2009

Semantics-Empowered Social Computing

Amit P. Sheth

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

Meenakshi Nagarajan

Wright State University - Main Campus

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., & Nagarajan, M. (2009). Semantics-Empowered Social Computing. *IEEE Internet Computing*, 13 (1), 76-80.

<https://corescholar.libraries.wright.edu/knoesis/328>

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.

Semantics-Empowered Social Computing

Amit Sheth and Meenakshi Nagarajan • *Kno.e.sis Center, Wright State University*

The social Web — also called Web 2.0 — which stems primarily from user-generated content (UGC) and the Semantic Web, a collection of machine-understandable documents and data, will soon merge into the social Semantic Web. Currently hailed as Web 3.0, this social Semantic Web will use rich domain knowledge and document-level metadata to organize and analyze social media content. Vital to its success will be how much the Semantic Web can enrich the social Web, which includes not only data or Web pages and the links between them but also people, the connections among them, and the connections that people make with data.

In the Semantic Web vision, data on the Web is made more meaningful through labels (via marking up, tagging, or annotating) that follow an agreed-upon reference model, be it a common nomenclature, dictionary, taxonomy, folksonomy, or ontology that represents a specific domain model. Annotations with these vocabularies make Web-based documents and data machine-understandable as well as easier to integrate and analyze. When applications use an ontology, rules that range from simple to complex, whether they're explicitly stated or inferred from the ontology's class properties and relationships, allow powerful reasoning over annotated data.

Today, communities in varied domains such as life sciences, healthcare, finance, and music have begun to provide ontologies with associated knowledge or instance bases to richly describe their domains. Services that allow the use of populated ontologies for annotation and applications that can exploit annotations and rules have been available since the early 2000s and are becoming increasingly common.¹

Popular Web 2.0 technologies or social media software such as tagging, blogging, bookmark-

ing, social networking, image- and video-sharing sites and so on have allowed people to consume, produce, and share information easily, making this new class of UGC one of the richest forms of data available on the Web today.

On one hand, the social context surrounding the production, consumption, and sharing of UGC has opened several opportunities for enriching user interaction with content. But on the other hand, this same social aspect to content production has introduced new challenges in terms of the content's informal nature.

In this article, we discuss some of the challenges in marking-up or annotating UGC, a first step toward the realization of the social Semantic Web. Using examples from real-world UGC, we show how domain knowledge can effectively complement statistical natural language processing techniques for metadata creation.

Using Background Knowledge for Semantic Metadata Creation

User-generated textual content in social media has unique characteristics that set it apart from the traditional content we find in news or scientific articles. Due to social media's personal and interactive communication format, UGC is inherently less formal and unmediated. Off-topic discussions are common, making it difficult to automatically identify context. Moreover, the content is often fragmented, doesn't always follow English grammar rules, and relies heavily on domain- or demographic-specific slang, abbreviations, and entity variations (using "skik3" for "SideKick 3," for example). Some UGC is also terse by nature, such as in Twitter posts, which leaves minimal clues for automatically identifying context. All of these factors make the process of automatically identifying what a social media snippet is actually about much harder. Conse-

quently, an important challenge that Web 3.0 applications will face is the process of automatically creating accurate markups or annotations from UGC to common referenced models.

We believe the role of ontologies and knowledge bases in creating markups will be even more important than they were before the growth of the social Web. Not only can they act as common reference models, but they'll also play a key role in inferring semantics behind UGC while supplementing well-known statistical and natural language processing (NLP) techniques.

First, let's examine a few examples using data from online user-generated textual content to show the challenges well-known tasks such as named entity identification² will face and how background knowledge can help.

Ambiguity in Entity Mentions

Consider the following post on a music group's discussion board: "Lily I loved your cheryl tweedy do...heart Amy." A human would know that the poster (Amy) is praising artist Lily's impression of an entity Cheryl Tweedy. Assuming that the end goal is to annotate artist and track/album mentions, the task here is to decide whether the entities Lily, cheryl tweedy, and Amy (identified using statistical NLP named entity identification techniques) are of interest.

This task is slightly complicated here given that the poster "Amy" shares a first name with a popular recording artist, "Amy Winehouse," and shouldn't be marked as an artist. A domain model such as MusicBrainz (<http://musicbrainz.org>), for example, will state that "Amy Winehouse" and "Lily Allen" are different artists from different genres – pop and jazz, respectively. It will also tell us that "Cheryl Tweedy" is a track by artist "Lily Allen." Thus, in spite of capitalized first letters, high string similarity with the art-

Figure 1. A syntactic parse of a user-generated comment from MySpace.

(a) A singular proper noun tag assigned to the ambiguous entity "Amy" and (b) a verb, non-third-person-singular present assigned for the word "smile."

ist's first name, and the singular proper noun (NNP) tag assigned by an NLP parser (see excerpt of parse in Figure 1a), there's no additional support from the knowledge base for "Amy" referring to the artist. Obtaining such additional evidence from the knowledge base might also be more economical than rigorous statistical NLP techniques that disambiguate the mention of "Amy." Applications that index and retrieve information, for example, could take this into consideration so as to not markup "Amy" as an artist or return this post for search queries about the artist "Amy."

Identifying Entities

In another post, "Lils smile so rocks," the poster could be seen as praising the artist Lily's "smile" (her facial expression). A knowledge base, however, will tell us that "Smile" is also a track by "Lily Allen" (with a high string similarity between "Lily" and "Lils") and is a possible entity of interest. This can be considered as a form of support toward "Smile" be-

ing a named entity of interest in spite of its verb (VBP) part of speech tag (see Figure 1b) and lack of first letter capitalization.

Similarly, in the tweet, "Steve says: All Zunes and OneCares must go, at prices permanently slashed!", it's safe to conclude that "Steve" here is referring to "Steve Ballmer," Microsoft's CEO, given that a knowledge base mentions Zunes and OneCares as Microsoft products and Steve Ballmer as the company's CEO. (Tweets are the user-generated posts on twitter.com.)

Off-Topic Noise

Another characteristic of content on social media is the tendency for users to digress to multiple topics. Removing off-topic noise is important for understanding what the content is about.

Consider the following post from a social network forum in which the user is talking about a project using "Sony Vegas Pro 8" but digresses to other topics. The keywords "Merrill Lynch," "food poisoning," and "eggs"

User post and advertisements generated when content has off-topic keywords

Topic :Illustrator CS3
Main Post :I need this in order to familiarize myself with it prior to beginning grad school My Masters will be in Graphic Design

Online Master's Degrees
Learn at Your Own Pace & Earn Your Fully Accredited Degree Online.
www.vgu.edu

University of Phoenix
Online degrees save time and gas. Learn about your career options.
UoPhx.info

Masters
Earn your Masters Online at Northeastern University (NEU).
OnlineMBA.Neu.Edu

Ads by Google

Advertisements generated after off-topic keywords are eliminated

illustrator cs3
Try Illustrator CS3 Today
Instant Download. Newest Version. In Stock. Buy Direct. Full Support!
www.CS3deals.com

Illustrator 10 Tutorials
Move your career forward with an accredited online degree!
www.CourseAdvisor.com

\$990 Adobe CS3 Premium - Free Priority Shipping \$790
Adobe CS3 Standard
www.WebCommunitySoftware.co

Adobe Illustrator Schools
Advance your career or hobby at top Adobe Illustrator schools near you.
www.ComputerTrainingSchools.c

Ads by Google

Figure 2. Contextual advertisements. The top half shows a user post and ads generated when content has off-topic keywords, and the bottom half shows ads after we eliminate off-topic keywords.

are clearly off-topic in this context:

I NEED HELP WITH SONY VEGAS PRO 8!! Ugh and i have a video project due tomorrow for merrill lynch :(all i need to do is simple: Extract several scenes from a clip, insert captions, transitions and thats it. really. omgg i can't figure out anything!! help!! and i got food poisoning from eggs. its not fun. Pleassssse, help? :(

In addition to association strengths between words (derived from a corpus), a knowledge base of computer software (generated from http://computers.shop.ebay.com/Computers-Networking__W0QQ_sacatz58058, for example) will readily tell us that none of the off-topic keywords are relevant to the discussion about "Sony Vegas Pro."

The presence of off-topic noise especially affects the results of content-analysis applications when a strong monetary value is associated with the content.³ Targeting advertisements against UGC on social networking sites is one such example. Advertisements in this medium have high visibility and also higher chances of being clicked, provided they're relevant to the user context. Figure 2 shows an example of the

targeted nature of advertisements delivered before and after removing off-topic noise in UGC.

Using Background Knowledge to Analyze User Comments

In recent work,⁴ we implemented a content-analysis system that mined music-artist popularity from user comments on MySpace artist pages. We designed

- an artist and music annotator to spot artists, albums, tracks, and other music-related mentions (such as labels, tours, shows, and concerts) in user posts, and
- a sentiment annotator to detect sentiment expressions and measure their polarities.

We backed the artist and music annotator with MusicBrainz, a knowledge base of musical artists, genres, albums, and tracks. The annotator compared artist or track mentions in user comments against artist entries and associated track entries in the knowledge base to gain more context. In addition to this, the annotator used results of a syntactic parse of the comment and corpus statistics to annotate a track or artist mention. The sentiment annotator used a syntactic parse of comments to extract adjectives and verbs as potential sentiment expressions. It then

consulted a slang dictionary (Urban Dictionary.com) to verify the expression's validity and ascertain polarity (positive or negative).

For both annotators, the combination of techniques proved to be more useful than using techniques in isolation. We aggregated positive and negative sentiments for all artists to generate a ranked list of the top X artists ordered by the number of positive sentiment comments (see Figure 3). By observing popularity trends over time and the patterns that stand out in the user activity of such online communities, we were also able to forecast what was going to be popular tomorrow.

With background knowledge and statistical and linguistic techniques, each providing different levels and types of support for UGC analysis, the important questions are what combination of these should applications use and when. This in turn will depend on the application's end goal and on the data with which it works. Blogs, for example, tend to be longer and have sufficient information to assess meaning behind the content. However, the analysis of tweets and forum messages might need more help from background knowledge, especially when there's insufficient support from corpus-based approaches. As more Web applications begin to combine domain knowledge with their existing content-analysis frameworks, this will become an important focus of investigations.

Aggregating Attention Metadata

User-generated textual content such as reviews, posts, and discussions are only one example of attention metadata – that is, any information generated as a result of a user's interest or attention to content. Other examples include

- descriptions, tags, and user-placed anchor links;

- page views and access logs;
- star ratings and diggs; and
- images, audio, video, and other multimedia content.

Today, applications that aggregate user activity typically operate with only one type of attention metadata. They might aggregate topical blogs (www.sifry.com/alerts/Slide0008.gif), visualize connections between people and the content produced within a network (www.neuroproductions.be/twitter_friends_network_browser/), or aggregate music listening (<http://lastgraph3.aeracode.org>).

Aggregating all known attention metadata for an object is more complicated because it involves multimodal information. In the music domain, for example, user interest that generates a song listen isn't the same as that which generates a video view or a textual comment. In a recent work,⁵ we used voting-theory principles to aggregate user activity from MySpace and Bebo comments, as well as LastFM listens and YouTube comments to measure overall artist popularity in the music community.

With the need to measure a population's pulse across all available information sources, we suspect this will be an important area of investigation.

A Newer Breed of Applications

Annotating UGC with common reference models will undoubtedly improve applications tasked with presenting a holistic view of all information available to a user. Content-delivery applications such as Zemanta (www.zemanta.com), for example, that match keywords to provide related information can utilize related concepts in the knowledge base to suggest additional content.

Perhaps the most interesting phenomenon on the social Web is that people aren't only connected to

Figure 3. System architecture. The use of domain knowledge helps analyze user-generated content for the task of popularity mining.

each other by means of a social tie (friends on social networks or referrals on LinkedIn) but are also connected via a piece of information. A user can link to someone's blog post, for example, follow someone's tweet, respond to a posting, tweet with other users from the same location, and so on. In addition to context derived from the content, a corpus, or a domain knowledge base, UGC also comes with a social context that includes the network in which it was generated. For certain types of data, such as tweets sent from a

cell phone, there is also a situational context, such as time and location, that becomes increasingly relevant to the analysis.

Tapping this machine-accessible people-content network and its associated social and situational contexts empowers a new breed of personalized socially aware systems.⁶

Imagine a scenario in which you're looking to get more information about a camera you heard described on the radio, but you don't remember the exact model number. However, you do remember the radio

host mentioned his blog post, which discussed a review he had read on his favorite gadget discussion forum for the same product. On the social Semantic Web, where all UGC is annotated, an intelligent search program would be able to sift through all of the host's blog posts and all the annotated gadget forum posts, look for the same camera object, and return matching pages to you.

Now, consider the following scenario in which an event-tracker system maintains a knowledge base of music events (including dates, times, and locations) along with artists and their work; it also continually tracks and annotates tweets related to the events. Now imagine a user tweets, "Hitting traffic jam. Looks like im missin lilys opening" from his iPhone (which also provides time and location information). Using situational context information and identifying "Lily" in the tweet, the system has enough support to asso-

ciate this message with the "Lily Allen concert" event in its knowledge base. The application can now alert users who have signed up for the same event and share similar location coordinates with a "watch out for a traffic jam" message.

The role of users in driving today's social media is undeniable. The wealth of user-generated information spans multiple content types, people networks, and people-content interactions. To effectively exploit this avalanche of information and build applications that enrich online user experiences, we must bring some level of organization to the otherwise loosely categorized content on the social Web.

We see great potential for a place where the social Web meets the Semantic Web, where objects are treated as first-class citizens, making it easier to search, integrate, and ex-

ploit the information surrounding them. Although there are important content-related challenges to be met, applications using this underlying semantic infrastructure will significantly enhance the business potential behind UGC as well as enrich user experience associated with social media. □

References

1. A. Sheth et al., "Managing Semantic Content for the Web," *IEEE Internet Computing*, vol. 6, no. 4, 2002, pp. 80–87.
2. D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Linguisticae Investigationes*, vol. 30, no. 1, 2007, pp. 3–26.
3. M. Nagarajan et al., *Monetizing User Activity on Social Networks*, tech. report, Kno.e.sis Ctr., Wright State Univ., 2008; <http://knoesis.wright.edu/library/download/MonetizingUserActivity-TechReport.pdf>.
4. J. Grace et al., *Artist Ranking through Analysis of Online Community Contents*, tech. report, Kno.e.sis Ctr., Wright State Univ., 2007; <http://knoesis.wright.edu/library/publications/200711010839-submission.pdf>.
5. A. Alba et al., "Applications of Voting Theory to Information Mashups," to be published in *Proc. 2nd IEEE Int'l Conf. Semantic Computing (ICSC 08)*, IEEE CS Press, 2008.
6. P. Mika, *Social Networks and the Semantic Web (Semantic Web and Beyond)*, Springer-Verlag, 2007.

Amit Sheth is an IEEE fellow, LexisNexis Ohio Eminent Scholar, and director of the Kno.e.sis Center at Wright State University. Contact him via <http://knoesis.wright.edu/amit>.

Meenakshi Nagarajan is a researcher at the Kno.e.sis Center and a PhD candidate in computer science at Wright State University. Her research focuses on using contextual knowledge for the analysis of casual text from social media. Contact her via <http://knoesis.wright.edu/students/meena/>.

Call for Articles

IEEE Pervasive Computing

seeks accessible, useful papers on the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing. Topics include hardware technology, software infrastructure, real-world sensing and interaction, human-computer interaction, and systems considerations, including deployment, scalability, security, and privacy.

Further details:
pervasive@computer.org
www.computer.org/pervasive

Author guidelines:
www.computer.org/mc/pervasive/author.htm

IEEE pervasive COMPUTING
 MOBILE AND UBIGUITOUS SYSTEMS