Collective Knowledge Authoring

James Starz, Alden Roberts, and Brian Kettler

Lockheed Martin Advanced Technology Laboratories, 4301 N. Fairfax Drive, Suite 500
Arlington, VA 22203, USA
{james.c.starz, alden.b.roberts, brian.p.kettler}@lmco.com

Abstract. Most information systems support either a rigid schema enforced by software or a loose schema enforced by a select number of users. This paper investigates having the system enforce the use of a looser schema. Supporting this capability entails using multiple techniques for guiding users towards common semantics when authoring information.

Keywords: Semantic Web, Collaboration, Authoring

1 Introduction

Despite the vast amount of information available, today’s software systems are heavily dependent on information that is authored, aggregated, and organized by humans. This is seen by the massive number of documents (Word™, PowerPoint™, web pages) and hand authored records that reside in corporate intranets and on the Internet. Document-based information contains little structure, while the structured information found in software systems has very constrained semantic information. Emerging systems with no schema enforcement allow users to say nearly anything, severely limiting the utility of the information for exploitation and retrieval. There is significant opportunity to find a sweet spot between these three extreme view points.

Can we develop a new paradigm/framework in collective information authoring that lets end users richly author information so the maximum number of consumers can use that information? To reach this vision, a number of research areas must be investigated

- Identifying and integrating user generated semantic concepts to grow the semantic framework
- Determining the appropriate characteristics to suggest or enforce converging semantics
- Investigating unobtrusive ways to work with the user to reach semantic consensus
- Determining the correct occurrences to probe the user for additional semantic information.

A good example of a practical application related to this approach is Google base™ [1]. Over time, Google base™ has developed schemas to support common objects such as products, vehicles and jobs. While the list of available object schemas is extensive, Google base™ supports the capability to add new schema structures and modify existing schema structures for particular needs. This approach is quite flexible, but the information authored may not be easily queried by a user nor will it share semantics with similar, yet distinct, objects. Google base™ does support an evolving schema or schemas where the group's expressive needs can be supported over time. Although details are not public on how much user intervention is needed, much of this can be done with statistical approaches based on user activity. This automatic update of terminology is quite appealing and a significant feature of the system.

The real limitation with the approach is that if your perspective is not the perspective used by Google base™, or any similar system, your semantics won't align for input and query. For example, Google base™ has a schema specifically for events. Users may use that schema, but it is likely they would also query for concerts, meetings, or other types of events. If end users do not realize there is an event object, they will ultimately author information that has limited utility. There are many approaches to determining the semantic relationship between a query and existing information [2]. Research exists today to help guide the user towards common semantics. In many cases, users would be willing to leverage existing terms if they only knew of their existence. A dialog between the system and the user could have occurred to reconcile the difference. The trick is to do this in a way that is acceptable to end users.

In many cases, common semantics are not achievable due to subtle differences in meaning. This is an area where communities of authors could share terminology where appropriate and differ where necessary. The use of perspectives is a very powerful way to support the needs of different users. Some terms may be distinct to certain users and the same to others. There are techniques to support this at the knowledge base level [3], but support must be extended to the end user interface.

We believe that systems such as Wikipedia and Google base require significant user intervention to achieve common semantics. These systems have richer information sets than simple tagging of sites such as Flickr™. In the richer authoring environments, we contend that much of the effort to use consistent semantics could be automated. The next sections provide an example of the vision and a discussion of the rationale for our claims. We believe that in collective authoring environments users converge quickly towards a limited number of terms for the same attributes. We contend that machine intelligence can help them converge faster, but that they shouldn't need to converge to a single term in all cases.

2 Scenario Description

This use case describes the concepts and functionality investigated in the area of collective knowledge authoring. It supposes the use of a Google™ base-like system or a Semantic Wiki that allows user to add new semantics as they desire.

When users of reporting systems face novel situations that were not anticipated by system designers, users have historically "overloaded" existing fields that supported free text. Increasingly, underlying systems are "schema-less," allowing the users to modify the logical schema on the fly. While this is superior to legacy systems in many ways, this is problematic by the fact that multiple users are creating the overlapping schema with minimal forethought about the implications of their choices.

One can imagine the military challenge of intelligence collection and preparation for a cordon and search mission. Over a period of time, a team builds up information about an area so it can be displayed in a visualization program to help in the planning process. It is highly unlikely that systems will natively support the information relevant to cordon and search (i.e. description of facility, potential safehouse, family homes). Given the distributed information authoring task, it is highly unlikely that items will be referred to in the same way. For instance, one person may call a location an “abandoned house” while another may call it a “house that has zero residents.” These subtle differences will frequently occur without a pre-defined schema. The data is still valuable, but it cannot be easily used to answer queries or for analysis of the situation.

To remedy this problem, one can imagine that when a user authors additional information the system may alert the user that there is similar but disparate terminology being used. At this point, the user could suggest that the different terms are the same or different. Even with this user input, the impact of user feedback could allow various perspectives on the ontological information at hand. This would mean that users could use the terminology they are comfortable with though the system will leverage the equivalence of ontological elements.

Our hypothesis is that by monitoring the creation of new ontological terms, usage of those terms, queries of the data, and ingestion of new data sources that the system can guide users to use consistent terminology when applicable. Techniques leveraged will include semantic alignment, user-system dialog, and ontological perspectives.

3 Experimentation

To validate our hypothesis we needed to ensure that both semantics do not quickly converge in free-form environments and that semantic alignment tools could perform alignment over these disparate terms. To do this, we performed a small experiment where participants marked up documents with semantic content into an excel spreadsheet. Participants were allowed to construct ontological terms on the fly. A second phase of the experiment forced users to use existing spreadsheets of markup while marking up documents in a similar area of interest. The intuition was that Excel™’s autocomplete feature would be a simple way to encourage ontology term reuse. We measured the frequency that terminology was repeated across the various information authors.

Phase I of the experiment represented individuals marking up a document, while Phase II represents the markup of follow up articles using existing spreadsheets. As the number of users (and tags) increase the expectation is that overlap will increase among tag use. A major finding of the experiment was that there was extremely little syntactic overlap among users in all cases. This is partially attributed to the nature of the task, but it validates that in distributed situations individuals are likely to use different terminology for their situations. The results are shown in Table 1.

Table 1. Term uniqueness across article sets.

	Article Set 1		Article Set 2	
	#	% Unique	#	% Unique
Phase I Predicates	39		47	
Phase I Unique Predicates	36	92.31%	46	97.87%
Phase II Predicates	42		52	
Phase II Unique Predicates	40	95.24%	48	92.31%
Total Predicates	81		99	
Total Unique Predicates	69	85.19%	91	91.92%
Phase I Class Total	20		20	
Phase I Unique Classes	20	100.00%	16	80.00%
Phase II Class Total	14		12	
Phase II Unique Classes	13	92.86%	11	91.67%
Total Classes	34		32	
Total Unique Classes	29	85.29%	23	71.88%

To prove that semantic alignment could be applied to the disparate terminology, we attempted to identify semantic overlaps in terms used for Article Set 1/Phase I. Based on human analysis, we found four potential terms that could be rectified by semantic alignment processes (Table 2). If these terms were deemed identical, the number of unique terms would have dropped by over 10 percent.

Table 2. Reconcilable terms found in experimental data set.

Original Term	Equivalent Term
livesIn	Lives in
Recorded	Taped
hasBail	Has bail amount
worksFor	Works for

4 Conclusion

One of the key insights that were determined is that there is a loss of efficiency in the space of collective intelligence. In nearly all cases of collective knowledge authoring, there is significant investment in maintaining the knowledge or there is a significant loss of information. This can even be seen in the most prominent examples of collective intelligence. On Wikipedia™, the normalization and upkeep is non-trivial. It requires significant maintenance by particular users. On sites like Flickr™, there is often a convergence on tags used for images but there are many examples where numerous tags are appropriate and disjoint. In these cases, you could claim that a portion of the tagged information is never exploited.

A second major insight was the divergent terminology that appeared in our experiment. In tagging images, there have been studies that have shown a few tags are converged upon relatively quickly. The number of tags is usually quite small. In the world of semantic markup, there is much more flexibility in how people could represent equivalent items. In our limited experimentation we saw minimal convergence on semantic tags used. This was despite the fact that part of the experiment was seeded to encourage tag reuse. The amount of divergence was slightly surprising. We were able to detect cases where semantic alignment techniques could have easily aligned disparate semantic terms.

The emergence of the World Wide Web and Web 2.0 has brought collective intelligence to the forefront. Though there are many examples of how this technology has been extremely successful, it is not nearly as efficient as it could be. Nearly all of today's collective intelligence success stories require significant human maintenance and contain significant information loss. These challenges are only magnified when applied to formalized authoring of information. We have determined that there are opportunities to leverage machine automation for the process of collective knowledge authoring in free-form environments. In such situations, operators can enter any information they deem appropriate with the system attempting to reconcile disparate use of similar terms. We have shown that the approach would improve performance over situations without automation, but more work is necessary to compare this flexible approach with approaches that depend on structured schemas and ontologies.

References

1. Hsieh, W., Madhavan, J., Pike, R.: Data Management at Google. In: 2006 ACM SIGMOD International Conference on Management of Data, pp. 725--726. ACM, New York (2006)
2. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. In: VLDB Journal, vol 10. pp. 334--350. Springer, Heidelberg (2001)
3. Heflin, J., Pan, Z.: A Model Theoretic Semantics for Ontology Versioning. In: Third International Semantic Web Conference, Hiroshima, Japan, pp. 62--76. Springer, Heidelberg (2004)