Ten Years of Hyperlinks in Online Conversations*

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

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

ABSTRACT

Since social media sites have existed, a significant feature 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. Understanding linking behaviour is an important part of understanding the dynamics of online conversations, and identifying the changing interests of participants in these discussions. In this paper, we present a study of the hyperlinks posted on a large-scale message board dataset over a ten year period. We analyse the occurrences of hyperlinks and find that the practice of posting links has become more frequent over the lifetime of the message board. We examine the domains linked to by users, and see that they change greatly over the years. We focus in particular on the the increase of links to resources with associated structured data, and discuss the potential for using this data for enhanced analysis of online conversation.

Keywords

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

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]. On sites with limited post lengths such as Twitter, hyperlinks allow users to enhance content with detail and context that would otherwise not be possible.

As well as references to other documents, these links can be seen as social objects according to the object-centred sociality theory of Karin Knorr-Cetina [7]: 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.

Increasingly these online artifacts are being described by the data publishers using structured data as well as in a human-readable form. Some websites provide APIs to their data as web services, and others make information available in RDF (Resource Description Format) [8] through the Linking Open Data initiative[3]. 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 objects which are represented in structured data. Users are thereby providing an implicit machine-readable description of the topic of conversations. Hence, this readily available structured data has the potential to play an important role in researching social media. One example of research exploiting structured data in social media is the work of Cha et al. [4] which investigated the propagation of different types of YouTube videos in the blogosphere, based on category and creation date metadata.

In this paper, we investigate the practice of posting links in online conversation. We study how often users post links, where in conversations they post links, and what they link to. We examine how these habits have changed over a ten year period. In particular, we look at the increase of links to websites for which there is structured data available. Our analysis is based on data from the `http://boards.ie` bulletin board, the most popular Irish bulletin boards service, including ten years of conversations.

2. DATASET

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 [2]. SIOC is a vocabulary designed for enabling the representation and interconnection of online communities such as message boards, wikis and mailing lists. The FOAF (Friend-of-a-Friend) [3] and DC (Dublin Core) [9] vocabularies were also used to represent

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

[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

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

[3] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

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

elements of the dataset. Semantic Web technologies such as these are important for social media analysis because the use of structured common formats makes the task of extracting data much easier. The structure of the key elements of the data is shown in Table 1. Since the data is represented using the SIOC vocabulary, it is straightforward to extract links using the `sioc:links_to` property.

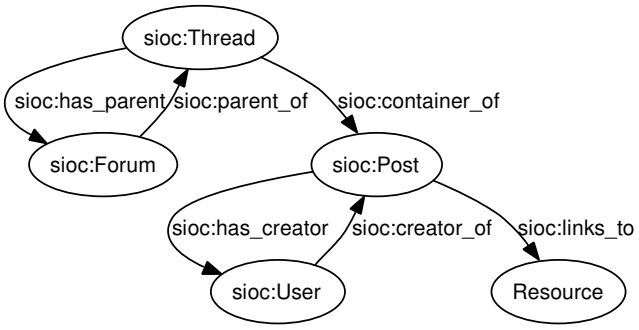

Figure 1: Partial structure of the SIOC dataset.

The data covers the first ten 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 more detail in Table 1. Note that some errors were encountered while processing the data, due to encoding issues, and these may lead to small inaccuracies in our analysis. The message boards are moderated so any spam or advertisement links have been removed from the website, and consequently from the dataset that we 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
Domains linked to	97,605

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

3. RESULTS

In this section, we present inital results of our analysis of the `boards.ie` SIOC Data Competition dataset. 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).

3.1 How often do users post links

We first investigated the prevalence of link posting in the message board. Figure 2 shows the percentage of posts containing links per year, and Figure 3 shows the average number of links in each of these posts. 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. These figures show that over the ten year span of the dataset, users were including links in their posts with an increasing frequency, and they were also including more links per post. Thus links are becoming more common in user conversations.

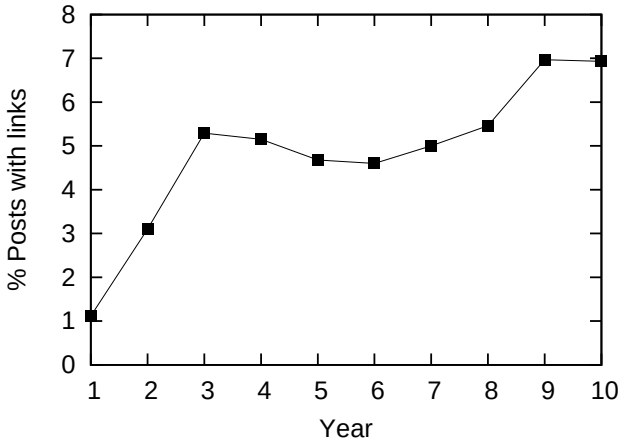

Figure 2: Percentage of posts in the dataset that contain links by year.

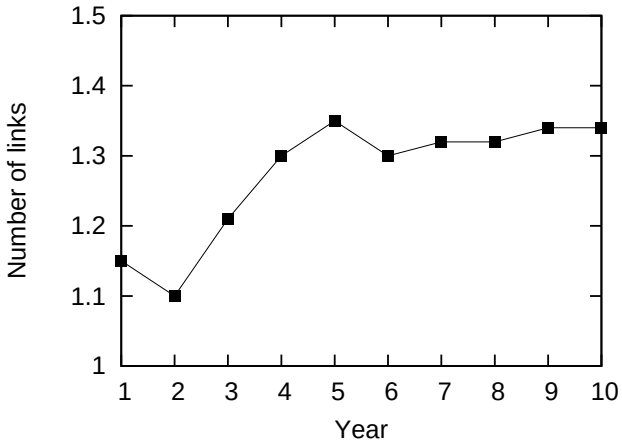

Figure 3: Average link count for posts that contain links by year.

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 on the Web 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 into a post 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 the content of posts and suggests related tags and webpages.

3.2 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 - 14.6% compared to 5.8% for other posts.

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

Table 2: Percentage of posts containing links.

This presence of a link in the first post of a thread suggests that the object linked to 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.

3.3 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 very 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.

3.4 Structured data in conversations

Several of the most popular domains linked to in the year 2007/2008 and listed in Table 4 are available online in a structured form, either via an API or as RDF data - specifically, `youtube.com`, `wikipedia.org`, `myspace.com`, `flickr.com`, `bbc.co.uk`, and `ebay.ie`. We identified all of the hyperlinks in the message board dataset which point to

[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 ten external domains linked to in 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 ten external domains linked to in 2007/2008.

one of these domains, extracted identifiers from the URLs, and attempted to retrieve the associated structured data for each resource.

Figure 4 shows the percentage of links in the data per year for which we could retrieve external data - *i.e.* where there is currently available an equivalent structured representation. Note that for the earlier years, the structured data was not necessarily available at the time the link was posted. For hyperlinks posted during the early years (1998 - 2005), even now very little structured data is available, but more recently the amount of structured data has grown rapidly. 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 considerably higher.

Figure 5 shows an example of a post which can be complemented by data from external sources (in this case LinkedMDB [5] (an RDF export of IMDB movie-related data) and DBpedia [1] (an RDF export of Wikipedia data). Rather than just a URL, this hyperlink represents a social object, in this case a movie, and the related structured data provides properties of the object, enabling us to find out that it is a comedy from 2001, titled "Amélie" and directed by Jean-Pierre Jeunet. The extra information contained in these sources provides additional context which can be useful for both analysis of online conversation, and for developing related applications.

Another approach for analysing online conversations is using Natural Language Processing (NLP) algorithms to extract entities, topics and relationships from textual content

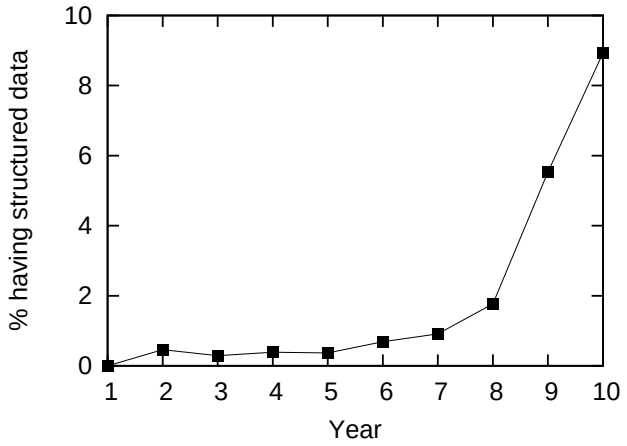

Figure 4: Percentage of links posted for which there is currently available structured data.

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.*

Subset of Structured Data Available:

Title	Amélie
Year	2001
Genre	Comedy
Genre	Romance
Actor	Audrey Tautou
Director	Jean-Pierre Jeunet

Figure 5: Example of a real message board post, and external structured data relating to it, from LinkedMDB and DBpedia.

generated by users. However when dealing with social media sites, performing NLP can be particularly difficult due to the typically informal nature of user posts, which tend to contain a lot of slang and context-dependant terms, with little attention given to spelling and grammar [6]. Thus, while NLP algorithms are potentially very useful tools for investigating social media sites, there are challenges particular to user-generated content which must be handled for named entity identification [6]. In addition, considering for instance microblogging sites such as Twitter, it appears that some posts contain no text apart from one link and perhaps a few hashtags, which make these approaches even more difficult to apply, and showcases the need to identify what object is described in these links. It could also be useful to combine semantic data with methods from NLP, for example to determine the sentiment of a post which mentions a particular entity.

4. CONCLUSION

Conversations in social media give a record of changes in Web usage over time, allowing us to study trends such as the move towards more collaboratively-created websites and more structured data, and also propose future trends such as automatic link suggestions. 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. This enables us to retrieve information about what the links mean, *i.e.* regarding what are the *objects* (as referred to in the object-centred view) being discussed in a conversation. Future work will focus on using the meaning of the structured data to characterise social objects in online conversations and exploiting this information to better understand online communities and communication.

5. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*. Springer, 2007.

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

[3] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.97. *Namespace Document 1 January 2010* `http://xmlns.com/foaf/spec/`, 2010.

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

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

[6] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. Context and Domain Knowledge Enhanced Entity Spotting In Informal Text. In *Proceedings of the 8th International Semantic Web Conference (ISWC 2009)*. Springer, 2009.

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

[8] F. Manola, E. Miller, and B. McBride. RDF primer. *W3C Recommendation* `http://www.w3.org/TR/rdf-primer/`, 2004.

[9] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery. *Internet Engineering Task Force RFC*, 2413, 1998.