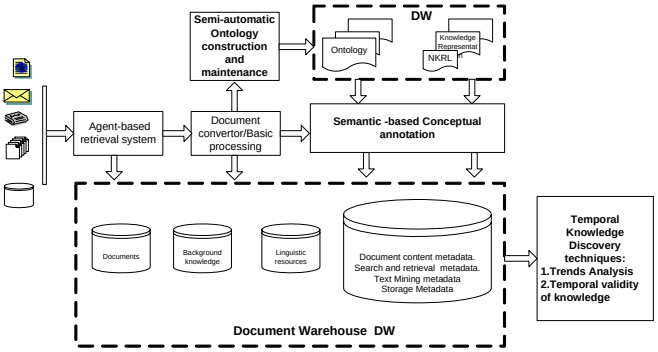
PARMENIDES White Paper: Discover hidden information from your texts!

Parmenides

Discover hidden information from your texts!

Information overload is a well known issue in the knowledge industry. At the same time most of this information becomes available in natural language which makes it difficult to find and analyze in ways that are domain specific and useful for end users. From the business perspective, there is an increasing need for decision makers to be able to access and digest the available information in order to understand developments in their markets and act accordingly.

The Parmenides system offers a novel approach to Text Mining, via an integrated suite of state-of-the-art tools and algorithms. The transformation of textual data into an XML-based representation, enriched with morphological, syntactical and semantic information, provides the basis for the application of advanced mining algorithms to extract useful, hidden knowledge from the text.



The PARMENIDES Architecture

The Parmenides Approach

The sheer volume of organizational knowledge, as expressed in textual documents, as well as the constant inflow of novel information via text-based channels (e-mail, RSS feeds, etc.) is increasing at such a scale that traditional Knowledge Management approaches are bound to fail. The modern Knowledge Worker simply does not have the time to absorb such an impressive amount of information.

The Parmenides system can cope with a wide variety of input formats (pdf, html, mail, etc) and, differently from all the existing Text Mining approaches can preserve and exploit the structural information of the original document, thanks to the unique internal representation format.

The application of state-of-the-art mining algorithms allows the discovery of previously unforeseeable trends and patterns hidden in a large text collection. Temporal discovery is a unique feature of the Parmenides system. The discovered trends and facts are stored into compact, information-rich structures, which are deeply interlinked with domain ontologies, either provided by the users, or automatically discovered by the system.

The system has been applied to a number of different domains including Press Releases, Company Project Reports and Security-Related News. The benefits of the system for all mission-critical Predictive text Analytics tasks are enormous. Textual data can be directly incorporated in knowledge-based strategic planning, in order to support everyday decision making.

Text Intelligence at its best!

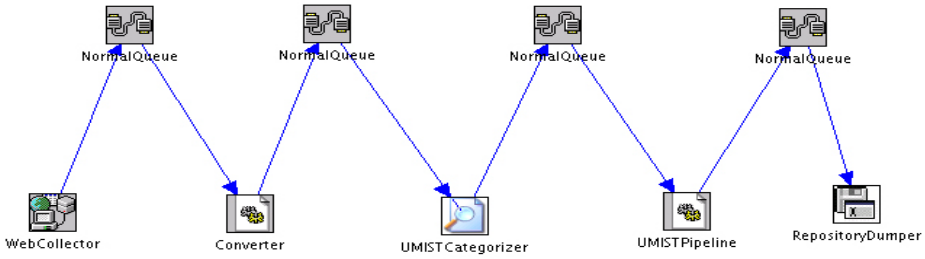
The Parmenides Document Lifecycle

Parmenides consists of components for the construction and maintenance of a Document Warehouse (DW). The document lifecycle begins with its collection and conversion to an XML-based annotation scheme. The documents can be of various types: web pages, e-mails, word processor document, etc. An agent-based document collector gathers these from various sources and converts them into XML for further processing. The utilized XML scheme is called *Common Annotation Scheme* **(CAS**).

The conceptual annotator extracts temporal semantic information from the documents, by applying information extraction and lexical chaining techniques. During this process, the Domain Knowledge is provided by Domain Ontologies, through their classification and organization of the concepts of the domain. The semantic annotations are stored inside the document with the use of the proper CAS structures, thus enriching the document with metadata.

Document Workflows

The analysis of documents is organized in *workflows* (user-defined pipelines of analysis tasks), which can be fully customized by the user. New analysis components can be added and others can be removed, according to the information that the user seeks to retrieve. At the end of each workflow, the documents are stored in the DW. The DW contains both documents and metadata based upon the extracted semantic contents of the documents.



Document Collection and Conversion

The Parmenides document lifecycle involves gathering the documents that are of interest and converting them to a format suitable for analysis and indexing.

The Parmenides architecture supports collection and conversion of various file types, including Portable Document Format (PDF), Word Processor (DOC) and web pages (HTML). With regard to web pages, a web crawler periodically collects web pages that interest the user. The pages are then converted to the Parmenides XML schema, the Common Annotation Scheme (CAS). Certain semantics of the document can be extracted, such as the date of the news story. In addition, irrelevant parts of the web page, i.e. headers and footers, are removed. In order to achieve this, a component that can be installed as an extension to the Mozilla Firefox browser has been built. With the use of this component, the user can customize the CAS converter. More specifically, the user can define which parts of a web page should be converted to CAS. XPath is used to uniquely identify each part of the page. To add page numbers to additional pages, copy the text box that contains the page number on this page, paste the text box on the additional pages, and then update the page number field.

Parmenides
School of Informatics
University of Manchester
Sackville Street
Manchester M60 1QD
United Kingdom

Phone:
+44 161 306-3309

Fax:
+44 161 306-3324

E-Mail:
babis.theodoulidis@manchester.ac.uk

Visit us at:
www.crim.co.umist.ac.uk/parmenides

Semantic Annotation

During the semantic annotation process, documents already converted in the CAS format get automatically conceptually annotated through consecutive steps of NLP and IE analysis. The current state of the implemented system incorporates four levels of analysis:

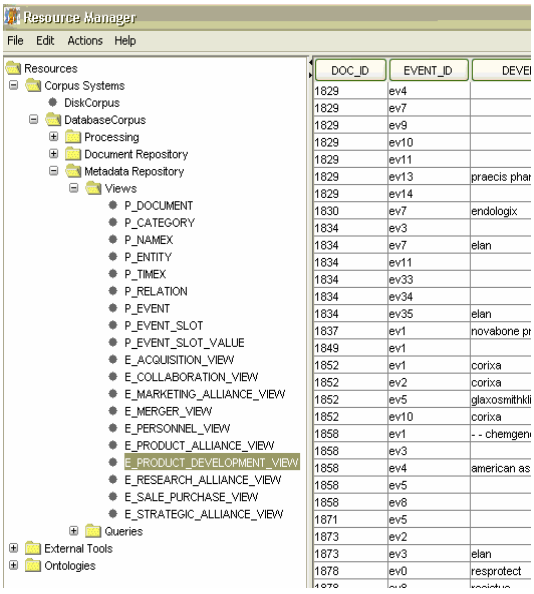
- *Tokenization* consists of breaking natural language text in atomic textual constituents (*tokens*), with an appropriate orthographic value.
- *Categorization* categorizes each document is with the use of the Support Vector Machines (SVM) categorization algorithm. SVM is first trained on positive and negative sample documents from a number of predefined categories. During the training process, a list of terms is created for each category, containing positive and negative weights. The document is assigned to the categories that produce a positive total weight.
- *POS Tagging and Sentence Splitting* assigns a morphosyntactic label (part-of-speech tag) to every token. In addition, tokens are grouped together in higher level syntactic structures: the *sentences*. Sentences are particularly valuable for later analysis steps, such as Named Entity Recognition, Event Extraction and Topic Discovery.
- *Ontological lookup* consists of the interaction with the domain-specific ontology. For every *phrase* (single or multi token textual span) of the input text, the respective ontological information is collected. In this way, textual spans are mapped to concepts or instances in the ontology and *all* information gets appended as additional *inline* annotation in the CAS representation of the input document.
- *Rule-based Information Extraction* uses human-authored linguistic rules to recognize *basic semantic elements* of interest (name instance expressions and respective entities, time expressions) *event instances* with appropriate participants (*slots*) and *relations* between various elements.

All the modules are grouped together under the **Cafetiere** Environment as described in and can be also run independently if included in appropriate workflows through the *Parmenides Resource Manager System*.

Document Warehousing

The semantic information gathered during the analysis of the documents is accommodated in the Parmenides Metadata Repository. The Parmenides Metadata Repository allows for warehousing and retrieval of the documents' semantics, as well as the documents themselves.

An interface for visually building queries is also provided by the Parmenides Resource Manager System. The user can use the tables of the repository to form queries in order to retrieve specific semantics. For example, a query could be targeted to the products that have been developed in the last 6 months and the developing company has participated



Parmenides
School of Informatics
University of Manchester
Sackville Street
Manchester M60 1QD
United Kingdom

Phone:
+44 161 306-3309

Fax:
+44 161 306-3324

E-Mail:
babis.theodoulidis@manchester.ac.uk

Visit us at:
www.crim.co.umist.ac.uk/parmenides

Ontology Management and Evolution

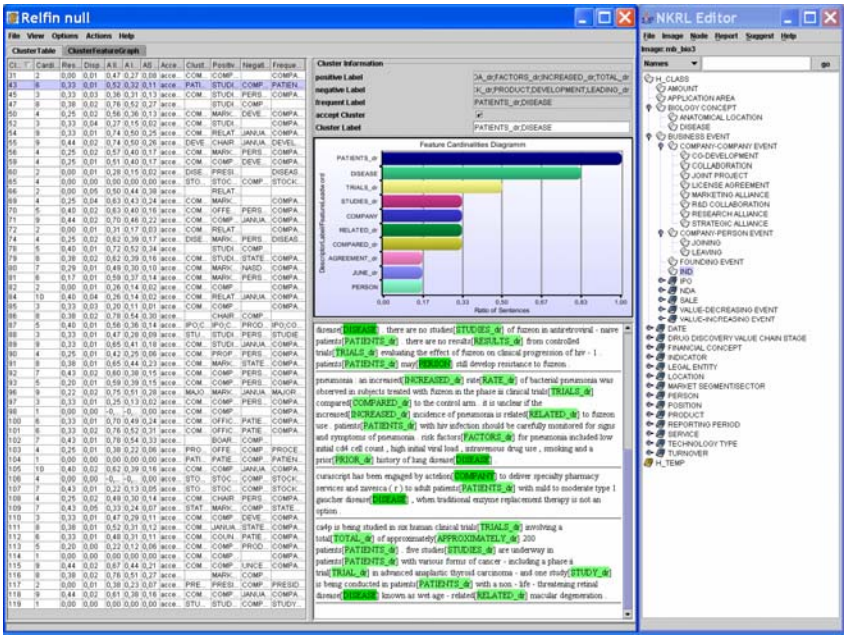
Ontologies and text annotations can be used for easy human access to huge or complicated text corpora. An ontology describes the relationships among concepts and maps them to their (textual) representations. An annotation is a semantic label that may or may not consist of terms in the ontology. The PARMENIDES approach provides an integrated environment for linguistic pre- processing, text annotation and ontology management and evolution.

The **Ontology Editor** is used for ontology creation and management. It allows the specification of multiple PS-NKRL ontologies, term extraction from XML or text documents and features an intelligent drag & drop functionality of ontology components. The PS-NKRL ontology format allows the specification of unconstrained object, event and slot types similar to OWL-Lite. An event can take an object instance as slot filler (but not vice versa) and it includes temporal semantics either through date primitives or with slot-types (Timex). The term extraction is an integrated moduleset within the ontology editor and it is based on the C/NC Algorithm that has been extended with a non-statistical extraction algorithm and it has been optimised for speed.

Furthermore, the ontology editor includes a complete Java API for access from arbitrary external Parmenides modules and it can export ontologies in HTML format, appropriate for reports and in DOM tree format.

RELFIN is used for the discovery of topics for ontology enhancement and for the assignment of labels to document regions. Both tasks require a lot of context knowledge which should be provided by a human expert or learned from the data. The RELFIN Learner and Annotator deliver functionality to distribute the work between the system and the human expert as follows:

- For *ontology enhancement*, additional terms are proposed as new concepts and groups of terms (concepts) are proposed for creating new links between them. Thus, the human expert is supported in creating added value by juxtaposing her background knowledge to corpus content for enhancing the ontology.
- For *text annotation*, labelled topic clusters are proposed and their labels are used for the annotation of the document regions in the cluster, while the expert can always supersede the systems selection of clusters and change the constructed labels.



Parmenides
School of Informatics
University of Manchester
Sackville Street
Manchester M60 1QD
United Kingdom

Phone:
+44 161 306-3309

Fax:
+44 161 306-3324

E-Mail:
babis.theodoulidis@manchester.ac.uk

Visit us at:
www.crim.co.umist.ac.uk/parmenides

Knowledge Discovery

One of the implications of the massive amounts of information is that a single person cannot possibly hope to be able to be aware, access and scan, let alone process and understand it, using traditional approaches. At the same time the notion that related information and hence knowledge lies 'hidden' in the literature and waiting to be discovered has become increasingly accepted in recent years. Knowledge discovery is a term used to describe this activity.

While the promise of knowledge discovery is appealing, in practice there are a number of difficulties. One of the most important difficulties is the 'analogical reasoning' step that is involved when matching the current problem with other 'similar ones' to which a solution exists. Yet another and more practical difficulty is that professionals hard pressed to solve their current problem now, have little time available to perform this 'analogy' step which often requires abstracting their problem to a more generic description on the basis of which similarity will be assessed.

The Parmenides approach to knowledge discovery is based on the use of analysis tools based on data mining that perform well understood functions for users in the context of problem solving. An example of these addresses the area of Press Release analysis where temporal reasoning allows the identification of trends in corporate agreements and collaborations.

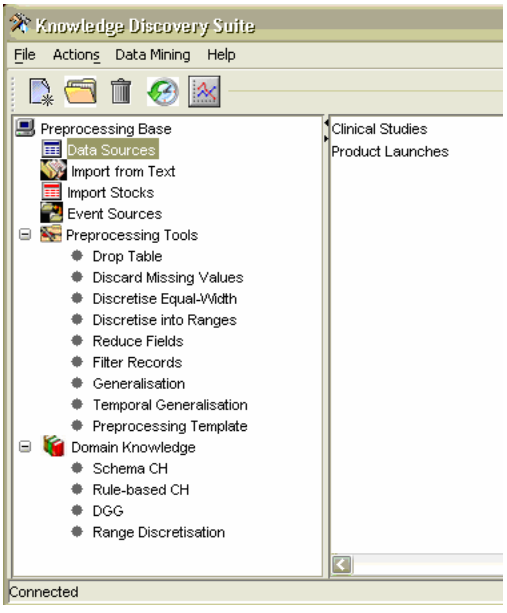
Knowledge discovery is an important area of research especially since it promises to enhance users' ability to identify useful information and solve problems in a more systematic manner than is currently possible. A number of research issues still need to be addressed but advances in a number of related disciplines have begun to yield interesting results.

Knowledge Discovery Suite (KDS)

The PARMENIDES Knowledge Discovery Suite discovers relationships in the form of sequential patterns between the objects extracted from the analyzed documents of the domain as well as information derived from external sources. As a result, users are provided with additional data sources, which can be used in combination with the extracted events to enhance the analysis.

Since sequences usually encapsulate the notion of time, they form a natural way of representing information in many domains. The daily prices of stocks during a year can be represented as sequences of values. Discovered patterns in the stock sequences could support the prediction of future stock market prices. The buying behavior of a customer can be also predicted through analysis of the sequences of the items bought.

For each data mining task using KDS, a mining session is defined. The concept of mining session allows different configurations of the KDS to be stored and reused. Each configuration includes the selection of event sources to be mined, the mining algorithm parameters, value and temporal generalisations and the set of produced patterns.



Parmenides
School of Informatics
University of Manchester
Sackville Street
Manchester M60 1QD
United Kingdom

Phone:
+44 161 306-3309

Fax:
+44 161 306-3324

E-Mail:
babis.theodoulidis@manchester.ac.uk

Visit us at:
www.crim.co.umist.ac.uk/parmenides

Parmenides of Elea
(born c. 515 B.C.E.)

This painting by Arnold Hermann is based on the Parmenides statue found at the excavation site at Velia in Italy, formerly Elea

Parmenides Project
School of Informatics
University of Manchester
Sackville Street
Manchester M60 1QD
United Kingdom

Phone:
+44 161 306-3309

Fax:
+44 161 306-3324

E-Mail:
babis.theodoulidis@manchester.ac.uk

Visit us at:
www.crim.co.umist.ac.uk/parmenides

Parmenides Concept Monitor (PCM)

PCM monitors pattern evolution across time and detects interesting changes in patterns. It addresses the identification and categorization of changes based on an interestingness model.
PCM considers knowledge discovery as a series of mining sessions where data is accumulated during consecutive time periods. Each session reveals a set of patterns, some of them may be known from previous sessions, while others may be new. It is also possible that patterns may disappear.
Previously known patterns are not updated as in incremental mining approaches but instead, each pattern becomes a unique object in the rule base; its statistics become instances describing the pattern at each time point of its existence.
For the discovery of interesting pattern changes, heuristics may be applied to identify interesting pattern changes, where the interestingness of an observed pattern change is assessed according to one or more of the following indicators:

- Absolute or relative changes of the value of one or more statistical properties being monitored
- frequency of occurrence of patterns, i.e., clusters
- changes to the statistical properties of a cluster that differ stronger from past values than expected

EnVisioner

EnVisioner is fully scalable data mining system and has the power to provide a new understanding of your data by rapidly seeking, analysing and understanding the patterns in a data set. EnVisioner includes extensive data transformation, mining and visualisation features. Users can design their customised data manipulation, mining and reporting processes. Moreover, these processes can be stored, modified, executed or scheduled for execution at will. All this is achieved without compromising scalability or performance.
EnVisioner supports a spectrum of data mining functions, including Relevance analysis, Generalization, Classification, Clustering, Association rules, Temporal rules, Spatial rules, Sequential patterns and Time-series Analysis. In particular, EnVisioner provides capabilities to the user to take advantage of the temporal dimensions that exist in his data sets. More precisely, EnVisioner allows the user to:

- Extract association rules and monitor the evolution of their strength in time.
- Mine similarities in time series
- Discover Sequential Patterns
- Define generalizations for temporal attributes