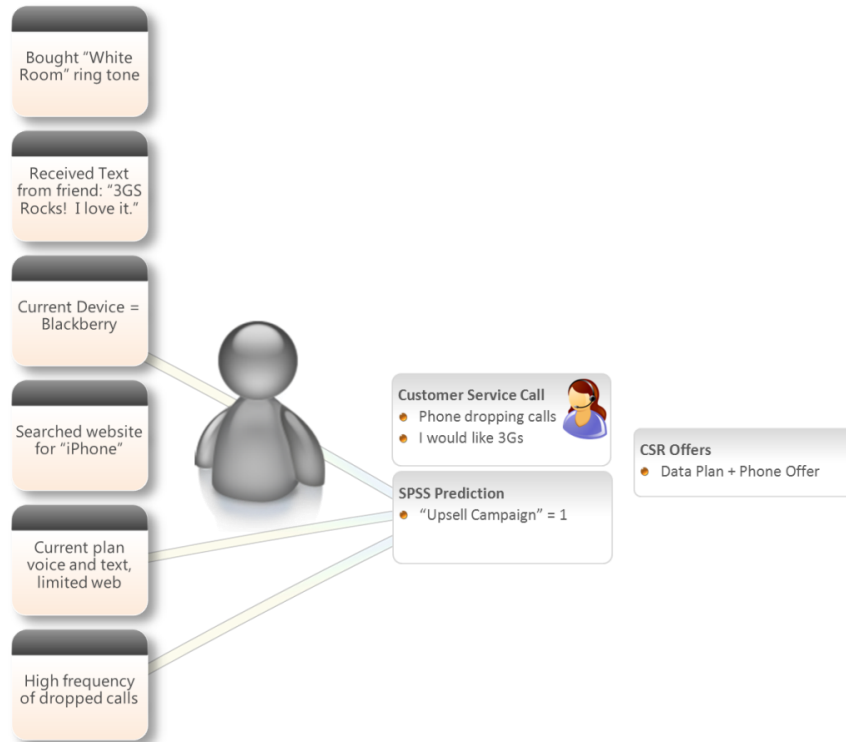


Figure 8 - Common Customer Service Call Center Retention/Cross-sell Offer Interaction

By leveraging semantic annotations in text analytics solutions, it is possible to identify the positive expression of sentiment exhibited by a friend towards the *iPhone* 3GS in this context. The website search logs also reinforce the likely user interest in the iPhone via the search for “iPhone” on CSP web properties. These data points can be used to augment the 'Predict' phase of the interaction cycle by providing new input attributes to predictive modeling. xPatterns leverages its domain expert to associate the iPhone 3GS which the CSP does not offer on its network, with the Google Android based smartphones the CSP does offer. The user's purchase of the "White Room" ringtone, a song by Cream of which Eric Clapton is a member, activates an association in the xPatterns music domain expert with Fender Guitars, for which there is currently a special edition Fender version of the myTouch Android-based phone, see Figure 9. This enriches the 'Act' phase of the modeling cycle, allowing a more personal, customer-specific offering to be made to the caller.

xPatterns' approach to self-constructing and maintaining semantic modules enables an enterprise to bridge the gap between structured and unstructured data assets, unlocking previously inaccessible business intelligence resources.

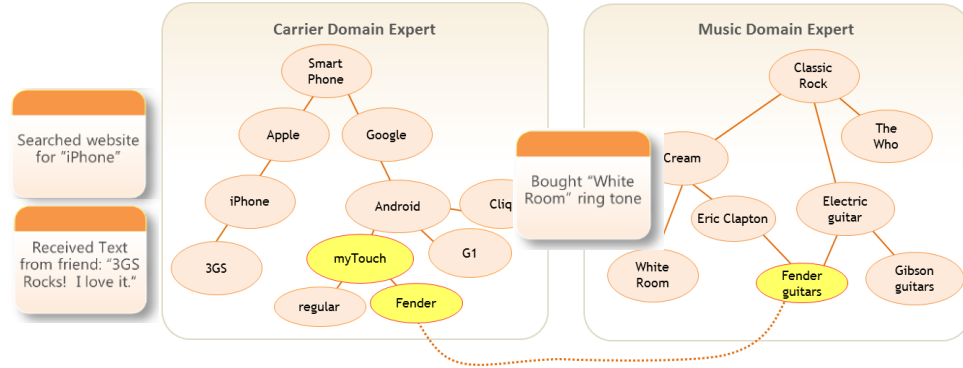


Figure 9 - xPatterns domain experts leveraged to reason with unstructured user data attributes

Using xPatterns technology, allows us to understand hidden meaning and relevance behind the user’s statements and behavior which offer an opportunity to make a uniquely personal and very insightful offer to the customer.

Conclusion

The challenge for enterprises today is the majority of their data is unstructured and dynamically changing, making it difficult to surface insights and take action upon it in real-time. The rapid growth of dynamic and real-time data feeds across multiple internal and external sources, vital to business intelligence, exacerbates the difficulty of extracting true meaning from unstructured text data. In an attempt to solve these issues, many enterprises suffer the poor ROI that conventional text analytics technologies provide, due to reliance on manually created and maintained ontologies, semantic rules, and templates.

xPatterns provides a cross-enterprise solution that efficiently integrates with current data handling and analysis systems. It enables the enterprise to index data from disparate data sources. While acting on the disparate data in a unified way, xPatterns continually responds to evolution of the data and feedback through customer action. xPatterns’ approach, coupled with a simple mechanism for leveraging semantic queries in structured analytics, enables an enterprise to bridge the gap between structured and unstructured data assets to unlock previously inaccessible business intelligence resources.