Abstract Number: 313

Knowledge Extraction from Mixed-Precision Information

Charles Parker[1], Nenad Krdzavac[2], Chuong Vo Phan[1], Kevin Petersen[2] and George M. Garrity[1,2]

[1]NamesforLife, LLC and [2]Michigan State University (East Lansing, Michigan)

Challenges of Human-Machine Communication

A fundamental barrier to effective human-machine communication is the lack of a shared, unambiguous language that is understandable to humans and precise enough for machine reasoning.

The knowledge of domain experts is aggregated from a variety of information sources, ranging from raw text or data to structured and normalized databases (Mixed Precision Information; MPI).

Contemporary knowledge extraction typically relies on interpretation (coding) of MPI by domain experts into a form compatible with a specific information system.

Human coding of information has significant drawbacks:

- Information loss due to system-specific coding standards.
- Inconsistent interpretation of MPI.
- Transitive closure of misinformation, which results in false equivalencies.
- Queries that return false equivalencies that are interpreted as knowledge.
- Incorrect knowledge that is carried forward to other information systems.

Autonomous systems must accommodate false knowledge while reasoning.

A Novel Approach to Extracting Knowledge

We introduce a novel standards-based method for extracting knowledge from MPI to provide knowledge workers and machine reasoners with verifiable interpretations of observational data.

Our approach combines semantic and semiotic methods to:

- represent information at multiple levels in concept hierarchies
- "slice" and aggregate concepts to represent information consistently for ambiguous human language *and* reasoners
- provide multiple entry points for information (term, concept, data)
- provide attachment points for reasoning over rules and axioms
- accommodate multiple interpretations of information

Our adherence to Knowledge Organization standards (SKOS-XL, OWL, RDF/S) provides integration points with existing triple stores, ontologies and thesauri.

We have demonstrated the effectiveness of this approach over a large corpus of scientific literature.

We are developing a standards-compliant semantic knowledge model to support abstract reasoning and predictive modeling.

Military Applications

Our company has developed a novel approach to knowledge modeling that bridges the gap between human and machine understanding of abstract concepts.

This technology enables abstract reasoning and sharing of Mixed-Precision Information that can benefit *Human/Autonomous System Interaction and Collaboration* (HASIC) through bi-directional human-agent interaction.

Benefits of Our Approach

- Common understanding by both human and machine.
- Compliant with existing ontology and thesaurus standards.
- Preserves precision of the original data.
- Utilizes data at appropriate levels of abstraction.
- Accommodates conflicting and ambiguous information.
- Supports defeasible inference of relations among concepts.
- Provides exact explanations for reasoning.
- Integrates data from multiple sources.

A Pilot Project for Knowledge Extraction

Our pilot project for Knowledge Extraction is based on previous work in which we assembled the historical literature (1800s-2016) for all named species and higher taxonomic groups of bacteria. This literature contains a wealth of observational data, but underlying information was not in a form that lent itself to integration with databases or ontologies. Predictive models require high quality input data, but not all data are of similar quality or suitable for computation without extensive cleaning, interpretation and normalization.

Bacteria are generally described using unstructured text that includes non-uniform descriptors. Descriptions are scattered throughout specialized databases and scientific, technical and medical literature. Integrating this data from such resources requires expertise in harvesting, modeling, interpreting, and validating these data.

Many properties of microbes (the phenotypes) consist of a combination of quantitative environmental conditions and qualitative growth and metabolic capabilities. Such terms are challenging to incorporate into query systems because of context-based interpretations, imprecision and conceptual overlap across multiple feature domains.

A Hybrid Knowledge Framework

We developed software to extract putative feature domain vocabularies from this corpus, resulting in the discovery of over 40,000 candidate terms used in descriptions of the 13,213 species of *Bacteria* and *Archaea* (N4L Database, March 1, 2016). We have subsequently developed this vocabulary into a thesaurus of phenotypic terms, which will ultimately conform to W3C SKOS-XL semantics and provide a link between the language of microbiology, the semantic web and our existing tools and services.

The thesaurus was developed in parallel with a formal ontology, which supports inference from observations of an organism under a set of environmental constraints. The ontology uses meta-modeling techniques to implement rule and constraint templates using these complex terms.

While developing these resources, we discovered a novel method for establishing semantic equivalence between concepts. This enables precise, consistent, verifiable reasoning over imprecise terms at multiple levels of abstraction [1].

Our thesaurus model complements ontology development by deconstructing ambiguous language into precise concepts for reasoners.

Knowledge Discovery

Initially, we developed a curated database containing descriptive information about all properly named bacteria and archaea and the supporting scientific literature (>50,000 sources) from which a candidate vocabulary of phenotypic terms was developed. Using these resources and some novel software methods, we could code raw text into assertions of microbial characteristics based on our ontology and thesaurus. These assertions are then supplied to a machine reasoner which can infer likely properties and establish the identity of the source organism. The reasoner interprets these assertions at appropriate levels of abstraction, correctly answers queries and produces new knowledge.

The reasoner has its own curation account, which it uses to supplement and direct the activities of human curators.

The reasoner examines all available assertions and uses the underlying ontology, axioms and rules to identify gaps and inconsistencies in information and to infer new knowledge about an organism.

Capabilities & Commercial Offerings

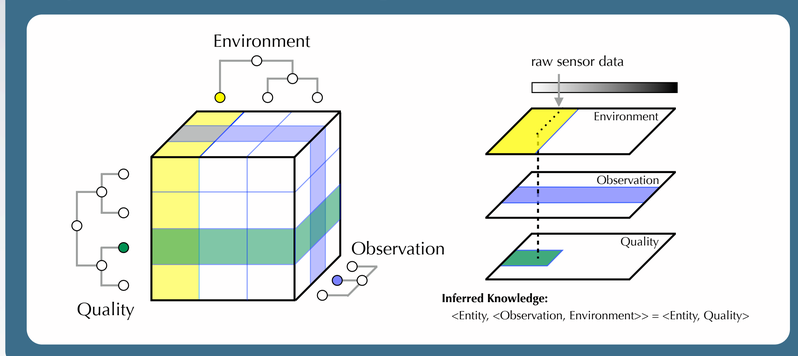
We have already developed a number of novel software components that overcome specific technical barriers in terminology management, text mining, Information Extraction, knowledge transformation, named-entity recognition, document classification and annotation.

Each component is implemented using W3C standards and recommendations (RDF/S, OWL2, SKOS-XL, XML, XSL, XSD, SPIN, SPARQL, OWL-RL, DOI, CORS) and commercially-compatible FOS frameworks (Java, Apache, PostgreSQL, Virtuoso, Jena/ARQ, SPIN Reasoner). We are integrating these components into a single software suite that supports a variety of Machine Learning and reasoning needs.

In the text-mining field, we have demonstrated our ability to reverse-engineer the diagnostic phrases that human indexers use to classify large corpora of technical documents, and to measure both the quality of previously-analyzed documents and the cohesion of individual document classes. Our software provides a novel way to navigate and bridge multiple classification systems.

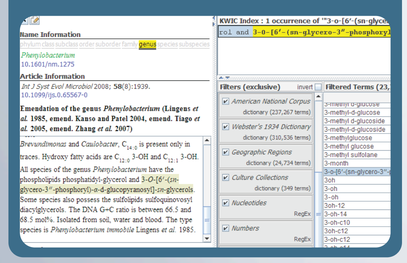
Encode Knowledge for Human and Machine Understanding

Our semantic equivalence method integrates observational data from multiple sources (e.g., sensor data, textual descriptions) at various levels of abstraction, resolves ambiguity and detects conflicting observations prior to resolving to labeled ontology concept identifiers suitable for reasoning.



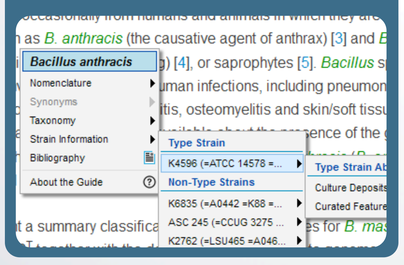
Mine for Novel Vocabularies

Discover complex entities in content and generate new vocabularies or resolve them to concepts in existing thesauri.



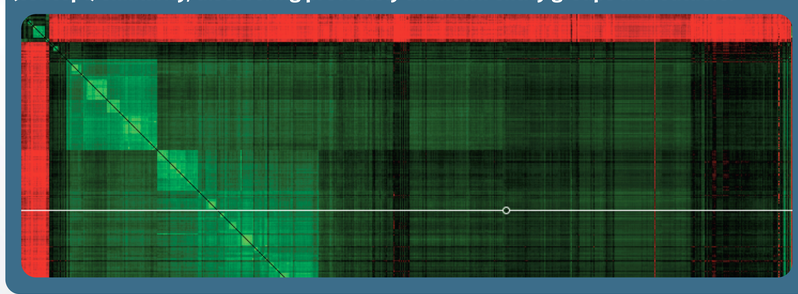
Enrich Documents

Annotate documents with authoritative information and embed actionable semantic links to additional resources.



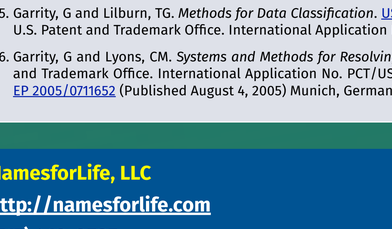
Discover and Label New Concepts

Our self-organizing heatmap clusters entities based on semantic (term) or semiotic (concept) similarity, discovering previously unknown entity groups and outliers.



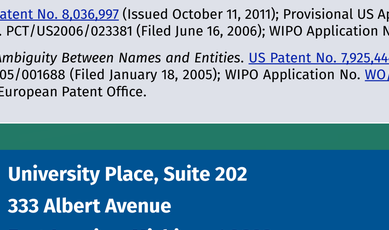
Identify Diagnostic Concepts

Extract and chart diagnostic concepts for arbitrary groups of resources.



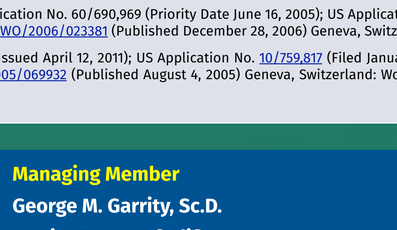
Measure Resource Similarity

A semantic similarity score provides a consistent measure for comparing resources over shared concepts.



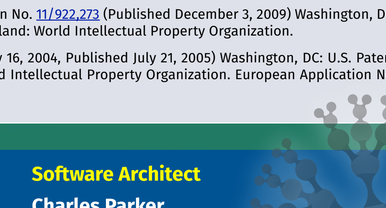
Multiple Classification with Explanations

Reverse-engineer human analyses, discover errors and perform multiple-classification automatically.



Visualize Group Quality

Multidimensional analysis of shared concepts within a group provides a means to visualize group cohesion.



Company Information

NamesforLife, LLC is an privately held company based in East Lansing, Michigan. It was founded in 2004 to create commercial applications of a technology developed at Michigan State University. The company provides taxonomic and analytical services, data, software and technology licensing for the publishing industry, life sciences research, commercial partners and Federal laboratories.

Funding for research and development was provided through the DOE SBIR/STTR program (DE-SC0006191). Funding for NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, the Michigan Economic Development Corporation, and the Michigan Universities Commercialization Initiative. Additional support for developing the Semantic Analysis Platform was provided through a commercial partnership with Fairview Research/IFI Claims.

Patent Portfolio

1. Parker, CT, Garrity, GM and Krdzavac, NB. *Systems and Methods for Establishing Semantic Equivalence Between Concepts*. Provisional US Application No. 61/880,244 (Priority Date September 20, 2013); US Application No. 15/022,870 (Published August 4, 2016) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2014/056808 (Filed September 20, 2014); WIPO Application Number WO/2015/042536 (Published March 26, 2015) Geneva, Switzerland: World Intellectual Property Organization. European Application No. EP 2014/084555 (Published July 27, 2016) Munich, Germany: European Patent Office.

2. Sayood, K, Way, S, Ozkan, UN and Garrity, GM. *Classification of Nucleotide Sequences by Latent Semantic Analysis*. Provisional US Application No. 61/677,316 (Priority Date July 30, 2012); US Application No. 13/954,925 (Published May 1, 2014) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2013/052797 (Filed July 30, 2013); WIPO Application No. WO/2014/022441 (Published February 6, 2014) Geneva, Switzerland: World Intellectual Property Organization.

3. Parker, CT and Garrity, GM. *Semiotic Indexing of Digital Resources*. US Patent No. 8,903,825 (Issued December 2, 2014); Provisional US Application No. 61/489,362 (Priority Date May 24, 2011); US Application No. 13/478,973 (Published January 10, 2013); Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2012/039168 (Filed May 23, 2012); WIPO Application No. WO/2014/022441 (Published November 29, 2012) Geneva, Switzerland: World Intellectual Property Organization. European Application No. EP 2012/0790213 (Published April 9, 2014) Munich, Germany: European Patent Office.

4. Parker, CT, Lyons, CM, Roston, GP and Garrity, GM. *Systems and Methods for Automatically Identifying and Linking Names in Digital Resources*. Provisional US Application Nos. 61/143,986 (Priority Date January 12, 2009) and 61/184,187 (Priority Date June 4, 2009); US Application No. 12/685,964 (Published August 5, 2010) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2010/020734 (Filed January 12, 2010); WIPO Application No. WO/2010/081133 (Published July 15, 2010) Geneva, Switzerland: World Intellectual Property Organization. European Application No. EP 2010/0729654 (Published November 16, 2011) Munich, Germany: European Patent Office.

5. Garrity, G and Lilburn, TG. *Methods for Data Classification*. US Patent No. 8,036,997 (Issued October 11, 2011); Provisional US Application No. 60/690,969 (Priority Date June 16, 2005); US Application No. 11/922,273 (Published December 3, 2009) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2006/023381 (Filed June 16, 2006); WIPO Application No. WO/2006/023381 (Published December 28, 2006) Geneva, Switzerland: World Intellectual Property Organization.

6. Garrity, G and Lyons, CM. *Systems and Methods for Resolving Ambiguity Between Names and Entities*. US Patent No. 7,925,444 (Issued April 12, 2011); US Application No. 10/759,817 (Filed January 16, 2004, Published July 21, 2005) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2005/001688 (Filed January 18, 2005); WIPO Application No. WO/2005/069932 (Published August 4, 2005) Geneva, Switzerland: World Intellectual Property Organization. European Application No. EP 2005/0711652 (Published August 4, 2005) Munich, Germany: European Patent Office.

NamesforLife, LLC
http://namesforlife.com
(517) 410-0525

University Place, Suite 202
333 Albert Avenue
East Lansing, Michigan 48823

Managing Member
George M. Garrity, Sc.D.
garrity@namesforlife.com

Software Architect
Charles Parker
https://www.linkedin.com/ctparker