

2010

## Continuous Semantics to Analyze Real-Time Data

Amit P. Sheth

Wright State University - Main Campus, amit@sc.edu

Christopher Thomas

Wright State University - Main Campus

Pankaj Mehra

Follow this and additional works at: <https://corescholar.libraries.wright.edu/knoesis>



Part of the [Bioinformatics Commons](#), [Communication Technology and New Media Commons](#), [Databases and Information Systems Commons](#), [OS and Networks Commons](#), and the [Science and Technology Studies Commons](#)

---

### Repository Citation

Sheth, A. P., Thomas, C., & Mehra, P. (2010). Continuous Semantics to Analyze Real-Time Data. *IEEE Internet Computing*, 14 (6), 84-89.

<https://corescholar.libraries.wright.edu/knoesis/776>

This Article is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact [library-corescholar@wright.edu](mailto:library-corescholar@wright.edu).



# Continuous Semantics to Analyze Real-Time Data

Amit Sheth and Christopher Thomas • Wright State University  
Pankaj Mehra • Inlogy

Order, unity, and continuity are human inventions, just as truly as catalogues and encyclopedias. — Bertrand Russell

**W**e've made significant progress in applying semantics and Semantic Web technologies in a range of domains. A relatively well-understood approach to reaping semantics' benefits begins with formal modeling of a domain's concepts and relationships, typically as an ontology. Then, we extract relevant facts – in the form of related entities – from the corpus of background knowledge and use them to populate the ontology. Finally, we apply the ontology to extract semantic metadata or to semantically annotate data in unseen or new corpora.

Using annotations yields semantics-enhanced experiences for search, browsing, integration, personalization, advertising, analysis, discovery, situational awareness, and so on.<sup>1</sup> This typically works well for domains that involve slowly evolving knowledge concentrated among deeply specialized domain experts and that have definable boundaries. A good example is the US National Center for Biomedical Ontologies, which has approximately 200 ontologies used for annotations, improved search, reasoning, and knowledge discovery. Concurrently, major search engines are developing and using large collections of domain-relevant entities as background knowledge, to support semantic or facet search.

However, this approach has difficulties dealing with dynamic domains involved in social, mobile, and sensor webs. Here, we look at how *continuous semantics* can help us model those domains and analyze the related real-time data.

### The Challenge of Modeling Dynamic Domains

Increasingly popular social, mobile, and sensor webs exhibit five characteristics. First, they're

spontaneous (arising suddenly). Second, they follow a period of rapid evolution, involving real-time or near real-time data, which requires continuous searching and analysis. Third, they involve many distributed participants with fragmented and opinionated information. Fourth, they accommodate diverse viewpoints involving topical or contentious subjects. Finally, they feature context colored by local knowledge as well as perceptions based on different observations and their sociocultural analysis.

### Minimizing the Need for Commitment

The formal modeling of ontologies for such evolving domains or events is infeasible for two reasons. First, we don't have many starting points (existing ontologies). Second, a diverse set of users or participants will have difficulty committing to the shared worldview we're attempting to model. Modeling a contentious topic might lead to rejection of the ontology or failure to achieve common conceptualization. On one hand, users often agree on a domain's concepts and entities, such as the lawmakers involved in drafting a bill, the bill's topic, an earthquake's spatial location, and key dates. On the other hand, users often contest the interpretation of how these entities are related, even taxonomically.

So, models that require less commitment are preferable. Models that capture changing conceptualizations and relevant knowledge offer continuous semantics to improve understanding and analysis of dynamic, event-centric activities and situations.

To build domain models for these situations, we must pull background knowledge from trusted, uncontroversial sources. Wikipedia, for instance, has shown that it is possible to col-

laboratively create factual descriptions of entities and events even for contentious topics such as abortion. Wikipedia articles show information agreed upon by most contributors. Separate discussion pages show how the contributors resolved disagreements to arrive at a factual, unbiased description. Such wide agreement combined with a category structure and link graph makes Wikipedia an attractive candidate for knowledge extraction. That is, we can harvest the wisdom of the crowds, or collective intelligence, to build a folksonomy – an informal domain model.

### Anticipating What We'll Want to Know

Traditional conceptual modeling is also inadequate for dynamic domains owing to their topicality. News, blogs, and microblog posts deliver descriptions of events in nearly real time. Twitter, for example, delivers information as short “tweets” about events as they unfold. Only a model with social media as its knowledge source will be up-to-date when modeling events that are unfolding in a similar medium. A domain model that doesn't significantly lag behind the actual events is crucial for accurate classification, which will result in maximum information gain.

The past few years have seen explosive growth in services offering up-to-date and, in many cases, real-time data. Leading the way is Twitter and a variety of social media services (see <http://gnip.com/sources>), followed by blogs and traditional news media. We want to be the first to know about change – ideally, before it happens, or at least shortly after. The paradigm for information retrieval is thus, “What will you want to know tomorrow?”

A recent paper showed success in predicting German election results using tweets.<sup>2</sup> However, there is more to elections than just the results. An event or situation can be

multifaceted and can be spatially, temporally, and thematically sliced and analyzed. For example, you could time-slice the 2009 Iranian election discussion on Twitter into events surrounding election campaign rallies and protests (starting 12 June), Mahmoud Ahmadinejad's victory speech (14 June), the decision to recount (16 June), Ayatollah Khamenei's endorsement of Ahmadinejad's win (19 June), Neda's brutal killing (22 June), and so on.

An approach to Web document search that can leverage billions of documents to deliver useful patterns<sup>3</sup> probably won't be very useful here. Our challenge involves extracting signals from thousands of tweets or posts (that is, a small corpus) containing informal text.<sup>4</sup> Furthermore, the discussion focus will often shift frequently, with new knowledge or facts generated along with the events. For example, regarding a natural disaster, the focus could shift from rescue to recovery. So, we're intrigued by the possibility of dynamic model extraction that can be tied to a situation's context and can keep up with context shifts (for example, response and rescue to recovery and, later, rehabilitation). We would like to use such an extracted model to organize (search, integrate, analyze, or even reason about) data relating to real-time discourse or relating to dynamic, event-centric activities and situations.

Traditional classification approaches based on corpus learning or user input can only react to domain changes. More recently, however, we find that social-knowledge aggregation sites such as Wikipedia quickly contain descriptions of events, emergent situations, and new concepts. For example, for some recent events such as US Representative Joe Wilson's “You lie!” outburst, the Mumbai terrorist attack, and the Haiti earthquake, anchor pages with significant details were available in less than an

hour to less than a day. Furthermore, these pages continued to evolve as the event or situation unfolded.

Technology lets us create snapshots of this evolution. So, if automatic techniques can tap such social knowledge to create a model, we can gain the ability to better understand the more unruly informal text that largely constitutes real-time data.

### Continuous Semantics

Previously, we outlined our vision of a comprehensive strategy for knowledge accumulation, using the notion of a *circle of knowledge life* (see Figure 1).<sup>5</sup> In this vision, continuous semantics is supported by knowledge that's dynamic and updated through automated techniques and user interaction with the knowledge. The classification and annotation of streaming data and users' choices regarding certain feeds or data items help update knowledge about the domain for which the users are requesting information.

### Wikipedia as an Underlying Corpus

Wikipedia, barring its news component, is an up-to-date collection of encyclopedic knowledge. When a page is updated because new information is available, the new information is integrated rather than simply added, as is usually the case with news streams.

How Wikipedia handles rapid coverage of new events makes it a good option for a knowledge repository from which to create models. Because Wikipedia is authored by humans for humans, its structure is intuitive and to some degree resembles a formal ontology's class hierarchy, even though many subcategory relationships in Wikipedia are associative rather than strict subclass or type relationships. For example, categories that contain the astronomer Carl Sagan are Cornell University faculty, cosmologists, search for extraterrestrial intelligence (SETI),

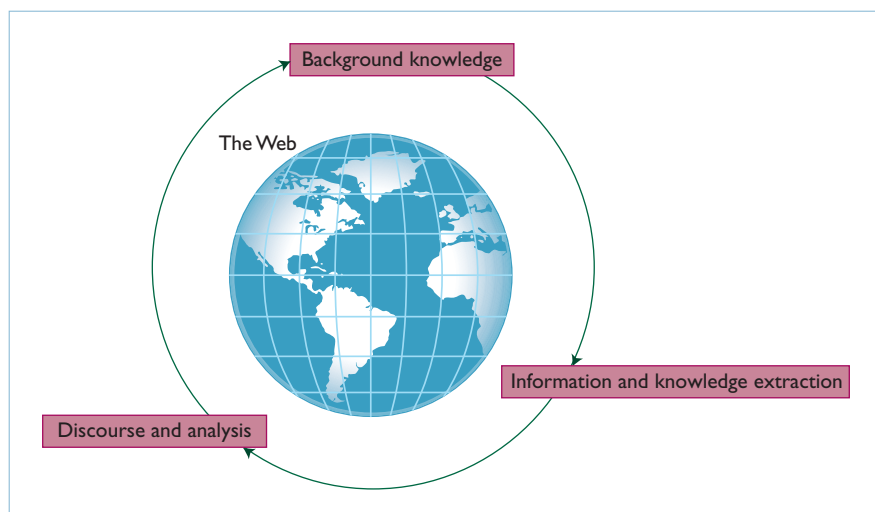


Figure 1. The circle of knowledge life on the Web to support continuous semantics. There is interdependence between the knowledge embedded in the content created by humans and through social processes. This knowledge can more easily be extracted by having algorithms focus on a domain and use known facts (background knowledge). The extracted knowledge can then be used to analyze new content. Being able to realize this cycle on a continuous, largely automated basis supports continuous semantics of real-time data.

American agnostics, and astrophysics. If we view this as a formal classification task, many of these categories are wrong. Carl Sagan wasn't literally SETI, no matter how involved he was in the movement. But he was a key figure in the search for extra-terrestrial life, so we don't object to this categorization in Wikipedia. A Wikipedia category list links to articles important to the category's topic, no matter whether an article's subject stands in a formal subclass or type relationship with that topic. Also, because articles describe particulars as well as generals, mapping categories and articles to classes and instances in a formally correct way is not straightforward.

So, we refrain from calling our resulting domain model an ontology. Ontologies used for reasoning, database integration, and so on must be logically consistent, well restricted, and highly connected to be of any use. In contrast, domain models for information retrieval and real-time data enhancement need only be comprehensive, focused, and up-to-date.

Simone Ponzetto and Michael

Strube described the creation of a more rigid taxonomic structure from the Wikipedia hierarchy.<sup>6</sup> They scrutinized Wikipedia's structure according to linguistic patterns indicating proper subclass and type relationships. Their intent thus complements ours. It carves out parts of Wikipedia that are formally more rigorous, whereas we use the knowledge created by a community to carve out the part that meets the user's current needs. In both cases, chipping away undesirable relations between entities is more reliable and more accurate than predicting new ones.

The Doozer project uses our approach to create focused models of evolving and fluctuating domains.<sup>7</sup> One of its key features is domain hierarchy creation.

### Dynamic Model Creation

An application that creates models on demand must have a significantly small runtime. Only a model that's created in seconds will be useful for semantic searching, browsing, or analysis of real-time content.

Here we briefly describe the steps in getting from a set of pertinent seed concepts to a comprehensive hierarchy that clearly focuses on the users' domain of interest. We employ an "expand and reduce" process that first allows exploration and exploitation of the concept space before reducing it to the concepts matching the domain of interest.

We look at a domain of interest from two levels:

- The *focus domain* is the actual point of interest – for example, Web 2.0 or cancer.
- The *broader focus domain* indicates the set of concepts immediately related to the focus domain and necessary to properly understand it – for example, social networking, Internet, and oncology concepts.

The expansion phase aims to maximize concept recall related to the domain of interest. It involves two steps. Step one is full text search – exploiting the knowledge space. First, we use a few words describing the focus domain to query the full text of Wikipedia. This produces the set of top-ranked articles.

Step two is link-based expansion – exploring the knowledge space. This step expands the set of top-ranked articles to a larger set of articles by including articles that appear closely related. It does this on the assumption that the more neighboring (linked) nodes two nodes in a Wikipedia article graph share, the more closely related those two nodes are.

The expanded set of concept terms (article titles) serves as input for the reduction phase (conditional pruning). For each term, we compute conditional probabilities describing its importance both for the domain  $p(\text{Term}|\text{Domain})$  and in the domain  $p(\text{Domain}|\text{Term})$ . We delete terms with a probability less than a given threshold. This probability is crucial

for the subsequent use of the created domain model during probabilistic document classification.

Finally, we impose a category hierarchy on the extracted concepts that is based on the Wikipedia categories.

### Using Dynamic Domain Models for Semantic Analysis of Real-Time Data

Here we show how we apply our approach, using Twitter and Twitris (<http://twitris.knoesis.org>), a system for spatio-temporal-thematic analysis that extracts social signals from tweets related to events and emergent situations.<sup>4</sup>

Figure 2 illustrates a continuous process of semantically analyzing real-time data using a dynamic model created by a system such as Doozer. This process starts with Twitter feeds related to a specific event – in this case, the Iranian election (see Figure 2a). The Twitris data collection component automatically identifies a collection of hash tags and keywords associated with that event and filters relevant tweets using the Twitter API (see Figure 2a). Thematic analysis by Twitris gives a set of  $n$ -grams or key phrases exemplified by the tag cloud in Figure 2b. Doozer uses key phrases to automatically and dynamically create a model from Wikipedia and other qualified sources such as Freebase (see Figure 2c). Twitris uses the domain model to semantically annotate and support semantic analysis of the original tweets (as in Figure 2a) and subsequent tweets (see Figure 2d). It does this by restricting Twarq<sup>8</sup> annotations of streaming data to the domain spanned by the model. Twitris can then identify new keywords and hash tags to expand or can modify semantic processing as the event evolves. This in turn leads to new key phrases for dynamic model extraction or updating.

However, by this time the underlying Wikipedia pages or other qual-

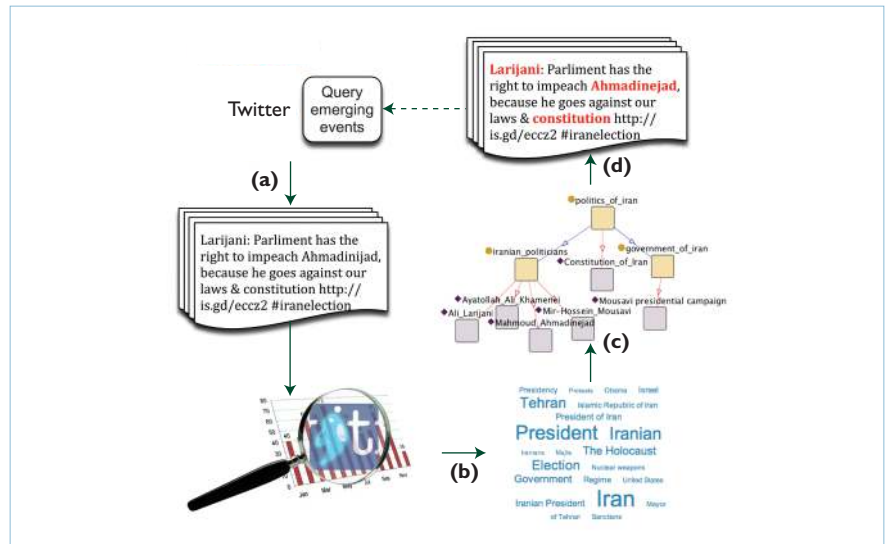


Figure 2. A pipeline for real-time data analysis using continuous semantics. (a) Real-time data can be queried or filtered to find event-specific content. (b) A system such as Twitris can analyze such content to extract social signals. (c) Domain models and background knowledge that can be dynamically and contextually created can allow more semantic analysis by identifying meaningful concepts. (d) Identification of new meaningful concepts can lead to continued processing of new real-time content using new concepts for further querying or filtering.

ified social knowledge sources might have been updated. This updating will yield new concepts in an evolved domain model that reflects the real-world changes being analyzed. Also, Twitris's thematic-analysis component can consider as new input the entities that are annotated using the Doozer output hierarchy. This creates a feedback loop between content analysis and model evolution.

Figures 3 and 4 show parts of Doozer-created models and how they can support semantic analysis. Figure 3 shows tweets mentioning locations in Iran and their mapping to locations in the model to allow for analysis of thematic elements with reference to different regions. Figure 4 shows a subgraph of the model representing Iranian politics and the mapping of entities to words and phrases in tweets (that is, semantic annotation of tweets).

Such semantic processing of real-time (textual) data shares the

technological underpinnings of the Semantic Sensor Web.<sup>9</sup> Combining the two easily leads to integrated semantic analysis of multimodal data streams. On-demand creation of semantic models from social knowledge sources such as Wikipedia offers exciting new capabilities in making real-time social and sensor data more meaningful and useful for advanced situational-awareness and situational-analysis applications. □

#### Acknowledgments

We acknowledge Meena Nagarajan's input and partial support from US National Science Foundation Award IIS-0842129 and a Hewlett-Packard Innovation Grant.

#### References

1. A. Sheth, D. Avant, and C. Bertram, *System and Method for Creating a Semantic Web and Its Applications in Browsing, Searching, Profiling, Personalization and Advertising*, US patent 6,311,194, to Taalee Inc., Patent and Trademark Office, 2001.
2. A. Tumasjan et al., "Predicting Elections with Twitter: What 140 Characters Reveal

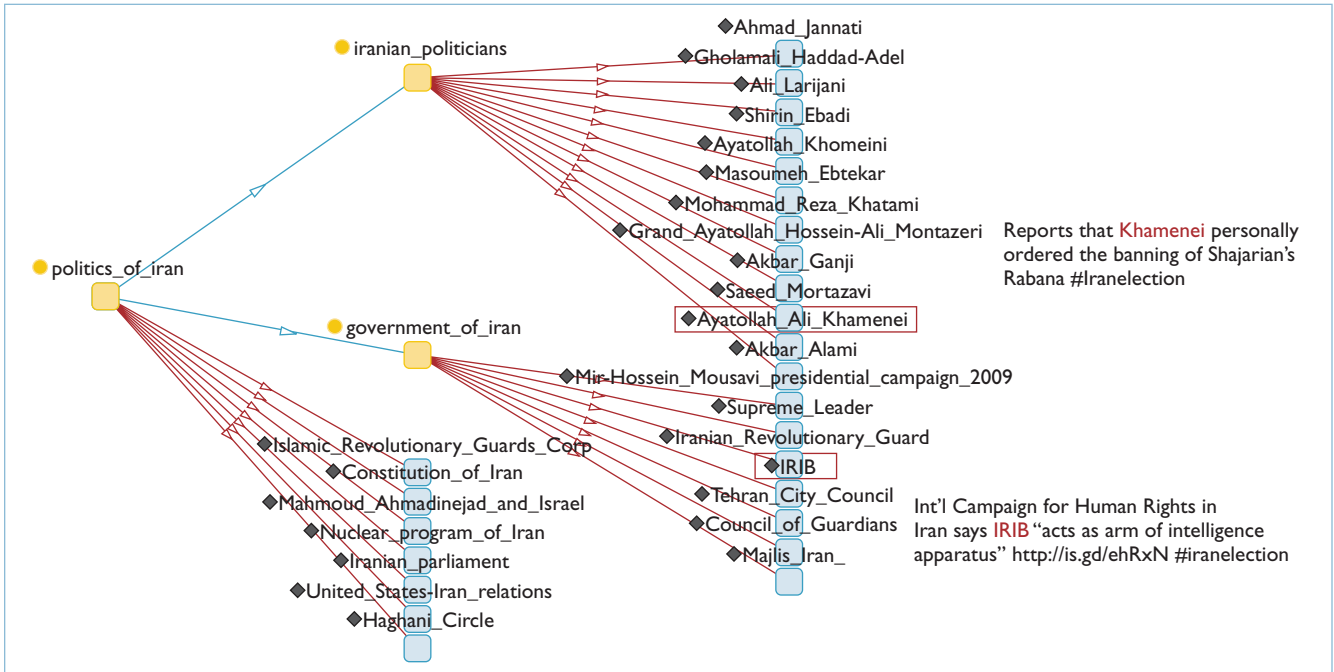


Figure 3. An excerpt of a model extracted from Wikipedia using Doozer to allow comprehension of relationships between locations mentioned in tweets. The concepts in this model can be used to annotate tweets or any other real-time textual content, and inherent relationships (for example, a town is in a region) can enable domain-specific semantics (in this example, spatial and geopolitical analysis).

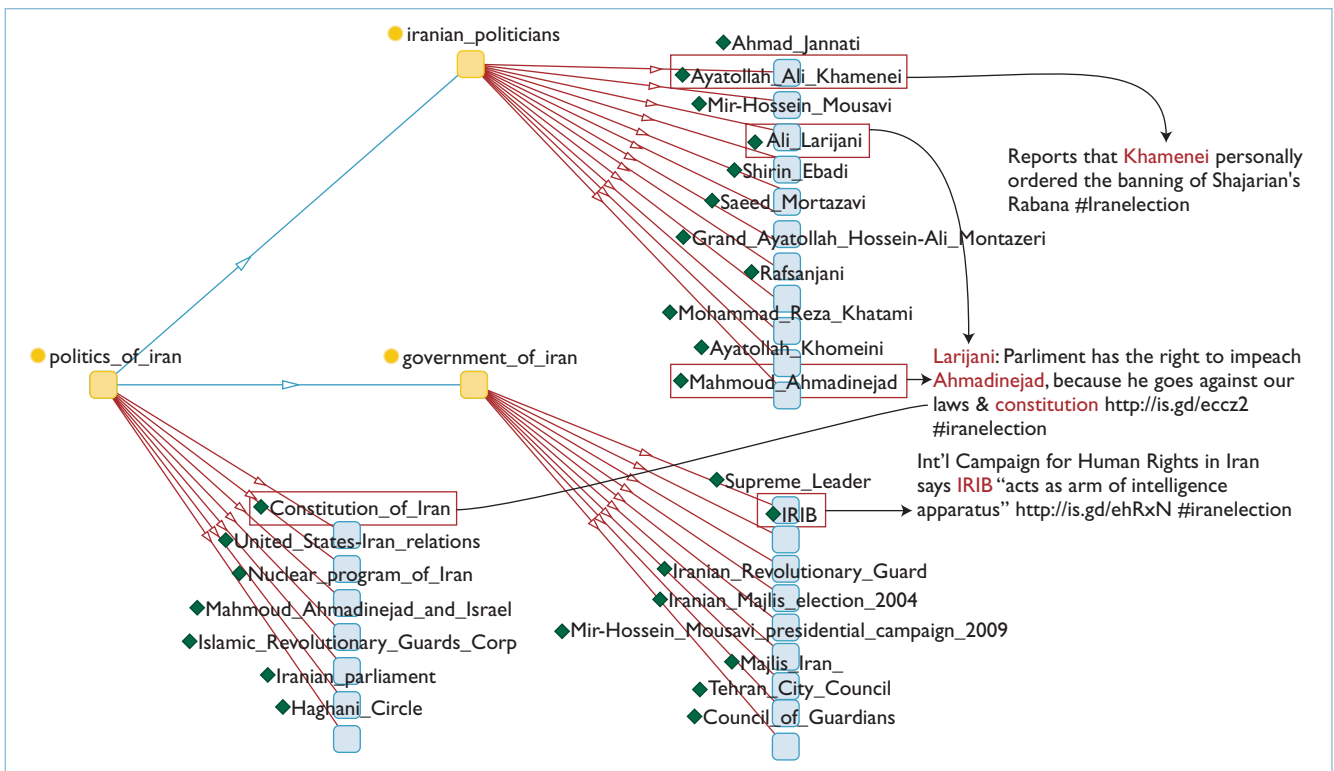


Figure 4. An excerpt of the model mentioned in Figure 3, focusing on the Iranian government and politicians. This example deals with the domain of political structure, including institutional and government aspects.

about Political Sentiment,” *Proc. 4th Int’l AAAI Conf. Weblogs and Social Media*, AAAI Press, 2010, pp. 178–185.

3. A. Halevy, P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data,” *IEEE Intelligent Systems*, vol. 24, no. 2, 2009, pp. 8–12.
4. M. Nagarajan et al., “Spatio-Temporal-Thematic Analysis of Citizen-Sensor Data – Challenges and Experiences,” *Web Information Systems Eng. – WISE 2009*, LNCS 5802, Springer, 2009, pp. 539–553.
5. C.J. Thomas et al., “What Goes Around Comes Around – Improving Linked Open Data through On-Demand Model Creation,” *Proc. WebSci10: Extending the Frontiers of Society On-Line*, Web Science Trust, 2010; <http://journal.webscience.org/397>.
6. S. Ponzetto and M. Strube, “Deriving a Large Scale Taxonomy from Wikipedia,” *Proc. 22nd Nat’l Conf. Artificial Intelligence (AAAI 07)*, AAAI Press, 2007, pp.

1440–1445.


7. C.J. Thomas et al., “Growing Fields of Interest – Using an Expand and Reduce Strategy for Domain Model Extraction,” *Proc. IEEE/WIC/ACM Intl Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT 08)*, IEEE Press, 2008, pp. 496–502.
8. P. Mendes, P. Kapanipathi, and A. Pas-sant, “Twarql: Tapping into the Wisdom of the Crowd,” *Proc. 6th Int’l Conf. Semantic Systems (I-Semantics 10)*, ACM Press, 2010, pp. 1–3; <http://doi.acm.org/10.1145/1839707.1839762>.
9. A. Sheth, C. Henson and S. Sahoo, “Semantic Sensor Web,” *IEEE Internet Computing*, vol. 12, no. 4, 2008, pp. 78–83.

Amit Sheth is the director of the Ohio Center of Excellence on Knowledge-Enabled Computing (Kno.e.sis) at Wright State

University. He’s also the university’s LexisNexis Ohio Eminent Scholar. Contact him via <http://knoesis.org/amit>.

Christopher Thomas is completing his PhD dissertation on dynamic model creation at the Ohio Center of Excellence on Knowledge-Enabled Computing (Kno.e.sis) at Wright State University. Contact him at [topher@knoesis.org](mailto:topher@knoesis.org).

Pankaj Mehra is the CTO and a cofounder of Inlogy. He coordinates industry participation in Stanford’s Collaborative Data Management Initiative and serves on the *IEEE Internet Computing* editorial board. Contact him at [pankaj.mehra@ieee.org](mailto:pankaj.mehra@ieee.org).

 Selected CS articles and columns are also available for free at <http://>



The magazine of computational tools and methods.

MEMBERS \$49 | STUDENTS \$25  
[www.computer.org/cise](http://www.computer.org/cise) | <http://cise.aip.org>

CiSE addresses large computational problems by sharing

- » efficient algorithms
- » system software
- » computer architecture