**Editor: Carole Goble** • *cag@cs.man.ac.uk*

# Semantic Filtering for Social Data

**Amit Sheth and Pavan Kapanipathi** • *Kno.e.sis Center, Wright State University*

Consumers of social data face information overload. Although information filtering can help, challenges specific to the short-text and real-time nature of social networks remain. Harnessing knowledge bases from crowd-sourced platforms such as Wikipedia can help build an effective information-filtering system.

One in three Web users looks for medical information on social networks (http://bit.ly/wiredarabspring), and more than 50 percent of users surveyed consume news on social networks (http://bit.ly/pewsnsnews). Twitter and Facebook were prominent platforms used for disseminating information and organizing protests during Arab Springs, Occupy Wall Street, and similar events.[1] Social data also plays a critical role in helping with coordination during natural disasters. Social networks have therefore not only changed the landscape for communicating and sharing information – they have also become a major source for users consuming information.

The popularity of social networks has led to an increase of user-generated content on the popular platforms. Facebook and Twitter together generate more than 5 billion microblogs per day. Because users consume information from these platforms, the overwhelming amount of content generated brings to mind Herbert Simon's famous quote: "a wealth of information creates a poverty of attention."[2] In turn, the growth in the volume of content has often drawn criticism of information overload from consumers. As users of the Web, it's important for us to realize that our dependence on information from these platforms will continue to grow, and hence, so should our focus on working towards making our lives easier in accessing the collected intelligence on these platforms, particularly by addressing the problem of information overload.

Researchers have addressed the challenge of information overload by developing information filtering systems that understand a group of users' interests and deliver relevant content to them. Although these filtering techniques have been adopted for filtering spam in emails and delivering relevant news and articles to interested users, leveraging these techniques on social networks and building an efficient information-filtering system presents distinct challenges, due to social networks' unique characteristics. Here, we consider how to address those challenges, using a crowd-sourced platform such as Wikipedia.

## The Challenges for Filtering Social Data

We observe two prominent characteristics of social networks that are distinct from the traditional information sources – the short-text and real-time nature of the medium. The ideal length of a post on any social networking website ranges between 60–140 characters (http://bit.ly/fastcompsnslength). This notion of creating short-text has let users quickly update their statuses with less effort. By real-time nature, we're referring to the fact that it takes only a brief amount of time to create and disseminate information across a network of users. Social networks' popularity in many domains has been credited to their real-time nature – particularly because users increasingly desire to consume and react to ongoing, real-world situations as they develop. During the 2008 Mumbai terrorist attacks, the 2011 Arab Spring, Hurricane Sandy in 2012, and the 2014 Kashmir Floods, information and news related to these events were communicated in real time using social networks, and user-generated content often became the primary mechanism for the dissemination of information rather

1089-7801/16/$33.00 © 2016 IEEE

than traditional media outlets. As a result of this shift, journalists are now advised to monitor social networks for the latest news, and the disaster-management community uses social networks as a real-time communication platform, monitoring them closely to coordinate and manage situations during disasters.

To process the textual content of social networks and address information overload, any information-filtering technique used needs to handle two technical challenges: a lack of context and a dynamically changing vocabulary.

**Lack of context.** The task of analyzing textual content from social networks is fundamental to building an information-filtering system. This is necessary for understanding user interests from the posts they like or share on social networks, and for filtering posts that are relevant to their interests. For instance, consider the following post by a user on Twitter:

> Example 1. "Great day for Chicago, Cubs beat Reds, Sox beat Mariners with Humber's perfect game."

The popular content-based assumption that users are interested in what they share, means we can infer that the author is interested in the Chicago Cubs, Cincinnati Reds, White Sox, and the Seattle Mariners. These topics we've identified then help us understand users' interests, to filter future posts. This is notable because existing topic identification[3] techniques perform significantly well on traditional and longer textual content such as blogs and news, but don't perform well on short-text because they lack context for processing.[4]

**Dynamically changing vocabulary.** Social networks can track topics such as a natural disaster, an election, or a sporting event. Changes in these topics are reflected by changes in the vocabulary used on social networks. For example, the 2014 Indian election had various subevents associated with it that were emphasized during different times as the event unfolded. These include the announcement of prime ministerial candidates, issues regarding corruption in the political parties, and polls in different states. This topic was represented by multiple terms on Twitter, such as *#modikisarkar*, *#NaMo*, *#VoteForRG*, and *#CongBJPQuitIndia*, which evolved over time.

Social network conversations during natural disasters also exhibit significant changes over time. Disasters have been shown to go through phases as the situation evolves (for example: mitigation, preparedness, recovery, and response), and the conversation and language of social networks reflect this. During Hurricane Sandy, in particular, the representative hashtags evolved from *#Frankenstorm* and *#Sandy* at the start, to *#StaySafe* and *#RedCross* during the disaster, and *#ThanksSandy* and *#RestoreTheShore* after the hurricane.

Because more than 50 percent of users on social networks are interested in keeping up-to-date with the most recent topics and resort to social networks to do so, it's important for filtering systems to adapt to changes happening in the real world. However, as social network platforms allow filtering based on keywords (or a combination of keywords), it becomes challenging for information-filtering systems to continuously monitor new, evolving keywords and filter relevant posts in real time.

## Collective Semantics and LOD to the Rescue

We can address these challenges by developing techniques that leverage knowledge bases to enrich the semantics of short-text. Semantics is the relevant information inferred from knowledge bases related to the content in short-text to facilitate better understanding and processing. For instance, the semantic enrichment of the post in Example 1 involves the use of information associated with the *Chicago Cubs*. This information can be facts, such as *Chicago Cubs are one of the Major League Baseball teams*, or *Jason Herward* and *Kris Bryant* are its players. Such information can be found in structured knowledge bases on the Web.[1]

Knowledge bases on the Web have grown in popularity due to the Linked Open Data (LOD) initiative and its focus on transforming the Web of hyperlinks into a Web of Data.[5] Although there's an abundance of knowledge bases on the Web, it's important to select a relevant knowledge base (also referred to as an ontology or knowledge graph) to deal with the unique characteristics of social networking platforms. The knowledge base must satisfy two prominent requirements: first, broader coverage — because social networks handle a large set of diverse users, the filtering system requires that the knowledge base have a broad coverage of topics; and second, near real-time updates — social networks are real time in nature and mirror changes and activities in the real world. To utilize a knowledge base for filtering social data, it's necessary for the knowledge base to be dynamically updated and also reflect the real world.

While knowledge bases on the LOD cloud, such as DBpedia, and Yago, encompass diverse topics and hence may satisfy the broad coverage requirements, these knowledge bases are updated infrequently — certainly not rapidly enough to cover a new, evolving event. Therefore, they're unsuitable for real-time filtering of social data. However, one of the prominent sources of information for the aforementioned knowledge bases is Wikipedia, which is dynamically updated and reflects unbiased views of the real world in near real time.[6] Wikipedia is an up-to-date collection of collaborative encyclopedic knowledge for most situations. In the remainder of this article, we discuss novel approaches that uses Wikipedia as a knowledge base by harnessing its semi-structure to address the challenges in filtering social data.

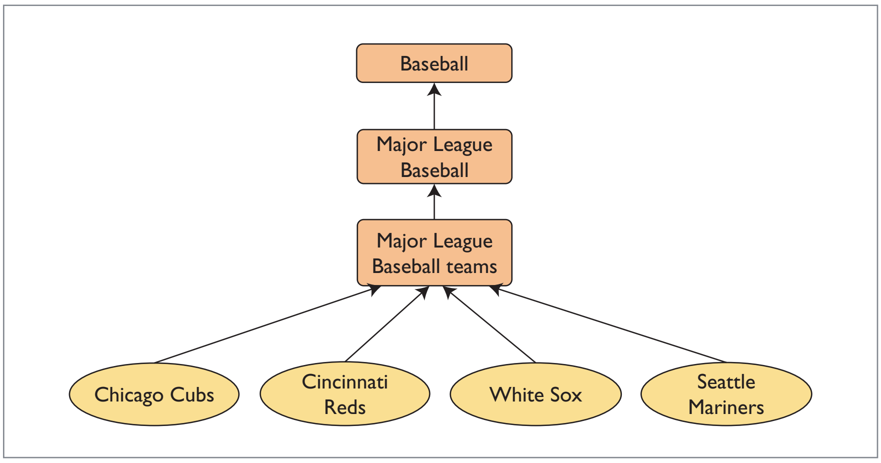

*Figure 1. Hierarchical context generated from Wikipedia category structure. This is based on topics mentioned in Example 1.*

***Table 1. Social network posts that can be filtered for the author of Example 1 using a content-based interest profile and Hierarchical Interest Graph (HIG).***

| Tweets for filtering | Content-only interest profile | HIG |
|---|---|---|
| The #**Cubs** winning World Series? 'Holy grail of baseball' http://usat.ly/1pbSvqS | √ | √ |
| Today in 1955, the **New York Giants baseball** signed **Willie McCovey** as an amateur free agent. | X | √ |
| **Sergio Romo** says **Bryce Harper** should switch jobs if he doesn't like how **baseball** is played today. | X | √ |

## Enhancing Context Using Hierarchical Interest Graphs

Because the lack of context is what makes processing short-text challenging, Wikipedia can be exploited to improve context and identify topics. As humans, we naturally infer that the topics mentioned in Example 1 are related to *Baseball*, and hence the author might be interested in it. A similar inference can be made by machines using Wikipedia's category structure, which is comprised of taxonomic knowledge that can be automatically extracted. Figure 1 portrays the relevant taxonomic knowledge from Wikipedia for topics mentioned in Example 1.

By using this hierarchical context from Wikipedia on social data, we can create an interest profile with a hierarchy — that is, a Hierarchical Interest Graph (HIG). The HIGs encompass not only explicitly mentioned interests (such as the *Chicago Cubs*, *Cincinnati Reds*, and so on) but also others that are implicitly inferred from Wikipedia (such as *Baseball* and *Major League Baseball*). We adapt the spreading activation algorithm[7] to score each of the inferred interests, where the scores represent the extent of users' interests.[8]

HIGs add context derived from a knowledge base; hence, they address the issue of the lack of context. By using the HIG that encompasses implicit interests — for example, *Baseball* and its related topics — we can broaden the coverage of social data filtered for the user. In Table 1, based on the profiles generated for the author of Example 1, we can see the tweets that can be filtered using the knowledge-enhanced HIG for this user (Figure 1) versus interests that are directly extracted from the content. Example 1 mentions the *Chicago Cubs*, which is also explicitly mentioned in the first tweet in Table 1; hence, this tweet can be filtered by both profiles because it captures the *Chicago Cubs* as the author's interest. However, the second and the third tweets in Table 1 are related to *Baseball* and can be filtered only using HIGs because they infer *Baseball and related topics* as the users' interests from the Wikipedia category structure. To evaluate our approach for identifying hierarchical interests, we performed a user study involving 37 participants, which concluded that approximately eight of the top 10 interests were relevant to the user, and around 60 percent of these interests were implicit. To demonstrate the applicability of our approach, we applied the hierarchy of interests generated to recommend tweets for users. By augmenting content with knowledge-based user profiles (HIGs) we can improve the performance of tweet recommendation systems by more than 40 percent in comparison to other existing content-based recommendation approaches.

## Harnessing Wikipedia's Evolving Knowledge for Continuous Filtering

Social networks reflect the evolving topics of the real world by changes in the representative vocabulary used in posts. As we mentioned, during the 2014 Indian election the representative hashtags used on Twitter included *#modikisarkar*, *#NaMo*, *#VoteForRG*, and *#CongBJPQuitIndia* over time. To keep track of the evolution of the topic, the keyword filter must be up-to-date with representative vocabulary.

The approach we outline utilizes hashtags as keywords to filter information. Hashtags are a common way to represent topics and activities on Twitter, and the increasing adoption of hashtags by users has forced Twitter to add hashtags as a standard feature. Hashtags were also adopted by other

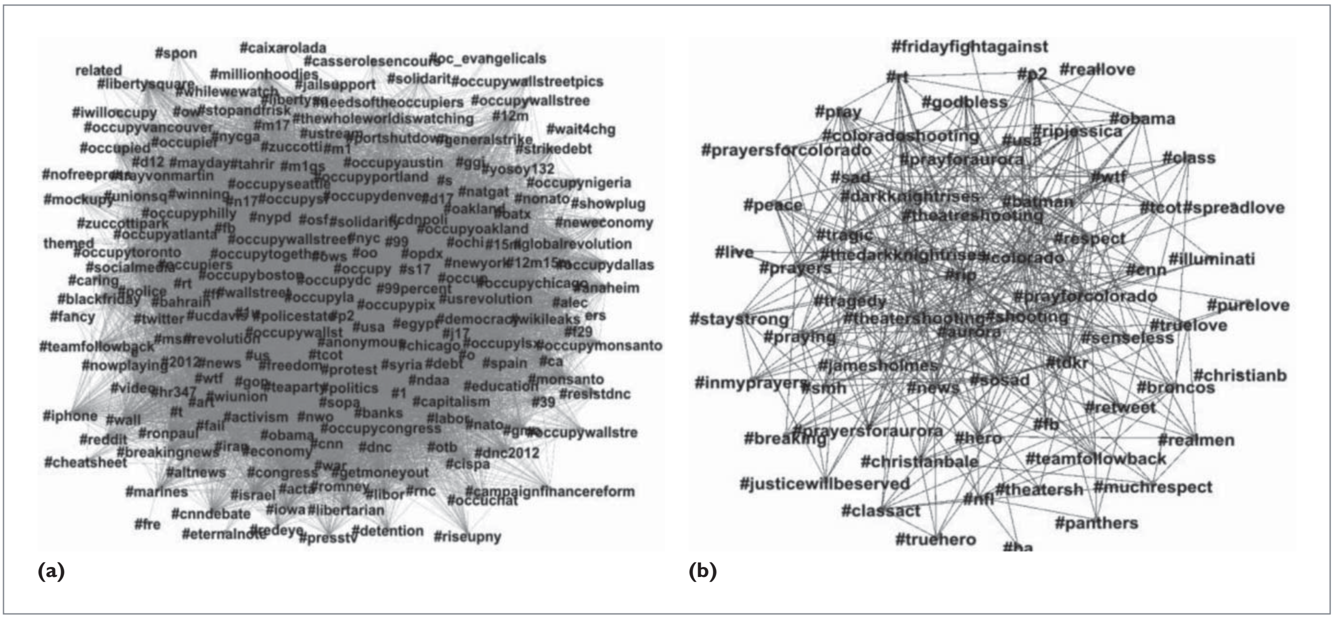

*Figure 2. Co-occurrence graph of hashtags relevant to (a) Occupy Wall Street and (b) the Colorado shooting. This is a hindsight analysis performed on manually curated data for Twitris.*

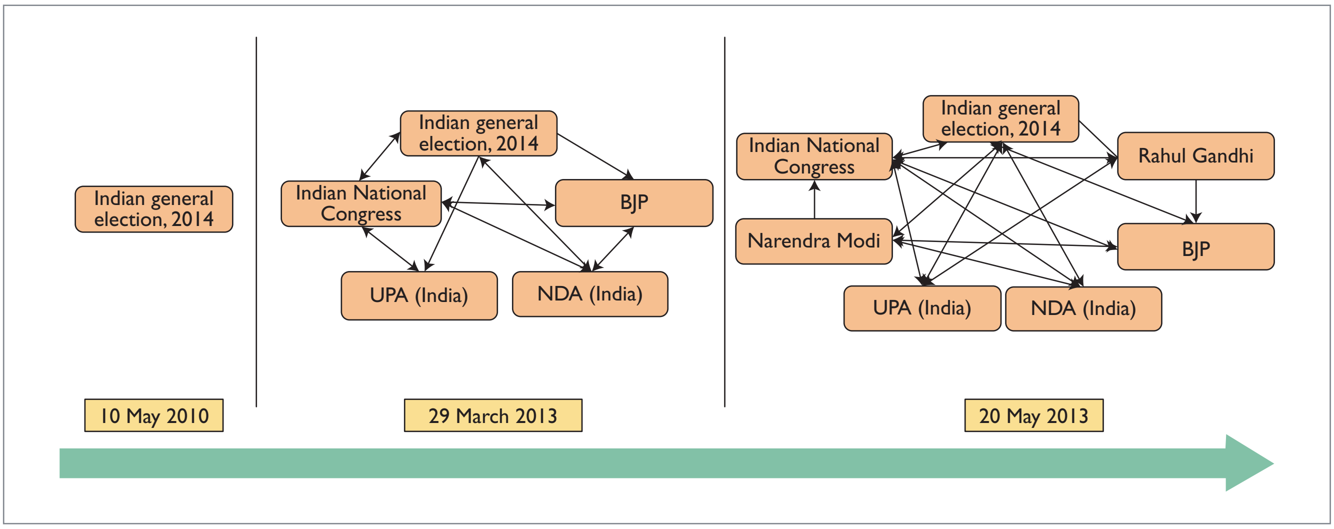

*Figure 3. Evolving Wikipedia hyperlink graph for the 2014 Indian general election as a topic, between 10 May 2010 and 20 May 2013.*

social networks such as Facebook and Instagram. We performed an analysis of hashtags extracted from 6 million tweets related to two dynamic topics — Occupy Wall Street and the Colorado shooting — and found that co-occurrence can be used as a starting point to continuously update hashtag filters. In other words, starting with a topic-relevant hashtag, such as *#ows* for Occupy Wall Street, we would be able to find other relevant hashtags (such as *#owsla* and *#owsny*). Figure 2 shows the co-occurrence graph of the hashtags relevant to the Colorado shooting and Occupy Wall Street.

Because detecting hashtags by co-occurrence alone can introduce noise, we take inspiration from the vision of continuous semantics[9,10] and leverage Wikipedia as an evolving knowledge base to determine semantically relevant hashtags. The Wikipedia hyperlink structure is beneficial, because it evolves to reflect the changes in the dynamic topics. (For example, in Figure 3 we can see the change of links on Wikipedia for the 2014 Indian general election between 10 May 2010 and 20 May 2013.) The hashtags detected are used to continuously, periodically update the filter that collects relevant posts.[10] A simulated real-time evaluation of this approach for two dynamic topics from 2012 — the US presidential election and Hurricane Sandy — showed that the top five hashtags detected were able to improve coverage by retrieving new tweets with a high mean average precision of 0.92.

Using Wikipedia for information filtering has its own limitations. The updates for topics such as natural disasters and elections are quicker and the timeliness is comparable to social networks such as Twitter. However, for topics such as terrorist attacks and civil protests, the information propagation on social networks and Wikipedia can be bottom up — that is, Twitter can reflect changes in the real world sooner than Wikipedia, and the lack of consensus on some topics (for example, whether democracy is suitable for a Middle Eastern country) can hamper the quality of timely knowledge that a Wikipedia-type knowledge source can provide. An analysis of Wikipedia updates and Twitter feeds shows that Wikipedia is comparably slower in receiving information.[11]

With the short-text on social networks, the network and the demographic attributes of users can enhance the performance of information-filtering systems by better understanding users. While the network dimension for filtering and recommending is well explored, inferring demographics of users on social networks has yet to be tackled, particularly due to the lack of demographic information shared on these platforms. In our recent work, we addressed the lack of location information[12] for users on Twitter, which is one of the important attributes of demographics. The methodology identifies the city-level location of users, accessing Wikipedia as the source of background knowledge. The use of Wikipedia is just the tip of the iceberg, and the potential of knowledge bases — including the broader LOD and open data for the task of social data filtering — has yet to be well explored.

## References

1. C. Bizer et al., "DBpedia — A Crystallization Point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, 2009, pp. 154–165.
2. H.A. Simon, "Designing Organizations for an Information-Rich World," *Computers, Comm., and the Public Interest*, M. Greenberger, ed., Johns Hopkins Press, 1971, p. 40.
3. C.-Y. Lin, "Knowledge-Based Automatic Topic Identification," *Proc. 33rd Ann. Meeting on Assoc. for Computational Linguistics*, 1995, pp. 308–310.
4. L. Derczynski et al., "Microblog-Genre Noise and Impact on Semantic Annotation Accuracy," *Proc. 24th ACM Conf. Hypertext and Social Media*, 2013, pp. 21–30.
5. C. Bizer et al., "Linked Data on the Web," *Proc. 17th Int'l Conf. World Wide Web*, 2008, pp. 1265–1266.
6. M. Ferron, and P. Massa, "Collective Memory Building in Wikipedia: The Case of North African Uprisings," *Proc. 7th Int'l Symp. Wikis and Open Collaboration*, pp. 114–123.
7. M.R. Quilian, "Semantic Memory," *Semantic Information Processing*, M. Minski, ed., MIT Press, 1968.
8. P. Kapanipathi et al., "User Interests Identification on Twitter Using a Hierarchical Knowledge Base," *The Semantic Web: Trends and Challenges*, LNCS 8465, Springer, 2014 pp. 99–113.
9. A.P. Sheth, C. Thomas, and P. Mehra, "Continuous Semantics to Analyze Real-Time Data," *IEEE Internet Computing*, vol. 14, no. 6, 2010, pp. 84–89.
10. P. Kapanipathi et al., *Continuous Semantic Crawling Events*, Kno.e.sis wiki, 2015; http://wiki.knoesis.org/index.php/Continuous_Semantic_Crawling_Events.
11. M. Osborne et al., "Bieber No More: First Story Detection Using Twitter and Wikipedia," *Proc. SIGIR 2012 Workshop on Time-Aware Information Access*, 2012; http://homepages.inf.ed.ac.uk/miles/papers/TwitterWikipedia.pdf.
12. R. Krishnamurthy et al., "Knowledge Enabled Approach to Predict the Location of Twitter Users," *The Semantic Web. Latest Advances and New Domains*, LNCS 9088, Springer, 2015, pp. 187–201.

**Amit Sheth** is the LexisNexis Ohio Eminent Scholar and executive director of Kno.e.sis — the Ohio Center of Excellence at Wright State University. His research interests include smart data, physical Internet of Things, cyber and social Big Data; and semantic-cognitive-perceptual computing. Sheth has a PhD in computer and information sciences from Ohio State University. He's an IEEE Fellow. Contact him at amit@knoesis.org or http://knoesis.org/amit.

**Pavan Kapanipathi** has recently completed his PhD at Kno.e.sis and will soon join as a research staff member at IBM T.J. Watson Research Center. His research interests include user modeling; the Semantic Web; recommender systems; and social data analysis. Kapanipathi has an MS in computer science from Wright State University. Contact him at pavan@knoesis.org.