Using Hyperlinks to Enrich Message Board Content with Linked Data*

Sheila Kinsella, Alexandre Passant
Digital Enterprise Research Institute
National University of Ireland, Galway
{firstname.lastname}@deri.org

John G. Breslin
School of Engineering and Informatics
National University of Ireland, Galway
john.breslin@nuigalway.ie

ABSTRACT

Since social media sites have existed, a major element of the conversations which take place on them has been links to external websites. Users share videos or photos they have seen, point to products or movies they are interested in, and use external articles as a reference in discussions. As websites publish more structured and machine-readable data, notably in RDF, an increasing number of the links posted to social media sites do not just point to a webpage but are also associated with a structured data source. The integration of such external data can give us an enhanced insight into the conversation, both for analysing communities and for building applications. This approach can be applied to any online social space, but in this paper we present the use-case of enriching bulletin boards. We investigate the hyperlinks posted on a message board over a 10 year period and show how the external structured information related to these links has grown. We use data aggregated from several external sites providing RDF data or APIs to show how enriching forums with external data can provide us with enhanced insight into online conversation. We discuss potential applications of this approach, for message boards and for social media in general.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Semantic Web, Web 2.0*

General Terms

Experimentation

Keywords

hyperlinks, Linked Data, message boards, online communities, online conversation, social media

*The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon 2).

I-SEMANTICS 2010, September 1-3, 2010 Graz, Austria

1. INTRODUCTION

Social media sites are some of the most popular sites on the Web, with Web users spending an increasing amount of time using them[1]. The conversations on these sites often revolve around links to other sites, notably in weblogs and bulletin boards, where such links are used to refer to external topics and related information. For example, videos from YouTube are a popular topic on many social sites[2]. These links can be seen as social objects according to the object-centred sociality theory of Karin Knorr-Cetina [15]: these are the artifacts around which communities are formed and people interact. Indeed, they frequently correspond to identifiable concepts such as movies, books or products. Understanding the characteristics of these objects (*e.g.* a movie genre or book author) would be a very useful aid in understanding online conversations, in order to effectively recommend related content and enhance the serendipitous value of Social Web applications.

Increasingly these online artifacts are being described by the data publishers using structured data as well as in a human-readable form, notably in the RDF format through the Linking Open Data initiative[3]. Figure 1 shows an example of a post which can be complemented by data from external sources (in this case LinkedMDB [10] and DBpedia [2]). The extra information contained by these sources provides additional context which can be useful for both analysis of online conversation, and for developing related applications. As structured data becomes easily available, the challenge is how to bridge the gap between "plain-text" data and structured content, a question that we tackle in this article.

Previous work on bringing structure to bulletin boards includes [6] which describes how SIOC can be applied to message boards in order to represent them in a structured format. While this and many approaches in the Linked Data and Social Semantic Web community have focused on representing meta-data about conversations in a unified way, with FOAF, SIOC, etc. we describe how these conversations can be enhanced with additional RDF data about what they contain, *i.e.* regarding what are the *objects* (as referred to in the object-centred view) being discussed therein. Our objective is to use the enhanced dataset to better understand

[1] `http://blog.compete.com/2009/02/09/facebook-myspace-twitter-social-network/`
[2] See `http://forumeter.com` which tracks videos posted in bulletin boards
[3] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

Thread Title: *Little Miss Sunshine*

Post Content: *it was a nice uplifting movie but not really as amazing as people make it out to be - as far as i'm concerned there* `[url='http://www.imdb.com/title/tt0211915/']` *are far better uplifting movies* `[/url]` *out there.*

External data:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpprop: <http://dbpedia.org/property/> .
@prefix imdb: <http://www.imdb.com/title/> .

imdb:tt0211915 foaf:topic dbpedia:Amélie .
dbpedia:Amélie dc:title "Amélie" ;
   dc:date "2001" ;
   dbpprop:starring dbpedia:Audrey_Tautou ;
   dbpprop:director dbpedia:Jean-Pierre_Jeunet .
```

Figure 1: Example of a real message board post, and external structured data relating to it.

users and communication in online communities. We also see our analysis as relevant for developers of applications for browsing or authoring in online communities.

Figure 2 shows the process of enriching the dataset. We make use of standard Semantic Web technologies to take existing Web data, integrate it with a social media dataset, and enable analysis or application development. In particular, we focus on data from the `http://boards.ie` bulletin board, the most popular Irish bulletin boards service, covering 10 years of conversations. We demonstrate how structured data can be used not only in novel applications (where such structured data is explicitly used) but in any context in which a community links to external data, even if such data is not explicitly structured. This allows us to envision a wide range of applications to understand the behaviours of users, and how they evolve over time.

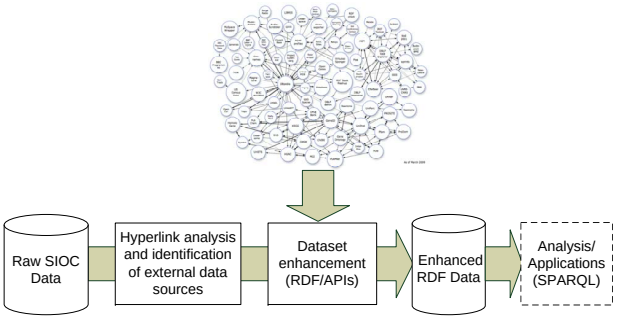


Figure 2: Process of enriching online conversations with structured data.

The contributions of this note are as follows:

(i) we study the behaviour of users in posting links on a message board over a 10 year period;

(ii) we present results of how much and what types of structured data were retrievable for the posts;

(iii) we describe examples of the novel analysis which can be performed on the enriched dataset.

We also discuss current and future applications of enriching social media with structured data.

2. TEN YEARS OF HYPERLINKS IN CONVERSATIONS

In 2008, the `boards.ie` SIOC Data Competition[4] was held and ten years of discussions from this message board site was made available in the SIOC (Semantically-Interlinked Online Communities) format [7]. The FOAF (Friend-of-a-Friend) and DC (Dublin Core) vocabularies were also used to represent elements of the dataset. The structure of the key elements of the data is shown in Figure 3. Since the data is already represented using the SIOC vocabulary, we can easily extract links using the `sioc:links_to` property.

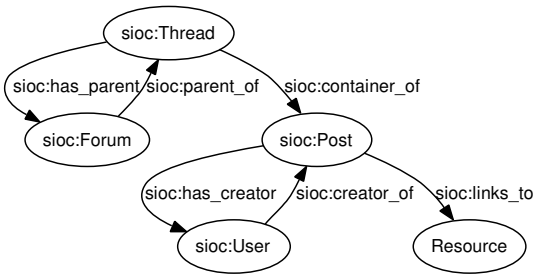


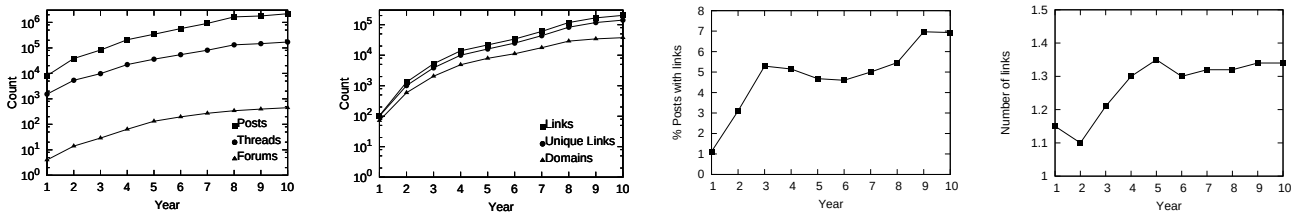
Figure 3: Partial structure of the SIOC dataset.

The data covers the first 10 years of existence of the message board, from February 1998 to February 2008. Over 130k users are represented in the data, along with over 7m posts which they have authored. The properties of the dataset are described in detail in Table 1. Note that some encoding issues were encountered while processing the data, resulting in the omission of some posts (0.65% of total). The long time span covered by the dataset allows us to investigate the increase in links with related structured data available, and the size of the dataset enables us to extract enough links to study the usefulness of the data. The message boards are moderated so any spam or advertisement links have been removed from the messages, and consequently from the dataset used in our analysis.

Start Date	12/2/1998
End Date	13/2/2008
Users	138,139
Forums	967
Threads	657,169
Posts	7,794,495
External Links	625,723

Table 1: Basic properties of the SIOC Data Competition dataset.

[4] `http://data.sioc-project.org/`



(a) (b) (c) (d)

Figure 4: Evolution of the dataset, by year: (a) Posts, threads, and forums; (b) Links, unique links and domains linked to; (c) Percentage of posts containing links; and (d) Average link count for all posts which contain links

Figure 4(a) shows the growth of the message board in terms of posts, threads and forums. Note that Year 1 corresponds to the first twelve month period covered by the dataset, *i.e.* 1998/1999 and Year 10 corresponds to the last, *i.e.* 2007/2008. Usage of the message board has grown steadily, and in 2007/2008, approximately 2.2m new posts were created. For the remainder of this section, we focus specifically on studying the properties of the links posted over the span of the dataset.

How often do users post links. Figure 4(b) shows the growth of the number of external hyperlinks, in terms of total number of links, unique links, and domains linked to. We exclude syntactically invalid links and links internal to the domain of the message board. Note that we did not resolve all links, therefore multiple distinct links may point to the same resource (or no valid resource). Figure 4(c) shows the percentage of posts containing links per year, and Figure 4(d) shows the average number of links in each of these posts. These figures show that over time, users are more frequently including links in their posts, and they are also including more links per post. Thus links are becoming more prevalant in user conversations.

This shows a shift in user behaviours - in the initial stages of the message board, user attention was held almost entirely based on the text that participants created, whereas now users much more frequently make use of external resources to enrich their posts. One probable factor is that nowadays, there is simply more content available to link to. If a fan starts a thread about their favourite band, it requires very little effort to embed a video from Youtube, whereas finding a video online several years ago would have been much more difficult. A technical aspect that may have impacted linking habits is that message board software has improved over the years, and the task of inserting a link has become easier. The increasing practice of augmenting posts with links to external information suggests that perhaps in the future, an automatic method for doing this could be popular. An example of a service performing such a task is Zemanta[5], an engine for blogs which analyses posts content and suggests related tags and webpages.

Where in conversations do users post links. It is not only the frequency of links that is notable in online conversation, but also the stage at which they occur. Table 2 shows the percentage of posts with links, for all posts, for first posts of threads, and for subsequent posts. In first posts, where a conversation is being initialised, there is a far higher rate of link posting. This suggests that the link is important for the entire conversation which follows. We argue that if we identify the *object(s)* described in that first link (movie, book, musicians, etc.), that would then enable us to have a better understanding of the focus of the conversation, and the main topic discussed in the rest of the thread.

Posts	% with link(s)
All posts	6.55%
First posts	14.61%
Subsequent posts	5.84%

Table 2: Percentage of posts containing links, divided according to whether they are first in a thread.

What do users post links to. We explored the domains which are most often linked to in the dataset and how their popularity changes over time. Table 3 shows the most popular sites linked to in 2002/2003, and Table 4 shows the most popular sites linked to in 2007/2008. We have labelled each site with a description of the type of content most commonly linked to on that site - note that this may not be the only type of content on the site, but it is the type that the message board's users usually linked to at this time.

The list of most popular domains during the early years of the message board is very different from the list of popular domains from recent years. Most notably, there has been a shift from unique, read-only sites created on Web hosting services, towards collaboratively-created read-write or content-sharing sites with many contributors. These collaborative sites make it easier for users to put content online without requiring technical skills. Collaborative and sharing sites also typically feature an innate structure, and they often have associated APIs or Linked Data to enable data reuse, thereby yielding much more semantically rich data than traditional sites. The additional information provided by this external structured data could be used to employ new methods for community detection and social network analysis.

3. ADDING STRUCTURED DATA TO CONVERSATIONS

In this section, we first characterise the structured data commonly available on the Web. We then detail the structured data which we obtained related to posts in the message board dataset.

[5]http://www.zemanta.com/

Rank and Domain	Dominant content type
1. bbc.co.uk	news media
2. komplett.ie	shop
3. ireland.com	news media
4. eircom.net	Web hosting
5. yahoo.com	news/discussion groups
6. rte.ie	news media
7. google.com	Web search
8. geocities.com	Web hosting
9. iol.ie	Web hosting
10. microsoft.com	technical support

Table 3: Top 10 domains of 2002/2003.

Rank and Domain	Dominant content type
1. youtube.com	video-sharing
2. wikipedia.org	collaborative encyclopedia
3. komplett.ie	shop
4. myspace.com	social networking/music
5. flickr.com	photo-sharing
6. bbc.co.uk	news media
7. rte.ie	news media
8. carzone.ie	shop
9. photobucket.com	media hosting
10. ebay.ie	shop

Table 4: Top 10 domains of 2007/2008.

3.1 Overview

While people generally link to plain HTML pages, in the past decade many websites have begun to provide structured data, including some of the popular sites from Table 4 such as Amazon, Wikipedia, Flickr and YouTube. As a side effect of this growing amount of structured data on the Web, the links which users post on social media sites are increasingly often to resources which are available as structured data. Users are thereby providing an implicit machine-readable description of the topic of conversations. We consider three different layers of structured data sources in online communities, ordered from the most expressive to the least expressive as follows:

1. **RDF Data:** Rich descriptions of content provided directly in RDF and back-ended by shared semantics using common ontologies [5]. Such data can be either directly published in RDF (often using the Linked Data principles [3]), embedded in Web documents (*e.g.* in RDFa), accessible via SPARQL endpoints or downloadable RDF dumps.

2. **API Data:** Less rich data can also be obtained by querying APIs of Web services. In that case, the description is generally provided using XML or JSON, and can be easily converted to RDF, though it requires additional effort. Such translation is generally done either (1) via XSLT transformations used to directly convert XML data to RDF/XML or (2) GRDDL to extract RDF data from microformatted XHTML documents or (3) by dedicated scripts that translate output data, *e.g.* JSON, in RDF, as done for instance by the Flickr-SIOC exporter or by SemanticTweet[6]. In that case, appropriate ontologies should also be identified to enable such translation.

3. **Tag Data:** This less rich data includes tags, plain text annotations whose meaning can be ambiguous. They are however an additional source of metadata that can be represented using ontologies, and methods exist to perform automatic grounding of tags to Semantic Web concepts [1], [12], [20]. Tag data is a subset of API data, but we consider the two separately in order to differenciate between structured API data such as genres and dates, which are easily mapped to RDF triples, and tags, which contain no structured information.

It is also worth noting that cases 2 and 3 involve matching simple text-strings (either tags or text output from the APIs) to resources, which implies that one has to deal with ambiguity and heterogeneity issues and that consequently the accuracy of the mappings may not be 100% correct.

3.2 Empirical investigation of hyperlinks

In this study we focused on the structured data sources that most commonly occur in the `boards.ie` dataset. We parsed identifiers from URLs in order to access corresponding data from the external sources. We made use of data from each of the three layers mentioned in Section 3.1. Table 5 shows the most commonly occurring domains in the whole dataset. Domains for which structured data is available are labelled according to whether they are available in RDF, or only via an API. In most of these cases the RDF data available is in fact provided by a third-party export or wrapper of the original website/API. The BBC website is a notable exception - they provide RDF data natively [16].

Rank and Domain	
1. **youtube.com (API)**	16. **ebay.co.uk (API)**
2. **wikipedia.org (RDF)**	17. yahoo.com
3. komplett.ie	18. **amazon.co.uk**
4. **bbc.co.uk (RDF)**	19. google.ie
5. **myspace.com (RDF)**	20. blogspot.com
6. rte.ie	21. play.com
7. carzone.ie	22. **ebay.com (API)**
8. google.com	23. bebo.com
9. photobucket.com	24. **guardian.co.uk (API)**
10. **flickr.com (API)**	25. adverts.ie
11. microsoft.com	26. overclockers.co.uk
12. eircom.net	27. geocities.com
13. **ebay.ie (API)**	28. dell.com
14. imageshack.us	29. ireland.com
15. **imdb.com (RDF)**	30. **amazon.com (API)**

Table 5: Top 30 domains linked to in the message board. Bold font indicates some content is available in RDF as Linked Data or else via an API.

From the domains in Table 5, we chose a subset which are most useful for analysis. We omitted `ebay.*` because product data is removed after 3 months, therefore data about the products linked to in our dataset is now inaccessible. We also omitted `bbc.co.uk` because while the BBC publishes useful structured data about their music and TV programs, the links occurring for this domain in the message board dataset are predominantly pointing to news articles, for which there is not yet structured information available. Finally, we omitted `guardian.co.uk` because it is not generally possible to

[6] `http://semantictweet.com/`

extract an identifier for a news article based on its URL, therefore we cannot use hyperlinks to construct a query to the Guardian API. We combined data from `amazon.com` and `amazon.co.uk` since they contain overlapping information about products.

Based on the remaining domains listed in Table 5, we enriched the message board dataset with data from the following sources:

1. **RDF data.** The original Boards.ie dataset was available as an RDF dump, and we also used a dump of DBpedia [2] to extract information about Wikipedia articles posted. We also retrieved data from SPARQL endpoints - LinkedMDB [10] (an export of IMDB movie-related data), DBTune[7] (a wrapper of Myspace music data), and the RDF Book Mashup [4] (a mashup of Amazon and Google Base book data).

2. **API data.** Some of the data we extracted is in the form of fixed categories which do not have URIs. The Amazon and YouTube APIs describe product groups and categories in simple text. The DBTune genres are assigned URIs from the Myspace Ontology[8] but this is currently a work in progress and the URIs do not resolve. We semi-automatically mapped these three taxonomies to DBpedia - specifically, we checked each item for a DBpedia resource with a matching label, then inspected the results and manually changed any incorrect mappings. An additional API that we used in order to gather more data was LibraryThing's service thingISBN[9], which takes an ISBN as input and returns additional associated ISBNs (*e.g.* other editions of a book).

3. **Tag data.** Much of the data that we extracted is contained in tags. The data sources that we retrieved tags from are Amazon, the RDF Book Mashup, Flickr and YouTube. We used these tags to supplement the semantic data we retrieved by mapping them to DBpedia URIs. We used a naïve method of obtaining corresponding URIs - specifically, for each tag we checked for a DBpedia resource with a matching label. In the cases where the tag matches a resource uniquely (*i.e.*, the resource does not redirect to a disambiguation page) we stored a triple relating the tagged resource to the DBpedia resource. This naïve method of mapping is certain to introduce noise to the data; however URIs which occur for multiple resources in a thread are likely to be relevance to that conversation. We were able to obtain resource mappings for 68% of all unique tags and 82% of all tag instances. Note that we have not evaluated the accuracy of these mappings and we suggest treating these particular relations as potentially useful, but uncertain.

We attempted to map concepts to DBpedia URIs for two main reasons. First, many Linked Data sources provide mappings to DBpedia, increasing the possibilities of interlinking. Secondly, the Wikipedia category structure enables us to expand a set of resources to include the categories to which they belong. This means that if a user links to very specific items, it is still possible to infer the broader topic. An example of this is shown in Figure 5, where we can see that "Amélie" is a film set in Paris, and hence also a film set in France.

[7] `http://dbtune.org/`
[8] `http://purl.org/ontology/myspace`
[9] `http://www.librarything.com/thingology/2006/06/introducing-thingisbn_14.php`

```
imdb:tt0211915 foaf:topic dbpedia:Amélie .
dbpedia:Amélie skos:subject
   category:Films_set_in_Paris .
category:Films_set_in_Paris skos:broader
   category:Films_set_in_France .
```

Figure 5: Example of using DBpedia category structure to infer a broader topic.

Figure 6 shows the percentage of links in the data per year for which we could retrieve external data - *i.e.* where there is currently available a structured representation. We focus on the popular domains which we identified as being likely to have a useful structured data source i.e. `youtube.com`, `wikipedia.org`, `myspace.com`, `flickr.com`, `imdb.com` and `amazon.*`. Note that for the early years (1998-2005), very little data is available, but more recently the amount of structured data has grown rapidly. In total, we could retrieve structured data for links in 24,264 posts, the majority of which occurred in the last year of the dataset. For 2007/2008, 9% of links posted were to resources available as structured data. It is very likely that by now the percentage of links with structured representations is much higher.

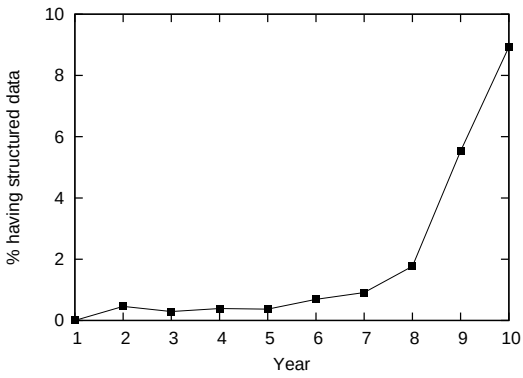


Figure 6: Percentage of links posted for which there is currently available structured data.

Table 6 summarises the six domains that we considered and the sources we used to gather structured data. Sources marked in bold are part of the Linking Open Data project and provide RDF data via wrappers or exporters of existing Web data sources. The remaining unbolded sources provide structured data in non-RDF formats, typically XML, which we transform into RDF. The table also contains figures regarding the quantity of data which we obtained. We found that users have linked to over 20k unique resources with structured represention. Of these, almost half (9.8k) are published as part of the Linking Open Data project. Table 6 also shows some examples of the relationships that we extracted. For these relationships alone we can create or retrieve over 100k RDF triples.

Resource Type and Domain	Structured data sources	Total Links	Distinct Items	Sample Relationships Extracted	Triples	Distinct Objects
Video `youtube.com`	YouTube API	10,543	7,773	Tag (mapped to DBpedia) Category Date published	52,385 7,773 7,773	12,590 15 7,773
Article `wikipedia.org`	**DBpedia**	9,248	6,240	Article about Resource	7,946	7,946
Music Artist `myspace.com`	**DBTunes**	4,383	1,793	Genre of Resource	4,458	103
Photo `flickr.com`	Flickr API	1,741	1,437	Tag (mapped to DBpedia) Date published	5,968 1,437	1,949 1,437
Movie `imdb.com`	**Linkedmdb**, **DBpedia**	957	540	Article about Resource Genre of Resource Director of Resource	542 886 248	542 46 216
Product `amazon.*`	Amazon API, LibraryThing	3,572	2,781	Tag (mapped to DBpedia) Product Group	17,339 1,582	3,979 25
(of which Book)	**RDF Book Mashup**, **DBpedia**	(1,540)	(1,192)	Article about Resource Author of Resource	126 1,144	111 518
	Total	30,444	20,564		109,607	

Table 6: External data sources, number of items extracted from hyperlinks, examples of relationships retrieved and the number of triples available to enrich the original dataset. Sources in bold are part of the Linking Open Data project.

4. UTILISING THE ENRICHED DATASET FOR NOVEL ANALYSIS

We have seen that taking into account external structured data can provide us with new related information, but what can we learn from it? Since the enriched data is stored in RDF, SPARQL provides an easy way to access the data for different kind of analysis. Consequently, all of the analysis that we performed and which we describe in this section made use of SPARQL queries (with aggregation — to be supported in SPARQL 1.1) and required minimal subsequent processing.

4.1 Analysis of post content

Making use of information from the Web of Data allows us to examine how posters describe a resource which they link to. We can access facts about resources, such as their title, creator or category, and compare these to the text accompanying the link.

We retrieved book and movie titles, and author and director names from structured data sources, where available, and checked whether the names occurred in the text content. The results are shown in Table 7. We observe that posters are far more likely to use the name of a movie than the name a of a book when linking to it. However they are more likely to mention the author of a book than the director of a movie.

The use of external structured data allows us to easily perform this sort of analysis of user behaviour, which otherwise would have to be performed by manually inspecting posts. We can also see from Table 7 that making use of links to structured data allows us to add new information that is related to the hyperlinks within the post, but was not explicitly mentioned in the post text. For example, if we wanted to detect mentions of movies in message board posts, searching for movie titles alone would result in missing hundreds of posts where there is a link to the movie on IMDB.

Books (1,540 links)		
Information Item	Location in Post	Times Mentioned
Book Title	Post	23.37%
	Anchortext	9.52%
	Post Title	0.68%
Book Author	Post	24.67%
	Anchortext	6.26%
	Post Title	0.91%
Movies (957 links)		
Information Item	Location in Post	Times Mentioned
Movie Title	Post	65.89%
	Anchortext	34.01%
	Post Title	9.24%
Movie Director	Post	4.81%
	Anchortext	1.20%
	Post Title	0.00%

Table 7: For links to books and movies, the percentage of times that the title/author/director was also mentioned in the post text.

4.2 Analysis of content sharing behaviour

By retrieving metadata about items posted, we can learn more about how different types of media are shared among social media users.

We retrieved the upload date of images and videos linked from users posts, and compared them to the dates when they were posted on the message board. This allows us to compare the typical age of different types of media posted. The distribution of the ages of images and videos are shown in Figure 7. We observe that posters are much more likely to post extremely new Flickr images than Youtube videos. The median age of a Flickr photo posted is only 2 days, but the median age of a Youtube video posted is 111 days. Reading the posts shows that this is because users often share their own, just uploaded photos, but when sharing videos they are more likely to share content they have found which has

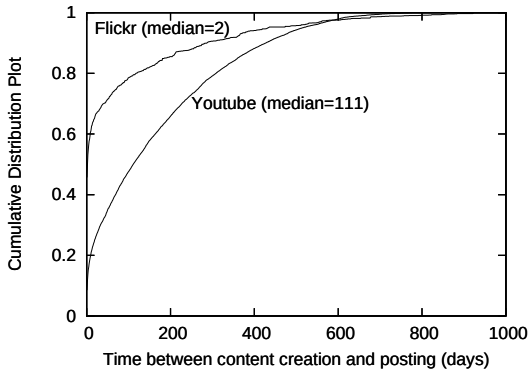


Figure 7: Age distribution of videos and images posted on the message board.

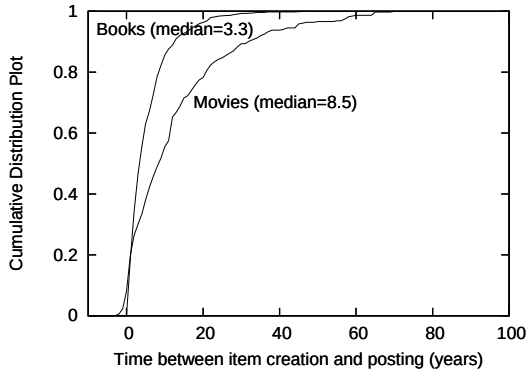


Figure 8: Age distribution of books and movies posted on the message board.

been created by someone else.

In Figure 8 we show a similar distribution plot, but in this case comparing books and movies, which require a time-scale of years rather than days. While all types of content items show a similar pattern of many links to new items, tailing-off to a few links to old items, they differ in the timescale involved. For example, approximately 80% of movie and book related discussions are about items which are over 1 year old, whereas links to user-generated content are typically days old. This suggests that studies of information propagation could benefit from considering the different properties of objects which are shared online - information which is often available on the Web in a structured format.

Related analysis has been performed as part of a study [9] of the spread of media content in the blogosphere. They combined a blog dataset [8] and metadata gathered from Youtube and compared propagation of videos for of different categories, finding for example that news videos spread much faster than music videos.

4.3 User profiling

We can aggregate all the structured data associated with posts of a certain user and build a set of related concepts. This enables us to mine user profiles based on their social contributions. As an example, Figure 9 displays the information extracted about a user profile based on the resources posted by this paricular individual, and the resources related to these (e.g. mapped from tags). We can also include the categories of these resources derived from the Wikipedia hierarchy, but have omitted these for clarity.

In addition, these profiles can be modelled using RDF(S)/ OWL, enabling their reuse in other applications than the system where they come from. One straightforward way to to so is to use the `foaf:topic_interest` property to link users to their interests. However, this does not consider the frequency regarding how these interests have been extracted, neither if they have been extracted explicitly (e.g. direct mapping to a structured resource) or implicitly (inferred from other categories or from tags).

Hence, to enable such a finer-grained representation of user-profiles, we designed a lightweight model, named the Interest Mining Ontology — available at `http://rdfs.org/imo/ns` —, consisting of a main class `imo:MinedInterest` representing an interest relationship between (1) a user (instance of `foaf:Person`) — with the `imo:person` property and (2) an interest (instance of `rdf:Resource`) — with the `imo:interest` property[10]. The model also relies on SCOVO[11] — Statistical Core Vocabulary [14] — and especially uses `scovo:Item` to represent the number of times an interest has been mined from a dataset. Finally, order to represent the mining more finely, our model includes two classes inspired by the APML[12] — Attention Profiling Mark-up Language — terminology in the realm of user interests:

- `ExplicitMining` – when the matching has been done with via a direct link to the resource or by direct translation (*e.g.* from a Wikipedia page to a DBpedia URI);
- `ImplicitMining` — when the accuracy of the mining is uncertain, *e.g.* when URIs have been mined from a tag set or using some disambiguation heuristics.

Then, considering our previous example, we can represent Bob's mined interests like so: (1) Science Fiction: mined one time using an explicit mapping and one time implicitly using a particular heuristic (in this case mapped from a tag) and (2) Animation: mined once via an implicit link (again, mapped from a tag). Bob's user profile represented with IMO is shown in Table 8.

5. DISCUSSION

Our approach of enriching social media with external data, via hyperlinks, could be integrated with existing methods which use Semantic Web capabilities to study online communities. Mika's work on ontologies [17] is probably one of the first studies that combined Semantic Web and social network analysis. He provided a social model for ontology mining from folksonomies, and showed how communities of users are formed based on the tag clusters that they use and some subsets from these clusters. More recently, Ereteo et al. [11] provided an ontology of social network analysis[13] and analyzed a social network of more than 60,000 users. However, they relied only on the interactions happening inside

[10]Moreover, we mapped the model to FOAF (and to `foaf:topic_interest`) using an online property chain, available in the model online.

[11]`http://sw.joanneum.at/scovo/schema.html`

[12]`http://apml.areyoupayingattention.com/`

[13]`http://ns.inria.fr/semsna/2009/06/21/voc`

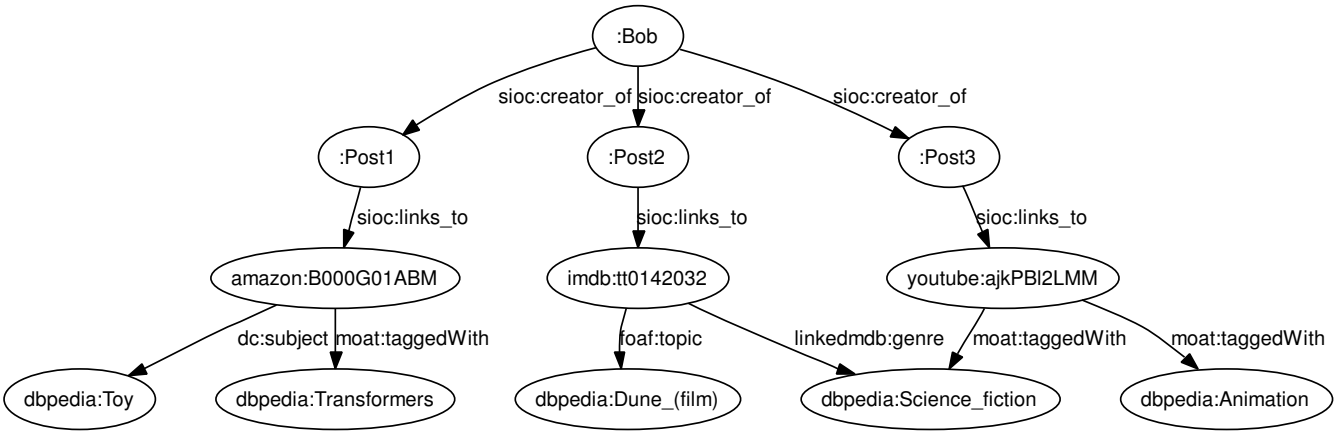


Figure 9: Example of how the links that a user shares can be used to link the user to his interests.

```
@prefix : <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix imo: <http://rdfs.org/imo/ns#> .
@prefix scovo: <http://purl.org/NET/scovo#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-
    syntax-ns#> .

[] a imo:MinedInterest ;
 imo:user :Bob ;
 imo:interest dbpedia:Science_Fiction ;
 imo:statItem [
  scovo:dimension imo:ExplicitMining ; rdf:value 1
 ] ;
 imo:statItem [
  scovo:dimension imo:ImplicitMining ; rdf:value 1
 ] .

[] a imo:MinedInterest ;
 imo:user :Bob ;
 imo:interest dbpedia:Animation ;
 imo:statItem [
  scovo:dimension imo:ImplicitMining ; rdf:value 1
 ] .
```

Table 8: Using IMO to represent one's interests.

the platform, hence not taking into account the additional expressiveness of linked content as we show in our work.

As well as being useful for understanding online communication, the work presented in this paper can also be used in various application contexts. The most simple example of this is that the user interface of a message board could be enhanced with external data, similar to what SuperTweet[14] does for Twitter posts. A related recent initative is Facebook's Open Graph Protocol[15], which allows external site owners to markup their content using Facebook-defined schemas, such that these enriched content items can then be used for metadata import into news feeds, profiles, etc. This will increase the amount of links to objects available online, as users will make use of this service or similar ones to easily share links with friends, instead of just referring to an object in plain text. It will also have the useful side-effect of increasing the amount of structured data available online, as publishers will format their data in such a way that it can easily be consumed by Facebook.

Taking the idea of enhanced interfaces a step further, content recommendation could be performed based on semantic links derived from information which is not explicitly mentioned in the post. For example when a user creates a post about a movie, it would be possible to suggest posts which mention movies having the same director. That way, we can mine additional relationships inside the bulletin board by using external data from the LOD cloud. In addition to recommending internal data, these links can be used for instance for targeted advertising, by providing adverts to related movies, books, etc. It could also be combined with other recommendations systems that use the same URIs to define their recommendations, as for instance dbrec[16].

Moreover, these links could also be used to interlink bulletin boards with other content from the Social Web sphere. We can imagine that semantically-enhanced microblog posts, using systems such as SMOB[17] or HyperTwitter[18], could be aggregated and displayed when browsing forums, so that in addition to the current conversation, users could see what is happening in real-time regarding the same topic.

One advantage of using RDF for representing social media data is that SPARQL queries can be used and reused for analysis of such datasets. The current version, SPARQL 1.0 [19] is limited by a lack of support for aggregation, an important requirement for data analysis in general. We therefore must depend on SPARQL extensions, specifically aggregation, to perform our analysis. The previously mentioned work of Ereteo et al. on analysis of social semantic data similarly relied on SPARQL extensions. San Martín and Gutierrez [18] also illustrated how SPARQL could be extended for Social Network Analysis purposes. It is worth noting that the extensions we used and some of those presented in these two recent works shall be included in the upcoming release of SPARQL 1.1 [13], such as aggregates and property paths (currently time-permitting) and that others are now available thanks to OWL2 [21], notably property chain axioms. In the future, the analysis we now present will be supported by SPARQL and the approach will therefore be easily replicated using W3C standards.

[14] http://www.cascaad.com/supertweet.php
[15] http://opengraphprotocol.org/
[16] http://dbrec.net
[17] http://smob.me
[18] http://semantictwitter.appspot.com/

6. CONCLUSION

In this paper, we present a series of steps that can provide additional useful information to consumers of social media contributions that contain hyperlinks, based on a trend towards more links to semantically-rich items on collaborative websites. We have studied ten years of hyperlinks posted on a large bulletin board site. Our analysis shows that links have become more frequent in postings throughout these ten years, and that the the amount of structured data available which relates to these links has sharply increased. A useful side effect of this trend is that there becomes available vast amounts of structured data relating to online conversations, without requiring any special input from the users. We have established that the data retrievable from external sources provides new information not available from text analysis and can give an insight into how people communicate online. We conducted our analysis on a bulletin board site, but the approach could be applied to any type of online conversations (e.g. blogs, real-time status updates). The approach has possible applications in recommendation or filtering of content, as well as providing the potential for social scientists to investigate individual online community activity or the connections between communities based on similar or shared sets of linked-to resources.

7. REFERENCES

[1] S. Angeletou. Semantic Enrichment of Folksonomy Tagspaces. In *ISWC 2008: Proceedings of the 7th International Semantic Web Conference*. Springer, 2008.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A Nucleus for a Web of Open Data. In *ISWC 2007: Proceedings of the 6th International Semantic Web Conference*. Springer, 2007.

[3] T. Berners-Lee. Linked Data. Design Issues for the World Wide Web, World Wide Web Consortium, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`.

[4] C. Bizer, R. Cyganiak, and T. Gauß. The RDF Book Mashup: from Web APIs to a Web of Data. In *SFSW2007: Proceedings of the 3rd Workshop on Scripting for the Semantic Web at ESWC 2007*, 2007.

[5] C. Bizer, R. Cyganiak, and T. Heath. How to Publish Linked Data on the Web. Technical report, 2007. `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`.

[6] J. Breslin, R. Kass, and U. Bojārs. The Boardscape: Creating a Super Social Network of Message Boards. In *ICWSM 2007: Proceedings of the 1st International Conference on Weblogs and Social Media*, 2007.

[7] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. In *ESWC 2005: Proceedings of the 2nd European Semantic Web Conference*. Springer, 2005.

[8] K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r Dataset. In *ICWSM 2009: Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. AAAI, 2009.

[9] M. Cha, J. Pérez, and H. Haddadi. Flash Floods and Ripples: The Spread of Media Content through the Blogosphere. In *ICWSM 2009: Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. AAAI, 2009.

[10] M. P. Consens. Managing Linked Data on the Web: The LinkedMDB Showcase. In *LA-WEB 2008: Proceedings of the 6th Latin-American Web Congress*. IEEE Computer Society, 2008.

[11] G. Ereteo, M. Buffa, F. Gandon, and O. Corby. Analysis of a Real Online Social Network using Semantic Web Frameworks. In *ISWC 2009: Proceedings of the 8th International Semantic Web Conference*. Springer, 2009.

[12] A. Garcia, M. Szomszor, H. Alani, and O. Corcho. Preliminary Results in Tag Disambiguation using DBpedia . In *CKCaR'09: Proceedings of the 1st International Workshop on Collective Knowledge Capturing and Representation at K-CAP 2009*, 2009.

[13] S. Harris and A. Seaborne. SPARQL 1.1 Query. *W3C Working Draft, W3C (October 2009)* `http://www.w3.org/TR/sparql11-query/`.

[14] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. SCOVO: Using Statistics on the Web of Data. In *ESWC 2009: Proceedings of the 6th European Semantic Web Conference*. Springer, 2009.

[15] K. Knorr-Cetina. Sociality with objects: Social relations in postsocial knowledge societies. *Theory, Culture and Society*, 14(4):1–30, 1997.

[16] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web – how the BBC uses DBpedia and Linked Data to Make Connections. In *ESWC 2009: Proceedings of the 6th European Semantic Web Conference*. Springer, 2009.

[17] P. Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In *ISWC 2005: Proceedings of the 4th International Semantic Web Conference*. Springer, 2005.

[18] M. San Martín and C. Gutierrez. Representing, Querying and Transforming Social Networks with RDF/SPARQL. In *ESWC 2009: Proceedings of the 6th European Semantic Web Conference*. Springer, 2009.

[19] A. Seaborne and E. Prud'hommeaux. SPARQL Query Language for RDF. *W3C recommendation, W3C (January 2008)* `http://www.w3.org/TR/rdf-sparql-query/`.

[20] L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web. In *ESWC 2007: Proceedings of the 4th European Semantic Web Conference*. Springer, 2007.

[21] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview. *W3C recommendation, W3C (October 2009)* `http://www.w3.org/TR/owl-overview/`.