Analyzing Clusters and Constellations from Untwisting shortened links on Twitter using Conceptual Graphs

Emma L. Tonkin[1], Heather D. Pfeiffer[2], and Gregory J. L. Tourte[3]

[1] UKOLN,
University of Bath, Bath, UK,
e.tonkin@ukoln.ac.uk
[2] Akamai Physics, Inc.
Las Cruces, New Mexico, USA,
hdp@cs.nmsu.edu
[3] School of Geographical Sciences,
The University of Bristol, Bristol, UK
g.j.l.tourte@bristol.ac.uk

Abstract. The analysis of big data, although potentially a very rewarding task, can present difficulties due to the complexity inherent to such datasets. We suggest that conceptual graphs provide a mechanism for representing knowledge about a domain that can also be used as a useful scaffold for big data analysis. Conceptual graphs may be used as a means to collaboratively build up a robust model forming the skeleton of a data analysis project. This paper describes a case study in which conceptual graphs were used to underpin an exploration of a corpus of tweets relating to the Transportation Security Administration (TSA). Through this process we will demonstrate the emerging model built up of the data landscape involved and of the business structures that underlie the technical frameworks relied upon by microblogging software.

Keywords: Conceptual Graphs, Twitter, Microblogging, Models

1 Introduction

The increasing prominence of Twitter as a social site in the last years has led to a great deal of interest in the way in which the site is used, as well as the technical enablers underlying the site and its applications. Of course, there is a significant existing body of literature describing various aspects of the site and the characteristics of its use, e.g., [11,12], so a researcher looking to analyse data taken from Twitter should naturally begin by reviewing that information. However, one aspect of Twitter is its apparent inconsistency across topics [30] and across cultures (see for example [39]). Another is the changing landscape of technologies and implementation decisions: as a developing platform seeking to marketize effectively, Twitter, like most services, evolves over time as a result

of various motivating factors. As a consequence there is a need for exploratory data analysis [22].

Typically, the exploratory analysis of data (EAD) involves the use of information visualisation tools, cluster analysis, data mining approaches and so forth [22], which permits domain experts to begin to develop an understanding of the dataset at hand. This permits them to develop testable hypotheses. However, as Perer notes, one difficulty with this approach is that it is typically somewhat scattershot—discoveries made in this way are typically opportunistic. Yet an entirely systematic approach risks undermining the knowledge-driven, insight-led research pattern of domain experts. Generally, Perer suggests, systematic approaches do not always suffice when faced with real problems. Thus, Perer suggests, a series of design goals should be considered when developing data exploration interfaces: most relate to the ability to track actions already taken, to see available actions not already taken, to annotate actions, to retrace existing steps taken, and so on. Particularly interesting is design goal 6: the need to share progress with other users.

We begin this paper by exploring knowledge representations through which information learned about the entities, agencies, interactions and underlying infrastructure of the Twitter environment can be stored and shared within a team to support EAD, explaining why we chose to use conceptual graphs for the purpose of supporting a text mining application. We explain the development philosophy underlying the EAD approach taken and its limitations, and we provide a brief introduction to the literature surrounding Twitter and findings resulting from a preliminary exploration of certain aspects of Twitter infrastructure.

2 Method

Involving each member of a research group into an iterative process of data model development requires both appropriate communication channels and sufficiently useful proxies (e.g. imagery, model diagrams, etc) on which to work. There are many candidates for this process, of course, ranging from pen and paper or whiteboard to a shared collaborative space online such as a wiki, Google Doc, or a version or revision control system such as Subversion. However, it is important to separate the collaborative space that is used from the actual representation that is employed within that workspace, and to recognise that such aids to teamwork, whilst innately prerequisite, typically provide neither formalism nor guidance. It is for this purpose that a formal knowledge representation structure becomes of importance; according to Davis et al [6], KR may be described in terms of five roles: a KR may act as a proxy through which via thought-experiment the effects of an action may be deduced; a KR represents a series of ways of thinking about an entity; a KR represents a formalism expressed in terms of sanctioned and recommended inferences; a KR can be seen as 'a computational environment for thinking' and as 'a medium of human expression'. KRs may be classified into five categories: pictorial, symbolic, linguistic, virtual and algorithmic [19].

Given these five categories of KR the internal structure of the data must be able to hold not only factual data, but the conceptual dependencies between the elements so that their relationships are defined within the data structure. These structures can hold scripts [33] of information that is represented textually and can be formed into a story. This story line can then be structurally stored into a commonsense database of records. This database structure could hold three of the categories—symbolic, linguistic, and algorithmic—by using language theory. The two other categories virtual and pictorial would be lacking because the text basis of the scripts. However the conceptual graphs structure, especially with time and space extensions [25], does not have this textual limitation, but does give the relational structuring between conceptual dependency. So can process all five categories.

Conceptual graphs (CGs) provide a formal visual approach to knowledge representation, closely linked to natural language [37] which have been found to be accessible by team participants from varying specialities in the past, including for example visual designers and managers [28], developers [16], engineers [4] and so forth. The graphical representation provided by the CG formalism is an aid to understanding that has in the past been shown to be effective in multidisciplinary team environments [4].

This graphical format of CGs can be represented in textual expression or as links to other types of conceptual information such as URL addresses to photos, videos, games, etc. The CGs as a set of partial models do not have to maintain truth as with other representation so they may contain conceptual relationship or dependencies that are in opposition to other graphs within the same model set. This is because partial models are snapshots in time [25]. When the final model is built all inconsistancies will be resolved.

2.1 Conceptual graphs in text mining

Text mining, an area that remains relatively youthful, is a research area based on the detection/discovery of interesting patterns within textual corpora. Whilst the majority of text mining applications are essentially focused on relatively simple representations—key words/phrases, or even in some cases 'bag of words' representations, the use of conceptual graphs in text mining problems is well represented in the literature. Cao [2] describes the use of conceptual graphs alongside fuzzy logic as a means of extending Semantic Web technologies to approach human expression and reasoning more effectively; conceptual graphs are here used as a means of representing natural language sentences. Montes-y-Gomez et al. [10], for example, describe the use of conceptual graphs to represent a series of text, permitting the detection of rare patterns and local deviations (occurring at specific contexts and generalization levels) within the textual corpus. Spasic et al. [38] identify Daraselia et al's [5] use of conceptual graphs as a representation of a number of ontological frames, permitting them to be queried or for further text mning work to be completed against them. Shehata et al (2006) describe the use of conceptual graph representations to capture in detail sentence-level semantics, in order to improve the quality of text retrieval and indexing [34].

In general terms, then, text mining and conceptual graphs are demonstrably viable companions. However, by no means should this be taken to mean that the problem of mining a research corpus such as the Twitter corpus described here reduces to the use of an existing software package or service. There is significant variation between corpora; Twitter, for example, limits users to a small number of characters per utterance, typically resulting in a telegraphic, abbreviated style.

2.2 Agile development

In this instance the proposal is to use conceptual graphs within the team to build up information about the various aspects of the dataset under investigation. We separated this work into two broad phases, the first one of which is exploratory in nature, and is intended to enable us to rapidly build up a basic model of the domain. For this purpose we use a variant of the agile software development methodology, conceptually linked to the Rapid Application Development models [15]. For the second, we link the conceptual graphs built up during the exploratory work in order to create a single composite knowledge representation, and explore its use as a basis structure on which to develop research questions about the dataset.

We begin by briefly reviewing literature relating to use of agile methodologies in exploration of scientific datasets, and move on to the development of fragmentary conceptual graphs though research findings.

2.3 Agile methodologies for scientific datasets

Agile software development methodologies are designed to prioritise certain aspects of the software development process. The agile manifesto [9] expresses the methodology's practices as follows:

- *Individuals and interactions* over processes and tools
- *Working software* over comprehensive documentation
- *Customer collaboration* over contract negotiation
- *Responding to change* over following a plan

and states that 'while there is value in the items on the right, we value the items on the left more'.

The use of agile development methodologies for the purpose of development of scientific software is a concept that has been explored elsewhere; for example, Lane [13] describes a theory-driven methodology that encodes scientific knowledge and natural processes within an implemented piece of software. The importance of the computational model is clearly stated by Lane [13]: computatoinal models, it is argued, are amongst other things able to clearly and rigorously lay out the components of the scientific theory under discussion, allow the derivation of testable predictions, and provide a useful mechanism to facilitate making sense of rich and dense datasets. Accepting the value of a computational model, it is therefore reasonable to consider the question of the quality, relevance and

accuracy of its implementation. To successfully resolve these queries it is necessary to establish appropriate tests—that is, what makes a model 'good', or 'accurate', in the context of our research?

2.4 Building and testing a conceptual model to underlie research

Due to the idiosyncratic composition of any research team it is reasonable to expect that the precise research interests/requirements of the individuals involved are likely to have an impact on the features highlighted, perhaps even on the inclusion of features. This is not uncommon; indeed, the process of mining a text is typically starkly reductive—reduction of entropy/compression of a text may be expected to have at its core a model of the aspects of that text most clearly of use. Features of the text that are not contained within that model may or may not survive the reduction. An obvious example of this is the previously mentioned bag-of-words model, that is, reduction of a text into its component words; the details of the syntax, the presentation, etc., cannot be expected to survive this process. If the research requirements of the team may be satisfied by the use of such a model, however, there is no pressing need to turn away from it. Thus the participants must be at the core of model development.

3 Mining a twitter corpus

The use of Twitter as a data source for various forms of data/text mining is well established. The data is usable for a variety of purposes, perhaps most easily classified according to the technologies used. Twitter is famously used as a resource for sentiment analysis and for opinion mining [20], with a variety of purposes in mind, including product/service/company profiling, marketing purposes, political analysis and opinion polling, for example, with activist aims in mind [20], but also for stock market prediction [1], disaster alerts [7], level of interest in news articles [26] and so forth. Explicitly topic-oriented mining is of use for various purposes, such as tracking public health trends [27,21], earthquake monitoring [31], news tracking [14,23] and so on. The very public nature of the service renders it of interest to spammers, and therefore another research topic in the text mining field is that of identifying and mining spam. Shekar et al. [35] demonstrate a mechanism for identifying spam from Twitter data through an initially manually input list of key terms, followed by the use of a Naive Bayesian algorithm and a J48 decision tree classifier. As well as straightforward use as a text corpus, Twitter's sharply time-based, turn-based and telegraphic nature introduces the need to consider issues that perhaps would not be as prominent where other types of information, such as perhaps academic papers or even blogs, are concerned.

Yet the content of users' tweets is certainly not the only aspect of Twitter that may be of interest, and the existing body of research certainly reflects this. Twitter's popular classification as a social network is well established despite senior Twitter executives' protestations that Twitter is 'a news network and

not a social network' [18]. Certainly the findings of Kwak et al [12] bear out the assertion that trending Twitter topics are, in the majority of cases, either 'headline or persistent news' in nature, whilst their topology analysis suggests that Twitter is not a pure social network, as the distribution of followers and low reciprocity does not closely resemble the typical social network. Yet other studies suggest that this is the effect of noise; as Huberman et al. [11], social interactions exist within twitter, as a sparse subset of the broadly declared set of friends and followers. It appears likely that Twitter, sharing aspects both of social networks and the emerging concept of a 'news network', must be modelled in such a way as to satisfy both definitions.

The technical infrastructure underlying Twitter's functionality is also of interest to researchers.

3.1 Infrastructure Underlying Twitter

A particularly important tool for Twitter users in the past has been the URL shortener [3]—a tool, often web based or built in to the application used by the individual to post their remarks to Twitter. These are conceptually simple: a URL is provided to the shortening tool, which assigns to it a unique key; when presented with that key, the tool will then present the browser with some form of redirect (often a 301) to return the user to the original long URL.

The primary benefit for Twitter users was simply that a shortened URL does not eat significantly into the limited space available for each tweet (Twitter's famous 140 characters or less), leaving the user with more space to present their own ideas or opinions. There are also secondary benefits, of course, such as relative opacity (i.e., it is not usually possible to guess at the destination of a shortened URL), making it possible for users to forward readers to unexpected URLs, providing potential for practical jokes and for malicious reuse as well as fulfilling the more general purpose of compressing information.

3.2 Rationale: Construction and maintenance—relative costs?

Since URL shorteners are not technically complicated, they are relatively easy to set up, and indeed a site that tracks URL shorteners in use has identified over a thousand individual services [42]. However, like many other such initiatives the attrition rate of URL shorteners over time appears to be quite high—according to `yi.tl`, the majority of shortening services identified have since closed. As we will discuss in this poster, the majority of shortened URLs from a given US-centric discourse during the spring of 2012 make use of one of a few major service providers, either directly or via aliases run by those providers. One clear advantage of making use of a URL shortener is the opportunity to gain information about the number of clickthroughs—how many people accessed the link that was posted, when, and from which broad geographic region. This is particularly useful to those for whom the distribution of links in a given venue forms part of a marketing strategy—a group in which Higher Education institutions

are increasingly likely to count themselves, as market forces penetrate ever more deeply.

This reasoning also leads the construction of URL shorteners in some domains—indeed, it is not uncommon for parent enterprises to sell social media analytics services or provide free or paid analytics services. Yet, as with sentiment analysis, much of this activity deals with short-term, transitory events. Such analysis is typically bound to a relatively brief timescale—a few hours to a few days. Little financial benefit may exist in long-term provision of a 'long tail' of older redirects.

3.3 Preservation of shortened URLs

Shortened URLs, once identified, can (if the underlying service is still available), trivially be resolved into the original destination URL. This is a useful step for many forms of analysis (e.g., content/contextual analysis of tweets on Twitter). The half-life of social services is often short, but a URL shortener is more intimately bound into our ability to follow a conversation than, for example, a news aggregation service might be. The loss of the news aggregation service potentially compromises our ability to identify the trigger for transitory interest in a given subject or resource. The loss of the redirect service means that the key resources referenced during a conversation can no longer be referenced, compromising our ability to understand the social or political context and underlying framing of the discussion. URL redirection increasingly offers a further challenge, for although the number of discrete services in popular use appears to be reducing, the penetration of these services into the user experience continues to increase. Twitter itself did not initially impose the use of a domain redirection service. Later, the service began to 'wrap' popular (frequently retweeted/referenced) URLs into Twitter's own domain redirection service, `t.co`. In late 2011 Twitter made this mandatory for all URLs [8]; therefore, any URL published through the Twitter service will be published in the form of a `t.co`/key alias. Since users' choice of URL redirection service typically relates to their choice of application (for example, HootSuite users will find that they are minting ow.ly URLs, which are inbuilt), this means that a user making use of HootSuite will have a characteristic 'fingerprint': `t.co` → `ow.ly` (→ previous source of link).

There are many reasons to look into URL redirection other than preservation, such as the need to identify spam [41], or an interest in conversation/discourse analysis and information propagation [29].

3.4 In chains: unwrapping the URL

The implementation of various services and applications leads to the 'wrapping' of existing URLs into one or more URL redirects. The effect is similar to taking a postcard, and placing it into an envelope addressed to the initial receiver care of an intermediary. Then that envelope is passed on to a courier service who insist on placing the mail into their own brand of envelope and addressing it to 'Original recipient, care of initial intermediary, care of the courier service's

posting office'. By this means, each agency is able to collect statistics about visitors to that URL.

For the user, this carries the penalty that URLs are both opaque and somewhat slower to resolve. It also implies that the user is providing considerable information about their interests and activities to each agency in the redirect chain. However, for the researcher at least, it provides us with additional information about the pathway that this information took on its way from the originator to the author of the tweet.

3.5 Backtracking the trackers

A simple URL redirect tracker was developed for the purpose of tracking each step of URL redirection, using Perl's LWP libraries to extract information about each step of domain resolution. This 'traceroute' application is able to generate information about a shortened URL by backtracking through each step and documenting each redirect. A sample result is given (see Table 1).

Table 1. A sample HTTP response chain

Short URL	Response Code	Redirect	Chain ID
`t.to/example`	301	`ow.ly/example`	1
`ow.ly/example`	301	`bbc.in/example`	2
`bbc.in/example`	301	`news.bbc.co.uk/example`	3

An aggregate view of the redirect landscape is shown in Fig. 1. As is visible from this graph representation, there are many redirect services in use. To begin to build up the CG model for analysis of the trackers of the retweets the definitional graph for the knowledge representation of the data from Table 1 and extended data is found in Fig. 2. This REDIRECT graph indicates that a shortened URL can be redirected into a modified URL. From the type hierarchy associated with this definitional graph we see that during a join of factual data later in the representation processing URLs that are either 'Short URL' concepts or 'Modified URL' concepts can be joined. However, as our literature review has indicated, many of these are 'vanity' domains, so it is inaccurate to think of each redirect as a separate service. Rather, it is suggested that many are simply aliases of an existing commercial service. The question of identifying individual business entities within this group is one that can be solved quite simply on a technical level, through the use of domain analysis tools to identify the site operator. The use of WHOIS information presents difficulties since domains registered through separate registrars/various countries have quite different recordkeeping conventions and access regulations. Instead, service-level information such as IPs may be used as a rough indicator, with results such as those shown in Fig. 3.

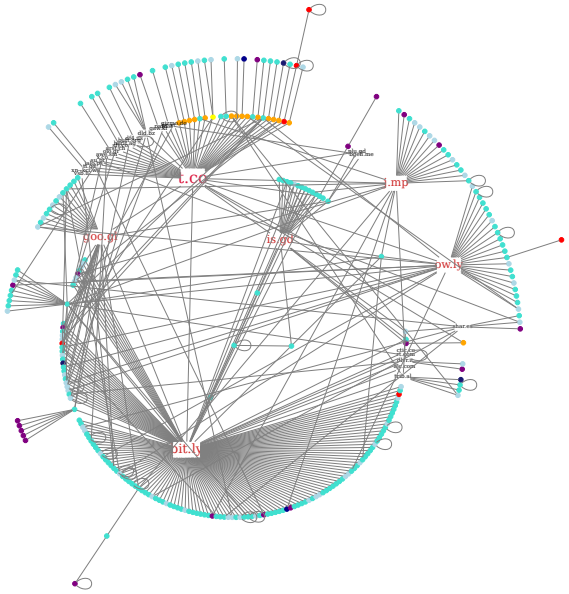


Fig. 1. The TSA web of redirects

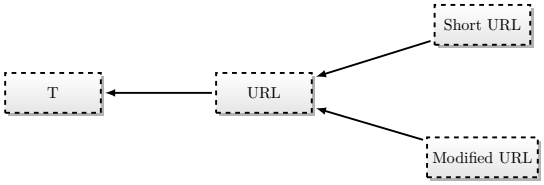
4 Results

4.1 Service model

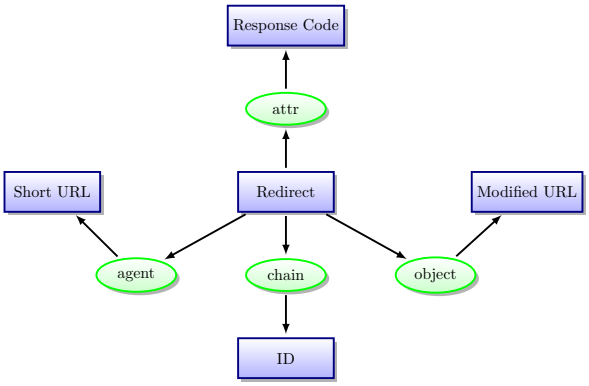
Network architecture relies on mapping relationships between conceptual entities such as businesses, service names and API endpoints and their representations on the network, i.e. IP addresses, ports, protocols and attributes. There is a great deal of information to record here (see Fig. 3); however, Fig. 4 presents a small part of the network with the essential information, prerequisite to further analysis—i.e., how the URLs are associated with their IPs. In Fig. 5, this basic conceptual information is given in a definitional CG. The underlying network (and, consequentially, business) relationships upon which the system depends can be instantiated into multiple factual CGs, in which Fig. 6 is an example.

4.2 URL redirection information model

We presented in Fig. 2(b) a CG definitional graph representation of the information to be used to create a partial model for a URL redirection operation. This includes the URL originally provided (the short URL), the object that constitutes the direct object of that redirect (modified URL), and the agent responsible for the redirect (the redirect). Additional information typically retrieved during the URL resolution process is also indicated in this CG, such as the response code provided by the redirect service during the lookup process (instance metadata) and the position of this redirect object within the chain of redirects, which again differs by instance.



(a) Type hierarchy for REDIRECT schema



(b) CG definitional graph for REDIRECT schema

Fig. 2. CG schema for REDIRECT

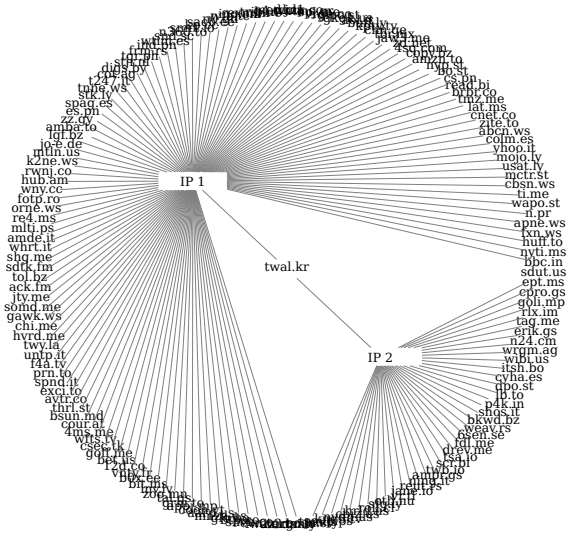


Fig. 3. The `bit.ly` infrastructure constellation

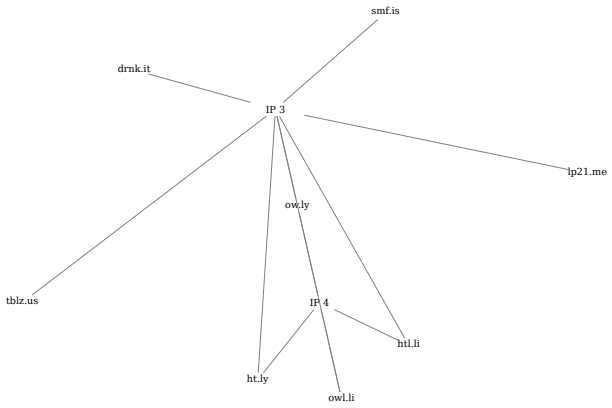


Fig. 4. A simple constellation of redirect services

Fig. 5. The service relationship definitional CG

Fig. 6. Instantiation of service relationship CG with single linkage from network

This CG graph can be instantiated with actual factual data producing, for example, the graph seen in Fig. 7. This information is processed from a bank of tweets, and can later be joined with the instantiated SERVICE CG already encountered (Fig. 6) using the Type Hierarchy from Fig. 2(a) to produce a graph with the service relationship from one CG tied to the instantiation of a redirection graph. This creates a partial model CG that has the original shortened URL linked to the service providing the actual disk space/web hosting service (see Fig. 8) by joining on the modified URL of the instantiated redirect CG.

4.3 Contextual representation

This example, containing both instance data and modelled generalities, is contextual to the resolution process and outcome. By continuing on with the process discussed in the previous sub-section, the generation of the representation of Fig. 4 as a CG partial model is produced in Fig. 9. The ability to represent contextual information is a strength of conceptual graph theory and represents a core requirement for analysis of social data such as Twitter. This consideration becomes particularly important if the information is to be treated as elements of a broader discourse rather than orphaned utterances [17]. Extension to the

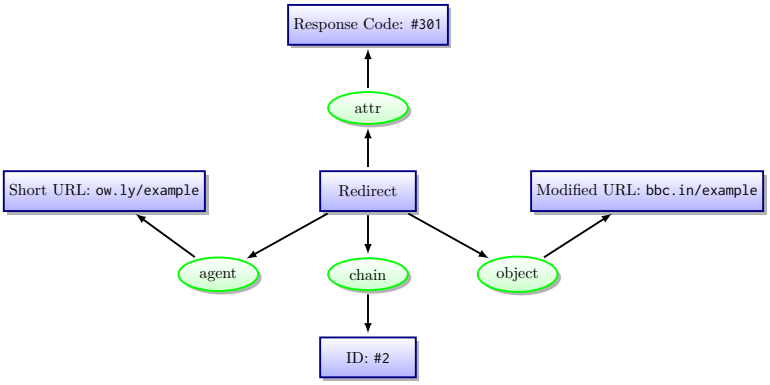


Fig. 7. Instantiation of REDIRECT definitional graph

Fig. 8. Join of fact SERVICE CG with fact REDIRECT CG

model to deal with temporal and sequential aspects of Twitter discourse may be of benefit.

4.4 Individual and chained utterances

In reality there is a time-based aspect to the data in its originating context, that of conversations or interaction chains published on Twitter. There is also a mapping of the landscape of shared references upon which Twitter's message-passing depends. If the original redirect definitional graph also stored the relevant time information from the tweets [25], then a time line chart could be generated showing the impact on not only Twitter, but on the services providing storage. We wish to include this in future.

Social network graph representations are typically designed as directed graphs, showing self-declared relationships between individuals (i.e., 'friend', 'colleague', or—in the case of Twitter's '@'-reference, 'referent'. In a subset of cases, SNG representations are used that permit temporal reasoning (i.e., progression of the system's development through time). Consider for example Tang et al's proposed temporal distance metrics, designed to quantify the speed of information diffusion processes [40] in a manner that is sensitive to local and global network characteristics, Shekhar and Oliver's review of the challenges inherent in modelling time-aggregated graphs [36], or Santoro et al's judgment that '[m]ost instruments—formalisms, concepts, and metrics—for social networks analysis fail to capture their dynamics' [32].

Many research questions—particularly those linked to information propagation, reaction, etc. through Twitter—benefit from accurate and detailed mod-

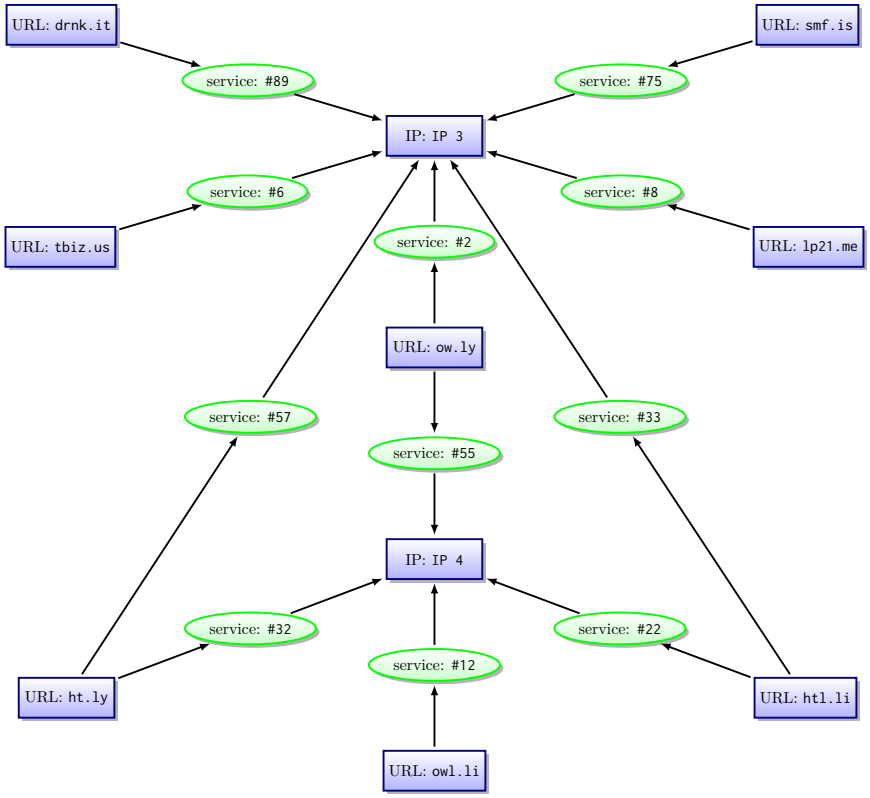


Fig. 9. CG storage of simple constellation of redirect services

elling of temporal precedence. Research into the attractions of Twitter to its users are likely to focus on the reactions of its user community to different types of input. Investigation of the attractions of Twitter as asocial news service (in comparison to a microblogging platform), will often focus on broad-grained metrics such as the overall proportion of tweets in any given locality that contain or refer to news items in some manner or another. However, a number of problems also exist that take a broader view of the information proliferation landscape, of which Twitter remains only a proportion, albeit at present an influential one. The relative significance of traditional media, 'new media' an social network services in information proliferation is an interesting subject and one which will undoubtedly continue to attract attention as the role of services in reflecting or even setting public opinion comes under scrutiny. Businesses continue to offer services intended to manage public opinion on social networking sites; mapping the territory is an important step in evaluating any such claim.

Much of the descriptive language from Santoro's paper is of direct relevance to our model; for example, Santoro et al [32] separate the concept of 'journey'

from that of 'path'; that is, a type of path through a graph that includes waiting times at intermediate stages in travel through the graph. They also identify recent papers proposing temporal versions of the typical social-network metrics of proximity, betweenness, closeness and so forth.

4.5 Evaluation of the CG representation as an EAD research tool

The exploratory analysis of a Twitter dataset described during this paper used CG representations as a backbone for representing information gathered about entities, agencies, interactions and infrastructure. This paragraph provides a brief review of this addition to the loose EAD methodology of visualisation, analysis and mining that we typically apply in the early stages of getting to grips with a large dataset. As we expected, conceptual graphs provided an accessible mechanism for knowledge representation within the team context. One team member was already familiar with the conceptual graph structuring. Another found that they were not intuitively readable, but was able to read them given appropriate guidance. It does seems necessary to have training before use.

5 Conclusion

Analysis of large dataset can be a tedicious process. However new tools enhance the ability to process these datasets. These tools must be flexible while at the same time have a solid knowledge representation. We have discovered that graphical representation and graphics operators make building of the underlying models (and partial models) easier to visualize. Because conceptual graphs are both built on logic and graphical operators they can be used for this stable representation. They are also built such that time and space structure and process is built directly into the representation [25]. Microblogging creates many data records that are both similar and different at the same time. In particular re-tweets on Twitter can grow at a very fast rate and they are time dependent. Therefore the underlying representation needs to be easily to implement and fast to process [24]. The basic CG definitional and instantiated graphs for this case study has given us a good start on an over all processing graph set for discovering the data clustering and topology of the constellations from shortened links and service provider on Twitter. We can also use CGs as a teaching tool for learning how to define context with social network relationships.

References

1. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011), http://www.sciencedirect.com/science/article/pii/S187775031100007X
2. Cao, T.H.: Fuzzy conceptual graphs for the semantic web. In: proceedings of 2001 BISC International Workshop on Fuzzy Logic and the Internet, FLINT'2001. Berkeley, CA, USA (August 2001)

3. Carmody, T.: A Tangled Web of Shortened Links. A study of link shortening reveals hidden strands of the Web. Retrieved May 16, 2011 (2011), http://www.technologyreview.com/news/423170/a-tangled-web-of-shortened-links/
4. Carpenter, S.: Developing a measure to elicit and compare mental models of processes. Tech. rep., The University of Alabama in Huntsville (2007)
5. Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., Mazo, I.: Extracting human protein interactions from medline using a full-sentence parser. Bioinformatics 20(5), 604–611 (2004), http://bioinformatics.oxfordjournals.org/content/20/5/604.abstract
6. Davis, R., Shrobe, H.E., Szolovits, P.: What is a knowledge representation? AI Magazine 14(1), 17–33 (1993), http://www.aaai.org/ojs/index.php/aimagazine/article/view/1029
7. De Longueville, B., Smith, R.S., Luraschi, G.: "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 International Workshop on Location Based Social Networks. pp. 73–80. LBSN '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1629890.1629907
8. dev.twitter.com: The t.co URL wrapper (2012), https://dev.twitter.com/docs/tco-url-wrapper, retrieved May 15, 2012
9. Gelperin, D.: Exploring agile. In: Proceedings of the 2008 international workshop on Scrutinizing agile practices or shoot-out at the agile corral. pp. 1–3. APOS '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1370143.1370144
10. Montes-y Gómez, M., Gelbukh, A.F., López-López, A.: Detecting deviations in text collections: An approach using conceptual graphs. In: Proceedings of the Second Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence. pp. 176–184. MICAI '02, Springer-Verlag, London, UK, UK (2002), http://dl.acm.org/citation.cfm?id=646402.691915
11. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. CoRR abs/0812.1045 (2008)
12. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772751
13. Lane, P.C.R., Gobet, F.: A theory-driven testing methodology for developing scientific software. Journal of Experimental & Theoretical Artificial Intelligence 24(4), 421–456 (2012), http://www.tandfonline.com/doi/abs/10.1080/0952813X.2012.695443
14. Lerman, K., Ghosh, R.: Information contagion: an empirical study of the spread of news on digg and twitter social networks. CoRR abs/1003.2664 (2010)
15. Maurer, F., Martel, S.: Extreme programming. rapid development for web-based applications. Internet Computing, IEEE 6(1), 86–90 (January–February 2002)
16. Mishne, G.: Source code retrieval using conceptual graphs. Master of logic thesis, Institute for Logic, Language and Computation, University of Amsterdam (2003)
17. Moulin, B.: Temporal contexts for discourse representation: An extension of the conceptual graph approach. Applied Intelligence 7, 227–255 (1997), http://dx.doi.org/10.1023/A:1008224616031
18. Oricchio, R.: Is Twitter A Social Network? (2010), http://www.inc.com/tech-blog/is-twitter-a-social-network.html, retrieved September 16, 2012

19. Owen, R., Horváth, I.: Towards product-related knowledge asset warehousing in enterprises. In: Proceedings of the Fourth International Symposium on Tools and Methods of Competitive Engineering. pp. 155–170. HUST Press (2002)
20. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010)
21. Paul, M., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: International AAAI Conference on Weblogs and Social Media (2011), https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880
22. Perer, A., Shneiderman, B.: Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In: Proceedings of the 13th international conference on Intelligent user interfaces. pp. 109–118. IUI '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1378773.1378788
23. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 181–189. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1857999.1858020
24. Pfeiffer, H.D.: The Effect of Data Structures Modifications on Algorithms for Reasoning Operations Using a Conceptual Graphs Knowledge Base. Dissertation, New Mexico State University (Dec 2007)
25. Pfeiffer, H.D., Hartley, R.T.: Temporal, spatial, and constraint handling in the conceptual programming environment, cp. J. Exp. Theor. Artif. Intell. 4(2), 167–182 (Apr 1992), http://dx.doi.org/10.1142/S0218001490000125
26. Phelan, O., McCarthy, K., Smyth, B.: Using twitter to recommend real-time topical news. In: Proceedings of the third ACM conference on Recommender systems. pp. 385–388. RecSys '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1639714.1639794
27. Quincey, E., Kostkova, P.: Early warning and outbreak detection using social networking websites: The potential of twitter. In: Electronic Healthcare. Springer Berlin Heidelberg (2010)
28. Ribière, M., Matta, N., Cointe, C.: A proposition for managing project memory in concurrent engineering. In: in International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'98) (February 1998)
29. Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., Almeida, V.: On word-of-mouth based discovery of the web. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 381–396. IMC '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2068816.2068852
30. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th international conference on World wide web. pp. 695–704. WWW '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1963405.1963503
31. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772777

32. Santoro, N., Quattrociocchi, W., Flocchini, P., Casteigts, A., Amblard, F.: Time-varying graphs and social network analysis: Temporal indicators and metrics. CoRR abs/1102.0629 (2011)
33. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures. Lawrence Erlbaum, Hillsdale, NJ (1977)
34. Shehata, S., Karray, F., Kamel, M.: Enhancing text retrieval performance using conceptual ontological graph. In: Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on. pp. 39–44 (December 2006)
35. Shekar, C., Wakade, S., Liszka, K., Chan, C.C.: Mining pharmaceutical spam from twitter. In: Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on. pp. 813–817 (November–December 2010)
36. Shekhar, S., Oliver, D.: Computational modeling of spatio-temporal social networks: A time-aggregated graph approach. In: Proceedings of the 2010 specialist meeting on Spatio-Temporal Constraints on Social Networks (2010), http://www.ncgia.ucsb.edu/projects/spatio-temporal/docs/Shekhar-position.pdf
37. Sowa, J.F.: Conceptual structures: information processing in mind and machine. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1984)
38. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics 6(3), 239–251 (2005), http://bib.oxfordjournals.org/content/6/3/239.abstract
39. Sullivan, J.: A tale of two microblogs in china. Media, Culture & Society 34(6), 773–783 (2012), http://mcs.sagepub.com/content/34/6/773.short
40. Tang, J., Musolesi, M., Mascolo, C., Latora, V.: Temporal distance metrics for social network analysis. In: Proceedings of the 2nd ACM workshop on Online social networks. pp. 31–36. WOSN '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1592665.1592674
41. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 243–258. IMC '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2068816.2068840
42. Yi.tl: Url shorteners (2012), http://yi.tl/pages/urlshorteners.php, retrieved May 15, 2012