Proceedings of the 1st Workshop on

Making Sense of Microposts (#MSM2011)

Big things come in small packages

at the 8th Extended Semantic Web Conference (ESWC 2011)
Heraklion, Greece
30th of May 2011

edited by
Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, Mariann Hardey

Preface

The 1st Workshop on Making Sense of Microposts (#MSM2011) was held in Heraklion, Greece, on the 30th of May 2011, during the 8th Extended Semantic Web Conference (ESWC 2011).

Posting information about on-going events on Twitter and Facebook; checking into and contributing new information about Points of Interest on Foursquare; updating one's status on a variety of social networking sites, using a variety of devices, while stationary or on the move, are now so commonplace that such platforms are often the first point of call for searching for and sharing information covering a wide range of events, topics and personal or work-related interests.

As a result, enormous quantities of small user input are being piped into the data streams of the Web, leading to a rate of growth which has never before been witnessed. We refer to such small user inputs as Microposts. The #MSM2011 workshop discussed emerging to fairly advanced work on the research this has engendered. Micropost data, which spans disparate, heterogeneous topics, therefore requires new techniques for information extraction and the leveraging of semantics from Microposts, to glean the knowledge contained, and new tools that make optimal use of the semantics encoded in Microposts'. The discussions also looked at studies related to Microposts, both social and from a more technically oriented perspective, that should contribute to building appealing new systems based on this type of data.

#MSM2011 is unique in targeting both Semantic Web researchers and other fields, both within Computer Science, such as Human-Computer Interaction, and outside Computer Science, particularly the Social Sciences, in order to harness the benefits different fields bring to research involving Microposts. #MSM2011 recognises the importance of maintaining a focus on the end user of Microposts' – ranging from the mainstream mobile phone or computer user with little to no technical expertise, to the Semantic Web expert – in order to ensure that appealing and usable tools are developed, that harness the particular benefits of Semantic Web technology.

Many hearty thanks to all our contributors and participants, and also the Programme Committee whose feedback resulted in a rich collection of papers, posters and demos, each of which adds to the state of the art in leading edge research. We are confident that this is the start of a series of workshops that will continue to target the rich body of information generated by the many and varied authors in the online world.

Matthew Rowe KMi, The Open University, UK
Milan Stankovic Hypios / Université Paris-Sorbonne, France
Aba-Sah Dadzie The University of Sheffield, UK
Mariann Hardey The University of Durham, UK
#MSM2011 Organising Committee, May 2011

Introduction to the Proceedings

Out of a total of 19 paper submissions, 7 full and 2 short papers were accepted, around which two main discussions were held. This was in addition to a poster and demo session, to exhibit practical application in the field, and foster further discussion of the ways in which data extracted from Microposts is being reused. The accepted submissions cover an array of topics; we highlight these here.

Information Diffusion and Influence

Three submissions to the workshop addressed the topics of Information Diffusion and Influence within Microposts. Weller *et al.*'s paper, '*Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences*', analysed the flow of information through tweets at scientific conferences, by assessing the diffusion of URLs and retweets. '*Making Sense of Location Based Micro-posts Using Stream Reasoning*' by Celino *et al.* proposes a framework to identify so-called *mavens* – experts on a given POI – utilising stream reasoning to handle the deluge of Microposts and enable effective identification. A demo by Huron *et al.*, titled '*Polemical Video Annotation by Twitter*', models arguments and discussions between users on Twitter during video broadcasts. The application enables contentious points in videos to be identified, and leads to further information exchange and debate.

Entity Extraction and Semantics

Microposts often refer to entities within their content; identifying such entities enables effective tracking of mentions and consensus concerning opinion. However, the limited length of Microposts makes detecting and extracting such references challenging. '*Extracting Semantic Entities and Events from Sports Tweets*' by Choudhury and Breslin presents an approach to this problem in the form of named entity recognition from sports tweets, testing various features for the detection task – i.e. linguistic features, statistical analysis and domain knowledge. Entity extraction is utilised in '*Follow me: Capturing Entity-Based Semantics Emerging from Personal Awareness Streams*' by Cano *et al.*, to first detect entities in users' personal awareness streams – derived from status updates coupled with the context of the user – before using such entities to suggest concepts that correlate with the context of the user.

The interestingly titled '*Does Size Matter? When Small is Good Enough*' by Gentile *et al.* presents the novel experiment of truncating emails from Micropost size (i.e. 140 characters) up to the full size of each email in a given corpus, and then performing text classification over the truncated emails. Results are compared with the classification using the full emails, showing that truncated emails provide a sufficient summarisation for accurate classification. In '*Discovering the Dynamics of Terms' Semantic Relatedness through Twitter*', by Milikic *et al.*, the semantic relatedness of terms in Microposts is assessed. Their approach measures the dynamic aspect of semantic relatedness over time under the hypothesis that the relation between terms is incorrectly assumed to be static.

Politics and Sentiment

Microposts enable opinion to be expressed to a global audience with relative ease. As a result, platforms that provide functionality to publish Microposts are often central to emotive discussions, such as political uprisings. The workshop accepted three papers that explore work in this area. Skilters *et al.*'s paper, '*The Pragmatics of Political Messages in Twitter Communication*', assesses political messages on Twitter; their analysis reveals a link between retweet popularity and offline political consensus. '*Automatic Detection of Political Opinions in Tweets*' by Maynard and Funk presents an approach to political opinion detection and analysis in tweets. The authors conjecture that a middle ground is required between sophisticated NLP techniques that function over rich review corpora and more naïve, simplistic, weighted lexicon-based approaches. They present work that attempts to fills this gap and evaluate their work over a large corpus of political tweets from the 2010 UK General Election.

To gauge sentiment in tweets, Nielsen's paper, '*A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs*', proposes a new sentiment word list. Detection of sentiment in tweets normally utilises a weighted semantic lexicon, looking up individual terms and returning their valence. Existing lexicons are available for this task, therefore Nielsen presents his new word list and compares its performance in sentiment detection against, among others, the ANEW semantic lexicon. Results show comparable performance.

Workshop Awards

Two awards were made, sponsored by the EU FP7 *WeGov* project[1]. Best paper nominations were sought from the reviewers, and a final decision agreed by the Chairs, based on the nominations and review scores. The best poster/demo award was based on nominations by participants during the workshop.

Additional Material

The call for participation and all paper, poster and demo abstracts are available on the #MSM2011 website[2]. The full proceedings are also available on the CEUR-WS server, at: `http://ceur-ws.org/Vol-718`.

[1] `http://www.wegov-project.eu`
[2] `http://research.hypios.com/msm2011`

Programme Committee

Harith Alani KMi, The Open University, UK
Sofia Angeletou KMi, The Open University, UK
Uldis Bojars University of Latvia, Latvia
David Beer University of York
John Breslin NUIG, Ireland
A. Elizabeth Cano The University of Sheffield, UK
Óscar Corcho Universidad Politécnica de Madrid, Spain
Guillaume Ereteo INRIA, France
Fabien Gandon INRIA, France
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Jon Hickman Birmingham City University, UK
Robert Jäschke University of Kassel, Germany
Jennifer Jones University of the West of Scotland, UK
Jelena Jovanovic University of Belgrade, Serbia
Philipe Laublet Université Paris-Sorbonne, France
Pablo Mendes Kno.e.sis, Wright State University, USA
Eric T. Meyer Oxford Internet Institute
Alexandre Passant DERI, Galway, Ireland
Danica Radovanovic University of Belgrade, Serbia
Yves Raimond BBC, UK
Harald Sack University of Potsdam, Germany
Bernhard Schandl University of Vienna, Austria
Elena Simperl University of Innsbruck, Austria
Raphaël Troncy Eurecom, France
Mischa Tuffield Garlik, UK
Victoria Uren The University of Sheffield, UK
Claudia Wagner Joanneum Research, Austria
Shenghui Wang Vrije University, The Netherlands
Ziqi Zhang The University of Sheffield, UK

Multidisciplinary Steering Committee

Alexandre Monnin IRI, Centre Pompidou, Paris, France
Danica Radovanovic University of Belgrade, Serbia

Table of Contents

Section III: Politics and Sentiment

Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences

Katrin Weller[1], Evelyn Dröge[1] and Cornelius Puschmann[2],

Heinrich-Heine-University Düsseldorf,
[1] Dept. of Information Science & [2] Dept. for English Language and Linguistics,
Universitätsstr. 1, 40225 Düsseldorf, Germany
{katrin.weller, evelyn.droege, cornelius.puschmann}@uni-duesseldorf.de

Abstract. This paper investigates Twitter usage in scientific contexts, particularly the use of Twitter during scientific conferences. It proposes a methodology for capturing and analyzing citations/references in Twitter. First results are presented based on the analysis of tweets gathered for two conference hashtags.

Keywords: Twitter, microblogging, tweets, citation analysis, informetrics.

1 Introduction

With its enormous gain in popularity, the microblogging service Twitter has already become the subject of different scientific studies. [1] were among the first to investigate why and how people use Twitter: "From our analysis, we find that the main types of user intentions are: daily chatter, conversations, sharing information and reporting news. Furthermore, users play different roles of information source, friends or information seeker in different communities". Studies from different fields of research exist that focus on specific application areas, for example Twitter in politics and elections [2], in organizational informal communication scenarios [3] or during natural disasters [4]. Within this paper, we investigate Twitter usage in *scientific* contexts and consider Twitter as a means for scientific communication. The scientific use of Twitter has received some attention in previous work, e.g. [5], [6], [7], [8]. Our paper suggests refinements of analyzing datasets based on tweets collected during *scientific conferences* and present our results from applying novel forms of *intellectual* tweet content analyses. Our overall aim is to better understand how scientists use Twitter and whether traditional patterns of scientific communication are being mapped to microblog communications or whether entirely new practices emerge. Therefore we consider information flows as an aspect of *citation analysis* within scientific Twitter communication. Scientific communication in its classical form of *publications* and *citations* has long been a subject to analyses in the fields of scientometrics and informetrics. Informetric citation analysis distinguishes *citations* from *references* [9]: A citation is a formal mention of another work in a scientific publication – viewed from the *cited* work's perspective. A

reference is the same mention of a work but viewed from the *citing* work's perspective (typically in form of a reference section in a publication). Thus, citations and references are two sides of the same coin[1]. Our paper investigates whether and how similar information flows exist in microblog communications by comparing two types of citations on Twitter: *URLs* pointing to external resources and *retweets* (RTs) that cite other users' tweets (a more detailed definition will be given in section 2.2). For both types, we propose methodological approaches to performing citation analysis on conference tweets data and present first results for these approaches based on a test dataset collected via the hashtags of two scientific conferences. This paper should thus primarily be viewed as exploratory research in the field of informetrics for microblogging. It may provide a basis for future work on developing novel informetric indicators or for the development of applications that make use of these indicators, e.g. for identifying and ranking popular tweets, popular twitterers or external resources, as well as for displaying user networks based on co-ciation or bibliographic coupling. This work thus also relates to webometrics [10], a sub-discipline of informetrics that discusses metrics for information exchange and communication on the Web. Recently, new Web 2.0 tools that enable novel forms of social interaction have brought about a range of new aspects that can be measured and evaluated (e.g. relating to access and usage, Web publication behavior, user interrelations). [10] explains that measuring Web 2.0 services offers new ways for data mining; it can help to gain insights to "patterns such as consumer reactions to products or world events". [11] provide an overview on Web 2.0 services (including microblogging) that may be of interest for new scientometric indicators by measuring publication impact based on social mentions.

2 Identification of Scientific Microblogging Activities

Twitter is a tool which is not dedicated to one particular application scenario and thus includes users with various backgrounds and different motivations. It is difficult to identify scientific tweets or twitterers for analyses.[2] In the next section we will discuss the challenges of gathering data about scientific Twitter usage in order to explain why our datasets are purely based on hashtags from scientific conferences and thus to indicate some limitations of our current approach.

2.1 Basic Problems in Identifying Scientific Microposts

Currently, there are no reliable statistics about how many scientists use Twitter (and more specifically, how many of them do so for science-related communication). Empirical studies (quantitative and qualitative designs) that investigate scientists' motivations for using Twitter are still missing. Presumable reasons for using Twitter might be timely access to novel information sources and spontaneous creation of

[1] We will use 'citation' as the broader term for both citations and references.

[2] In a fundamental consideration one may furthermore discuss the proper definition for what exactly counts as a scientist or a scientific publication, but this is not a focus of our work.

networks based on shared interests (e.g. via hashtags), as well as general benefits of informal communication as identified by [3]. There is also no general definition of scientific tweeting. It may for example refer to the following aspects:

- *Any tweet with scientific content or linking to scientific content*: The scientific Twitter data could be a set of tweets with actual scientific contents. This, however, is almost impossible to achieve, as it would require either manual identification of tweet contents or elaborated computer-linguistic automated methods as well as an elaborated definition for 'scientific contents'. Another interesting subset of Twitter is the number of tweets that include links to purely *peer-reviewed* scientific publications on the Web [11]. Yet, currently tweets with links to scientific publications are also difficult to collect automatically.
- *Any tweet published by a scientist*: Analyses of scientific microblogging may be entirely based on its users. Such approaches are frequently applied in analyses of scientific blogging, while the definition of 'scientist' in this context may be narrow (only including members of universities) or broad (including also, for example, teachers and science journalists). In analyzing Twitter based on users, one always depends on the biographical information provided by the twitterers. Furthermore, a selection of users will have to be made manually. [12] have for example manually identified 28 twittering scientists (using a snowball system) to analyze their citation behavior. [13] has identified twitterers with academic background by examining the list of followers of the Chronicle of Higher Education's Twitter account. To our awareness, there are so far no studies that analyze Twitter accounts belonging to scientific groups or institutions.
- *Any tweet with a science-related hashtag*: Finally, one may identify scientific tweets based on hashtags. In still rather rare cases, scientists announce particular hashtags for their projects or topics of interest. One example is the hashtag "#altmetrics" which is introduced by [14] for work on measuring scholarly impact on the Web. More frequently, we find specific hashtags for scientific conferences, some of them officially proposed by the organizers (e.g. "#websci10"). So far, most studies on scientific microblogging have used datasets collected via conference hashtags. For example, [6] and [7] have gathered sets of conference tweets to perform automatic analyses on measures such as the number of tweets, the most active twitterers and the dynamics of the conference. [15] are developing automatic methods for extracting semantic information from conference tweets. [5] and [8] have performed manual/intellectual categorizations of tweet contents. This paper is the first to focus on Twitter citations in the context of scientific conferences.

2.2 Citation Analysis on Twitter

[12] define Twitter citations as "direct or indirect links from a tweet to a peer-reviewed scholarly article online" and distinguish first- and second-order citations based on whether there is an "intermediate webpage between the tweet and target resource". In their sample of tweets collected from 28 academics they discovered that of all tweets including an URL, 6% fit into their definition of twitter citations, i.e. they linked directly or via an intermediate page (like a blog post) to a peer-reviewed

article. We suggest that linking to a peer-reviewed publication is only one possible dimension of citing with Twitter and want to discuss the following alternatives:

- All *URLs* included in tweets may be counted as a form of reference. Analyses may focus on the types of resources that are referenced in URLs. URLs in tweets act as *external citations* (where the tweet includes a reference and the external source receives a citation).
- *Retweets* can be interpreted as a form of inter-Twitter citations (*internal citations*). A user who retweets another one publishes a reference, the retweeted user gets a citation. In general, users retweet for different reasons like information diffusion or use retweets as a "means of participating in a diffuse conversation" [16]. Yet, retweet analyses are not easy to perform, due to the lack of format standardization.
- *@mentions* of usernames within tweets also sometimes resemble references, e.g. in tweets like "Just read an interesting paper by @sampleuser". Yet, they can currently not be automatically distinguished from other @messages and will thus have to be excluded from current analyses.

In the following section, we will exemplarily analyze and compare some test sets of hashtag-based conference tweets with regard to the first two types of Twitter citations, namely URLs in tweets (external citations) and retweets (internal citations).

2.3 Data Collection

For our study we have adapted the conference hashtag principle[3] to gather a collection of tweets. During our previous work [5] we collected tweets from four scientific conferences; we selected two smaller conferences (<500 participants) and two major conferences (>1.000 participants), with one small and one larger conference on topics from (digital) humanities and one small and one larger conference in the field of computer sciences. In [5] we performed intellectual analyses of tweets in these conference datasets. In this paper we now continue this work and perform the additional manual analysis of URLs included in tweets. For this purpose we have chosen the two major conferences investigated in [5], namely the World Wide Web Conference 2010 (WWW 2010, hashtag #www2010) and the Modern Language Association Conference 2009 (MLA 2009, hashtag #mla09), as we expected to find discipline-specific differences there. Table 1 presents an overview of the key information about the selected conferences and their respective hashtags. It is necessary to point out that this approach inevitably leads to loss of data: there may be tweets about the conferences without these particular hashtags or with misspelled hashtags (e.g. #www10). While typical misspellings may be considered for data collection, tags without any referencing hashtag cannot be collected for events like conferences. As we could not guarantee to capture all spelling variants for the conferences in our dataset[4], we deliberately concentrated on the main hashtag for each conference in order to achieve uniform preconditions for each set. For the same reason, we did not

[3] We intend to broaden the approach and want to analyze and compare additional datasets based on identified scientific twitterers in future work.

[4] This is mainly due to limitations to retrieve tweets older than a few days via the Twitter API, as there were no tweet archives available for all possible spelling variants.

include hashtags for associated or co-located events (e.g. #websci10 for the Web Science Conference co-located with WWW 2010). Tweets were collected for a period starting two weeks before and ending two weeks after the conference (Table 1).[5]

Table 1. The test dataset for tweets with conference hashtags #mla09 and #www2010.

Hashtag	**#www2010**	**#mla09**
Conference	World Wide Web Conference (WWW 2010)	Modern Language Association Conference (MLA 2009)
Conference location	Raleigh, NC, USA	Philadelphia, PA, USA
Conference dates	26.-30. April 2010	27.-30. December 2009
Discipline	Computer science	Linguistics, literature, (digital humanities)
No. of tweets from two weeks before until two weeks after the conference	3,358 [during period: 13. April 2010-14. May 2010]	1,929 [during period: 15. Dec. 2009-14. January 2010]
Total no. of unique twitterers (average no. of tweets per twitterer)	903 (∅ 3.72)	369 (∅ 5.23)
Total no. of tweets during actual conference days only	2,425 [26.-30. April 2010]	1,206 [27.-30. December 2009]

4 Analysis of URLs in Tweets

Within our two datasets of #www2010 and #mla09 tweets, we identified all tweets that include an URL as a link to a website[6] as an external citation. Within Twitter, URLs are often shortened with so-called URL shorteners (such as Bit.ly). Shortened URLs were resolved to create a list of all URLs included in the datasets. Multiple appearances[7] of exactly the same URLs[8] could be identified and counted (Table 2).

A basic categorization scheme was developed to classify types of websites that the URLs included in tweets are pointing to. Each URL within the dataset was classified by hand according to the following scheme:

[5] We may now principally analyze data for this entire period or for the actual conference days only. If not indicated otherwise, all numbers in the following sections refer to the broader period from two weeks before until two weeks after the conference dates.

[6] URLs were detected by the character strings "http://", "https://" and "www." (followed by additional text, not a blank space). Expressions like 'Amazon.com' or 'Twitter.com' are more difficult to detect automatically and were deliberately left out, as one may not definitely state that these should act as links to Websites, they may also be interpreted as proper names of companies or products.

[7] URLs may appear more than once per dataset. This may in some cases be due to retweets, in other cases different users may post the same URL independently.

[8] As we worked with automatic techniques, only exact character string matches were identified as being multiple appearances of the same URL. For more precise results and for subsequent studies, we suggest to also check URLs with different strings pointing to the same resources, e.g. "http://twapperkeeper.com/hashtag/mla09" and "http://twapperkeeper.com/mla09".

- *Blog*: This category is used for all kinds of blogs and blog posts as well as other private commentaries on personal websites.
- *Conference*: This category is used for the official conference websites.
- *Error*: If a URL could not be accessed, it was marked with this category.
- *Media*: This category was applied for all types of multimedia data, e.g. photos, videos, other types of visualizations and graphics.
- *Press*: This refers to *non-scientific* publications, e.g. articles in online newspapers or journals (in contrast to category "blog", websites in this category have to belong to a journalistic source).
- *Project*: This category is used for (official) websites by projects (e.g. the website of a research group or of a scientific project) and project results (e.g. a particular tool or platform).
- *Publication*: This includes scholarly publications, e.g. an article in a scientific online journal (these may be open access publication or intermediate pages that link to paid content). In contrast to category press, URLs in this category should refer to a publication following scientific criteria, i.e. they should be peer-reviewed, follow scientific guidelines and be published by a scientific journal or publishing house or be accepted for a scientific conference.[9] The category also comprises lists of publications, e.g. tables of content from a proceedings volume, a scientist's website with his personal list of publications.
- *Slides*: This category is used for links to presentation slides, either on presentation sharing platforms like Slideshare, on personal, institutional or conference websites.
- *Twitter*: This category comprises links to subpages of Twitter, e.g. Twitter profiles, as well as Twitter-related websites such as Twapperkeeper.
- *Other*: Not specified, everything that does not belong to the categories above. In future work, URLs classified as "Other" should be investigated in more detail in order to refine the categorization scheme.

A considerable number of all conference tweets in our dataset includes links. Within the #www2010 set, 39.85% of tweets included URLs, within the #mla09 set there were 27.22% tweets with at least one URL. Within the total collection of 1,460 URLs from #www2010, 574 unique URLs have been identified. Thus, each URL appears 2.54 times at an average (for the #mla09 set: 2.77 times).

Table 2. Different ways to count URL citations in conference tweets.

	#www2010	**#mla09**
Number (and %) of tweets including at least one URL	1,338 (39.85%)	525 (27.22%)
Number of total URLs	1,460	551
Number of unique URLs	574	199

Of course, there are highly cited URLs and those that appear only once, resulting in a left skewed distribution as depicted in Figure 1. For #mla09 120 URLs (60.3% of

[9] We are aware that this definition needs refinements and additional qualitative analysis about different notions of 'scholarly publication' across scientific disciplines.

unique URLs) and for #www2010 312 URLs (54.36% of unique URLs) appear only once in the dataset. Table 2 sums up the different ways to count URL citations.

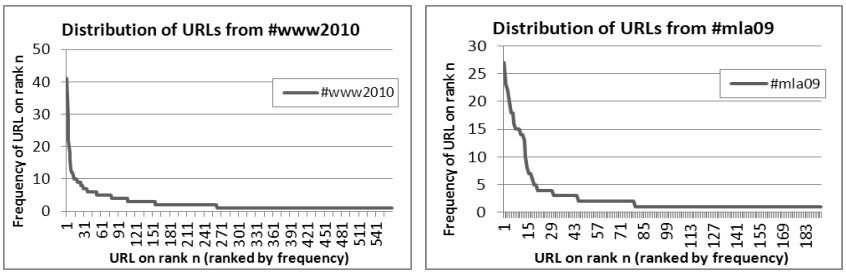


Fig. 1. How often do URLs appear in the dataset?

Table 3. Most popular URLs for # www2010.

URL	Frequency	Category
http://blog.marcua.net/post/566480920/twitter-papers-at-the-www-2010-conference	41	Blog
http://www.danah.org/papers/talks/2010/WWW2010.html	35	Publication
http://kmi.tugraz.at/staff/markus/www2010/www2010_roomstream.html	29	Twitter
http://xquery.pbworks.com/rtp-meetup	22	Error
http://www.elon.edu/e-web/predictions/futureweb2010/carl_malamud_www_keynote.xhtml	22	Conference
http://www.elon.edu/e-web/predictions/futureweb2010/default.xhtml	18	Conference
http://futureweb2010.wordpress.com/schedule/	16	Conference
http://www.slideshare.net/haewoon/what-is-twitter-a-social-network-or-a-news-media-3922095	13	Slides
http://events.linkeddata.org/ldow2010/	12	Conference
http://opengraphprotocol.org/	12	Project
http://www.websci10.org/program.html	12	Conference

Table 4. Most popular URLs for # mla09.

URL	Frequency	Category
http://amandafrench.net/2009/12/30/make-10-louder/	27	Blog
http://www.briancroxall.net/2009/12/28/the-absent-presence-todays-faculty/	23	Blog
http://nowviskie.org/2009/monopolies-of-invention/	22	Blog
http://chronicle.com/article/missing-in-action-at/63276/	20	Error
http://www.profhacker.com/?p=4448	18	Press
http://www.samplereality.com/2009/11/15/digital-humanities-sessions-at-the-2009-mla/	18	Blog
http://chronicle.com/blogpost/the-mlathe-digital/19468/	16	Press
http://www.profhacker.com/2010/01/09/academics-and-social-media-mla09-and-twitter/	15	Press
http://academhack.outsidethetext.com/home/2010/the-mla-briancroxall-and-the-non-rise-of-the-digital-humanities/	15	Blog
http://www.samplereality.com/2010/01/02/the-mla-in-tweets/	15	Blog

Table 3 and 4 list the top ten most frequent URLs from the #www2010 and the #mla09 dataset. Such analyses could help to identify the most influential conference

contents or those conference aspects that receive high attention (particularly if URLs link to papers or presentation slides presented at the respective conference). In case of the MLA 2009 conference this will not work directly: all of the top ten[10] URLs refer to press reports or blog posts about the conference in general.

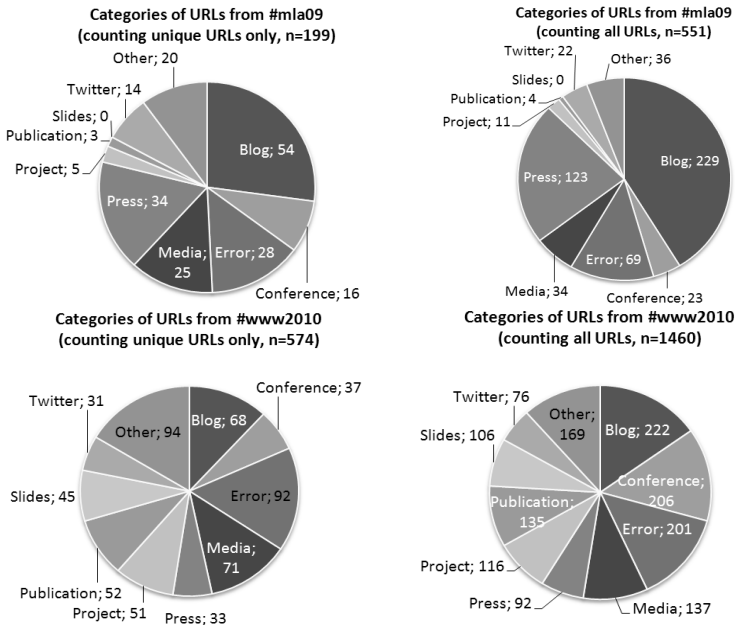


Fig. 2. Analysis of URL categories for unique URLs and all aggregated URLs.

In a general analysis of URL categories, great differences can be found between the profiles of the two test datasets (Figure 2). Twitterers during #mla09 had a general preference of linking to blog posts (27.14 of unique URLs, 40.29% of total URLs are categorized as "Blog") and press articles (21.61 unique URLs; 16.7% of total URLs). They did not link to any presentation slides (0 times category "Slides") and hardly to any scientific publications (3 unique URLs).[11] For #www2010, the percentage of links to publications and slides is clearly higher, but blogs still play an important role. Furthermore, 14.11% of overall URLs (6.45% of unique URLs) link to conference-related websites (e.g. video lectures from the event). At the time of our study (May-August 2010/ March 2011), a high number of URLs were no longer accessible or could not be identified due to misspellings (category "Error").[12]

[10] The URL on rank no. 4 had to be classified as "Error" as the URL cannot be opened, but as it is located at http://chronicle.com it can also be assumed to have been a "Press" link.

[11] More qualitative research is needed in order to explore these discipline specific behavior. The majority of researchers in humanities' discipline may not be using presentation slides.

[12] The process of identifying URLs and resolving shortened URLs is error prone and can hardly be re-enacted with consistent results at different points of time.

5 Analysis of Retweets

While external citations might become useful for detecting highly cited publications, presentations or projects, analyses of RTs are promising for identifying influential persons (or those receiving high attention) during a conference. So far, we have analyzed retweets with respect to cited and citing persons and to highly cited tweets.

Table 5. Different ways to count retweets (RTs).

	#www2010	**#mla09**
Automatically detected RTs: Number and percentage of RTs in entire conference dataset	1,121 (33.38% of 3,358)	414 (21.46% of 1,929)
∅ RTs per twitterer (automatically detected RTs, entire conference dataset)	1.24	1.12
Retweets including at least one URL	530	207
Manually detected RTs: Number and percentage of RTs in entire conference dataset	1,318 (39.25% of 3,358)	514 (26.65% of 1,929)
Manually detected RTs: Number and percentage of retweets in subdataset of tweets during actual conference days	828 (34.13% of 2,426)	269 (30.6% of 1,206)

Counting retweets automatically may lead to some loss of information. Not all RTs start with the characteristic "RT @user"-label at the beginning of a tweet. Some may also be indicated with "via @user", others simply copy a message without standardized identification mark. Within our analyses, we have also manually classified tweets as retweets. [13] Table 5 shows the different counts for retweets, among them the different values for retweets that were automatically detected via the "RT @user"-label and manually identified retweets. We did not yet distinguish simple RTs from "encapsulated retweets" [16]. There is a slightly higher percentage of retweets during the WWW 2010 conference than the MLA 09. For both conferences, a significant number of additional non-standard retweets could be identified manually: of 1,318 manually identified RTs for #www2010 85% have also been detected automatically (80% for #mla09 retweets). For #www2010, the percentage of RTs is slightly lower during the actual conference dates compared to the entire dataset with an included period before and after the conference; for #mla09 it is slightly higher during the conference days.

Retweets can help to identify highly cited persons within a network. In future work we intend to analyze the networks based on retweets more closely. So far, we have identified the persons who publish the most retweets and the persons who are often retweeted during a conference (based on automatically identified RTs). Typically,

[13] We automatically counted tweets *starting* exactly with the string "RT @"; these counts do not include tweets where a "RT @" appears at other positions within the tweet text. Manually identified RTs should comprise all tweets that include copied tweets, whether or not they are labeled "RT @user". Yet, the manual identification of RTs is not always error-free and depends on the definitions for labeling a tweet as RT. We aimed to include all tweets with "RT @user", "via @user" or "@user" at some position in the tweet and/or identical text strings.

these are not the same persons within one conference (Table 6): the top 3 persons who publish retweets are themselves rather rarely retweeted (#mla09: newfacmajority 1 RT received, ryancordell 3, jcmeloni 5; #www2010: laterribleliz and uncpublichealth have not received any RTs, olgag has received 18). For both conferences, the three users who received the most retweets do all belong to the top 10 most active users with the most tweets in the dataset (#www2010: boraz on rank 1 with 173 tweets, apisanti on rank 7 with 54 tweets, futureweb2010 on rank 2 with 129 tweets; #mla09: samplereality on rank 1 with 150 tweets, briancroxall on rank 3 with 61 tweets, nowviskie on rank 9 with 45 tweets). Future work should include qualitative analyses to find out more about these persons backgrounds and motivations.

Table 6. Top 3 of highly citing and highly cited twitterers during #www2010 and #mla09.

#www2010 RTs given	#www2010 RTs received	#mla09 RTs given	#mla09 RTs received
laterribleliz (46)	boraz (85)	newfacmajority (25)	samplereality (49)
uncpublichealth (42)	apisanty (61)	ryancordell (20)	briancroxall (35)
olgag (30)	futureweb2010 (51)	jcmeloni (13)	nowviskie (33)

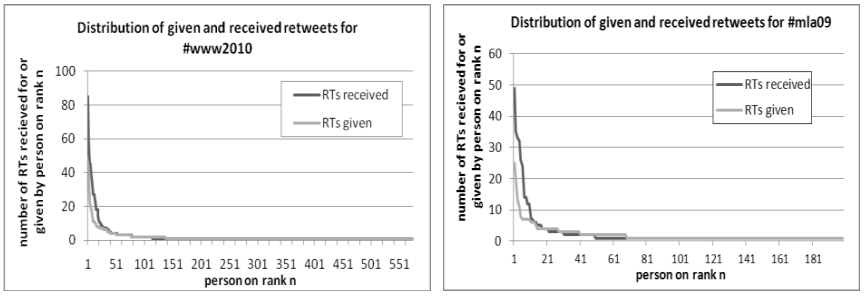


Fig. 3. Distribution of given and received retweets for #www2010 and #mla09.

In future, we intend to describe types of users based on the percentage of received and given RTs. For #mla09, there are 199 persons who have published at least one of the 413 retweets but only 89 persons who have 'received' at least one of those retweets[14]. Figure 3 shows the distribution of given and received retweets. It is furthermore possible to identify particular tweets that were highly cited. Within the manually collected retweets we have identified the most highly cited original tweets. Table 7 and 8 show the top 3 most cited tweets[15] for #www2010 and #mla09. Most of the highly cited tweets do also include URLs – thus, external and internal citations are interwoven in Twitter. For #mla09, the top 5 RTs all include an URL. The URL contained in the most frequent RT is also the most frequent one in Table 4, RT no. 2 includes the URL on rank 5 from Table 4, the URL in RT no. 3 is on rank 3.

[14] For #www2010 there are 574 users who have published at least one retweet and 239 who have received at least one.

[15] Here, only those tweets are summed-up that include the same text and refer to the same user.

Table 7. Top 3 retweets for #www2010 (manually detected retweets).

Tweet text and ID	From User	RTs
a delegação brasileira presente na #www2010 acaba de receber a notícia: a cidade do Rio de Janeiro sediará a Conferência #WWW2013 (ID: 13206448810)	w3cbrasil	24
twitter roomstreams for every conference room at #www2010 can be found at http://bit.ly/bRfE69 #302C (ID: 12881760468)	mstrohm	16
Summary of Twitter papers presented at #www2010 http://is.gd/bRqBF (ID: 13268676873)	alisohani	11

Table 8. Top 3 retweets for #mla09 (manually detected retweets).

Tweet text and ID	From User	RTs
Hey, guys, I've blogged about "the amplification of scholarly communication": Twitter, #MLA09, @briancroxall, & such: http://bit.ly/7SRgqZ (ID: 7221520139)	amanda-french	18
New at ProfHacker: "Academics and Social Media: #mla09 and Twitter," by @GeorgeOnline (and a bunch of you): http://wp.me/pAGUw-19K (ID: 7566711357)	profhacker	17
"Monopolies of Invention:" text of my #MLA09 talk on labor & IP issues in humanities collaboration: http://is.gd/5Gckz (ID: 7185970970)	nowviskie	16

6 Conclusion and Outlook

We have shown that scientists use two types of Twitter citations during scientific conferences. Users cite external sources in form of URLs and quote statements within Twitter via RTs. This is a first indication that citations/references in Twitter do not exactly serve the same purposes as classical citations/references. Future work should investigate more closely *why* users cite something on Twitter and compare the reasons with those that have been detected for classical citations. Furthermore, both types of Twitter citations may act as webometric resources: RTs may help to identify the most popular twitterers; URLs could be counted to measure impact of referenced publications or presentation slides. Both types appear with similar frequency within one dataset, but differences could be identified for the behavior of participants from the two different conferences. Future work will have to show, whether these differences indicate discipline-specific characteristics. Plans for successive work are the inclusion of additional conference datasets as well as the creation of datasets based on scientific twitterers, the analysis of citation patterns over time and the inclusion of qualitative work (e.g. intense content analyses and interviews with users).

Acknowledgements

Many thanks to Julia Verbina and Parinaz Maghferat for their contributions to data collection. Thanks to Bernd Klingsporn for advice and support and to Wolfgang G.

Stock for critical remarks. Thanks to our anonymous reviewers for helpful ideas. Financial support from the Heinrich-Heine-University Düsseldorf for the Research Group "Science and the Internet" is greatly acknowledged.

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: Understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis at ACM SIGKDD, San Jose, California, pp. 56--65. New York: ACM (2007)
2. Gaffney, D.: #iranElection: Quantifying Online Activism. In Proceedings of the Web Science Conference (WebSci10), Raleigh, NC, USA (2010)
3. Zhao, D., Rosson, M. B.: How and why people twitter: The role that microblogging plays in informal communication at work. In Proceedings of the 2009 SIGCHI International Conference on Supporting Group Work, pp. 243--252. New York: ACM (2009)
4. Vieweg, S., Hughes, A. L., Starbird, K., Palen, L.: Microblogging during two natural hazards events. In CHI 2010 – We are HCI: Conference Proceedings and Extended Abstracts of the 28th Annual CHI Conference on Human Factors in Computing Systems, Atlanta, GA, USA, pp. 1079--1088. New York: ACM (2010)
5. Dröge, E., Maghferat, P., Puschmann, C., Verbina, J., Weller, K.: Konferenz-Tweets: Ein Ansatz zur Analyse der Twitter-Kommunikation bei wissenschaftlichen Konferenzen. In Proceedings of ISI 2011: Internationales Symposium der Informationswissenschaft 2011, Hildesheim, Germany, pp. 98--110. Boizenburg: VWH (2011)
6. Ebner, M., Reinhardt, W.: Social networking in scientific conferences: Twitter as tool for strengthen a scientific community. In Learning in the Synergy of Multiple Disciplines. European Conference on Technology Enhanced Learning, Nice, France. Berlin: Springer (2009)
7. Letierce, J., Passant, A., Decker, S., Breslin, J. G.: Understanding how Twitter is used to spread scientific messages. In Proceedings of the Web Science Conference (WebSci10): Extending the Frontiers of Society On-Line, Raleigh, NC, USA (2010)
8. Ross, C., Terras, M., Warwick, C., Welsh, A.: Enabled backchannel: Conference Twitter use by digital humanists. Journal of Documentation, 67(2), 214--237 (2011)
9. Stock, W.G.: Information Retrieval. München, Wien: Oldenbourg (2007)
10. Thelwall, M.: Bibliometrics to webometrics. Journal of Information Science, 34(4), 605--621 (2008)
11. Priem, J., Hemminger, B. M.: Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. First Monday, 15(7) (2010)
12. Priem, J., Costello, K. L.: How and why scholars cite on Twitter. In Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, Pittsburgh, PA, USA, Article No. 75. New York, NY: ACM (2010)
13. Young, J. R.: 10 High Fliers on Twitter: On the microblogging service, professors and administrators find work tips and new ways to monitor the world. The Chronicle of Higher Education, 31, A10 (April 10, 2009)
14. Priem, J., Taraborelli, D., Groth, P., Neylon, C.: Alt-metrics: A Manifesto. Retrieved January 13, 2011, from http://altmetrics.org/manifesto (2010)
15. Stankovic, M., Rowe, M., Laublet, P.: Mapping tweets to conference talks: A goldmine for semantics. In Proceedings of the Third Social Data on the Web Workshop SDoW2010, collocated with ISWC2010, Shanghai, China (2010)
16. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In R. H. Sprague (Ed.), Proceedings of the 43rd Conference on System Sciences (HICSS 10), Honolulu, Hawaii, USA. Piscataway, NJ: IEEE (2010)

Making Sense of Location-based Micro-posts Using Stream Reasoning

Irene Celino[1], Daniele Dell'Aglio[1], Emanuele Della Valle[2,1], Yi Huang[3], Tony Lee[4], Stanley Park[4], and Volker Tresp[3]

[1] CEFRIEL – ICT Institute, Politecnico of Milano, Milano, Italy
[2] Dip. di Elettronica e dell'Informazione – Politecnico di Milano, Milano, Italy
[3] SIEMENS AG, Corporate Technology, Muenchen, Germany
[4] Saltlux, Seoul, Korea

Abstract. Consider an urban environment and think to its semi-public realms (e.g., shops, bars, visitors attractions, means of transportation). Who is the maven of a district? How fast and how broad can such maven influence the opinions of others? These are just few of the questions BOTTARI (our Location-based Social Media Analysis mobile app) is getting ready to answer. In this position paper, we recap our investigation on deductive and inductive stream reasoning for social media analysis, and we show how the results of this research form the underpinning of BOTTARI.

1 Introduction

In the last few years, we have been witnessing the increasing popularity and success of Location-based Services (LBS), especially of those with a Social Networking flavour. Twitter, Facebook Places, foursquare, Gowalla are only a few examples of applications; those services bring a wide range on useful information about tourist attractions, local businesses and points of interests (POIs) in the physical world.

Although these services are enormously popular, users still suffer from a number of shortcomings. The overwhelming information flow coming from those channels often confuses users; it is also very difficult to distinguish between a fair personal opinion and a malicious or opportunistic advice. This might be the reason why users primarily link to people they know personally since there is no clear way find out those who are trustable in an on-line social network.

In this paper, we present our collaborative effort to the design and development of the BOTTARI application, a Location-based Service for mobile users that exploit Social Media Analysis techniques to identify the "mavens" of a specific geographical area, who can be considered as experts of the POIs in this area. BOTTARI was conceived by Saltlux, a Korean Knowledge Communication Company. The application is still under development and it will be made available to Korean users in the Seoul area.

BOTTARI exploits hybrid Stream Reasoning both on heterogeneous social network data [1] and geo-location data. The hybrid reasoning engine combines deductive and inductive techniques. Since the input data are huge and change in real-time, the reasoning engine works by processing streaming data. The hybrid reasoning engine is developed on top of the LarKC platform [2], a pluggable architecture to build applications with Semantic Web technologies.

The remainder of the paper is organised as follows. Section 2 explains the concept of stream reasoning and delineates the system architecture. Section 3 describes the BOTTARI app. Section 4 details some user questions in terms of queries to our stream reasoner. Finally, Section 5 concludes the paper.

2 System Architecture

Continuous processing of information flows (i.e. **data streams**) has widely been investigated in the database community. [3]. In contrast, continuous processing of data streams *together with rich background knowledge* requires semantic reasoners, but, so far, semantic technologies are still focusing on rather static data. We strongly believe that there is a need to close this gap between existing solutions for belief update and the actual need of supporting decision making based on data streams and rich background knowledge. We named this little explored, yet high-impact research area **Stream Reasoning** [4]. The foundation for Stream Reasoning has been investigated by introducing technologies for wrapping and querying streams in the RDF data format (e.g., using C-SPARQL [5]) and by supporting simple forms of reasoning [6] or query rewriting [7].

We are developing the Stream Reasoning vision on top of LarKC [8]. The LarKC platform is aimed to reason on massive heterogeneous information such as social media data. The platform consists of a framework to build workflows, i.e. sequences of connected components (plug-ins) able to consume and process data. Each plug-in exploits techniques and heuristics from diverse areas such as databases, machine learning and the Semantic Web.

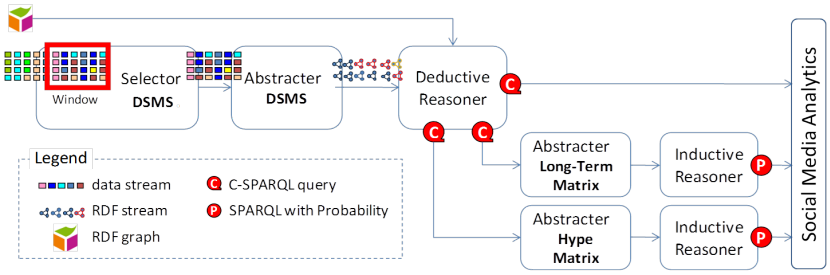


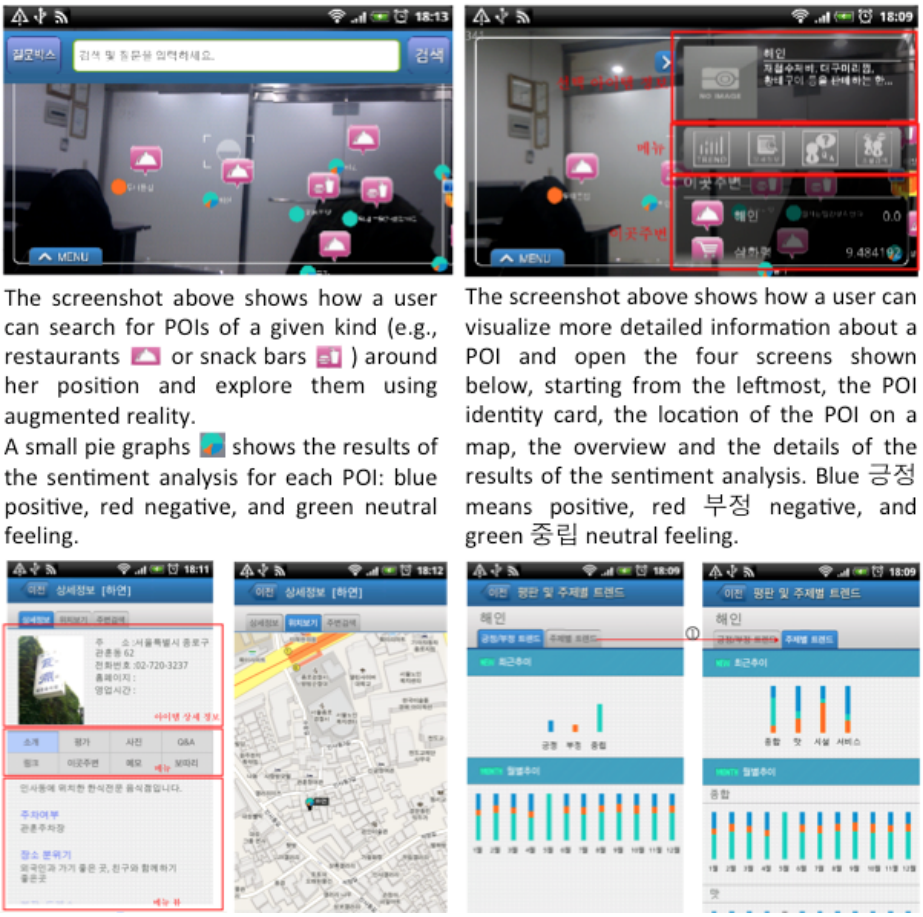
Fig. 1. Architecture of our Stream Reasoner

We built our Stream Reasoning system by embedding a deductive reasoner and an inductive reasoner within the LarKC architecture (see Figure 1). First, BOTTARI pre-processes the micro-posts by extracting information[5] whether a micro-post expresses a positive or a negative feeling of its author about a certain POI. After BOTTARI data arrives to the stream reasoner as set of data streams, a selection plug-in extracts the relevant data in each input stream in form of windows. A second plug-in abstracts the window content from fine grain data streams into aggregated events and produces RDF streams. Then, a deductive reasoner plug-in is able to register C-SPARQL queries, whose results can be of immediate use (cf. Section 4) or can be processed by other two sub-workflows. Each sub-workflow is constituted by an abstracter and an inductive reasoner, which uses an extended version of SPARQL that supports probabilities [9].

3 The BOTTARI mobile app

The BOTTARI mobile app is a location-based service that exploits the social context to provide relevant contents to the user in a specific geographic location.

[5] This technology is a Saltlux trade secret.



The screenshot above shows how a user can search for POIs of a given kind (e.g., restaurants or snack bars) around her position and explore them using augmented reality.
A small pie graphs shows the results of the sentiment analysis for each POI: blue positive, red negative, and green neutral feeling.

The screenshot above shows how a user can visualize more detailed information about a POI and open the four screens shown below, starting from the leftmost, the POI identity card, the location of the POI on a map, the overview and the details of the results of the sentiment analysis. Blue 긍정 means positive, red 부정 negative, and green 중립 neutral feeling.

Fig. 2. Some screenshots of the BOTTARI Android application

The purpose of the BOTTARI service is to provide recommendations on local context information to users through an augmented reality interface. BOTTARI gives detailed information on local POIs, including trust or reputation information. In Figure 2 , we provide some sample screenshots on how the BOTTARI mobile application will look like once completed.

The input data for the BOTTARI service come from public social networks and location based services (Twitter, local blogs and Korean news), are converted in RDF streams and are then processed and analysed by the system described in Section 2. The RDF-ized data are modelled with respect to the ontology represented in Figure 3, which is an extension to the SIOC vocabulary [**?**]. Our model takes into account the specific relations of Twitter (followers/following, reply/retweet); it adds the geographical perspective by modelling the POIs; it includes the "reputation" information by means of positive/negative reviews.

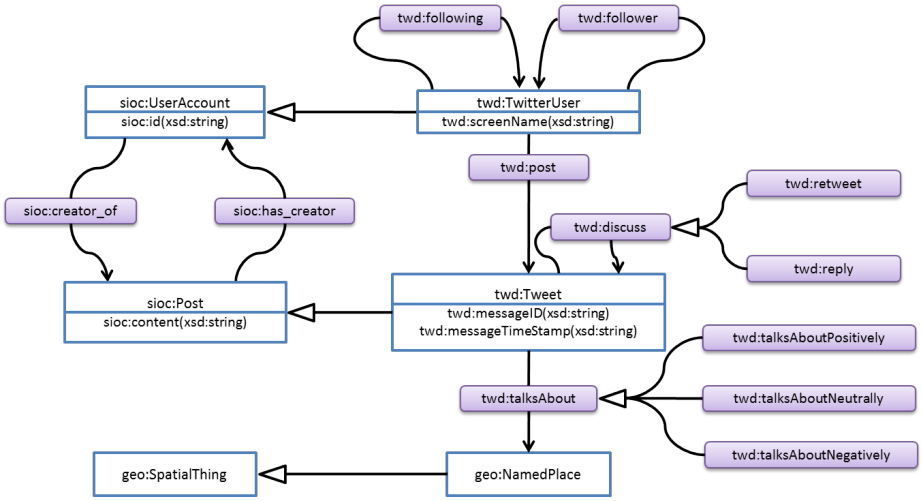


Fig. 3. Ontology modelling of BOTTARI data

4 Computing Answers to User Questions

The hybrid Stream Reasoning solutions we are developing is able to answer questions like: Who are the opinion makers (i.e., the users who are likely to influence the behaviour of their followers with regard to a certain POI)? How fast and how wide are opinions spreading? Who shall I follow to be informed about a given category of POIs in this neighbourhood?

In the rest of the section we show how to issue the three queries above using C-SPARQL and SPARQL with probabilities.

Who are the opinion makers?

Lines 1 and 3 of the following listing tell the C-SPARQL engine to register the continuous query on the stream of micro-posts generated by BOTTARI considering a sliding window of 30 minutes that slides every 5 minutes. Line 2 tells the engine that it should generate an RDF stream as output reporting who are the opinion makers for a certain POI and if they are rating it positively or negatively.

```
1.  REGISTER STREAM OpinionMakers COMPUTED EVERY 5m AS
2.  CONSTRUCT { ?opinionMaker a twd:opinionMaker ;  twd:discuss [ ?opinion ?poi ] . }
3.  FROM STREAM <http://bottari.saltlux.com/posts>  [RANGE 30m STEP 5m]
4.  WHERE {
5.          ?opinionMaker a twd:TwitterUser ;
6.                        twd:posts [ ?opinion ?poi ] .
7.          ?follower sioc:follows ?opinionMaker;
8.                    twd:posts [ ?opinion ?poi ] .
9.          FILTER ( cs:timestamp(?follower) > cs:timestamp(?opinionMaker)
10.                 && ?opinion != twd:talksAbout )
11. }
12. HAVING ( COUNT(DISTINCT ?follower) > 10 )
```

The basic triple pattern (BTP) at lines 5 and 6 matches micro-posts of the potential opinion makers with a POI. The variable `opinion` can match one of the properties `talksAbout`, `talksAboutPositively`, or `talksAboutNegatively`. The BTP at lines 7–8 looks up the followers of the opinion makers. The

FILTER clause at line 9 checks whether the micro-posts of the followers, which talk about the same POI, occurs after those from the opinion makers. At line 10 the query filters out actions of type `twd:talksAbout` and concentrates on micro-posts clearly discussing a POI in a positive or negative way. Finally, at line 12 the clause HAVING promotes the true opinion makers which have at least ten followers who expressed the same opinion about the POI after them.

How fast and wide opinions are getting spread?

Using the RDF stream computed by the previous query, the query in the following listing informs about how wide the micro-posts of an opinion maker are getting spread in half an hour. To do so, it considers the reply and re-tweet relationships among tweets (i.e., tweets linked by the `discuss` property in BOTTARI data model). Being `discuss` a transitive property, the C-SPARQL engine uses the materialization technique presented in [6] to incrementally compute the transitive closure of `discuss`.

```
1.  REGISTER STREAM OpinionSpreading COMPUTED EVERY 30s AS
2.  SELECT ?user ?opinionMakerTweet count(?aPositiveTweet) count(?aNegativeTweet)
3.  FROM STREAM <http://bottari.saltlux.com/posts> [RANGE 30m STEP 30s]
4.  FROM STREAM <http://bottari.saltlux.com/OpinionMakers [RANGE 30m STEP 30s]
5.  WHERE {
6.        ?user a twd:opinionMaker ;
7.              twd:post ?opinionMakerTweet .
8.        { ?aPositiveTweet a twd:Tweet ;
9.                          twd:discuss ?opinionMakerTweet ;
10.                         twd:talksAboutPositively ?poi .
11.       } UNION {
12.         ?aNegativeTweet a twd:Tweet ;
13.                         twd:discuss ?opinionMakerTweet ;
14.                         twd:talksAboutNegatively ?poi .
15.       }
```

Lines 1, 3 and 4 tell the C-SPARQL engine to register the continuous query on the stream of micro-posts generated by BOTTARI and on the streaming results of the opinion makers query. In both cases, a sliding window of 30 minutes, which slides every 30 seconds, is considered. The BTP at lines 6–7 matches the micro-posts of the opinion makers. The BTP at lines 8–10 and the BTP at lines 12–14 look up other micro-posts that, respectively, positively and negatively discussed those of the opinion makers. Line 2 asks the engine to generate a variable binding reporting how many positive and negative micro-posts are discussing the micro-posts of the current opinion makers.

Who shall I follow?

Let us consider now a specific BOTTARI user named Giulia. In the following listing we show a query that asks for the mavens Giulia should follow to be informed about attractions for kids, even among people she does not know. The system uses the social network of Giulia and the last window in the stream (generated by the query in the first listing) to determine such predicted probability.

```
1.  SELECT ?user ?prob
2.  FROM STREAM <http://bottari.saltlux.com/OpinionMakers [RANGE 30m STEP 30s]
3.  WHERE{
4.    ?opinionMaker a twd:opinionMaker ;
5.                  twd:discuss [ twd:talksAboutPositively ?poi ] .
6.    ?poi skos:subject twd:attractionsForKids .
7.    :Giulia twd:following ?opinionMaker.  WITH PROB ?prob
8.    FILTER ( ?prob > 0.8 && ?prob < 1 )
9.  } ORDER BY ?prob
```

The BGP at lines 4–6 matches the opinion makers that have been recently expressing positive opinions about attractions for kids. The triple patter at line 7 matches BOTTARI users that Giulia is following. Note that the `following` relationship may have not been asserted yet, the construct WITH PROB extends SPARQL by letting it query an inducted model. The variable `?prob` assumes the value 1 for the user she follows already and assumes the estimated probabilities between 0.8 and 1 for users she may be recommended to follow (cf. line 8). The ORDER BY clause is used to return users sorted by decreasing probability. The query answer includes pairs of users and predicted likelihood (e.g. `:Alice` with probability 0.99, `:Bob` with probability 0.87).

5 Conclusions and Future Works

In this paper we presented BOTTARI, a location-based mobile application which is able to supply contents and personalized suggestions to the users. We explained the processing of new recommendations, based on the elaboration of data streams generated by microblogging platforms like Twitter and foursquare. The computation is defined as a workflow combining Semantic Web and machine learning techniques and it is executed on top of the LarKC platform.

Our future work will focus on the development of the first stable version of the BOTTARI application and its release as Android app. The initial release will focus on Korea and will be evaluated by following a user-centered approach: a set of users will try out the application, supplying us feedbacks via a survey with questions about the system and its accuracy in providing suggestions.
This work was partially supported by the EU project LarKC (FP7-215535).

References

1. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Huang, Y., Tresp, V., Rettinger, A., Wermser, H.: Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics. IEEE Intelligent Systems **25**(6) (2010) 32–41
2. Cheptsov, A., et al.: Large Knowledge Collider. A Service-oriented Platform for Large-scale Semantic Reasoning. In: Proceedings of WIMS 2011. (2011)
3. Garofalakis, M., Gehrke, J., Rastogi, R.: Data Stream Management: Processing High-Speed Data Streams. Springer-Verlag New York, Inc. (2007)
4. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a Streaming World! Reasoning upon Rapidly Changing Information. IEEE Intelligent Systems **24**(6) (2009) 83–89
5. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: C-SPARQL: a Continuous Query Language for RDF Data Streams. Int. J. Semantic Computing **4**(1) (2010) 3–25
6. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental Reasoning on Streams and Rich Background Knowledge. In: Proc. of ESWC2010. (2010)
7. Ren, Y., Pan, J.Z., Zhao, Y.: Towards Scalable Reasoning on Ontology Streams via Syntactic Approximation. In: Proc. of IWOD2010. (2010)
8. Fensel, D., et al.: Towards LarKC: a Platform for Web-scale Reasoning. In: Proc. of ICSC 2008. (2008)
9. Tresp, V., Huang, Y., Bundschus, M., Rettinger, A.: Materializing and querying learned knowledge. In: Proc. of IRMLeS 2009. (2009)
10. Berrueta, D., et al.: SIOC Core Ontology Specification. W3C Member Submission, W3C (2007)

Polemical video annotation by Twitter

Samuel Huron[1], Yves Marie Haussonne[1],
Yves-Marie L'Hour[1], Alexandre Monnin[1]

[1] Institute for Research and Innovation,
4 rue Aubry le Boucher 75004 Paris

{samuel.huron, yves-marie.haussonne, yves-marie.lhour, alexandre.monnin}
@centrepompidou.fr

Abstract. In this paper we present a method to enhance video metadata by using microposts generated through social interactions during live events. Our goal is to make visible the audience "polemical activity" (the exchange of arguments, counter-arguments and references) elicited by the talk, and use it as a tool to browse the video record. To achieve it, we design a new interface and service that makes a synthetic view of microposts interaction.

Keywords: micropost, annotation, video, social interactions, live, polemic

1 Introduction

During a public event, more and more social Web tools are used to post real-time information (e.g.: Twitter, Foursquare, Facebook). In most cases, users can follow the production of microposts during the event thanks to tagging systems - for instance, the hashtags on Twitter.

For live video streams broadcasts, various webservices already offer Web pages with an embedded video player and interfaces for reading and writing microposts. This design pattern is interesting because it contextualizes the production and consumption of microposts during the talk. Despite the undeniable contextualization offered by this kind of interfaces (video and tweets), their use is not accurate in all cases. Like asynchronous or distant users may encounter difficulties to link the purpose of the talk with existing microposts, and due to the heterogeneous nature of microposts, it is difficult to generate a synthetic overview from the polemical activity. After the event, the memory of the social interactions is lost (especially on Twitter), and it is hard to retrieve the video sequence in relation to a given micropost.

2 Polemic tweet device

In our experiment we tried to address these issues by making a device to qualify and quantify micropost interaction. Before the event we are sharing with the audience a flyer (Fig.1) which present a simple "polemical syntax" to express formally the polemical position adopted in a micropost during the talk. During the event we are recording videos stream and microposts containing the live event hashtag and propose a polemical twitter client (Fig.2). After the event

we produce a special interface to browse and represent the aggregation of microposts synchronised with the video recording (fig.3).
This experiment was done on a 1 hour and 11 minutes Clay Shirky's talk the 31st January 2011. We have harvested 440 tweets with the "#rsln" hashtag including 97 tweets with the polemical syntax.

Fig. 1

Fig.1 The distributed flyer to announce the polemical syntax

Fig. 2

Fig. 2 Visual extract of the twitter client[1]

Fig. 3

Fig.3 Time stamped bar chart to synchronize tweets with the video timeline [2]

3 Benefits for social annotation practices

The polemic tweet device leverages the analysis and synthesis possibilities offered by classical linear tweeter interface. Timeline provides a graphical representation of tweets flow, revealing the reference content, immediately laying the emphasis on timecoded hot spot. Moreover it offers unprecedented access to the tweets for *a posteriori* analysis.

Last but not least, in addition to deepening the critical dimension of a discussion with the present or distant audience, the polemical tweeter annotation device allows engaging the user to take position through the polemical syntaxes and enable to avoid the determination of his position by post processing methods (like natural language algorithm or a Mechanical turk form).

Here are some additional benefits of the device for social annotations practices:

- It encourages public and audience to participate to the debate, to take position and to formalize arguments;
- It induces a new subjective social annotation level, in complement to the so-called objective indexation or quotation level in so far as it enables participants to become aware of their subjective approach to an issue;
- It offers new opportunities to study feedback loops on the microposts, particularly by identifying group emergences;

[1] http://amateur.iri.centrepompidou.fr/live/client.php
[2] **http://amateur.iri.centrepompidou.fr/live/rsln/polemicaltimeline.php**

References

1. Bermingham, A., Smeaton, A.F.: Classifying sentiment in microblogs: is brevity an advantage?. In: Proceedings of the 19th ACM international conference on Information and knowledge management ACM, New York, NY, USA (2010)
2. Boyd, D., Golder, S., Lotan .: "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter,". In: Hawaii International Conference on System Sciences, pp. 1-10 (2010)
3. Letierce J., Passant A., Decker S.: Using Social Media to Spread Science and to Engage Readers in Conversation. In: 11th International Conference on Public Communication of Science & Technology (2010)
4. Diakopoulos, N. A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th international conference on Human factors in computing systems (2010)

Extracting Semantic Entities and Events from Sports Tweets

Smitashree Choudhury[1], John G. Breslin[2]

[1] DERI, National University of Ireland, Galway, Ireland
[2] School of Engineering and Informatics, National University of Ireland, Galway, Ireland
smitashree.choudhury@deri.org, john.breslin@nuigalway.ie

Abstract. Large volumes of user-generated content on practically every major issue and event are being created on the microblogging site Twitter. This content can be combined and processed to detect events, entities and popular moods to feed various knowledge-intensive practical applications. On the downside, these content items are very noisy and highly informal, making it difficult to extract sense out of the stream. In this paper, we exploit various approaches to detect the named entities and significant micro-events from users' tweets during a live sports event. Here we describe how combining linguistic features with background knowledge and the use of Twitter-specific features can achieve high, precise detection results (f-measure = 87%) in different datasets. A study was conducted on tweets from cricket matches in the ICC World Cup in order to augment the event-related non-textual media with collective intelligence.

1. Introduction

Microblogging sites such as Twitter[1], Tumblr[2] and Identi.ca[3] have become some of the preferred communications channels for online public discourse. All of these sites share common characteristics in terms of their real-time nature. Major events and issues are shared and communicated on Twitter before many other online and offline platforms. This paper is based on data obtained from Twitter because of its popularity and sheer data volume. The amount of content that Twitter now generates has crossed the one billion posts per week mark from around 200 million users, covering topics in politics, entertainment, technology and even natural disasters like earthquakes and tsunamis. Extracting useful information from this constant stream of uninterrupted but noisy content is not trivial.

[1] http;//www.twitter.com/
[2] http://www.tumblr.com/
[3] http://www.identi.ca/

The extraction of useful content such as entities, events and concepts needs to address many conventional IR-related issues as well as some Twitter-specific challenges. Nevertheless, the results can be useful in many real-world application contexts such as trend detection, content recommendation, real-time reporting, event detection, and user behavioural and sentiment analysis, to name a few. In the present study, we tried to detect named entities and interesting micro-events from user tweets created during a live sports event (a cricket match). The application of these results aims to augment sports-related multimedia content generated elsewhere on the Web.

Making sense of social media content is not trivial. There are many social media-specific challenges in capturing, filtering and processing this content. Some of the typical issues are as follows:

- Tweets are 140 characters in length, forcing users to use short forms to convey their message. Many routine words are shortened such as “pls” for “please”, “forgt” for “forgot”, etc. We need a special dictionary to understand this constantly-evolving community-specific lingo.
- There is a lack of standard linguistic rules. Due to the lack of space, language rules are avoided when necessary, and as a result conventional information extraction techniques do not work as expected.
- The use of slang words, abbreviations and compound hashtags are community driven rather than based on any dictionary or knowledge base.

The goal and objective of this paper is to classify the tweets mentioning the named entities and interesting events occuring during a live game. Despite knowing that the content generated during an event includes discussions and opinions about the event, detecting the discussed entities and interesting sub-events is challenging. As an example, consider a tweet “O'Brien goes ARGH!!!” which actually means that a player called (surname) O'Brien got out. Manual observation says that this tweet contains one named entity (the player's name) and one interesting event (getting out), but text processing applications fail to detect them due to the lack of context rules. We propose various approaches including linguistic analysis, statistical measures and domain knowledge to get the best possible result. For instance, instead of simple term frequency measures, we represent each player and possible interesting events with features drawn from multiple sources and further strengthen their classification score with various contextual factors and user activity frequency (tweet volume).

Our contribution includes:

- Detecting named entities based on various feature sets derived from tweets and with the help of background knowledge such as event websites and Wikipedia.
- Developing a generic framework to detect interesting events which can be easily transferred to other sports events.

Figure 1 shows a visual illustration of the steps followed in this work.

The rest of the paper is organised as follows: section 2 presents our methodology and approaches to address the issues of feature selection and classification; section 3 describes the evaluation and results of the study. Related work is discussed in section 4, followed by conclusions in section 5.

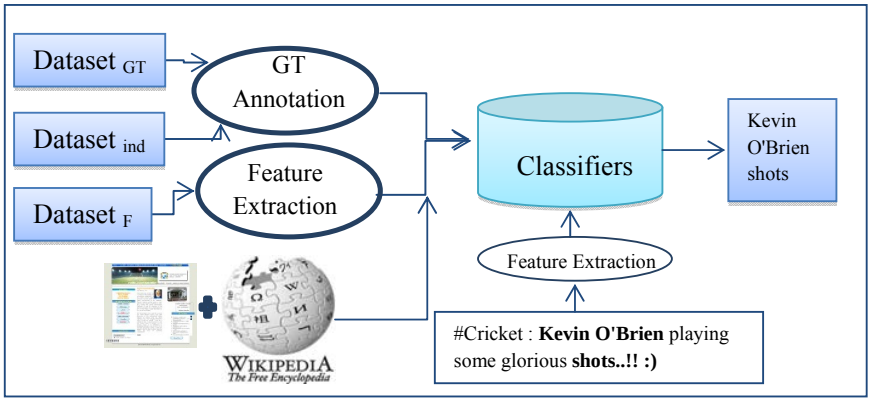


Fig.1. Overview of various steps followed.

2. Methodology

Our goal is to build classifiers which can correctly detect the players' named entities and the interesting micro-events within a sports event. We started by crawling tweets during the time of the cricket matches using the Twitter API. Since we can crawl tweets with keywords, we collected some related keywords and various hashtags (ICC cricket world cup, #cwc2011, cwc11, cricket, etc.) as a seed query list. Despite our filtered and focused crawling, many users use the popular hashtags and keywords to spam the stream to get attention. Including these tweets due to the mere presence of hashtags or keywords may bias the analysis, so a further round of de-noising is performed following a few simple heuristics as described below:

1. Messages with only hashtags.
2. Similar content, different user names and with the same timestamp are considered to be a case of multiple accounts.
3. Same account, identical content are considered to be duplicate tweets.
4. Same account, same content at multiple times are considered as spam tweets.

Using the above heuristics, we were able to identify and remove 1923 tweets from the dataset of 20,000 tweets. Our goal is not to eliminate all noise but to reduce it as much possible in order to get a proportionally higher percentage of relevant tweets.

The next step is to divide the datasets into two parts ($D_{Feature}$ and $D_{GroundTruth}$). $D_{GroundTruth}$ is manually annotated and $D_{Feature}$ is used for feature extraction. Each event and entity is considered as a target class and is represented with a feature vector. Details of the feature vector are described in sections 2.3 and 2.4.1.

Once the players are represented with the feature vector, the next step is to classify the tweets to say whether it contains any mention of a player or not. If the classification is positive, then matching is performed based on the player's full name. Each player is considered as a target class. Let $P = \{p_1, p_2, \ldots p_n\}$ be a set of players and let $FV(p_i)$ be a set of features used to represent the player. Let $M = \{m_1, m_2, \ldots m_n\}$ be a set of tweets belonging to a single game. We then train the classifier:

$$f(pi, mi) = \begin{cases} 1 = if\ mi\ makes\ a\ reference\ to\ a\ player\ pi \\ 0 = if\ mi\ makes\ no\ reference\ to\ a\ player\ pi \end{cases}$$

where p_i is the player's feature and m_i is the input tweet. Similar classification is performed for the micro-event detection task.

2.1 Dataset

We collected three datasets for training, testing and feature selection. Dataset (D_F) is a collection of 20,000 messages collected during the first round matches of the ICC World Cup. Dataset D_{GT} is a subset of D_F and consists of 2000 tweets. Dataset $D_{independent}$ (D_{ind}) (independent of training) is a set of 1500 messages from one game played between Ireland and England. Dataset D_{GT} and D_{ind} are manually annotated with a label of the player's name for any player entities and with "yes", "no" or "others" for the presence or absence of interesting events. Three students with a knowledge of the game were asked to annotate D_{GT} and D_{ind}. To increase the quality, we gave them information regarding the matches they were looking at and also regarding the team players. To maintain the quality of annotations, we considered that two out of three annotators had to agree for a label. The results showed that all three agreed on labels in 86% of cases while agreement between two occurred 94% of the time.

2.2 Background Knowledge

Since the main event (a game between two teams) is a pre-scheduled event, we obtained the background knowledge - in terms of the team names, venue, date, starting time, duration, and player details (names) - from the game website. We also collected various concepts common to cricket games from Wikipedia as a list of context features. The list consists of domain terms such as "crease", "field", "wicket", "boundary", "six", "four", etc. All of this background information was collected manually.

2.3 Feature Selection for Entity Detection

We developed a player classifier which captures a few general characteristics and language patterns from the tweets. Each feature is given a binary score of 1, 0.

2.3.1 Terms Related to a Player: The vector consists of name-related features. These are: full name, first name only, last name only, initials, etc. One more feature which we considered to be useful was the nickname of the player. However, since correlating nicknames to player names proved difficult, we could not include that feature. Table 1 below shows a few examples of the feature subset.

Table 1: Features related to a player.

Player	**Name-Related Feature**
Kevin Peterson	<Kevin Peterson, Peterson, KP>
Sachin Tendulkar	<Sachin Tendulkar, Sachin, Tendulkar, SRT>

2.3.2 Terms Related to the Game: While studying the tweets, we realised that a player's name alone and its variations will lead to low precision as there may be many irrelevant discussions mentioning the player's name. In order to increase the quality and precision, we added a context feature where the game-related key terms appear within a window of four words. These key terms are manually prepared, which has been discussed in the background knowledge section. Examples of such occurrences are given below in Table 2. If we find these rules existing in the message, the feature score becomes 1.

Table 2: Tweets with the context feature.

#Cricket : **Kevin O'Brien** playing some glorious **shots..!! :)**
Captain Afridi goes this time**, wicket** for **Jacob Oram.**
First **SIX** of the tournament for **Afridi!!!** #cwc2011

As tweets are highly informal, capitalisation is infrequent, but when it does occur we count it as a feature and score accordingly. Many players are now addressed and mentioned via their Twitter account, so the presence of a player's username (@<player>) or hashtag (#<player>) are also counted as Twitter-specific features. Finally, a player's feature vector looks like:

$$FV(p_i) = \langle full\_name, first\_name\_only, last\_name\_only, initials, initial{+}lastname, context\_word, capitalisation, player\_mention, player\_hashtag \rangle$$

2.4 Micro-Event Detection

An event is defined as an arbitrary classification of a space/time region. We target events which are expected to occur during a certain time frame (i.e. the match duration), but location is not an issue here as we know the venue of the match and we are not interested in fine-grained locational information such as field positioning within the stadium. We made a few assumptions regarding an event's characteristics, namely that (1) they are significant for the results of the game, and (2) many users (the audience) will be reacting to these events via their tweets. The methodology options available for detecting game-related micro-events from tweets are: (1) statistical bursty feature detection; and (2) feature-based event classification. We combined both approaches to get the best possible result.

2.4.1 Event Feature Selection

Interesting events that arise during a game are not pre-scheduled, but there is the possibility that these events can occur at any moment of time during the game. We manually selected these events from the Wikipedia "Rules of Cricket" pages. There are two broad categories ("scoring runs" and "getting out") and 12 sub-categories of micro-events. Through our observation of tweets, we saw that most tweets referred to the "out" event by itself while not bothering too much with the specific "out" types such as "bowled", "LBW" or "run-out", though they are occasionally mentioned. Based on this, we restricted our classification task to three major possible events, i.e. "out", "scoring six", and "scoring four". Each event is represented with a feature vector which consists of keyword features related to the event.

Keyword Variations: An event is represented by various key terms related to the event. The logic of including such variations is that users use many subjective and short terms to express the same message - "gone", "departed", "sixer", "6", etc. – when caught up in the excitement of the game. These features are again extracted from the D_F dataset.

Linguistic Patterns: Like the player classifiers, the event classifier also includes contextual features and linguistic patterns to detect the events. The presence of such a pattern gets a score of 1 for the feature, otherwise 0. A few of the examples are shown below:

Table 3: Mentions of interesting events during a match.

#sixer from **#kevinobrien** for #ireland against #england #cricket
Kevin O'Brien OUT ! Ireland 317/7 (48.1 ov) #ENGvsIRE #cricket #wc11
Crap **O'Brien goes** ARGH!!!

2.4.2 Tweet Volume and Information Diffusion

We cannot say from a single tweet that an event has occurred. In order to make our detection reliable, we take crowd behaviour into account. Based on the assumption that interesting events will result in a greater number of independent user tweets, we computed two more features to add to the event feature vector: (1) the tweet volume; (2) the diffusion level. Tweet volume is the level of activity while the event is being mentioned, taken during a temporal interval tm_i where $i = \{1 \ldots n\}$ and the duration of each tm_i is two minutes (can be any duration depending the requirement). We used a two-minute interval for simplicity but it can be of any temporal size. If the number of messages is higher than a threshold of average plus 1 α, we mark the feature as 1, otherwise set it to 0.

The second feature is the level of information diffusion that takes place during the time interval tm_i. It is presumed that more and more users will be busy sharing and communicating the event through their own tweets rather than reading and forwarding others. This means that there will be less retweets (RTs) during the event interval compared to the non-event intervals. This assumption has been confirmed from our observations of the data that the immediate post-event interval has a lesser number tweets than the non-event intervals. The same assumption is also proved in the study [2]. The feature is marked the same way as the tweet volume feature.

3. Evaluation and Results

Our evaluation started with the dataset D_{GT} which is manually labelled both for players and interesting events. We first ran the players classifier and the results are shown in Figure 2. The objective of the evaluation is to judge the effectiveness of the proposed approaches to detect players' named entities and game-related micro-events against the manually-annotated datasets D_{GT} and D_{ind}. We also tested the weight of various features in classification (positive) and found that a combination of any name feature with the context feature (game-related term) is the best performing feature compared to any other combinations (Figure 5).

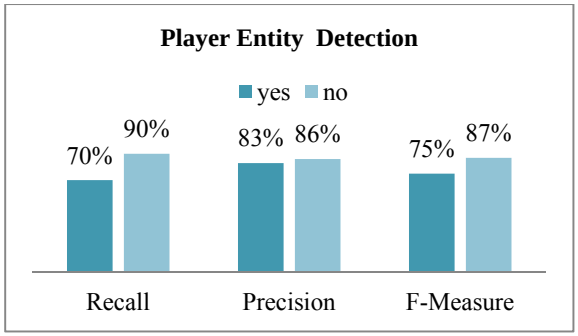


Fig. 2: Recall and precision of the player detection classifier.

Like the player classifier, we ran the same evaluation for micro-event detection but in two different stages: (1) classification with only linguistic features, and (2) classification with all features. With linguistic features only (Figure 3), recall is very low at 70% and precision is 74%. This may be due to the noise in tweets. Many event-related keywords are also used in normal conversations like "out", "over", etc.

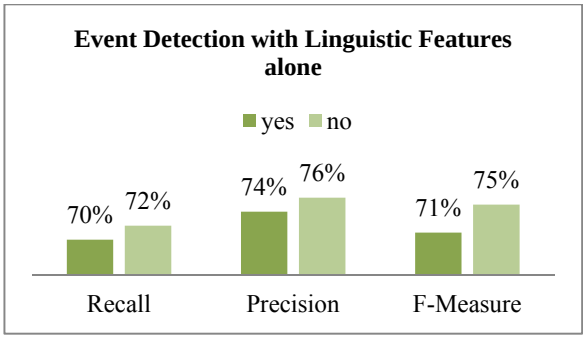


Fig. 3: Event detection performance with linguistic features only.

However, when we included the tweet volume and information diffusion level scores, both recall and precision further increased to 86% and 85% respectively, as shown in Figure 4.

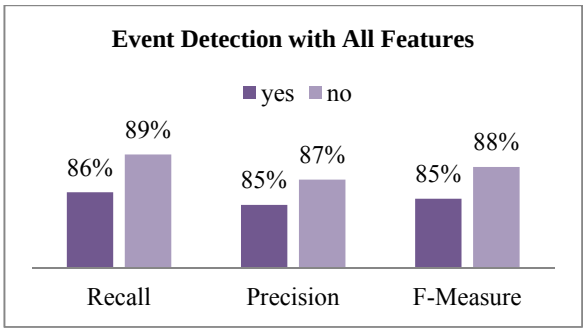


Fig. 4: Event detection performance with all features combined.

The results show that irrespective of any features, performance for the "no" labels is always better than for the "yes" labels. We assume this result may be due to the

greater number of negative samples available in the data compared to the positive samples.

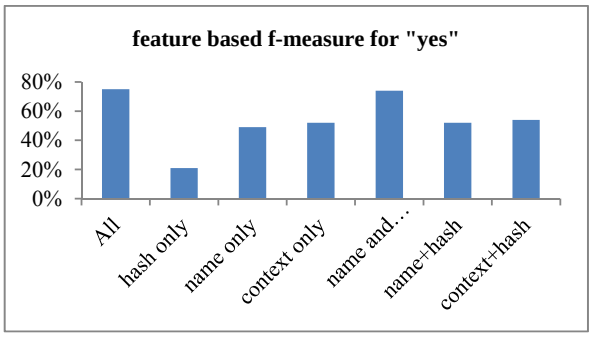


Fig. 5: Individual feature performance in player classification.

One question we were interested in answering was can the classifiers be used on other data which is independent of the training and the testing data? To explore this proposition, we ran the classifier on the independent dataset D_{ind} collected from a different game involving two different teams (England vs. Ireland). For this experiment, we tagged the content with part-of-speech tagging using the Stanford NLP tagger[4]; in the feature space, we replaced the player's name with a proper noun placeholder. A summary of the results for both players and event detection is shown Figure 6.

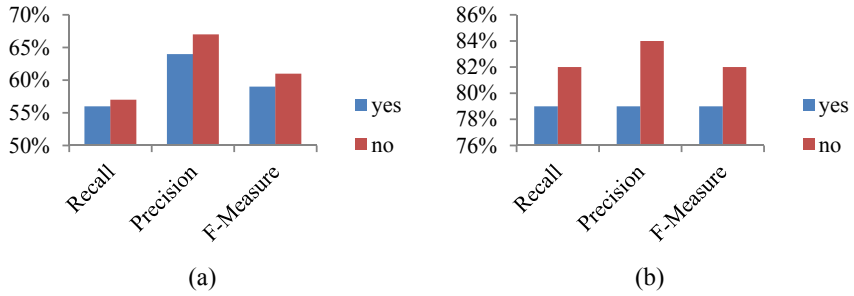


Fig. 6: (a) Player detection and (b) event detection in dataset D_{ind}.

As expected, the player classifier scored poorly compared to the event classifier, as the player classifier is heavily dependent on the players' names and their variations. Even if we replace the names with proper noun placeholders, many player mentions are only by first or last name, and other names could not be identified as proper nouns

[4] http://nlp.stanford.edu/links/statnlp.html

by the part-of-speech tagger. However, the event detection results are good, and the F-measure is above 80% as the features are more generic in nature.

4. Related Studies

Twitter is one of the most popular social media sites with hundreds of thousands of users sending millions of updates every day. It provides a novel and unique opportunity to explore and understand the world in real time. In recent years, many academic studies have been carried out to study issues such as tweet content structures, user influence, trend detection, user sentiment, the application of Semantic Web technologies in microblogging [1], etc. Many tools exist for analysing and visualizing Twitter data for different applications. For example, [3] analyses tweets related to various brands and products for marketing purposes. A news aggregator called "TwitterStand" is reported in [4] which captures breaking news based entirely on user tweets.

The present study addresses the research question of identifying named entities mentioned in microblog posts in order to make more sense of these messages. Therefore, the focus of our discussion in this section will be on various related studies concerning entity and event recognition in social media scenarios, especially in microblogs. Finin et. al [7] attempted to perform named entity annotation on tweets through crowdsourcing using Mechanical Turk and CrowdFlower. Similar research in [8] reported an approach to link conference tweets to conference-related sub-events, where micro-events are pre-defined as opposed to the sports domain where interesting events unfold as and when the event proceeds. Researchers in [2] built a classifier based on tweet features related to earthquakes and used a probabilistic model to detect earthquake events. Authors in [5] used content-based features to categorise tweets into news, events, opinions, etc. Tellez et al. [6] used a four-term expansion approach in order to improve the representation of tweets and as a consequence the performance of clustering company tweets. Their goal was to separate messages into two groups: relevant or not relevant to a company. We have adopted many lightweight techniques to identify named entities and micro-events during a sports event so that we can later use these results to address existing problems related to conceptual video annotation.

5. Conclusion

We presented approaches to identify named entities and micro-events from user tweets during a live sports game. We started with a filtered crawling process to collect tweets for cricket matches. We arranged three datasets (D_F, D_{GT}, D_{ind}); D_{GT} is a subset of D_F. D_{GT} and D_{ind} are manually annotated with player names and "yes" or "no" for players and events respectively, while D_F was used to extract the feature set. Classifiers built on these features were able to detect players and events with high precision. The generic features of our event detection classifier were applied to an independent dataset (D_{ind}) with positive results. Our future work includes transferring

the algorithm to other sports areas as well other domains such as entertainment, scientific talks and academic events.

Acknowledgments

This work was supported by Science Foundation Ireland under grant number SFI/08/CE/I1380 (Líon 2).

References

1. A. Passant, T. Hastrup, U. Bojars, J.G. Breslin, "Microblogging: A Semantic Web and Distributed Approach", The 4th Workshop on Scripting for the Semantic Web (SFSW 2008) at the 5th European Semantic Web Conference (ESWC '08), Tenerife, Spain, 2008.
2. T. Sakaki, M. Okazaki, Y. Matsuo. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors", Proceedings of the 19th World Wide Web Conference (WWW2010), Raleigh, NC, USA, 2010.
3. B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth", Journal of the American Society for Information Science and Technology, 2009.
4. J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, J. Sperling. "Twitterstand: News in Tweets", Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51, Seattle, WA, USA, November 2009.
5. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas, "Short Text Classification in Twitter to Improve Information Filtering", Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), pp. 841–842. New York, NY, USA, 2010.
6. F.P. Tellez, D. Pinto, J. Cardiff, P. Rosso, "On the Difficulty of Clustering Company Tweets", Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC '10), pp. 95–102. New York, NY, USA, 2010.
7. T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, M. Dredze, "Annotating Named Entities in Twitter Data with Crowdsourcing", Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10), 2010.
8. M. Rowe, M. Stankovic, "Mapping Tweets to Conference Talks: A Goldmine for Semantics", Proceedings of the 3rd International Workshop on Social Data on the Web (SDOW 2010) at the 9th International Semantic Web Conference (ISWC 2010), 2010.

Capturing Entity-Based Semantics Emerging from Personal Awareness Streams

A.E. Cano, S. Tucker, F. Ciravegna

Department of Computer Science,
University of Sheffield,
Sheffield, United Kingdom
{firstinitial.surname}@dcs.shef.ac.uk

Abstract. Social activity streams provide information both about the user's interests and about the way in which they engage with real world entities. Recent research has provided evidence of the presence of emergent semantics in such streams. In this work, we explore whether the online discourse of user's social activities can convey meaningful contextual information. We introduce a user-centric methodology based on tensor analysis for deriving personal vocabularies given an entity-based context. By extracting entities (e.g. location, organisation, people) from the user's stream content, we explore the data structures that emerge from the user's interrelationship with these entities. Our experimental results revealed that the simultaneous correlation of entities leads to the identification of concepts which are relevant to the user given a specific context. This methodology is relevant for mobile application designers (1) in fostering user entity-based ontologies for merging user context in pervasive environments, (2) for personalising entity-based recommendations.
Keywords: linked data streams, social awareness streams, microblogging, context

1 Introduction

The past few years have seen the launch of different social networking platforms that allows a user to expose their online presence, create groups and build bridges for communicating within their online social spheres. The high usage of these platforms has generated an enormous amount of personal information online creating unprecedented opportunities for a wide range of research related to knowledge management, user contextualisation, and the Semantic Web.

In this paper, we focus on the analysis of a user's social activity streams (a.k.a personal awareness streams [32]) generated from different social networks.We consider a user's social activity stream as a historical dataset from which context-sensitive items can be derived. Users produce data streams, not only providing information regarding the physical world (e.g. location, surrounding things) but also regarding their digital environment (e.g. adding new friends, microblogging). Therefore, we see the user's social activity streams as virtual sensors that could provide valuable information not only about the user interests but also about the user's physical contextual situation.

This paper sets out to explore whether the use of aggregations of personal awareness streams can convey meaningful contextual information given a set of different

entities that the user has interacted with within their online discourse during a timeline. In this paper, we introduce the Concept Selection Induced from Social Stream Aggregations (CSISSA) methodology, which captures entity-related information (e.g. organisations, locations, people, links) emerging from a personal awareness stream aggregation. This methodology is based on a three-mode network of social awareness streams (a.k.a. Tweetonomy [32]) and lightweight associative resource ontologies [20]. CSISSA applies tensor analysis for performing a simultaneous correlation of the given entities. Computing the decomposition of the tensor yields to conceptual structures that characterise a user given a context.

In this work we investigate the way in which a user refers to entities in the content of the message he generates. These entities are interlinked to others through, for example text and hashtags. We explore if this entity-based interrelationship can yield emerging conceptual structures that can aid in the user modelling. Our experimental results suggest that a key factor for successfully deriving relevant concepts for a given context is the user's microblogging verbosity, and the use of common vocabularies referring to the entities involved in the context.

The contributions of this paper are as follows: we study personal awareness stream aggregations as a source of information for deriving users' relevant concepts given an entity-based context. We present a novel approach which enables the explicit declaration of the context in which a user needs to be analysed. Our model abstracts the semantics of the vocabularies introduced by the user in his social activity stream by means of the derivation of lightweight ontologies. We make use of tensor analysis for building a user's entity-based context.The encapsulation of an entity-related lightweight ontology constitutes a slice of a tensor. The decomposition of this tensor reveals concepts relevant to the user in the analysed context. We believe that entity-based user modelling could aid in the future integration of user context to pervasive environments.

2 Background

In this section we start by defining concepts from principal component analysis (PCA) and then we give a brief introduction to tensor analysis. We will follow the typical conventions, and denote matrices with upper case bold letters (e.g. $\mathbf{X}$, row vectors with lower-case letters (e.g. $\mathbf{v}$), and tensors with calligraphic font (e.g., $\mathcal{X}$).

Principal Component Analysis (PCA) PCA [8] helps to identify patterns in data by expressing this data in such a way that it highlights a limited number of "components" that capture most of the information contained in the observed variables. By performing an orthogonal linear transformation, PCA finds the best linear projections which minimize least squares cost. For a given matrix $\mathbf{X}$ with zero mean (i.e. the mean of the distribution has been subtracted from the data set), PCA can be computed by obtaining the Singular Value Decomposition (SVD)[8][2] of $\mathbf{X}$; according to which $\mathbf{X} = \mathbf{U}_{svd} \times \mathbf{\Sigma}_{svd} \times \mathbf{V}^T_{svd}$; then $\mathbf{Y} = \mathbf{U}_{svd} \times \mathbf{\Sigma}_{svd}$ and $\mathbf{U} = \mathbf{V}_{svd}$. For example, if $\mathbf{X}$ is a user's status-keywords matrix taken from a user's stream aggregation dataset, then the $\mathbf{Y}$ and $\mathbf{U}$ matrices can be interpreted as the status-concept matrix $\mathbf{Y}$, and the keywords-concept matrix $\mathbf{U}$.

A user's post can be further analysed by considering not only keywords but also other resources (e.g. location, people) embedded on its content; forming a multidimensional set of parameters. An example for such analysis could study those topics that emerge from a user's posts generated during the morning hours at the office (location×time×keywords). A mathematical abstraction for the representation of a higher way structured data is a Tensor.

Tensor Analysis Tensors[12] are multidimensional M-ways or Mth-order arrays which generalize the notion of vectors(1-way or first-order array) and matrices (2-ways or second-order arrays). Tensors of order greater or equal to three are called higher-order tensors. In order to identify patterns that emerge from the simultaneous correlation of a set dimensions it is necessary to decompose a tensor. Tensor decomposition can be considered as a higher-order generalisation of SVD and PCA. In this paper we will use the Tucker decomposition approach.

Tucker Decomposition The Tucker decomposition was first introduced by Tucker in 1963 [30]. Given a tensor $\mathcal{X} \in R^{I_1 \times \ldots \times I_N}$ PCA is performed so as to decompose tensor $\mathcal{X}$ into a core tensor $\mathcal{G} \in R^{R_1 \times .. \times R_N}$ multiplied by a set of matrices $\mathbf{U}^{(i)} \in R^{I_i \times R_i}$. Therefore the Tucker decomposition of a three-order tensor $\mathcal{X}$ can be expressed as.

$\mathbf{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} =$
$\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \equiv [[\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]]$

One of the approaches for computing a Tucker decomposition of a three-order vector is to start with a first approximation obtained by applying a Higher Order SVD (HOSVD) [16] and then apply the alternating least squares algorithm (ALS) [15].

3 Related Work

Mika [20][28] explores how community-based semantics, in the form of lightweight associative ontologies, emerge from folksonomies. He introduces the semantic-social networks model which consists of a tripartite graph of people, concept and instance associations. Wagner and Strohmaier [32] introduce the Tweetonomy model, which is a formalisation of social awareness streams. This model adopts a theoretic approach similar to the one presented by Mika. However, the Tweetonomy model presents a more complex and dynamic structure than folksonomies. Strohmaier et al[18] and Körner et al[13], study quantitative measures for tagging motivation. In their study they found empirical evidence that the emerging semantics of tags in folksonomies are influenced by individual user tagging practices.

Tensor decompositions have a long history and have been applied in different research communities. In particular the Tucker decomposition has been used in chemical analysis [4], psychometrics [9] and computer vision [31]. Tensor analysis has also been applied in web search; Kolda et al [11] propose a method called Topical HITS (TOPHITS) which can be considered as an extension of Kleinberg's HITS (Hypertext Induced Topic Selections) algorithm [10]. TOPHITS analyses a semantic graph that combines anchor text with the hyperlink structure of the web. In order to avoid losing edge type information when modelling the adjacency structure of a semantic graph as a matrix, they modelled it as a three-way tensor containing both the hyperlink and an-

chor text information. Their tensor decomposition leads to triplets of vectors containing authority, hub scores for the pages, and topic scores for the terms.

Rendle and Thieme [25] apply tensor factorisation for personalised tag recommendation and learning. They introduce a model based on Tucker decomposition to explicitly model the pairwise interaction between users, items and tags. More similar to our work is the approach of Wetzker et al [33]. They follow a user-centric tag model for deriving mappings between personal tag vocabularies (a.k.a personomies [6]) and the corresponding folksonomies. Our approach differs from previous work in that rather than building the tensor as a three-way tensor of items-users-tags, we generate a three-way tensor in which each slice is a lightweight associative "resource" ontology; which allows to store multiple stream qualifiers in the tensor.

The analysis of user-generated content extracted from social media sites is an active research area. Qualitative and quantitative studies have been carried out for leveraging the "wisdom of crowds" [22]. Some of this research has focused on questions related to network and community structure. For example, Krishnamurthy et al [14] present a characterisation of Twitter social network, which includes patterns in geographic growth and user's social activity. In their work, they suggest that frequent updates might be correlated with high overlap between friends and followers. Java et al [7], present an analysis of Twitter and suggest that the differences in users' network connection structures can be explained by the following types of user activities: information seeking, information sharing and social activity.

Other work has presented a systematic analysis of the content of posts in social networks. Recent work [21], introduces the term "Social Awareness Streams" for referring to this aggregation of short status messages. They proposed a characterisation of these messages via a human coding of tweets into nine categories including "Information sharing" and "Self promotion". By extrapolating from these categories, they induced two types of users the "informers", who post about non-personal information, and the "meformers" which mostly post about themselves. Stankovic et al [17], study conference related tweets. They map tweets to talks an subevents that they refer to. Using linked data they derive additional knowledge about event dynamics and user activities.

Data structures emerging from the Social Web have been studied in the Information Retrieval and Semantic Web communities. Research in this area includes the study of content and link analysis algorithms and ontology learning algorithms. Heymann et al [5] present an algorithm for hierarchical taxonomy generation from social tagging systems. For generating a taxonomy of tags, they apply graph centrality in a cosine similarity graph of tags. Ramage et al [23], apply labelled Latent Dirichlet Allocation (LDA) [24] for mapping content of the Twitter feed into four dimensions including style and substance. Schmitz [26] introduces a subsumption-based model for inducing faceted ontologies from Flickr tag vocabulary. Our work was inspired mainly by Mika's [20], and Wagner and Strohmaier's [32] work. We apply the Tweetonomy formalisation for obtaining personal awareness stream aggregations. Our work differs from existing work (1) through our focus on deriving person-based lightweight ontologies from personal awareness stream; which enrich concepts and reveal structures that are meaningful to the owner of the stream; (2) we study the content of the messages not only in terms of traditional resources as hashtags, and links, but also in terms of entities (e.g loca-

tion, people, organisations); (3) we present a methodology based on tensor analysis that allows the definition of entity-based context for deriving person-based ontologies.

4 Social Stream Aggregation and Entity-Based Concept Induction

Our interest is to enable a way in which a user's social activity streams can be analysed in order to discover concepts that can aid in profiling him. These concepts are revealed as a combination of featuring dimensions. Example of these dimensions include e.g. a user's interests, user location, user's tendencies in favouring a position in a discussion etc. The following subsection presents the definition of three different social networks modelled as tripartite social awareness streams.

4.1 User's Social Stream Aggregation

Following the Tweetonomy model suggested by Wagner and Strohmaier[32], we describe a social awareness stream as a sequence of tuples S, according to the following definition:

Definition 1. *A tweetonomy is a tuple* $\mathtt{S} := (\mathtt{U}_{\mathtt{q1}}, \mathtt{M}_{q2}, \mathtt{R}_{q3}, \mathtt{T}, ft)$*, where*

- $\mathtt{U}$,$\mathtt{M}$,$\mathtt{R}$ *are finite sets whose elements are called users, messages and resources.*
- *Each of these sets are qualified by* $\mathtt{q1}$,$\mathtt{q2}$*, and* $\mathtt{q3}$ *respectively (explained below).*
- $\mathtt{T}$ *is the ternary relation* $\mathtt{T} \subseteq \mathtt{U} \times \mathtt{M} \times \mathtt{R}$ *representing a hypergraph with ternary edges. The hypergraph of a tweetonomy* $\mathtt{T}$ *is defined as a tripartite graph* $\mathtt{H}(\mathtt{T}) = \langle \mathtt{V}, \mathtt{E} \rangle$ *where the vertices are* $\mathtt{V} = \mathtt{U} \cup \mathtt{M} \cup \mathtt{R}$*, and the edges are:* $\mathtt{E} = \{\{u, m, r\} \,|\, (u, m, r) \in T\}$*. Each edge represents the fact that a given user associates a certain message with a certain resource.*
- f_t *is a function that assigns a temporal marker to each ternary edge.*

In this study we will focus on user-centric social streams generated in Facebook, Foursquare and Twitter, according to the following qualifiers:

- The way a user can be related to a message is represented by the qualifier $\mathtt{q1}$. For this analysis we only consider the authorship relationship: $\mathtt{U}_a$ (the author of the message).
- The qualifier $\mathtt{q2}$ represents the types of messages. This is a comment or a status in Facebook; a broadcast message, direct message, re-tweeted message in Twitter; a broadcast message (shout) in Foursquare are considered to be the same type. For this experiment we don't differentiate between types.
- The qualifier $\mathtt{q3}$ for resources considers: $\mathtt{R}_k$ (keywords), $\mathtt{R}_h$ (hashtags), $\mathtt{R}_{li}$ (URLs), $\mathtt{R}_{mlo}$ (message-emitted location), $\mathtt{R}_o$ (organisations - entities recognised as an organisation), $\mathtt{R}_p$ (people -entities recognised as a person), $\mathtt{R}_l$ (location - entities recognised as a location).

We focus on a user given the streams he has produced within a window of time. Given the tuples $\mathtt{T}_{\mathtt{facebook}}, \mathtt{T}_{\mathtt{foursquare}}, \mathtt{T}_{\mathtt{twitter}}$, we define the sets $\mathtt{U}, \mathtt{R}, \mathtt{M}$ as:

$\mathrm{U} = \mathrm{U}_{\mathtt{facebook}} \cup \mathrm{U}_{\mathtt{twitter}} \cup \mathrm{U}_{\mathtt{foursquare}}$,

$\mathrm{R} = \mathrm{R}_{\mathtt{facebook}} \cup \mathrm{R}_{\mathtt{twitter}} \cup \mathrm{R}_{\mathtt{foursquare}}$,

$\mathrm{M} = \mathrm{M}_{\mathtt{facebook}} \cup \mathrm{M}_{\mathtt{twitter}} \cup \mathrm{M}_{\mathtt{foursquare}}$

We are interested in extracting the concepts emerging from the streams produced by a user:

$\breve{\mathtt{u}} \in \mathtt{U} : \breve{\mathtt{u}} \in \mathrm{U}_{\mathtt{facebook}} \wedge \breve{u} \in \mathrm{U}_{\mathtt{twitter}} \wedge \breve{\mathtt{u}} \in \mathrm{U}_{\mathtt{foursquare}}$

In order to do so we consider a user stream aggregation defined as a tuple:

$\mathtt{S}_{\mathtt{a}}(\mathtt{U}') = (\mathtt{U}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', ft)$, where

$\mathtt{Y}' = \{(\mathtt{u}, \mathtt{m}, \mathtt{r}) \mid \mathtt{u} \in \mathtt{U}' \vee \exists \mathtt{u}' \in \mathtt{U}', \tilde{\mathtt{m}} \in \mathtt{M}, \mathtt{r} \in \mathtt{R} : (\mathtt{u}', \tilde{\mathtt{m}}, \mathtt{r}) \in \mathtt{Y}\}$

and $\mathtt{U}' \subseteq \mathtt{U}$ and $\mathtt{Y}' \subseteq \mathtt{Y}$. $\mathtt{S}_{\mathtt{a}}(\mathtt{U}')$, consists of all messages related with a user $\mathtt{u}' \in \mathtt{U}'$ and all the resources and users related with these messages.

4.2 Lightweight Associative Ontologies

An ontology, is a shared, formal conceptualization of a domain [3][1]. It is a data structure which is an advancement in conceptual modelling over taxonomic structures [28]. A lightweight ontology can be considered as an evolving classification structure created by users [27], which can be considered to be closer to a thesaurus (i.e. a structure organising topics). We want to derive a set of concepts from a simultaneous correlation among the resources q3 (e.g. keywords, hashtags, links) extracted from a user stream aggregation. In order to obtain this correlation, we start identifying those bipartite graphs (two-mode graphs) that could be of any interest to our analysis.

Consider for instance the association between keywords and location; which can be obtained as a combination of location×message ($\mathbf{R}_l\mathbf{M}$) and keywords×messages ($\mathbf{R}_k\mathbf{M}$). Where the location×messages (bipartite graph $\mathbf{R}_l\mathbf{M}$) is defined as:

$\mathrm{R}_l\mathbf{M} = \langle \mathtt{R_l} \times \mathtt{M}, \mathtt{E_{rm}} \rangle = \{(r, m) \,|r \in \mathtt{R_l} \wedge \exists u \in U : (u, m, r) \in E\}$,

$w : E \rightarrow R, \forall e = (r, m) \in \mathtt{E_{rm}}$

and the keywords×message (bipartite graph $\mathbf{R}_k\mathbf{M}$), is defined as:

$\mathrm{R}_k\mathbf{M} = \langle \mathtt{R_k} \times \mathtt{M}, \mathtt{E_{rm}} \rangle = \{(r, m) \,|r \in \mathtt{R_k} \wedge \exists u \in U : (u, m, r) \in E\}$,

$w : E \rightarrow R, \forall e = (r, m) \in \mathtt{E_{rm}}$

These bipartite graphs represent the adjacency or affiliation matrices: $\mathbf{R}_l\mathbf{M}$; which links the resources (of type location) to the messages in which this resource has been mentioned by this user. In the same way, $\mathbf{R}_k\mathbf{M}$; links the resources (of type keyword) to the messages in which this resource has been mentioned by at least one user. Each link (edge) can be weighted following a local or global weighting function in order to condition the data to be analysed (see Fig. 1).

Finally, the association between keywords and location is expressed as $\mathbf{R}_k\mathbf{R}_l = (\mathbf{R}_k\mathbf{M})(\mathbf{R}_l\mathbf{M})^{\mathbf{T}}$. We can now encapsulate the information that associates locations with keywords only in terms of keywords by multiplying $\mathbf{R}_k\mathbf{R}_l$ with its transpose, i.e. $\mathbf{O}(\mathbf{R_k}\mathbf{R_l}) = (\mathbf{R_k}\mathbf{R_l})(\mathbf{R_k}\mathbf{R_l})^{\mathbf{T}}$. This matrix, known as co-affiliation matrix, can be considered as a lightweight associative location ontology [20] based on overlapping sets of keywords.

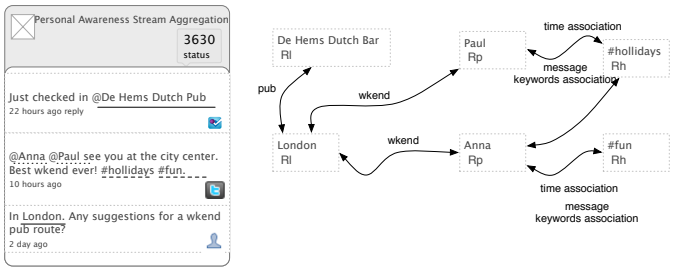


Fig. 1. A personal awareness stream on the left yields the semantic graph on the right, formed of resources of type location, people and hashtags. The edges in the graph are labelled with the resources that link the entities.

4.3 Concept Selection Induced from Social Stream Aggregations (CSISSA)

In this paper we propose the **Concept Selection Induced from Social Stream Aggregations** technique. This technique obtains a set of concepts derived from the simultaneous analysis of the correlation of different stream qualifiers. It is based on the analysis of Sp3way tensors [29] in which each slice consists of a dense matrix formed by the product of a sparse matrix and its transpose. The motivation for using this class of tensors arises from the need of simultaneously storing multiple stream qualifier matrices.

Given P lightweight ontologies characterising a user's social streams consisting of N messages; we define a tensor $\mathcal{O} \in R^{N \times N \times P}$ consisting of frontal slices of the form $\mathbf{O}_p = \mathbf{B}_p \mathbf{B}_p^T$ with $p = 1, ..P$, where $\mathbf{B}$ is a bipartite graph deriving the lightweight ontology $\mathbf{O}_p$; see Figure 2.

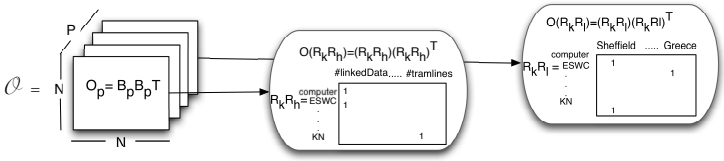


Fig. 2. Lightweight ontology tensor O.

The computation of a Tucker decomposition (presented in subsection 2) of $\mathcal{O}$ yields to an approximation of the form

$\mathrm{O} \approx \mathcal{G} \times_1 \mathbf{K} \times_2 \mathbf{K}' \times_3 \mathbf{C} =$

$\sum_{i=1}^{N} \sum_{i=1}^{N} \sum_{p=1}^{P} g_{iip} \mathbf{m}_i \circ \mathbf{m}'_i \circ \mathbf{c}_p \equiv [[\mathcal{G}; \mathbf{K}, \mathbf{K}', \mathbf{C}]]$

The output has the property $\mathbf{K} \approx \mathbf{K}'$, the rows of these matrices contain feature vectors that encapsulate a compilation of the different similarities expressed in the frontal matrices. $\mathbf{K}$ and $\mathbf{K}'$, can be regarded as keyword $\times$ keyword-group matrices highlighting those keywords that are more relevant to the similarities expressed in all $\mathbf{O_p}$. The

matrix $\mathbf{C}$ represents an index $\times$ index-group matrix which highlights $\mathbf{O_p}$ matrices. Finally the tensor $\mathcal{G}$ expresses how groups (keywords-group and index-group) relate to each other. The frontal matrix K highlights those concepts.

5 Deriving Relevant Concepts with CSISSA

The analysis with CSISSA is carried out on a user's social stream aggregation $\mathtt{S_a}(\mathtt{U}')$, this aggregation is built upon the messages the user has posted in different social networks. These messages are saved in a data store as the user generates them, and can be retrieved in windows of time of n days, this is: $\mathtt{S_a}(\mathtt{U}')\,[t_s, t_e] = (\mathtt{U}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', f_t)$, where $f_t : \mathtt{Y}' \rightarrow N, t_s \leq f_t \leq t_e$ and $|t_e - t_s| = n$ days.

The retrieved messages need to be pre-processed; 1) Stop words, punctuation and numbers from the message content are removed; 2) From the message content, entities of type: Location, Person and Organisation are extracted. Qualifiers of type: keywords, hashtags and geocodes (when provided) are also extracted. This section presents a concrete example in which CSISSA can be applied.

5.1 Recurrent Entity-Concept Analysis

Consider the problem of finding a temporal correlation among certain entities to which a user is engaged with, through the messages he has posted within a window of time; and from these entities induce a set of concepts to which they can be linked (this can be applied in temporal user profiling and event detection). The selection of the correct bipartite graphs to take part on the three-order tensor depends on the situation from which the entity-based context needs to be extracted. For example, considering the entities: Hashtag and Location; we define the following lightweight ontologies:

- **Lightweight Associative Keyword Ontology** Given a keyword $\times$ message matrix $\mathbf{R_kM} = w_{ij}$, where w_{ij} is computed following a term frequency-inverse document frequency (tf-idf) weighting function [19]. We define the lightweight associative keyword ontology $\mathbf{O}(\mathbf{R_kM})$ as $\mathbf{O}(\mathbf{R_kM}) = (\mathbf{R_kM})(\mathbf{R_kM})^\mathbf{T}$.
- **Lightweight Associative Hashtag Ontology**, we define the hashtag $\times$ message matrix $\mathbf{R_pM}$ following as well a (tf-idf) weighting function. The $\mathbf{O}(\mathbf{R_hM})$ is defined as $\mathbf{O}(\mathbf{R_hM}) = (\mathbf{R_hM})(\mathbf{R_hM})^\mathbf{T}$.
- **Lightweight Associative Location Ontology**, we define the places $\times$ message matrix $\mathbf{R_lM}$ following as well a (tf-idf) weighting function. The $\mathbf{O}(\mathbf{R_lM})$ is defined as $\mathbf{O}(\mathbf{R_lM}) = (\mathbf{R_lM})(\mathbf{R_lM})^\mathbf{T}$.
- **Ligthweight associative time ontology**, first, we obtain the hour $\times$ message affiliation matrix $\mathbf{HM} = v_{ij}$ where $v_{ij} = 1$ if the time message m_j was produced during the hour h_i and $v_{ij} = 0$ otherwise. We define the ligthweight associative time ontology $\mathbf{O}(\mathbf{HM})$ as $\mathbf{O}(\mathbf{HM}) = (\mathbf{HM})(\mathbf{HM})^\mathbf{T}$.

To analyse the correlation of these entities and derive the related concepts, it is necessary to encapsulate the previous ontologies in terms of keywords (see section 4.2); i.e to obtain $\mathbf{O}(\mathbf{R_kR_h})$, $\mathbf{O}(\mathbf{R_kR_l})$, $\mathbf{O}(\mathbf{R_kH})$. These ontologies will form the slices of

the tensor $\mathcal{O}$. The computation of a Tucker decomposition of the $\mathcal{O}$ tensor will reveal a ranked vector of concepts. By decomposing each of the tensor slices, it is possible to derive the entities relevant to the decomposition.

Table 1, presents the relevant concepts, and the highlighted entities derived from the Tucker decomposition of a tensor built from the stream aggregation of one of the users we followed in our evaluation (see section 6). This analysis reveals concepts that are recurrently relevant to the user. In this case, these results expose the correlation of the locations: Sheffield, London and Washington with the user's work related concepts during working hours.

Table 1. Concepts in the context of Hashtags-Places-Time

Emerged Concepts	linkeddata, semanticweb, talis, data.ac.uk, wrt, link, quality, astonbusinessschool, environment, funded
Hash tags	#linkeddata, #semanticweb, #talis, #astonbusinessschool, #linkquality, #ldal, #sheffield, #isko, #informationextraction, #unsupervisedclustering
Places	London, Sheffield, Washington
Time	[9:00am-5:00pm], [7:00pm-11:00pm]

6 Evaluation and Conclusions

CSISSA was evaluated on the grounds of the relevance of a concept induced by a given contextual need. A contextual need was expressed by a pair of contexts, e.g. Location-Time, Hashtag-Location. CSISSA provides a set of relevant concepts computed by the simultaneous correlation of the entities involved in a given context. For testing this technique, we "followed" a set of four "active" microbloggers. Three of them technology oriented user, and one of them an active blogger in education. The stream aggregations were recorded from 1st of July until the 25th September 2010 , and entities where extracted using Open Calais services [1].

In the absence of a gold standard, evaluating the concepts that emerge from a user's social aggregation given a context is a difficult task; it requires consulting the author of the social stream whose context-induced concepts are being mapped. For evaluating the effectiveness of CSISSA, each user was presented with a contextual need, and a set of concepts derived by CSISSA. The users were asked to mark each concept as relevant or irrelevant to the given context. Although CSISSA allows the simultaneous correlation of n-entities, which define the context; we performed the evaluation on a maximum of two entities at a time. The evaluated contexts are: hashtag-time, location-people, and organisation-people. For example, by deriving concepts related to hashtag and time for one of the users, the question was: *In terms of the association between the hashtag #linkeddata, and the timeslots ([12pm-5pm], 8pm), which of the following concepts do*

[1] Open Calais, http://www.opencalais.com/

you consider relevant?. For the hashtag-time context, three different hashtags where evaluated, and in the same way for the other two contexts.

As it is well known, acquiring the relevance judgement of all the ranked concepts in terms of precision/recall is a time-consuming and expensive process [19]. Mainly because the ranked vector can consist of hundreds of concepts that a user would not be willing to evaluate. Therefore, we have decided to use the Mean Average Precision (MAP) metric [19]. MAP measures the mean of the precision scores obtained after each relevant concept is retrieved, using zero as the precision for relevant concepts that are not retrieved. The MAP value represents the average under the precision-recall curve for a set of queries. MAP values were averaged for the three cases of each context. The results are depicted in Figure 3 a), which shows a generalized MAP performance of the relevancy of the concepts judged by each user given a context using CSISSA.

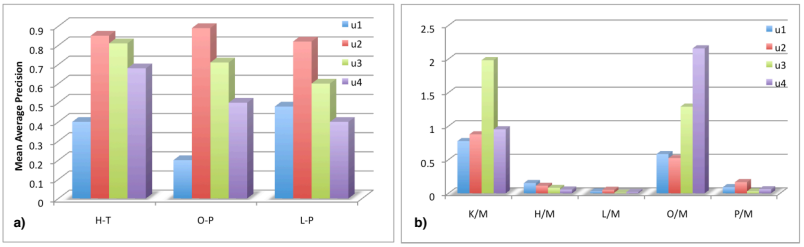


Fig. 3. a) Mean average precision (MAP) performance by user and contextual information need including HashTag-Time, Organisation-People, and Location-People, for the top 15 concepts. b) Normalised Lexical (Number of keywords(K)/ Number of Messages(M)), Topical (Hashtag), Spatial, Organisation-Entity, People-Entity Diversity.

These results suggest that higher lexical diversity (K/M) leads to better MAP results (see Figure 3 b)), this is an expected result since CSISSA explores the way in which an entity is linked to another one through keywords. We expected to discovered relevant concepts first if the user exposed a correlation between contexts, and second if this correlation was able to be expressed by keywords.

However, although the microblogging verbosity provided a better basis for deriving meaningful concepts, the relevance of the concepts given a context depended highly on the user's patterns of correlating the entities through keywords. In our experiments a fairly naive approach was taken by not considering the ambiguity in which user's can relate two entities with a keyword. Future work considers the introduction of concept disambiguation for tackling this issue.

CSISSA enabled to model users' generated patterns in their social activity streams given an entity-based context. These patterns expose the implicit association in which the user interlinks entities. The concepts derived with CSISSA suggests their applicability in user modelling, and the awareness of user intentions. A main implication of our work is that personal awareness streams can be used effectively to model context

by leveraging the user's entity affiliations. We believe that our approach can also help in merging user contexts in pervasive environments.

During the evaluation, one of the users did not remember to have tweeted about a particular topic, until we showed him the tweet, this suggest the necessity of introducing relevance-decay functions in our calculations. We also noticed that many of the users' streaming topics' relevance was in many cases volatile; further research is necessary to address these issues. We are also planning to test this technique on a bigger corpus, and to compare this technique against other baselines e.g. topic analysis.

7 Acknowledgements

This work has been supported by the European Commission as part of the WeKnowIt project (FP7-215453), and partially supported by CONACyT, grant 175203

References

1. Borst, Pim, Akkermans, Hans, and Top, Jan. Engineering ontologies. *International Journal of Human-Computer Studies*, 46(2-3):365–406, February 1997.
2. G. H. Golub and C. F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October 1996.
3. Gruber, Thomas R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.
4. R. Henrion. N-way principal component analysis theory, algorithms and applications. *Chemometrics and Intelligent Laboratory Systems*, 25(1):1–23, 1994.
5. P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, April 2006.
6. A. Hotho, R. JÃschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, chapter 31, pages 411–426–426. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2006.
7. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
8. I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.
9. V. M. I. Kiers HA. Three-way component analysis: principles and illustrative application. *Psychol Methods*, 6(1):84–110, 2001.
10. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:668–677, 1999.
11. T. Kolda and B. Bader. The TOPHITS model for higher-order web link analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security*, 2006.
12. T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
13. C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 521–530, New York, NY, USA, 2010. ACM.

14. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.
15. P. M. Kroonenberg and J. D. Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45:6997, 1980.
16. B. D. M. L. De Lathauwer and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21:12531278, 2000.
17. M. R. M. Stankovic and P. Laublet. Mapping tweets to conference talks: A goldmine for semantics. In *Proceedings of Social Data on the Web workshop, ISWC 2010. Shanghai, China*. ISWC 2010, 2010.
18. R. K. M. Strohmaier, C. Körner. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, 2010.
19. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
20. P. Mika. Ontologies are us: A united model of social networks and semantics. In *In Proceedings of ISWC 2005*, 2005.
21. M. Naaman, J. Boase, and C. H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192, New York, NY, USA, 2010. ACM.
22. O'Reilly, Tim. O'Reilly Network: What Is Web 2.0, September 2005.
23. D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.
24. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Morristown, NJ, USA, 2009. Association for Computational Linguistics.
25. S. Rendle and L. S. Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90, New York, NY, USA, 2010. ACM.
26. P. Schmitz. Inducing ontology from flickr tags. In *WWW 2006, May 22-26, 2006, Edinburgh, UK*. IW3C2, 2006.
27. L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *In Proc. of the 4th ESWC*, pages 624–639, 2007.
28. Technology, Knowledge, P. Mika, and H. Akkermans. Towards a new synthesis of ontology. Technical report, 2004.
29. W. K. T.M. Selee, T. Kolda and J. D. Griffin. Extracting clusters from large datasets with multiple similarity measures using imscand. In *CSRI Summer Proceedings*, 2007.
30. L. R. Tucker. *Implications of factor analysis of three-way matrices for measurement of change*. C. W. Harris, University of Wisconsin Press, 1963.
31. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *In Proceedings of the European Conference on Computer Vision*, volume 1, pages 447–460, 2002.
32. C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.
33. R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak. I tag, you tag: Translating tags for advanced user models. In *WSDM '10: Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 71–80, New York, NY, USA, 2010. ACM.

Does Size Matter? When Small is Good Enough

A.L. Gentile*, A.E. Cano, A.-S. Dadzie, V. Lanfranchi, and N. Ireson

Department of Computer Science,
University of Sheffield,
Sheffield, United Kingdom
{a.l.gentile,a.cano,a.dadzie,v.lanfranchi,n.ireson}@dcs.shef.ac.uk

Abstract. This paper reports the observation of the influence of the size of documents on the accuracy of a defined text processing task. Our hypothesis is that based on a specific task (in this case, topic classification), results obtained using longer texts may be approximated by short texts, of micropost size, i.e., maximum length 140 characters. Using an email dataset as the main corpus, we generate several fixed-size corpora, consisting of truncated emails, from micropost size (140 characters), and successive multiples thereof, to the full size of each email. Our methodology consists of two steps: (1) corpus-driven topic extraction and (2) document topic classification. We build the topic representation model using the main corpus, through k-means clustering, with each k-derived topic represented as a weighted number of terms. We then perform document classification according to the k topics: first over the main corpus, then over each truncated corpus, and observe the variance in classification accuracy with document size. The results obtained show that the accuracy of topic classification for micropost-size texts is a suitable approximation of classification performed on longer texts.

Keywords: Short Messages; Email Processing; Text Processing; Document classification.

1 Introduction

The advent of social media and the widespread adoption of ubiquitous mobile devices has changed the way people communicate: fast, short messages and real time exchange are becoming the norm. This phenomenon was first manifested with the introduction of SMS (Short Messaging Service) capabilities on mobile phones. Despite the technical restrictions the size limit of 160 characters imposed, SMS was quickly adopted by users, thanks to ease of use and very short delivery time. The widespread adoption has had significant impact on the language used and the way people communicate; as pointed out by Grinter and Eldridge [8], users tend to adapt media to make themselves understood. In the case of SMS this meant modifying language to condense as much information as possible into 160 characters.

* to whom correspondence should be addressed

At the same time Instant messaging (IM) services such as MSN[1], Yahoo[2] and Jabber[3] rose in popularity, offering another platform with low barrier to entry and use, for real time, text-based *chatting* and communication. Newer social media services and applications such as Twitter[4] adopted this interaction paradigm (restricting even more, messages to 140 character chunks), evolved to support real-time communication within social networks. FourSquare[5], Facebook[6] and MySpace[7] posts, while using relatively longer feeds, also follow the general trend of using small chunks of text, i.e., microposts, to carry out (asynchronous) conversations.

While a large amount of this information exchange is social, micropost services are also used to exchange information in more formal (working) environments, especially as collaboration crosses wide geographical borders, bandwidth increases and the cost of electronic services decreases [10, 11]. Twitter, for instance, is currently one of the most widely used methods for exchanging up to date information about ongoing events, and topical discussion in professional and social circles [23, 25]. However, while the usage of Twitter and similar services in the workplace is increasing, it is sometimes perceived negatively, as they may be seen to reduce productivity [22], and/or pose threats to security and privacy.

The impact of text-based SMS and IM has however been such that where restrictions to use are in place, alternatives are sought that obtain the same benefits. Individuals in such environments often adopt the same communication patterns in alternative media, e.g., both desktop-based and mobile email usage often follow the same pattern. Further, empirical evidence suggests that even where IM and social media services are available, individuals may employ email as a short message service for communication via, e.g., mailing lists. This is often done in order to reach a wider audience that includes both the initiator's personal networks and other individuals with shared interests and who may be potential sources of expertise. Because mailing lists are in essence based on communities of practice (CoPs) with shared, specialised interests [20], both detailed and quick, short requests posted to mailing lists tend to receive quick replies from colleagues and more distantly related members of a network or CoP. Such email exchanges converge to a rapidly evolving conversation composed of short chunks of text.

The aim of this paper is twofold. We first consider a corpus of emails exchanged via an internal mailing list (over a period of six months), and perform statistical analysis to determine if email is indeed used as a short messaging service. Secondly, we analyse the content of emails as microposts, to evaluate to what degree the knowledge content of truncated or abbreviated messages can be compared to the complete message. Further, we wish to determine if the knowl-

[1] `http://explore.live.com/windows-live-messenger`
[2] `http://messenger.yahoo.com`
[3] `http://www.jabber.org`
[4] `http://twitter.com`
[5] `http://foursquare.com`
[6] `http://www.facebook.com`
[7] `http://www.myspace.com`

edge content of short emails may be used to obtain useful information about e.g., topics of interest or expertise within an organisation, as a basis for carrying out tasks such as expert finding or content-based social network analysis (SNA).

We continue the paper with a review of the state of the art in section 2. We then describe, in section 3, the corpus we employ, followed by our experimental methodology (section 4) and the results of the text classification experiments used to extract and compare the knowledge content of different size emails (sections 5.1 and 5.2). We conclude the paper in section 6, and discuss briefly the next stages of our research.

2 Related Work

Expertise identification and knowledge elicitation, key components of effectiveness and competitiveness in formal organisations, are often achieved via informal networks or CoPs [5, 20]. Email is a common tool for quick exchange of information between individuals and within groups, both on a social basis, but especially also in formal organisations, both for co-located and dispersed communication [6]. Email content, and addressee and recipient, often provide clues about the existence of CoPs and the interests and expertise of participants [2]. Quantitative data from email traffic (e.g. frequency of exchange) is useful in inferring social networks, and mining email content complements this by supporting the exploration and retrieval of organisational knowledge and expertise.

Exchange Frequency In the panorama of work on extracting social networks from email, the frequency of email exchange has been widely used as the main indicator of relevance of a connection. In some cases the effort is on determining frequency thresholds [24, 7, 1, 3], while in others time-dependent threshold conditions are defined to detect dynamic networks [4, 15]. Diesner et al. [6] construct a social network via weighted edges over a classical dataset, the *Enron* corpus[8], a large set of email messages made public during the legal investigation of the Enron corporation. They reported the emergence of communication subgroups with unusually high email exchange in the period prior to the company becoming insolvent in 2001, when email was a key tool for obtaining information especially across formal, inter-organisational boundaries. Diesner et al. [6] also observed that variations in patterns of email usage were influenced by knowledge about and reputation of, in addition to, formal roles within the organisation.

Content-Based Analysis Email content analysis has been used for different purposes: determining expertise [20], analysing the relations between content and people involved in email exchange [2, 12, 17, 26], or simply extracting useful information about names, addresses, phone numbers [16]. Schwartz et al. [20] derived expertise and common interests within communities from email exchange. While acknowledging the value of the results obtained, Schwartz et al. [20] note the risk to privacy in mining emails.

[8] `http://www.cs.cmu.edu/~enron`

Campbell et al. [2] exploit addressee and recipient information, in addition to information obtained from clusters of emails created through supervised and unsupervised keyword extraction, to create networks of expertise. McCallum et al. [17] recognise the contribution of Machine Learning (ML) and Natural Language Processing (NLP) to SNA, in order to retrieve the rich knowledge content of the information exchanged in such networks, and better interpret the attributes of nodes and the types of relationships between them. By running their experiments on the Enron email dataset and that of an employee in a research institution, [17] highlight a phenomenon that is becoming increasingly common – the blurring of the lines between inter-communication on purely professional and social levels. This underlines the importance of the analysis of the content of email documents in the derivation and verification of roles (a significant attribute of nodes) and relationships within communication networks, when used for expertise determination or topic extraction, for instance.

Keila et al. [12] investigate the use of domain-specific terms and the relationships between these and roles or activity in organisations, using the Enron email dataset. They conclude that e-mail structure and content is influenced by users' overall activity, e.g., when engaged in unusual activities. They, as do [6], who reported the emergence of communication sub-groups, observed alterations in patterns in email usage in the lead up to the failure of Enron, with similarity influenced by organisational roles. Zhou et al. [26] perform textual analysis of the Enron dataset to discover useful patterns for clustering in a social network. They found that individuals communicate more frequently with others who share similar value patterns than with those exhibiting different ones. They however could not draw definite conclusions about whether or not individuals who communicate more frequently with each other share similar value patterns.

Laclavík et al. [16] observe that enterprise users largely exploit emails to communicate, collaborate and carry out business tasks. They design a pattern-based approach to information extraction (IE) from and analysis of enterprise email communication, and exploit the data obtained to create social networks. The test sets (one in English containing 28 emails, and a second in Spanish with 50) consist of mainly formal emails exchanged between different enterprises. Their experimental design follows the classic IE approach: they automatically extract information such as names, telephone numbers and addresses from the email corpus, and compare results against a gold standard, the same email corpus, manually annotated. The results obtained indicate that emails are a valid means for obtaining information across formal, inter-organisational boundaries.

The work we present in this paper, on the other hand, makes use of a test set containing more informal email exchange in an internal mailing list for an academic research group, for a pilot, exploratory set of experiments. Rather than carrying out a classic IE evaluation task, we wish to determine if relatively short and informal texts can be used to aid the understanding of the content of the conversations carried out via email, and depict the variety of topics discussed using this communication medium.

Corpora The Enron corpus is a preferred test set in this field. The original corpus contains 619,446 messages belonging to 158 users, but [14], among others, suggest that cleaning is needed, for a number of reasons, including the fact that some of the folders are computer-generated (such as "discussion threads" and "notes inbox"), others contain duplicate email messages (such as the folder "all document"), and yet others constitute delivery failures and repeated attempts to deliver the same message.

Depending on the task being performed, accurate cleaning is required to avoid misleading results; [6, 12, 17] all perform cleaning and merging of data to increase the accuracy and reliability of the results of analysis. While the Enron corpus is valued as a widely available test set that aids replication of experiments in the field, we do not use it at this stage in our research. The main reason for this is that our experiments currently examine the usage of email as a tool for sharing information within a fixed community, as an alternative to social publishing services, and explore phenomena observed in such environments. The internal mailing list we use as a starting test set meets this requirement. A statistical analysis of our corpus is provided in section 3.

3 Email Corpus

The corpus used for analysis and knowledge content extraction is an internal mailing list of the OAK Group[9] in the Computer Science Department of the University of Sheffield. The mailing list is used for quick exchange of information within the group on both professional and social topics.

We use all emails sent to the mailing list in the six month period from July 2010 to January 2011, totalling 659 emails. For each we extracted the email body: the average length of which is 351 characters (just shorter than 2.5 microposts), with a standard deviation of 577 characters. We refer to this corpus as *mainCorpus*. Detailed statistics on document length are shown in Fig. 1. The percentage of messages of micropost size (up to 140 characters) constitutes more than 35% of the whole corpus. Considering emails up to two micropost sizes increases the percentage to ∼65%. Very few emails (around 4%) are really long (above 1000 characters).

These statistics indicate that the corpus largely consists of *micro-emails* – which we define as short email messages exchanged in rapid succession about a topic. We carried out a number of experiments on this corpus, to understand the knowledge content of the (micro-)emails. Future work will consider how this corpus varies from other email corpora of the same type (mailing lists) and what generic assumptions could be made about the existence and use of micro-emails.

4 Dynamic Topic Classification Of Short Texts

One of our main goals is to evaluate to what degree the knowledge content of a shorter message can be compared to that of a full message. Our hypothesis

[9] `http://oak.dcs.shef.ac.uk`

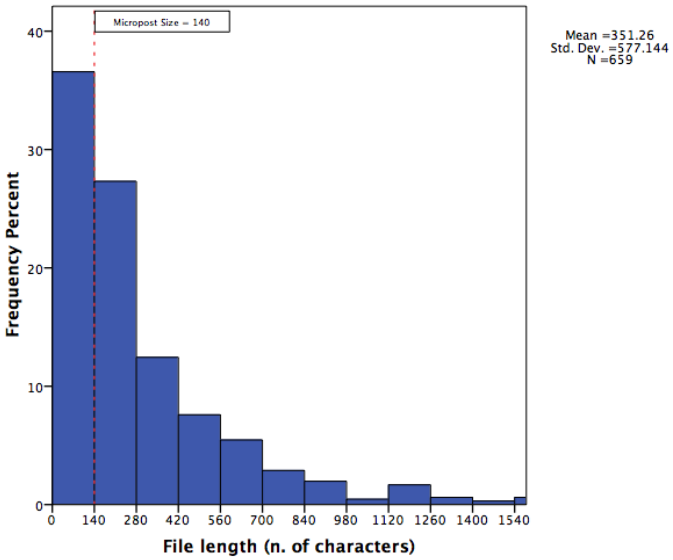


Fig. 1. Email length distribution

is that based on a specific task, results obtained using short texts of micropost size approximate results obtainable with longer texts. The task chosen for the evaluation is text classification on non-predefined topics. The test bed is generated by preprocessing the email corpus (see section 3) to obtain several fixed-size corpora, as detailed in section 5.1. The overall method consists of two steps:

corpus-driven topic extraction: a number of topics are automatically extracted from a document collection; each topic is represented as a weighted vector of terms;

document topic classification: each document is labelled with the topic it is most similar to, and classified into the corresponding cluster.

4.1 Topic Extraction: Proximity-based Clustering

Given a document corpus D, we represent it using a vector space model; each document in the corpus $d = \{t_1, ..., t_v\}$ is represented as a vector of weighted terms (using tf-idf weights). Using an inverted index of the documents we generate clusters of terms. Each cluster in $C = \{C_1, ..., C_k\}$ is represented as a weighted vector of terms $C_k = \{t_1, ..., t_n\}$, selecting n terms t for each cluster with highest tf-idf weight. Each cluster ideally represents a topic in the document collection. To obtain the clusters we apply a K-Means algorithm [9], using as feature space the generated inverted index of document terms (i.e., for each term we define which document contains it). Starting with k random means (centroids), each vector is assigned to the nearest centroid. By minimising the euclidean distance between each point and the centroids the process is repeated until convergence is reached.

4.2 Email Topic Classification

We then use cosine similarity to determine similarity between documents and clusters; we will explore further, in the next stage of our research, alternative similarity functions and their impact on the results obtained. For each document we calculate the similarity $sim(d, C_i)$ with each cluster. The labelling process $labelDoc : D \rightarrow C$ consists of mapping each document d to the topic C_i, which maximises the similarity $sim(d, C_i)$. The complete procedure is shown in Fig. 2.

labelDoc procedure

Input: Collection of documents $\{d_1, ..., d_{|D|}\}$, set of clusters $C = \{C_1, ..., C_k\}$, term representation for each cluster $C_k = \{t_1, ..., t_n\}$

Step 0: Obtain a document d_i's feature vector, of tf-idf weighted terms.

Step 1: Apply cosine similarity between a document d_i's feature vector and the k clusters and generates a vector of similarities $S_i = \{S_{i0}, ..S_{ik}\}$, $S_{ij} = sim(d_i, C_j)$.

Step 2: Label d_i with the highest weighted cluster in S_i.

Output: All classified documents.

Fig. 2. *labelDoc*: Topic classification procedure

5 Experiments

5.1 Dataset Preparation

In this experiment we artificially generate comparable corpora starting from the mailing list described in Section 3. The notions of *comparable corpora* and the strongly related alternative, *parallel corpora*, are very common in multi-language IE. Parallel text corpora contain the same documents with different content representation. An example is parallel language resources [19, 27], where a corpus consists of a set of documents, each of which is an exact translation of the original document in a different language. Comparable corpora [21, 13], however, do not contain document-level or sentence-level alignment across corpora, but talk about the same important facts. An example of comparable corpora for different languages is the multi-lingual Wikipedia [18], where the same articles are available in different languages, but which are not necessarily literal translations of each other, e.g., an article may be richer in one language than in another.

We produce comparable corpora using the following process: starting form *mainCorpus* we generate different corpora, each containing documents of fixed maximum length, by chunking the email body in multiples of 140 characters. We generated 8 comparable corpora, as shown in Table 1.

Table 1. Automatically generated comparable corpora.

Corpus Name	Maximum text length of each document
*corpus*140	email body truncated at length 140 if longer than 140 characters, full text otherwise
*corpus*280	email body truncated at length 280 if longer than 280 characters, full text otherwise
*corpus*420	email body truncated at length 420 if longer than 420 characters, full text otherwise
*corpus*560	email body truncated at length 560 if longer than 560 characters, full text otherwise
*corpus*700	email body truncated at length 700 if longer than 700 characters, full text otherwise
*corpus*840	email body truncated at length 840 if longer than 840 characters, full text otherwise
*corpus*980	email body truncated at length 980 if longer than 980 characters, full text otherwise
mainCorpus	full email body

5.2 Experimental Approach

As described in section 4, the set of topics for categorising the initial documents is not predefined, but corpus-driven. We use the *mainCorpus* for topic extraction. Since the k in the k-means clustering approach must be approximated, we repeated the clustering process several times. We applied the procedure *labelDoc* over the *mainCorpus*, varying each time the input clusters, from 3 to 15. The cardinality of clusters providing the widest distribution of classified documents on *mainCorpus* was 10; we therefore selected this as the optimal number of clusters for the final experiment on document classification. The main keywords in each cluster are shown in Fig. 3.

Using the 10 clusters obtained from the main corpus, we apply the *labelDoc* procedure to the different comparable corpora, including *mainCorpus*. Results obtained for the classification of *mainCorpus* are considered as the gold standard, and used for comparing results of all the other corpora.

5.3 Results and Discussion

We evaluate the performance of the topic classification using standard Precision (P), Recall (R) and F-Measure (F). Given the number of classes for classification (10) we calculate P, R, and F by micro-averaging results on the classification confusion matrix. Results for all text size corpora are shown in Table 2.

As expected, it is recall rather than precision with a bigger decrease as text length is reduced. If we relax the limitation of 140 characters and consider the next size corpus (280) the drop in performance is much lower.

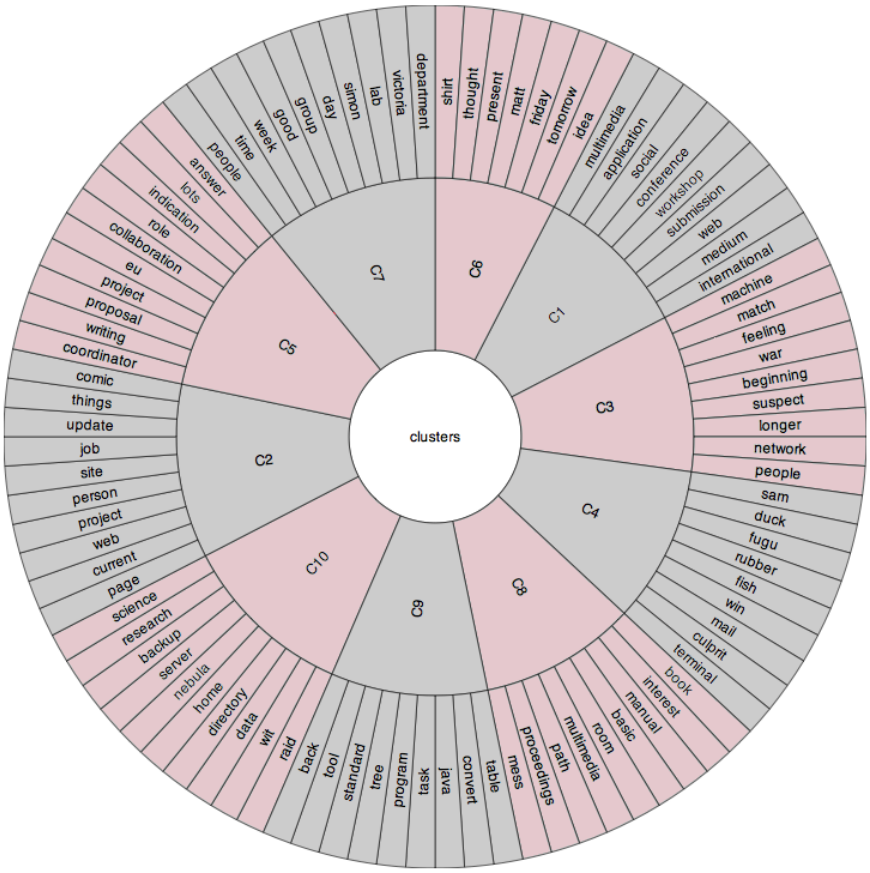


Fig. 3. Visualisation of topic clusters.

Table 2. Precision, Recall and F-measure values for topic classification for each corpus

	Precision	Recall	F-Measure
*corpus*140	0.86	0.66	0.74
*corpus*280	0.93	0.88	0.90
*corpus*420	0.95	0.94	0.95
*corpus*560	0.98	0.97	0.97
*corpus*700	0.99	0.98	0.99
*corpus*840	0.99	0.99	0.99
*corpus*980	0.99	0.99	0.99

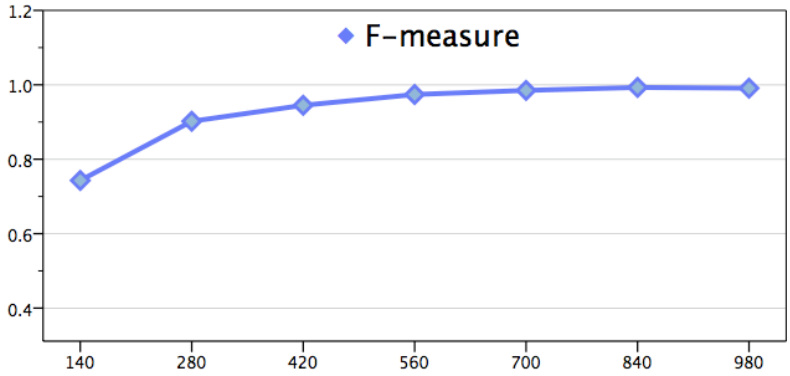


Fig. 4. F-Measure trend on different text size corpora

Considering the results obtained for *mainCorpus* as the upper boundary for classification, the trend of F-Measure over the different size corpora in Fig. 4 shows the impact of using shorter texts for topic classification. As expected, the trend increases monotonically with the size of texts, which means that reducing the text size directly affects the classification performance. What is interesting is that the performance is not significantly affected by reduction in text size.

At one micropost size there is a drop in F-Measure, to 74%. However with an increase to only two micropost sizes this improves significantly, to 90%. The larger drop at one micropost size may be explained by the method of truncation we use; among others, where a greeting exists this takes up a fair portion of the first micropost block. We are currently exploring the use of a sliding window to determine how best to chunk the e-mail content and identify the most salient region(s) of each, as a way of improving recall.

6 Conclusions

We have presented in this paper exploratory work on the usage of email as a substitute for online social publishing services. We explore how this kind of data may be exploited for the knowledge discovery process and how document size influences the accuracy of a defined text processing task. Our results show:

1. that a fair portion of the emails exchanged, for the corpus generated from a mailing list, are very short, with more than 35% falling within the single micropost size, and ∼65% up to two microposts;
2. for the text classification task described, that the accuracy of classification for micropost size texts is an acceptable approximation of classification performed on longer texts, with a decrease of only ∼ 5% for up to the second micropost block within a long e-mail.

These results are indicative of the convenience in communication using microposts in different environments and for different purposes. Because the research at this stage is still exploratory we refrain from generalising to other datasets.

However, our test corpus, which contains emails talking about both formal work and social activities, is not atypical in the workplace (see, for instance, [17]). We therefore believe that this work does provide a starting point from which to carry out more extensive analysis, using other standard email corpora such as the Enron corpus, in addition to other enterprise mailing lists similar to the corpus we analyse in this paper. This will allow us to explore what generic assumptions could be made on the creation and use of micro-emails.

A second hypothesis we wish to examine is whether enriching the micro-emails with semantic information (e.g., concepts extracted from domain and standard ontologies) would improve the results obtained using unannotated text. We also plan to investigate the influence of other similarity measures.

One area we wish to explore more fully is the application to expert finding tasks, exploiting dynamic topic extraction as a means to determine authors' and recipients' areas of expertise. For this purpose a formal evaluation of topic validity will be required, including the human (expert) annotator in the loop.

Acknowledgements A.L Gentile and V. Lanfranchi are funded by the Siloet project. A.E. Cano is funded by CONACyT, grant 175203. A.-S. Dadzie is funded by SmartProducts (EC FP7-231204). A.L. Gentile, A.-S. Dadzie, V. Lanfranchi and N. Ireson are also funded by WeKnowIt (EC FP7-215453).

References

1. L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187 – 203, 2005.
2. C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: 12th international conference on Information and knowledge management*, pages 528–531, 2003.
3. M. D. Choudhury, W. A. Mason, J. M. Hofman, and D. J. Watts. Inferring relevant social networks from interpersonal communication. In M. Rappa et al., editors, *Proc., 19th International Conference on World Wide Web*, pages 301–310, 2010.
4. C. Cortes, D. Pregibon, and C. Volinsky. Computational methods for dynamic graphs. *Journal Of Computational And Graphical Statistics*, 12:950–970, 2003.
5. A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *CEAS 2004: Proc., 1st Conference on Email and Anti-Spam*, 2004.
6. J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the Enron email corpus "It's Always About the People. Enron is no Different". *Compututational & Mathematical Organanization Theory*, 11(3):201–228, 2005.
7. J. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.
8. R. E. Grinter and M. A. Eldridge. y do tngrs luv 2 txt msg? In *Proc., 7th European Conference on Computer Supported Cooperative Work*, pages 219–238, 2001.
9. J. Hartigan. *Clustering Algorithms*. John Wiley and Sons, New York, 1975.
10. J. D. Herbsleb, D. L. Atkins, D. G. Boyer, M. Handel, and T. A. Finholt. Introducing instant messaging and chat in the workplace. In *Proc., SIGCHI conference on Human factors in computing systems*, pages 171–178, 2002.

11. E. Isaacs, A. Walendowski, S. Whittaker, D. J. Schiano, and C. Kamm. The character, functions, and styles of instant messaging in the workplace. In *Proc., ACM conference on Computer supported cooperative work*, pages 11–20, 2002.
12. P. S. Keila and D. B. Skillicorn. Structure in the Enron email dataset. *Computational & Mathematical Organization Theory*, 11:183–199, 2005.
13. A. Klementiev and D. Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *Proc., main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 82–88, 2006.
14. B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In J.-F. Boulicaut et al., editors, *ECML 2004: Proc., 15th European Conference on Machine Learning*, pages 217–226, 2004.
15. G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
16. M. Laclavik, S. Dlugolinsky, M. Seleng, M. Kvassay, E. Gatial, Z. Balogh, and L. Hluchy. Email analysis and information extraction for enterprise benefit. *Computing and Informatics, Special Issue on Business Collaboration Support for micro, small, and medium-sized Enterprises*, 30(1):57–87, 2011.
17. A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.
18. A. E. Richman, P. Schone, and F. G. G. Meade. Mining wiki resources for multilingual named entity recognition. *Computational Linguistics*, pages 1–9, 2008.
19. E. Riloff, C. Schafer, and D. Yarowsky. Inducing information extraction systems for new languages via cross-language projection. In *COLING '02: Proc., 19th international conference on Computational linguistics*, pages 1–7, 2002.
20. M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.
21. R. Sproat, T. Tao, and C. Zhai. Named entity transliteration with comparable corpora. In *Proc., 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 73–80, 2006.
22. TNS US Group. Social media exploding: More than 40% use online social networks. `http://www.tns-us.com/news/social_media_exploding_more_than.php`, 2009.
23. T. Turner, P. Qvarfordt, J. T. Biehl, G. Golovchinsky, and M. Back. Exploring the workplace communication ecology. In *CHI '10: Proc., 28th international conference on Human factors in computing systems*, pages 841–850, 2010.
24. J. Tyler, D. Wilkinson, and B. Huberman. E-Mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005.
25. J. Zhang, Y. Qu, J. Cody, and Y. Wu. A case study of micro-blogging in the enterprise: use, value, and related issues. In *CHI '10: Proc., 28th international conference on Human factors in computing systems*, pages 123–132, 2010.
26. Y. Zhou, K. R. Fleischmann, and W. A. Wallace. Automatic text analysis of values in the Enron email dataset: Clustering a social network using the value patterns of actors. In *HICSS 2010: Proc., 43rd Annual Hawaii International Conference on System Sciences*, pages 1–10, 2010.
27. I. Zitouni and R. Florian. Cross-language information propagation for arabic mention detection. *ACM Transactions on Asian Language Information Processing*, 8:17:1–17:21, 2009.

Discovering the Dynamics of Terms' Semantic Relatedness through Twitter

Nikola Milikic[1], Jelena Jovanovic[1], Milan Stankovic[2]

[1]University of Belgrade, Jove Ilica 154, 11000 Belgrade, Serbia
[2]STIH, Université Paris-Sorbonne, 28 rue Serpente, 75006 Paris, France

nikola.milikic@gmail.com, jeljov@gmail.com, milstan@gmail.com

Abstract. Determining the semantic relatedness (SR) of two terms has been an appealing topic in information retrieval for many years as such information is useful for various tasks ranging from tag recommendation, over search query refinement to suggesting new web resources for the user to discover. Most approaches consider the SR of terms as static over time, and disregard the eventual temporal changes as imperfections. However, detecting and tracing changes in SR of terms over time may help in understanding the nature of changes in public opinion, as well as the change in the usage of terms in common language and jargon. In this paper, we propose an approach that makes use of microposts data in order to establish a dynamic measure of SR of terms, i.e., a measure that accounts for the changes in SR over time. We propose different scenarios of use (in online advertising and organizational knowledge management) which demonstrate the applicability of our approach in real life situations. We also provide a demo application for visualizing the change in micropost-based SR of terms.

Keywords: Semantic relatedness, dynamic measure of semantic relatedness, microposts, Twitter

1 Introduction

Many research papers (such as in Wagner [1]) claim that Twitter and similar micro-blogging services have become a valuable source of knowledge, and have tried to extract this knowledge and use it for various purposes, such as the creation of dynamic domain models suitable for semantic analysis and annotation of real-time data [2], modeling of users' interests and finding experts [3], etc. However, from our point of view, the real-time nature of Twitter and Twitter-like services has not really been explored to its full extent, yet.

Most approaches exploit the mass of data that users generate on real-time services as their most valuable feature. We believe that there is a significant value in the fact that tweets (and microposts in general), posted frequently and massively, represent the moment in which they are created and the characteristics of that moment. Therefore, we have been exploring how these real-time services can support the detection of changes

in semantics of terms, by enabling one to observe the changes of a term's use over time. We focus particularly on the semantic relatedness (SR) of terms which is also subject to temporal changes.

The scope of meaning of a certain term is always defined in a social circle in which that meaning emerges, is agreed upon and accepted. Knowing that social systems are dynamic, it is difficult to neglect the natural changes (i.e., evolution) in socially agreed upon meaning of terms. If the meaning of a term is changing over time, so is the relatedness of that term to other terms whose dynamics in the given time period might be different. The most basic illustration of this is the term *totalitarian regime*. It is reasonably close to the terms identifying particular totalitarian governments and dictators of particular countries. However, this proximity should decrease if the totalitarian regime in a country is replaced by a democratic government – which happens more and more often in recent times.

Although tendencies in the public expressions can easily be detected through search query frequencies and trending topics on Twitter, the nature of tendencies and their mutual relationships are not directly evident from such observations. We could imagine having three trending topics on Twitter: *Egypt*, *revolution*, and *Britney Spears*. Although a human may grasp that it is more likely that the revolution is happening in Egypt and not that Britney Spears is leading a revolution, for a computer, this is far less obvious. The change of SR, however, could indicate the rationale for the raising public interest in a particular term. For instance, we could see that the recent popularity of the term *Egypt* might have been related to the temporary increase of SR of terms *Egypt* and *revolution*; and that it had nothing to do with the raise of popularity of the term *Britney Spears*. Spotting the change in SR of terms could thus help to give meaning to the observed trends in Web content, and enable machines to grasp this meaning and take advantage of it in many real life scenarios.

In this paper we present our initial research on using real-time services, in general and Twitter in particular, to detect the changes in SR of terms. We also explore the scenarios where reacting to those changes might be beneficial. In Section 2, we present the state of the art in research on SR of terms as well as in using Twitter to detect tendencies and make use of them. Section 3 introduces our measure of SR based on micropost data – Normalized Micropost Distance, whereas Section 4 gives some suggestions on how the relevancy of the change in SR of terms could be detected. We present our application for testing the proposed approach in Section 5, and consider the potential usage scenarios in Section 6. In Section 7, we give some interesting examples of changes in terms' SR which we have observed by using our application. Section 8 concludes the paper with propositions of future work that will help give maturity to our initial research.

2 State of The Art

The problem of determining semantic relatedness of terms has been studied for decades, in various contexts and using different approaches. Semantically related terms have been used to help users choose the right tags in collaborative filtering systems [4]; to

discover alternative search queries [5]; for query refinement [6]; to enhance expert finding results [7]; for ontology maintenance [8] [9], and in many other scenarios.

Different techniques and different sources have been used and combined to develop measures of semantic relatedness (MSRs). These measures could be split into three major categories: 1) net-based measures, 2) distributional measures and 3) Wikipedia-based measures [10]. In what follows we briefly examine each category of MSRs.

Net-based measures make use of semantic (e.g., hyponymy or meronymy) and/or lexical (e.g., synonyms) relationships within a network (graph) of concepts to determine semantic proximity between the concepts. For example, Burton-Jones et al. [11] exploit the hypernym graphs of WordNet[1]; Safar et al. [6] use Gallois lattice to provide recommendations based on domain ontologies, whereas Ziegler et al. [12] and Resnik [13] use the ODP taxonomy[2]. This category also includes measures that rely on the graph structure of concepts to determine semantic relatedness of those concepts. Shortest path is among the most common of such measures. It is often enhanced by taking into account the informational content of the nodes in the graph [14].

Distributional measures rely on the distributional properties of words in large text corpora. Such MSRs deduce semantic relatedness by leveraging co-occurrences of concepts. For example, the approach presented in Salton et al. [15] uses co-occurrence in text of research papers, pondered with a function derived from the tf-idf measure to establish a notion of word proximity. Co-occurrence in tags [4] and in search results [17] is also commonly used. In Strube et al. [18], the authors introduced Normalized Web Distance (NWD) as a generalization of Normalized Google Distance (NGD) MSR and investigated its performance with six different search engines. The evaluation (based on the correlation with human judgment) demonstrated the best performance of Exalead-based NWD measure, closely followed by Yahoo!, AltaVista, Ask and Google; only Live Search and Clusty showed significantly lower results.

As its name suggests, the third category of MSRs – Wikipedia-based measures – makes use of Wikipedia as the resource for computing semantic relatedness and often combines the features of the previous two MSR groups. For example, [18] relies on the graph of Wikipedia categories, whereas Waltinger et al. [10] rely on co-occurrence of words in the text of Wikipedia pages, combined with the information about the categories of pages in Wikipedia to compute semantic relatedness.

In Waltinger et al. [10], the authors report on a comparative analysis of a large number of MSRs (at least 4 algorithms from each major category of MSRs were included in the study, resulting in sixteen algorithms in total). The most important results could be summarized as follows: 1) small, hand-crafted and structured resources (e.g., WordNet) are inferior to large and semi-structured (i.e., Wikipedia) or even unstructured resources (i.e., plain text); 2) the distributional MSRs (especially measures like Latent Semantic Analysis) perform significantly better than the net-based measures and those using explicit categorical information; 3) MSRs that use the Web as a corpus were inferior to those operating on smaller but better controlled training corpora (e.g., Normalized Distance based on Wikipedia significantly outperformed NGD).

Most of the existing approaches do not take into account the dynamic nature of semantic relatedness between terms. An exception would be the work presented in

[1] http://wordnet.princeton.edu/
[2] http://www.dmoz.org/

Nagarajan et al. [22] where authors take the approach of identifying 'strong descriptors' of an event by querying Google Insights to get the terms the event's name was queried with the most (referred as 'seed keywords'). Afterwards, they query Twitter to get the tweets containing seed keywords and extract the strong descriptors from them. However, this approach does not measure SR between two specific terms, but rather identify terms relevant to the name of an event being examined. Other approaches even take the stability of their measure over time, to demonstrate the solidity of their approach [17].

On the other hand many approaches exist for extracting meaning from Twitter [20][1]. Some of them make extensive use of Twitter dynamics, like the approach for detecting events through peaks of word popularity [20]. Most related to our work is the approach presented in Song et al. [21] which relies on spatio-temporal characteristics of topics mined from Twitter data, for the calculation of semantic relatedness among topics. The temporal aspect of a topic is determined by the frequency of its occurrence in Twitter data streams over a given time period, whereas the spatial aspect refers to the regional distribution of messages mention the given topic over the same time period. Although this approach looks promising, its usefulness for measuring SR of topics has not been fully proved yet.

3 Normalized Micropost Distance

Inspired by the work of Cilibrasi et al. [16] on establishing Normalized Google Distance (NGD) as a MSR of terms based on Google search result, we propose a similar measure – Normalized Micropost Distance (NMD) – based on the results of searching the content (i.e., microposts) of real-time (Twitter-like) services. By leveraging micropost streams of real-time services, this measure should reflect the change in terms' SR more quickly than the standard web search results that are not updated in real-time. The basic assumption behind our approach is that Google's Search API results tend to be stable and based on content with a lower frequency of change, and as such would not be as good in indicating the changes in the SR of terms as could be search results that are based on real-time content..

NGD uses the frequencies of appearance of two terms in the Google index, as well as the frequency of their mutual appearance to quantify the extent to which the two terms are related. The basic assumption behind this measure is that terms that co-occur more frequently would be more related. Similarly, the proposed NMD measure can be calculated using the formula (1).

$$NMD(x,y)_t = \frac{\max\{\log f(x)_t, \log f(y)_t\} - \log f(x,y)_t}{\log M - \min\{\log f(x)_t, \log f(y)_t\}} \quad (1)$$

The formula allows one to calculate the NMD of two terms x and y for the time interval t. $f(x)_t$ and $f(y)_t$ represent the number of results returned for the term x and y, respectively, within the time interval t, when searching the content (i.e., microposts) of a real-time, Twitter-like service. The terms in the formula (x and y) may also be compound terms. Calculating the value of this formula for the same terms over different

time intervals is essential for determining the dynamics of their relationship, as we further explain in the following two sections.

4 Detecting the Significance of Change

The notion of NMD defined above is useful for measuring the difference in SR of two terms, but will not, by itself help to detect changes worthy of notice, and distinguish them from small and frequent variations. We suggest two complementary ways to perform this detection.

First, calculating the standard deviation of NMDs over a longer period of time would give a good ground to judging the significance of the identified changes. Standard deviation of NMDs can be calculated using the formula (2). The given formula represents the standard deviation of NMDs over a sample of N observations in which NMDs were calculated in time intervals *i* that are of the same length.

$$\sigma(NMD(x,y)) = \sqrt{\frac{\sum_{i=1}^{N}(NMD(x,y)_i - avg(NMD(x,y)))^2}{N}} \quad (2)$$

Detection of a change in terms' SR (measured using NMD) that is greater than the standard deviation σ could be an indicator of a significant change.

In addition to this indicator, one could observe the stability of change over several consecutive time instances to make sure that the change is not of a too short breath. However such a criterion may not be generally applicable and is specific to each use case, as even short changes might matter in some use cases, while in others only a change that spans several days would be significant.

5 Demo Application

In order to test the proposed approach of using micropost streams to calculate SR of terms, we have developed a simple web application that makes use of Twitter Search API[3] for computing NMD. The application, entitled Tweet Dynamics, currently in private beta, demonstrates how the NMD measure can be utilized, visualized and interpreted. Application is built in Java programming language using Tapestry Web Framework[4]. Javascript plotting library for jQuery named Flot[5] is used for plotting the result diagram.

The application's home page presents a user with a simple interface (Figure 1) which allows her to input the number of days and two keywords that NMDs should be calculated for. By clicking on the button 'Calculate', NMD calculation process is

[3] http://search.twitter.com/api/
[4] http://tapestry.apache.org/
[5] http://code.google.com/p/flot/

invoked. Application then queries the Twitter API to get all posts containing the first keyword, then posts containing the second keyword, and at the end to get all the posts containing both keyword. This process is repeated for the given number of days. With that data, NMDs are being calculated according to the formula (1).

The result of calculation is shown in a diagram (Figure 2) where each day is presented as a dot on the diagram line. One can easily perceive a trend of SR between two keywords during the past days.

Although, for the purpose of calculating standard deviation, our application keeps the computed values of NMD, the value of standard deviation is not shown on Figure 2 since we do not yet have a significant sample of values (e.g., dating from at least a month ago) and thus taking into account the currently available value of standard deviation would not be methodologically sound. Once a significant sample is present, the user would see a second line representing the standard deviation, so he/she could spot when the change in NMD becomes significant.

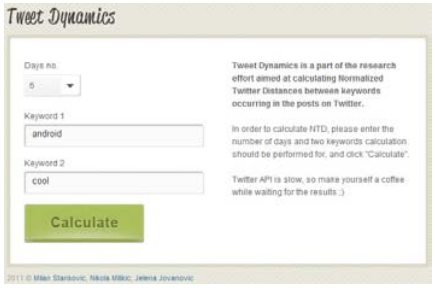


Figure 1 - Tweet Dynamics home page. User can enter two terms and a number of days to observe.

Figure 2 – The diagram illustrates the dynamics of SR of the terms *android* and *cool*, measured using NMD

Although some Web actors have access to the total history of tweets, most of interested parties have quite limited access to the Twitter Search API, which allows up to 1500 results per query. For terms of high frequency this can be a limiting factor since it makes it impossible to estimate their full frequency, and compare it with other high-frequency terms. A workaround that we use is to sample the tweets within short time intervals in which the number of tweets per terms is lower than the imposed limit. This however involves the risk of hitting the limit of 150 requests to the Twitter API per hour.

Another limitation of using Twitter's Search API is the restriction on the temporal range of tweets that can be returned as a search result. In particular, according to the API's documentation,[6] a post returned as a search result must not be 'too old', which in practice brings down to a number of six days, meaning that the oldest post returned as a result of a query is six days old. This restriction highly limits the ability to test our application on the microposts generated during a longer time span and detect trends in SR between keywords related to certain events or periods of year. If we had access to data spanning a longer period of time, we would have been able to test our results by

[6] http://apiwiki.twitter.com/w/page/22554756/Twitter-Search-API-Method:-search

comparing them with various indicators such as survey results, sales changes of a product etc.

6 Scenarios of Use

In this section we present two usage scenarios aiming to illustrate the potential benefits of the suggested dynamic MSR in real life settings. The first scenario assumes the usage of Twitter content stream for the calculation of NMD, whereas the second one relies on the micropost exchanged in a (internal) micro-blogging tool of an organization.

Scenario 1: Adapting Online Advertising Campaigns to the Changes in Term Relatedness

Optimization of the keyword choice for online advertising campaigns has become a vivid market with more and more players in the field. Using the information about keywords similarity and relatedness, combined with prices of keywords in advertising services, such as AdWords, it is possible to find a combination of keywords that costs less, but drives the same or bigger amount of relevant traffic. Such services, however, do not take advantage of keywords that become occasionally relevant. For instance, let us consider the situation happened at this year's SXSW[7] conference held at Austin, Texas, USA. Many new iPad applications were showcasted at the conference and a rumor appeared, and lately became truth, that iPad 2 would start selling on the second day of the conference. This trend would be noticed if NMD was measured for the words 'ipad' and 'sxsw' A company selling iPad accessories, would in such an occasion have a clear interest to alter the keywords for their AdWords campaign for promoting its products and add the word 'sxsw', thus getting new relevant traffic. Once the NMD for the two words goes up again, the advertising campaign can again be changed to avoid driving the traffic that became less relevant.

Responding to changes in terms' relatedness over time, for advertising campaigns means not missing out relevant traffic, and as such is of high importance for this market. Web marketing tools such as KeywordDiscovery.com do offer the possibility to discover relevant keywords and include them in marketing campaigns, but do not reflect the change in this relevancy. Changes in relevancy might open completely new possibilities for advertising campaign optimization, and using our notion of NMD, these changes may even be taken into account in an automated or semi-automated way.

Scenario 2: Facilitating Discovery of Relevant Resources in Organizations

Many organizations, especially larger ones, maintain organizational vocabularies and use them for the annotation of different kinds of documents and other digital assets. Such a vocabulary often results from a collaborative work of domain experts and a knowledge engineer. Therefore, it tends to reflect the experts' view of the subject

[7] http://sxsw.com/

domain, and the terms it defines reflect the jargon used by these experts. However, this jargon does not necessarily overlap with the everyday language used by the employees within the organization. As a consequence, employees would experience difficulties in formulating their requests for different kinds of organizational resources using the organization's official vocabulary. This indicates the need for harmonizing the official and the actual vocabularies within an organization. Furthermore, each organization evolves and many organizations need to go through continuous changes in order to respond to the constantly changing conditions in their environment. To properly address the evolving work practices in the organization, the organization's vocabulary has to evolve as well, and it should evolve to be comprehensible and usable by the employees (i.e., it should incorporate the terminology used by the employees). This is where the suggested dynamic MSR applied over the messages exchanged in the organization's Twitter-like communication channels (e.g., Yammer[8]) can help. In particular, the proposed MSR can be used for extracting terms related to certain tasks, projects, organizational positions, etc., in order to use them for evolving the organization's vocabulary. This would increase the usability of the vocabulary and consequently improve the search and discovery of organizational resources.

The suggested dynamic MSR can also be applied for facilitating people search within an organization by enabling the deduction of terms that best describe each employee. Previous studies exploring the practice of people tagging in organizations [23][24] have confirmed that people do perceive such a practice beneficial as it allows for, e.g., finding out who is working on a certain project/task, or identifying experts in a particular topic. However, the main obstacle for applying this practice in workplace lies in the very act of directly tagging (labeling) a person; many participants in the cited studies were reluctant to directly tag their colleagues as they were worried about potentially inadvertent effects those tags might cause. With the proposed dynamic MSR applied to the messages exchanged within the organization's micro-blogging and/or social streaming application, an organization would be able to identify the terms (tags) related to each employee. These terms would still reflect the community's perception of any particular employee, while freeing people from the unnecessary cognitive burden of inadvertently affecting their colleagues.

7 Example Diagrams

In order to test the use of the formula (1) on the data gathered from Twitter in several consecutive days for detection of the change in SR between two terms, we chose several examples of term pairs whose popularity we, as humans, were able to perceive from the news. The testing was done using our Tweet Dynamics application (cf. Section 5).

Since, unfortunately, catastrophic events were happening in Japan at the time of writing this paper[9], we used keywords 'japan' and 'nuclear' and calculated their NMDs for 5 days starting from March 8, 2011.

[8] https://www.yammer.com/

[9] On March 11, 2011, a strong earthquake struck Japan which triggered a failure of the cooling system of the reactor at Japan's Fukushima nuclear power plant, causing a huge explosion at the power plant the day after, on March 12.

Figure 3 - NMD diagram for terms 'japan' and 'nuclear' for the 5 days period

By looking at the diagram (Figure 3), one can observe that by March 11, there was a small relatedness between the terms 'japan' and 'nuclear' because the earthquake happened suddenly; thus the value of NMD (shown on Y axis) is higher. On the day of the earthquake (March 11th), one can see that the NMD significantly decreased, i.e., SR of the terms increased, as many people tweeted about the danger of explosion at the nuclear power plant. That trend continued in the following days.

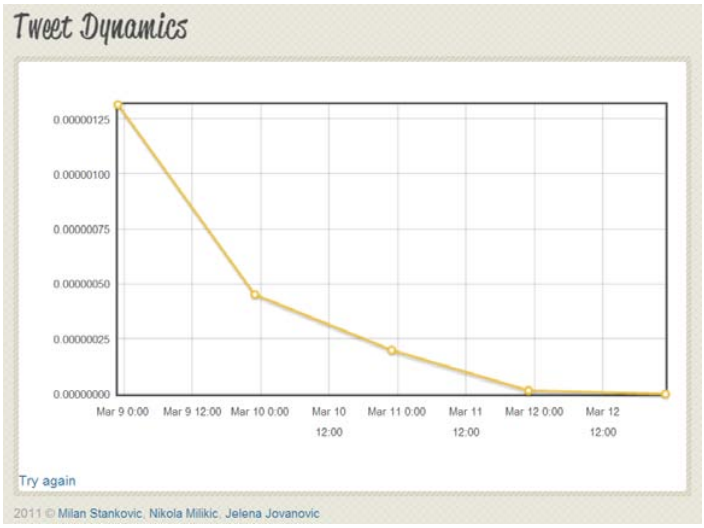


Figure 4 - NMD diagram for terms 'ipad' and 'sxsw' for the 5 days period

Our second example is about terms 'ipad' and 'sxsw' (already mentioned in Section 6). iPad started selling unexpectedly during the SXSW on the 12th of March. From the diagram, it is obvious that there was a rumor about it some days ago, as the NMD decreased exponentially towards the first day of sales, to reach its lowest value on the 12th of March. It is easy to think of potential benefits that the owners of iPad-related content might derive from this newly related term, by including it in their advertising campaigns, and using it for positioning their content. The relevant NMD diagram is shown on Figure 4.

As already mentioned, there is a big limitation of using Twitter Search API, because it limits the number of search results to a maximum of 1500. If we had access to the whole corpus of messages posted in this period, we would have been able to measure the change in relatedness more precisely. However, in the case of terms that are usually rather non-related, the importance of the change is still noticeable even with such limitations imposed.

8 Conclusions and Future Work

This paper presents our initial work on using data streams from Twitter or Twitter-like services for the detection of changes in semantic relatedness of terms. In particular, being inspired by the work of Cilibrasi & Vitanyi [16] on using Google search results for computing semantic relatedness of terms, we have introduced Normalized Micropost Distance (NMD). It makes use of micropost streams of Twitter-like services to compute semantic relatedness of two terms for a given time period. We have also suggested how our approach can be leveraged in two real-life scenarios that differ both in the application domain (online advertising and organizational knowledge management) and the data source to be used for the computation of the NMD measure (Twitter and organization's internal micro-blogging service).

An important challenge to attack in our future work is the detection of good candidate term pairs, i.e., pairs where a change is likely to happen. Our NMD measure allows one to measure the change in semantic relatedness, and follow it over time, but does not directly help in identifying which term pairs are likely to be the subject of change without calculating the NMD values for all possible term pairs. Having such a possibility is important in light of the need for computational efficiency and of the limits imposed by Twitter and other major players on Real-time Web. The detection of candidates for NMD calculation is dependent of the actual usage scenario, as each real-life scenario is related to a specific subject domain characterized by its specific language and important topics. Accordingly, for each scenario, there would be a list of terms to watch. With such a list available, it would be enough to identify the candidate terms that, when coupled with the watched terms could form pairs for which the calculation of NMD might lead to the detection of significant relatedness. We believe that looking at trending topics on Twitter, as well as in recent news articles, might help in finding good candidate terms for a Web marketing scenario (as presented in Section 6). Our intention is thus to explore this research question and deliver a system that could take a number of terms to watch, and provide a list of terms that have recently become more related to one or more of the watched terms.

Another equally important direction of our future work is a comprehensive evaluation of the proposed dynamic measure of semantic relatedness of terms. For that purpose we intend to use Twitter's Streaming API[10], and in particular its "Gardenhose" access level which offers the proportion of the Twitter's public data stream (currently, around 10%) that could form a statistically significant sample. This approach would help us overcome the mentioned limitations of using Twitter Search API. Besides that, since Google recently started including real–time updates coming from Twitter, it could serve us as an important source of data. But since, at the time of writing this paper, these data were not accessible through Google Search API, we need to wait for this feature to become programmatically available. Using this data stream, we intend to do an evaluation study that would consist of a comparative analysis of our approach and the approach we found as the most related to our work, namely the approach reported in Song et al. [21].

References

[1] Wagner, C. (2010). Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams. PhD Symposium at 7th Extended Semantic Web Conference (ESWC2010) Heraklion, Crete, Grece: Springer. Retrieved March 10, 2011, from http://www.springerlink.com/index/R4463T1333777N11.pdf.

[2] Sheth, A., Thomas, C., & Mehra, P. (2010). Continuous Semantics to Analyze Real-Time Data. IEEE Internet Computing 14, 6 (November 2010), 84-89.

[3] Stankovic, M., Rowe, M., & Laublet, P. (2010). Mapping Tweets to Conference Talks: A Goldmine for Semantics. *in Proceedings of the 3rd Social Data on the Web Conferece, SDOW2010, collcoated with International Semantic Web Conference ISWC2010*. Shanghai, China.

[4] Sigurbjörnsson, B., & Zwol, R. van. (2008). Flickr tag recommendation based on collective knowledge. Proceeding of the 17th international conference on World Wide Web - WWW '08, 327. New York, New York, USA: ACM Press. doi: 10.1145/1367497.1367542.

[5] Mei, Q., Zhou, D., & Church, K. (2008). Query suggestion using hitting time. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 469. New York, New York, USA: ACM Press. doi: 10.1145/1458082.1458145.

[6] Safar, B., & Kefi, H. (2004). OntoRefiner, a user query refinement interface usable for Semantic Web Portals. *Proceedings of Application of Semantic Web technologies to Web Communities, Workshop ECAI'04* (pp. 65-79). Retrieved January 25, 2011, from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:OntoRefiner,+a+user+query+refinement+interface+usable+for+Semantic+Web+Portals#0.

[7] Macdonald, C., & Ounis, I. (2007). Expertise drift and query expansion in expert search. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*, 341. New York, New York, USA: ACM Press. doi: 10.1145/1321440.1321490.

[8] Cross, V., "Semantic Relatedness Measures in Ontologies Using Information Content and Fuzzy Set Theory," In Proc. of the 14th IEEE Int'l Conf. on Fuzzy Systems, (2005), pp. 114–119.

[9] Gasevic, D., Zouaq, A., Torniai, C., Jovanovic, J., Hatala, M., "An Approach to Folksonomy-based Ontology Maintenance for Learning Environments," IEEE Transactions on Learning Technologies, 2011 (in press)

[10] Waltinger, U., Cramer, I., & Wandmacher, T. (2009). From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness. Cognitive Science.

[11] Burton-Jones, A., Storey, V., Sugumaran, V., & Purao, S. (2003). A heuristic-based methodology for semantic augmentation of user queries on the web. *Conceptual Modeling-ER 2003*, 476–489.

[10] http://dev.twitter.com/pages/streaming_api

Springer. Retrieved January 19, 2011, from http://www.springerlink.com/index/TP1URDMGDM3F0WP3.pdf.

[12] Ziegler, C.-N., Simon, K., & Lausen, G. (2006). Automatic Computation of Semantic Proximity Using Taxonomic Knowledge Categories and Subject Descriptors. *CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 465-474). Arlington, Virginia, USA: ACM New York, NY, USA. Maguitman, A. G., Menczer, F., Roinestad, H., & Vespignani, A. (2005). Algorithmic detection of semantic similarity. *Proceedings of the 14th international conference on World Wide Web* (p. 107–116). ACM. Retrieved January 25, 2011

[13] Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Arxiv preprint cmp-lg/9511007, 1. Retrieved January 24, 2011, from http://arxiv.org/abs/cmp-lg/9511007

[14] Matos, S., Arrais, J. P., Maia-Rodrigues, J., & Oliveira, J. L. (2010). Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC bioinformatics*, *11*, 212. doi: 10.1186/1471-2105-11-212.

[15] Salton, G. and McGill, M. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983

[16] Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370-383. doi: 10.1109/TKDE.2007.48.

[17] Gracia, J., & Mena, E. (2008). Web-Based Measure of Semantic Relatedness. In Proceedings of the 9th international conference on Web Information Systems Engineering (WISE '08), pp.136-150.

[18] Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, p. 1419). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. Retrieved February 22, 2011,

[19] Mendes, P. N., Passant, A., & Kapanipathi, P. (n.d.). Twarql: Tapping into the Wisdom of the Crowd. *Proceedings of the 6th International Conference on Semantic Systems* (p. 1–3). Graz, Austria: ACM. Retrieved March 14, 2011, from http://portal.acm.org/citation.cfm?id=1839762.

[20] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web* (p. 851–860). ACM. Retrieved March 6, 2011

[21] Song, S., Li, Q., and Zheng, N. (2010). A spatio-temporal framework for related topic search in micro-blogging. In Proceedings of the 6th international conference on Active media technology (AMT'10), Aijun An, Pawan Lingras, Sheila Petty, and Runhe Huang (Eds.). Springer-Verlag, Berlin, Heidelberg, 63-73.

[22] Nagarajan, M., Gomadam, K., Sheth, A., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. Web Information Systems Engineering-WISE 2009 pp. 539-553 (2009)

[23] Farrell, S., Lau, T., Wilcox, E., and Muller, M. "Socially Augmenting employee profiles with people-tagging," Proceedings of the 20th annual ACM symposium on User interface software and technology, Newport, Rhode Island, USA, 2007, pp. 91-100

[24] Braun, S., Kunzmann, C., & Schmidt, A. (2010). People Tagging & Ontology Maturing: Towards Collaborative Competence Management. In: David Randall and Pascal Salembier (eds.): From CSCW to Web2.0: European Developments in Collaborative Design, Selected Papers from COOP08, Springer, Berlin/Heidelberg.

The pragmatics of political messages in Twitter communication

Jurģis Šķilters[1], Monika Kreile[2], Uldis Bojārs[1], Inta Brikše[1], Jānis Pencis[1] and Laura Uzule[1]

[1] University of Latvia, Faculty of Social Sciences, Riga, Latvia
[2] University of Oxford, Oxford, United Kingdom
{jurgisskilters, uldis.bojars, janis.pencis}@gmail.com, monkru@mac.com, inta.brikse@lu.lv, laurauzule@inbox.lv

Abstract. The aim of the current paper is to formulate a conception of pragmatic patterns characterizing the construction of individual and collective identities in virtual communities (in our case: the Twitter community). We have explored several theoretical approaches and frameworks and relevant empirical data to show that the agents building virtual communities are 'extended selves' grounded in a highly dynamic and compressed, linguistically mediated virtual network structure. Our empirical evidence consists of a study of discourse related to the Latvian parliamentary elections of 2010. We used a Twitter corpus (in Latvian) harvested and statistically evaluated using the Pointwise Mutual Information (PMI) algorithm and complemented with qualitative and quantitative content analysis.

Keywords: Twitter, virtual identity, social science, political messages.

1 Introduction

In this paper, we explore the pragmatics of political messages in Latvian Twitter communication during the 2010 general election.

The results contain a topical analysis of election discussions as well as an analysis of hashtags and retweeted messages. The fast pragmatic dynamics in Titter communication can be observed through hashtags, showing a rapid reaction of Twitter users to the elections, while top retweets support the findings of content analysis with regard to political sentiment. Content analysis reveals the possibility of significant discrepancies in terms of the cognitive and physical distances between a group and its individual members in their identity generation processes. In view of the results, we propose a hypothesis that reveals correlations between a group and its individual members, the richness of topics, channels of communication, frequency of mention, and connotations and effects of messages.

2 Theoretical Background

We assume that the generation of identity takes place through two simultaneous and mutually interdependent social categorization processes – belongingness and

differentiation [3,4]. Our study undertakes to examine these two processes in action, constrained by two selection criteria: (1) Twitter messages only, and (2) messages relating directly to national politics. The homogeneity of format and topic draws attention to similarities and differences in content and in discourse strategies.

Twitter is a particularly fruitful resource for this type of analysis because its brevity constraint gives rise to an abundance of shortcut techniques including expressive lexis, the use of abbreviations and hyperlinks for proper names and keywords. Rigid information hierarchies reveal what users presume to be already known and/or shared by their in-group, and are a fertile soil for the investigation of presuppositions, cultural common ground, and cultural discrepancies [6,7,15]. This is especially prominent in Twitter discourse about politics, a topic where speakers generally exhibit willingness to report their opinions despite the fact that their perspectives are often conflicting. Although political opinions are usually articulated explicitly, belongingness to an identity group[1] may be partly implicit [19].

We focus on mechanisms of self-identification, formation and maintenance of in-groups and their differentiation from out-groups. The findings attempt to answer the following questions: 1) How are virtual political identities generated and maintained in a condensed public mode of communication? 2) What are the pragmatic instruments that help to achieve these processes?

Twitter can also help to understand implicit social categorization. Typically, research on social categorization is conducted using questionnaire or focus group methodologies, mainly addressing explicit political categorization. This study has incorporated some implicit factors of analysis, often crucial in political communication. Approaching human-generated digital content as empirical material for categorization analysis is not new (cp. [9]). Analysis of political messages on Twitter, although not directly focused on categorization, is also provided by several studies (cp. [22]). Several recent studies explore possible correlations between election outcomes and the level of Twitter activity of politicians (US Congress: [14], South Korea: [12]). This study, however, also analyzes political messages created by media organizations and other active users.

2.1 Collocations and concordance analysis

Co-occurrence statistics allow to quantitatively project some of a word's semantics grounded in users' categorization performance ([18]). Collocations show the relative most frequent (sometimes stereotypical, implicit) social categories in communication, but the research must be complimented with concordance analysis for semantic complexity. Of course, the output of such a combination of methods concerns the group (and not individual) patterns of social categorization, and pragmatic effects are related to statistical frequency of language used in communities and not to individual patterns of communication[2].

[1] We define identity as a continuous process where the sense of belongingness to a community interacts with the desire to be a unique individual. A community has an internal and an external structure (relationships within the group and relationships with other groups), and community identity can generate polarization effects.

[2] A pragmatic pattern is a typical way of using language in a linguistic community (e.g., in social media).

2.2 Political messages

Studies show that people frequently have difficulty explicitly articulating their ideology [19]. Thus self-report, focus groups, and questionnaires may often prove inadequate for analyzing political categorization. Ideological labels, moreover, may not correspond to subjective conceptions of beliefs, and undecided voters exhibit a much clearer opinion via implicit tasks than via explicit ones [19]. Political categories are distinguished above all by their extreme polarization (cp. [11,17]). On Twitter, initially informative messages are modified to become increasingly polarized [23].

2.3 The Latvian Parliamentary Election 2010

The *Saeima* (the parliament of Latvia) is elected using a proportional multi-partisan representation system for 100 seats. The 2010 election saw 13 competing political parties or their alliances. Candidates from 5 parties were elected: 33 seats for "Unity" (Unity), 29 seats for "Harmony Centre" (HC), 22 seats for the "Union of Greens and Farmers" (UGF), and 8 seats each for the National Association "All For Latvia!"-"TB/LNNK" (NA) and "For a Good Latvia" (FGL). The turnout for the 2010 elections was 63.12% or around 967 000 people.

3 Methodology and Design

The aims of this study are: (1) to build a feasible methodology using content and structural analysis of social media (in particular, Twitter) with respect to political communication; (2) to explore correlations between the election results and the representations of political parties and their candidates in Twitter communication; (3) to explore the identity generation of political actors in pre-election communication on Twitter.

We collected a dataset of tweets covering the election week, performed careful manual extraction work and numerous statistical comparisons. We also created custom tools for analyzing Latvian Twitter content including a concordance tool. We believe that this makes our results, in several respects, even more precise than, e.g., [22] who automatically translated their corpus of empirical data (German tweets) into English and only then processed it with LIWC (Linguistic Inquiry and Word Count).

3.1 Dataset

The dataset consists of one week of Twitter messages (from 28-Sep-2010 to 04-Oct-2010) from a subset of Latvian Twitter users, including 4 days before the election, the day of the general election (October 2) and 2 days following the election. The total size is 50'032 messages, consisting of: 50% regular tweets; 18% retweets; and 32% replies. There are no publicly available official data about the total number of Twitter participants in Latvia. According to local media experts, the estimate is approximately 40'000 users (November 2010).

In order to choose a topically relevant set of Twitter accounts, we started with a manually selected set that included (1) accounts of political parties and their

candidates to the Parliament (Saeima); (2) accounts of media organizations, political analysts, and other individuals who write about politics and the election; and (3) accounts of individuals most active in the Latvian Twitter-sphere. This formed an initial set of 179 accounts to follow. We enlarged the set of accounts by (1) retrieving tweets from the current set of accounts; (2) identifying new accounts mentioned in the tweets collected; (3) filtering out accounts not related to Latvia; and (4) repeating this process. The result is a total of 1'377 user accounts to collect tweets from.

We did not choose a random sample to avoid large amounts of redundant data consisting of ordinary discussions unrelated to our research interests - politics, identity generation, and the media. This intentionally selected dataset allows for a more precise analysis of the above research topics.

3.2 Tweet Processing and Analysis

Collected tweets are processed using the NLTK library [1]. The processing of tweets consists of: cleaning the dataset; saving the full tweet data for structure analysis; tokenizing tweets; replacing keywords, where we consolidate the various ways to write the same word or expression and replace it with a single keyword identifier.

Latvian is an inflected language in which the same word may appear in many forms. In the keyword replacement step, we collapse these forms into one keyword. We also replace different ways of writing the same expression (e.g. abbreviations and full names of party names). Since there was no stemming or lemmatizing software for Latvian that we could use, we created our own keyword replacement map for keywords related to elections.

Having processed the tweets, we performed: (1) content analysis in which we examined the text content of Twitter messages; and (2) structure analysis, in which we examine the metadata in tweets and associated with tweets. The main types of text processing performed in the content analysis phase are concordance lookup, word frequency analysis, and collocation (bigram) analysis. For collocation ranking, we used the Pointwise Mutual Information (PMI) metric [16].

4 Content Analysis

4.1. Representations of the candidates on Twitter

We made a list of all 1234 candidates competing for seats in the parliament, exploring their representations in selected tweets during the 4 days leading up to the election. Since only a small part of all candidates were represented in Twitter communication (in our dataset) four days before the election, we wished to compare our findings with publicity coverage of the candidates in other media in Latvia.

We identified 79 family names of the candidates occurring in collocations in the Twitter dataset, and 170 family names of the candidates occurring in the media monitoring dataset. We distinguish four groups of candidates: (1) those represented both in Twitter and print media and news agencies (44 candidates or 3.56% of all the candidates); (2) those who are represented mostly in Twitter (6.40%); (3) those who are represented mostly in print media and news agencies (7.37%); and (4) those who are mostly not represented in the media we studied (82.67% of all the candidates).

Further, we listed how many personal tweets, collocations and publications occur with every of the family names in various time periods (the average number of collocations of every family name of the candidates four days before the election is 4.68; later, we included only those (9) family names that are statistically significant with respect to their number of collocations ($n \geq 4.68$)). Almost all of these candidates (except one) were elected[1]. They also represent 4 out of the 5 parties elected to the parliament. We analyzed the split of the 100 elected candidates between four previously distinguished groups of candidates. Our calculations show that 32% of elected candidates correspond to the first group (represented in Twitter, print media, and news agencies); 5% correspond to the second group (mostly represented in Twitter); 42% correspond to the third group (mostly represented in print media and news agencies); and 16% correspond to the fourth group (mostly not represented in the media we studied). Based on all of the above, we have formulated a working hypothesis: (1) the more thematically varied and (2) the more frequent the communication, and (3) the more communication channels are used to mention a candidate, the higher the probability that he or she will be elected to parliament.

Table 1: Mentions of political parties (collocations on Twitter, publications in print media and news agencies) and the number of seats in the parliament.

Party	Collocations		Publications[2]		Seats
	28.09-01.10	28.09-04.10	27.09-01.10	1 year	
Unity	34	73	276	2799	33 c[3]
UGF	5	20	223	4837	22 c[6]
HC	10	31	232	4674	29
FGL	41	77	230	1647	8
NA	12	32	168	713	8
LP	3	30	0	0	0
FHRUL	3	7	118	1452	0
R	0	6	0	0	0
OPR	2	2	0	0	0
ML	0	1	0	0	0
DL	0	0	0	144	0
LCDU	0	0	0	0	0
PC	0	0	0	0	0

DL = “Daugava for Latvia”; FGL = “For a Good Latvia”; FHRUL = Union "For Human Rights in a United Latvia"; HC = “Harmony Centre”; LCDU = “Latvian Christian Democratic Union”; LP = “The Last Party”; ML = “Made in Latvia”; NA = National Association “All For Latvia!” – “TB/LNNK”; OPR = “For a Presidential Republic”; PC = “People's Control”; R = Social Democratic Alliance “Responsibility”; UGF = “Union of Greens and Farmers”; Unity = Union “Unity”.

[1] *Election of the 10th Parliament of the Republic of Latvia, October 2, 2010: list and statistics of the candidates. The website of the Central election committee.* Retrieved January 4, 2011 from http://www.cvk.lv/cgi-bin/wdbcgiw/base/komisijas2010.cvkand10.sak

[2] Publications in print media and news agencies for (1) the election week (27-Sep – 03-Oct); (2) one year (28-Sep-2009 - 03.10.2010). Dates differ from those in tweet collocations due to the source of press data.

[3] c = Formed the ruling coalition.

4.2 Representations of the parties on Twitter, in print media, and by news agencies prior to the election

For names of political parties (Table 1) we listed: (1) how many collocations occur with each name; (2) how many publications from print media and news agencies mention each name; and (3) the results each party has achieved in the election. Every party with an above-average number of collocations in Twitter communication before the election (8.46) is elected to the parliament. An exception is UGF, which was elected despite a below-average number of collocations. We assume that the latter was compensated in the long term by the highest number of publications in print media and news agencies. However, with the high ranking of mention on Twitter before the election (41 collocations), FGL obtained significantly fewer parliament places than "Unity" or other political parties with a lower ranking of mention on Twitter. Initially, it can be assumed that FGL was affected by relatively lower publicity rates in print media and news agencies; but in fact, FGL had conducted a more extensive advertising campaign than any other political party). Further investigation points to an important qualitative factor. A review of collocations of FGL and "Unity" in a detailed concordance analysis leads to the observation that the "Unity" collocations feature more positive connotations than the FGL collocations. This allows us to emphasize and modify our above hypothesis regarding the candidates: (1) the more thematically varied and (2) the more frequent the communication, and (3) the more communication channels a political party is mentioned in *positively*, the higher the probability that it will be elected to the parliament.

4.3 Identity-generation processes for political parties and individuals in Twitter communication

Two political parties – FGL and "Unity" - have significantly higher rankings of mention than other parties. Moreover, their candidates for the post of Prime Minister (Ainārs Šlesers (FGL) and Valdis Dombrovskis (Unity)) have similar rankings of mention. In spite of these similarities, the two have strikingly different election results ("Unity" won the election and got 33 seats in the parliament, with Valdis Dombrovskis approved as the prime minister, while FGL got only 8 seats in the parliament). This led us to investigate more closely the identity generation of these individuals and organizations through political categorization in pre-election tweets. First, we identified 10 collocations of significantly high ratings for the four name keywords. Secondly, we used concordance analysis to examine the semantics in each collocation.

We have listed in Table 2 what percentage of the topics bear positive, neautral or negative connotations and how many topics are covered by each of the keywords. As Table 2 demonstrates, the individual and the organization are categorized similarly in the case of Šlesers and his political party FGL: both are more related to negative topics than positive ones. The case of Valdis Dombrovskis and his political party "Unity" is different: the individual is mostly categorized in positive or neutral topics, while the political party is categorized in negative or neutral ones. This shows that the generation of identity of an organization and that of its individual members may

involve significant discrepancies in terms of cognitive versus physical distances[4]. In this case, the cognitive distance between Dombrovskis and "Unity" is bigger than the 'physical' one. This may be in part due to the fact that the "Unity" election campaign focused exclusively on Dombrovskis, promoting him as the principal benefit to the voters. Thus the individual became more cognitively important than the whole (an organization).

This allows us to expand our hypothesis regarding politicians and political parties as follows: (1) the more thematically varied and (2) the more frequent the communication, and (3) the more communication channels are used to mention a member of an organization (in this case, a politician) *positively*[5], the higher the probability that he or she will become cognitively more important than the organization (in this case, the political party) and cause a shift in the perception of the significance of the organization.

Table 2: Connotations of keyword topics.

	Positive	Neutral	Negative	No of topics
Politician (party)				
Dombrovskis (Unity)	27.59%	68.97%	3.45%	29
Šlesers (FGL)	25.00%	41.67%	33.33%	12
Political party				
Unity	0.00%	54.55%	45.45%	11
FGL	21.43%	21.43%	57.14%	14

5 Structural Analysis

In this section, we analyze Twitter messages by examining implicit and explicit metadata and structural information contained in tweets.

5.1 Hashtag Analysis

Hashtags were used in 2'238 tweets (4.47% of all tweets). In total, 750 different hashtags were used 2'668 times. Most hashtags were used just once. 29.06% of hashtags (218) were used more than once and 2.26% (17) were used at least 20 times.

The most popular hashtag was #velesanas ("election"), used in 459 tweets (17.2% of tweets containing hashtags). Other election-related hashtags that were used at least 20 times include #nobalsoju ("i voted"), #politsports (political sport), #pietiek

[4] In the Spreading-Activation Theory, assuming a correlation between the collocational structure of the corpus and the mental models of its users, collocational structure reflects the cognitive distance between conceptual entities such as political parties and individuals. Indirectly connected nodes are more distant than directly connected ones.

[5] Using manual concordance analysis, connotations are determined and generalized according to three categories (positive, neutral, and negative), determined individually for each tweet. Examples include: "Friends, tomorrow I shall vote for Dombrovskis, because I trust his professionalism …" (positive); "Šlesers doubts the objectivity of social media …" (neutral); "Dombrovskis: a protégé of corruption or a racketeer?" (negative).

("enough!"), #vēlēšanas (#velesanas with Latvian diacritics), #cieti ("solid" – a slogan of FGL), #twibbon (twibbons were used to show party support).

For the purposes of this paper, we limited Table 3 to hashtags related to politics. Most of the top 10 hashtags on election day were related to politics (9 out of 10) and appear in the table. Other days had less election-related tags, but also a lower hashtag usage activity in general. The #velesanas ("election") hashtag appeared the day before the election and had a remarkable spike in its usage on election day and the day following it, receding back to background level the day after that.

Table 3: Dynamics of top hashtags related to politics (30-Sep-2011 – 04-Oct-2011).

Hashtag	30-Sep	01-Oct	02-Oct	03-Oct	04-Oct
#ir	27	34	21	32	
#pietiek	7	10	16	10	5
#pll	5				
#politika	5				
#politsports	5				
#velesanas		12	346	94	5
#cieti		8	16		
#fail		9	5		9
#sleptareklama		5			
#nobalsoju			71		
#twibbon			60		
#vēlēšanas			35	11	
#velesanas2010			7		

Hashtags that retained popularity for at least 4 days in this 5 day period were the journalism tags #pietiek and #ir. Both refer to publications seen by top Twitter users as prestigeous and integral organizations for investigative journalism. The hashtag #sleptareklama ("hidden advertising") coincided with the appearance of controversial hockey-related advertisements that were suspected of containing hidden political advertising. A creative usage of a hashtag is its syntactic integration into a sentence: a notable example is using #ir, the magazine whose name literally means "is", as a verb: e.g., "There #is still time to form a new coalition".

Apart from the obvious purpose of attracting attention to major topics, hashtags carry the connotation of familiarity with the object of the tag, be it a topic, an individual, or an organization – at the very least, one must know what is worth tagging. Tags help to define group identity in two recursive ways: by highlighting issues considered important by the group, and by presenting the group as the kind of community where such issues are considered important.

5.2 Analysis of Retweeting

We considered a retweet any Twitter message that contains the string "RT @nickname" (17.68% of the selected dataset). Most retweets start with "RT @nickname", i.e. are marked as such and point to the original message. These results shows more uniformity of retweet formats than reported in [2], possibly a result of more officialized retweet functionality. For further analysis, we used retweets which

contained information about the original tweet (i.e. 90.46% of all retweets). An analysis of the top 20 most retweeted posts reveals that 70% of these posts are directly related to the elections; 10% are loosely related; 20% are unrelated.

There were 14 election-related messages among the 20 most retweeted messages. Of these, the majority (8 out of 14) were satirical tweets criticising a political party or a politician. Seven refer to FGL or its prominent members Ainārs Šlesers and Andris Šķēle. Other parties mentioned in these retweets were HC and FHRUL (one tweet each). The two most retweeted messages are related to the election.

5.3 Opinion leaders and in-group demarcation mechanisms

The content of top retweets and hashtags reveals that the opinion leaders in the Latvian Twitter-sphere, the in-group that enjoys the highest popularity and prestige, can be vaguely defined as a group of centrists who see themselves as positioned between two perceived polarities. The cognitive space, as regarded by the in-group, can be characterized thus: to the left are *krievi* ("the Russians"), the parties and their supporters commonly perceived as pro-Muscovite and representing the interests of the Russian-speaking population (HC, FHRUL). To the right are *nēģi* ("the parasites" – an imprecise translation of the word taken from a popular tweet criticising this group), the nationalist alliance (FGL, NA) that the Twitter opinion leaders see as outdated and highly corrupt, exploiting their privilege for personal gain. The in-group supports the political alliance "Unity" and particularly its leader, Valdis Dombrovskis, who was subsequently elected Prime Minister.

The fact that the in-group appears to take a centrist position is significant: their output is less polarizing than could be expected of a highly politicized group. Still, there is a clear demarcation of the in-group from both out-groups described above. This is achieved by the opinion leaders of the in-group through several group-identity-generating mechanisms and strengthened by the heightened emphasis on the social self [4], typical of both online communities and political discourse.

Manipulating cognitive distances is relatively easy in the dematerialized virtual space, which facilitates impressions of togetherness and mutual identification within the in-group, on the one hand, but also the distancing of the in-group from out-groups. Perhaps surprisingly, the brevity constraint of Twitter messaging, rather than complicating political categorization, can facilitate it: the format is well suited to the in-group's simplified tripartite view of the political space. Thus, through repeated tweeting of negative content containing the letters "PLL" or "PCTVL" (acronyms of the names of political parties on the two sides of the perceived spectrum), it is soon enough to write "PLL" or "PCTVL" to evoke a cognitive frame [8] associated with negative content. Clearly, the details of this content will be unique for each user; but as long as there is a basic understanding of a commonality of reference – in this case, of the negativity of the referents – a mention of a party acronym will effectively serve as an invitation to 'fill in the gaps' with each reader's own meaning [13].

Political jokes[6], abundant in top retweets, work in a similar manner. Provided that the humourous effect is usually achieved by inviting the audience to *frame-shift* through an unexpected element [7], political jokes on Twitter are doubly rewarding because they give the audience the feeling of belongingness through having understood the frame shift without surrounding linguistic context and through a very limited number of signs. Similarly to a hashtag, a retweet works recursively by simultaneously flaunting an individual's understanding (and hence his belonging to the in-group) and helping to define his individual identity through the content of what is understood and retweeted.

Our corpus shows that power and control are very much the preoccupation of Twitter users, and the independently formed, 'grass-roots' community of top tweeters quickly forms their own behaviour canons. This is typical of online communities, where a myriad of rules and expectations underlie seemingly free, chaotic communication [10]. A popular political message on Twitter is at once an expression of individual and group identity, an invitation to the in-group members to share the opinion expressed, and a warning about the consequence of deviating from the group's norms. By way of illustration, a message retweeted 15 times reads: "I heard that Šlesers won't vote for PLL *either*, because they're said to be thiefs" (our emphasis). In addition to cleverly poking fun at the politician by suggesting he will not vote for his own party, the message succeeds in conveying that the author will not vote for Šlesers, that he assumes that his in-group members will not do so, and that anyone who does vote for Šlesers will be seen as voting for a thief and undermining his or her in-group membership. In short, Twitter conformity mechanisms are just as compact as the medium itself.

Yet without a conforming audience, such successful guidance toward a rigidified, formal categorisation would not be possible (we may well judge the above message as successful, since it is on the list of top retweets). The tension between individual opinion and in-group identification (the personal vs. the interpersonal/social self) is resolved through a balance of stereotyping processes: just as the political parties and actors are stereotyped to fit into one of the few cognitive categories carved out for the occasion of the election, so the individual members engage in a certain degree of *self-stereotyping* [20]. Members will be more willing to overlook differences of opinion and concentrate on their commonalities (real or imagined) when membership is seen as beneficial, and particularly if the group is seen as working toward a common goal of some sort – in this case, victory in the parliamentary elections [4]. Because intra-group attraction on Twitter in the run-up to parliamentary election is ideational rather than interpersonal, the in-group *achieves* a high degree of political cohesion in part simply through *perceiving itself* as a cohesive unit.

[6] An example that is comparatively demure and reproducible in an academic paper refers to the leader of the party perceived as being in the "parasites" group: "A little boy falls. Šlesers helps him up. 'So, I guess now you will vote for me?' 'I only hurt my foot, not my head!'".

6 Results and Conclusion

We have formulated a correlation according to which three factors contribute to the efficiency of political messages in the electoral discourse – in particular, *for a given collocation bigram*: (a) the variety of thematic contexts of occurrence, (b) the frequency of mention, (c) positive connotations. (While there are other factors determining efficiency, this study has focused on popularity-oriented facets.) We have therefore extended the results stated by [12, 14] regarding the correlation between minority parties, Twitter activity, and election results. The dynamics of Twitter users' interest in the event (the election) can be observed through hashtag usage and the most retweeted messages. Top retweets, in turn, convey user sentiment toward political parties and individuals.

We have noted instances of discrepancy between attitudes toward individual politicians as opposed to attitudes toward political groups, and observed that frequent positive mention of individuals can lead to a heightened cognitive significance of this individual, causing the perception of the significance of the relevant organization to recede into the background.

We envision possible applications of this work in analysis tools correlating Twitter dynamics with the structure generated from the parameters: (a) the variety of occurrence contexts, (b) the frequency of mention, (c) positive connotations (generated semi-automatically). The items which fit into the highest ranking of such analysis results can be further analyzed manually and a variety of pragmatic effects (stereotyping, presupposition generation a.o.) might be observed.

Finally, we can hypothesize that the user of a microblogging resource such as Twitter extends the sphere of his or her cognitive processing by involving additional interactive structures of communication. Thus, if we assume that the social categorization in a community consists of (a) self-categorization as the most crucial and basic level of identity building, (b) interpersonal communities of individuals, and (c) large-scale social communities (e.g., national identity communities) including sub-communities [4], we could argue that self-categorization involves a substantial amount of extended cognitive processing offloaded onto the digital environment (in our case, Twitter). In this sense, the results provided by our study can complement research on the extented mind [5, 21]. A more detailed analysis of the extended self and offloading effects in cognitive processing is a topic for another study.

Acknowledgments

This work has been supported by the European Social Fund project «Support for Doctoral Studies at the University of Latvia».

References

1. Bird, S., Klein E., & Loper , E. (2009), Natural Language Processing with Python. O'Reilly.
2. Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS-43. IEEE: Kauai, HI, January 6.
3. Brewer, M. B., (1991). The social self: On being the same and different at the same time. Personality and Social Psychology Bulletin, 17, 475-482.
4. Brewer, M. B., & Gardner, W. (1996). Who is this "we"? Levels of collective identity and self representations. Journal of Personality and Social Psychology, 71, 83-93.
5. Clark, A. (2006). Pressing the flesh: Exploring a tension in the study of the embodied, embedded mind. Philosophy and Phenomenological Research.
6. Clark, H.H. et al. (1983). Common ground and the understanding of demonstrative reference. Journal of Verbal Learning and Verbal Behaviour, 22, 245-258.
7. Coulson, S. et al. (2006). Looking back: Joke comprehension and the space structuring model. Humor, 19, 229-250.
8. Fillmore, C. J. 1976. Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Vol. 280, pp. 20-32
9. Glushko, R.J., Maglio, P.P., Matlock, T., & Barsalou, L.W. (2008). Categorization in the wild. Trends in Cognitive Sciences, 12, 129-135.
10. Haythornthwaite, C. (2007). Social networks and online community. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.). The Oxford Handbook of Internet Psychology (pp. 121-137). Oxford: Oxford University Press.
11. Heit, E., & Nicholson, S.P. (2010). The opposite of republican: polarization and political categorization. Cognitive Science, 34, 1503-1516.
12. Hsu, C.-L., & Park, H.W. (2010). Sociology of hyperlink networks of Web 1.0, Web 2.0, and Twitter: a case study of South Korea. Social Science Computer Review, 0894439310382517, first published on September 21, 2010.
13. Langacker, R.W. (2000). Grammar and Conceptualization. Berlin/New York: Mouton de Gruyter.
14. Lassen, D.S., & Brown, A.S. (2010). Twitter: the electoral connection? Social Science Computer Review, 0894439310382749, first published on September 23, 2010.
15. Laurent, J.P. et al. (2006). On understanding idiomatic language: the salience hypothesis assessed by ERP's. Brain Research, 1068, 151-160
16. Manning, C.D., & Schutze, H. (2003). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, 2003.
17. Medin, D. L., Lynch, E. B., & Solomon, K. E. (2000). Are there kinds of concepts? Annual Review of Psychology, 51, 121-147.
18. Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. Cognitive Science, 34, 1388-1429.
19. Nosek, B.A., Graham, J., & Hawkins, C.B. (2010). Implicit political cognition. In B. Gawronski, & B.K. Payne (Eds.). Handbook of Implicit Social Cognition: Measurement, Theory, and Applications (pp. 548-564). New York: The Guilford Press.
20. Smith, E. R., & Henry, S. (1996). An in-group becomes part of the self:Response time evidence. Personality and Social Psychology Bulletin 22, 635-642.
21. Spivey, M., Richardson, D. & Fitneva, S. (2004). Thinking outside the brain: Spatial indices to linguistic and visual information. In J. Henderson & F. Ferreira (Eds.), The Interface of Vision, Language, and Action (161-189). New York: Psychology Press.
22. Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2010). Predicting elections with twitter: what 140 charcters reveal about political sentiment. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. 178-185.
23. Yardi, S., & Boyd, D. (2010). Dynamic debates: an analysis of group polarization over time on Twitter. Bulletin of Science, Technology & Society, 30(5), 316-327.

Automatic detection of political opinions in Tweets

Diana Maynard and Adam Funk

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield, UK
diana@dcs.shef.ac.uk

Abstract. In this paper, we discuss a variety of issues related to opinion mining from microposts, and the challenges they impose on an NLP system, along with an example application we have developed to determine political leanings from a set of pre-election tweets. While there are a number of sentiment analysis tools available which summarise positive, negative and neutral tweets about a given keyword or topic, these tools generally produce poor results, and operate in a fairly simplistic way, using only the presence of certain positive and negative adjectives as indicators, or simple learning techniques which do not work well on short microposts. On the other hand, intelligent tools which work well on movie and customer reviews cannot be used on microposts due to their brevity and lack of context. Our methods make use of a variety of sophisticated NLP techniques in order to extract more meaningful and higher quality opinions, and incorporate extra-linguistic contextual information.

Key words: NLP, opinion mining, social media analysis

1 Introduction

Social media provides a wealth of information about a user's behaviour and interests, from the explicit "John's interests are tennis, swimming and classical music", to the implicit "people who like skydiving tend to be big risk-takers", to the associative "people who buy Nike products also tend to buy Apple products". While information about individuals is not always useful on its own, finding defined clusters of interests and opinions can be interesting. For example, if many people talk on social media sites about fears in airline security, life insurance companies might consider opportunities to sell a new service. This kind of predictive analysis is all about understanding one's potential audience at a much deeper level, which can lead to improved advertising techniques, such as personalised advertisements to different groups.

It is in the interests of large public knowledge institutions to be able to collect and retrieve all the information related to certain events and their development over time. In this new information age, where thoughts and opinions are shared through social networks, it is vital that, in order to make best use of this information, we can distinguish what is important, and be able to preserve it, in order

to provide better understanding and a better snapshot of particular situations. Online social networks can also trigger a chain of reactions to such situations and events which ultimately lead to administrative, political and societal changes.

In this paper, we discuss a variety of issues related to opinion mining from microposts, and the challenges they impose on a Natural Language Processing (NLP) system, along with an example application we have developed to divulge political leanings from a set of pre-election tweets. While knowing that Bob Smith is a Labour supporter is not particularly interesting on its own, when this information is combined with other metadata, and information about various groups of people is combined and analysed, we can begin to get some very useful insights about political leanings and on factors that impact this, such as debates aired on television or political incidents that occur.

We first give in Section 2 some examples of previous work on opinion mining and sentiment analysis, and show why these techniques are either not suitable for microposts, or do not work particularly well when adapted to other domains or when generalised. We then describe the opinion mining process in general (Section 3), the corpus of political tweets we have developed (Section 4), and the application to analyse opinions (Section 5). Finally, we give details of a first evaluation of the application and some discussion about future directions (Sections 6 and 7).

2 Related Work

Sentiment detection has been applied to a variety of different media, typically to reviews of products or services, though it is not limited to these. Boiy and Moens [1], for example, see sentiment detection as a classification problem and apply different feature selections to multilingual collections of digital content including blog entries, reviews and forum postings. Conclusive measures of bias in such content have been elusive, but progress towards obtaining reliable measures of sentiment in text has been made – mapping onto a linear scale related to positive versus negative, emotional versus neutral language, etc.

Sentiment detection techniques can be roughly divided into lexicon-based methods [2] and machine-learning methods [1]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. A document's polarity is the ratio of positive to negative terms. Machine learning approaches make use of syntactic and/or linguistic features, including sentiment lexicons. Hybrid approaches are very common, and sentiment lexicons play a key role in the majority of methods. However, such approaches are often inflexible regarding the ambiguity of sentiment terms. The context in which a term is used can change its meaning, which is particularly true for adjectives in sentiment lexicons [3]. Several evaluations have shown that sentiment detection methods should not neglect contextual information [4, 5], and have identified context words with a high impact on the polarity of ambiguous terms [6]. Besides the ambiguity of human language, another bottleneck for sentiment detection methods is the time-consuming creation of sentiment dictionaries. One solution

to this is a crowdsourcing technique to create such dictionaries with minimal effort, such as the Sentiment Quiz Facebook application[1].

However, sentiment dictionaries alone are not enough, and there are major problems in applying such techniques to microposts such as tweets, which typically do not contain much contextual information and which assume much implicit knowledge. They are also less grammatical than longer posts and make frequent use of emoticons and hashtags, which can form an important part of the meaning. This means that typical NLP solutions such as full - or even shallow - parsing are unlikely to work well, and new solutions need to be incorporated for handling extra-linguistic information. Typically, they also contain extensive use of irony and sarcasm, which are also difficult for a machine to detect.

There exists a plethora of tools for performing sentiment analysis of tweets, though most work best on mentions of product brands, where people are clearly expressing opinions about the product. Generally, the user enters a search term and gets back all the positive and negative (and sometimes neutral) tweets that contain the term, along with some graphics such as pie charts or graphs. Typical basic tools are Twitter Sentiment[2], Twends[3] and Twitrratr[4]. Slightly more sophisticated tools such as SocialMention[5] allow search in a variety of social networks and produce other statistics such as percentages of Strength, Passion and Reach, while others allow the user to correct erroneous analyses. While these tools are simple to use and often provide an attractive display, their analysis is very rudimentary, performance is low, and they do not identify the opinion holder or the topic of the opinion, assuming (often wrongly) that the opinion is related to the keyword.

3 Opinion mining process

We have developed an initial application for opinion mining using GATE [7], a freely available toolkit for language processing. The first stage in the system is to perform a basic sentiment analysis, i.e., to associate a positive, negative or neutral sentiment with each relevant tweet. This is supplemented by creating triples of the form <*Person*, *Opinion*, *Political Party*>, e.g., <*Bob Smith*, *pro*, *Labour*> to represent the fact that Bob Smith is a Labour supporter. Given the nature of tweets, we have found that it is fairly rare to see more than one sentiment about the same thing expressed in a single tweet: if, however, two opposing opinions about a political party are mentioned, then we simply posit a neutral opinion at this stage.

Once the triples have been extracted, we can then collect all mentions that refer to the same person, and collate the information. For example, John may be equally drawn towards more than one party, not just Labour, but hates

[1] http://apps.facebook.com/sentiment-quiz
[2] http://twittersentiment.appspot.com/
[3] http://twendz.waggeneredstrom.com/
[4] http://twitrratr.com/
[5] http://socialmention.com/

the Conservatives. His opinion may also change over time, especially during the pre-election phase, or since the recent elections. We thus go beyond typical sentiment analysis techniques which only look at a static opinion at a fixed point in time. This is important because it enables us to make much more interesting observations about political opinions and how they are affected by various events.

4 The pre-election twitter corpus

For the development of our application, we used a corpus of political tweets collected over the UK pre-election period in 2010[6]. The Twitter Streaming API[7] was used to collect tweets from this period according to a variety of relevant criteria (use of hash tags such as #election2010, #bbcqt (BBC Question Time), #Labour etc., specific mention of various political parties or words such as "election", and so on). The tweets were collected in JSON format and then converted to xml using the JSON-Lib library[8]. The corpus contains about 5 million tweets; however it contains many duplicates. De-duplication, which formed a part of the conversion process, reduced the corpus size by about 20% to around 4 million tweets.

The corpus contains not only the tweets themselves, but also a large amount of metadata associated with each tweet, such as its date and time, the number of followers of the person tweeting, the location and other information about the person tweeting, and so on. This information is useful for disambiguation and for collating the information later. Figure 1 depicts a tweet loaded in GATE, with the text and some of the metadata (location, author, and author profile) highlighted. We should note that the method for collecting tweets is not perfect, as we find some tweets which are nothing to do with the election, due to ambiguous words (in particular, "Labour" which could also be a common noun, and "Tory" which could also be a person's name). For future work, we plan a more sophisticated method for collecting and pruning relevant tweets; nevertheless, this quick and dirty method enabled us to get the initial experiments underway quickly.

5 Application

The application consists of a number of processing modules combined to form an application pipeline. First, we use a number of linguistic pre-processing components such as tokenisation, part-of-speech tagging, morphological analysis, sentence splitting, and so on. Full parsing is not used because of the nature of the tweets: from past experience, we know it is very unlikely that the quality would be sufficiently high. Second, we apply ANNIE [8], the default named entity

[6] We are very grateful to Matthew Rowe for allowing us to use this corpus.
[7] http://dev.twitter.com/pages/streaming_api
[8] http://json-lib.sourceforge.net/

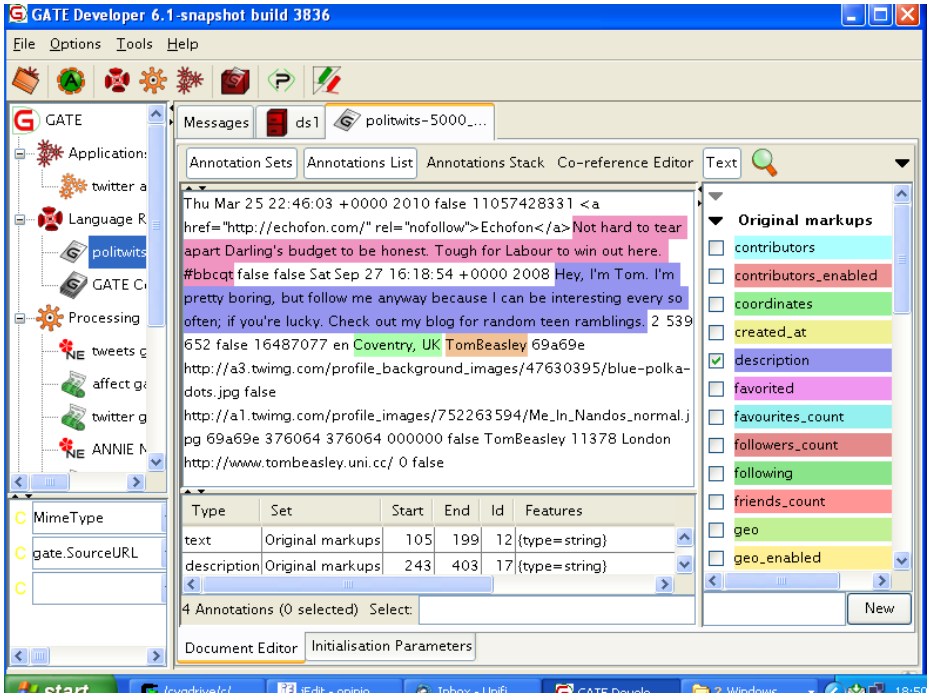


Fig. 1. Screenshot of a tweet in GATE, with relevant metadata

recognition system available as part of GATE, in order to recognise named entities in the text (Person, Organisation, Location, Date, Time, Money, Percent). The named entities are then used in the next stages: first for the identification of opinion holders and targets (i.e., people, political parties, etc.), and second, as contextual information for aiding opinion mining.

The main body of the opinion mining application involves a set of JAPE grammars which create annotations on segments of text. JAPE is a Java-based pattern matching language used in GATE [9]. The grammar rules create a number of temporary annotations which are later combined with existing annotations and converted into final annotations. In addition to the grammars, we use a set of gazetteer lists containing useful clues and context words: for example, we have developed a gazetteer of affect/emotion words from WordNet[10]. These have a feature denoting their part of speech, and information about the original WordNet synset to which they belong. The lists have been modified and extended manually to improve their quality: some words and lists have been deleted (since we considered them irrelevant for our purpose) while others have been added.

As mentioned above, the application aims to find for each relevant tweet, triples denoting three kinds of entity: Person, Opinion and Political Party. The

application creates a number of different annotations on the text which are then combined to form these triples.

The detection of the actual opinion (sentiment) is performed via a number of different phases: detecting positive, negative and neutral words (Affect annotations), identifying factual or opinionated versus questions or doubtful statements, identifying negatives, and detecting extra-linguistic clues such as smileys.

Because we only want to process the actual text of the tweet, and not the metadata, we use a special processing resource (the *Segment Processing PR*) to run our application over just the text covered by the XML "text" tag in the tweet. We also use this to access various aspects of the metadata from the tweet, such as the author information, as explained below.

5.1 Affect annotations

Affect annotations denote the sentiment expressed in the tweet, which could be positive, negative or neutral towards a particular party. These are created primarily by the gazetteer (sentiment dictionary), but the sentiment denoted can then be modified by various contextual factors. First, the gazetteer is used to find mentions of positive and negative words, such as "beneficial" and "awful" respectively. A check is performed to ensure that the part of speech of the gazetteer entry matched and the word in the text are the same, otherwise no match is produced. This ensures disambiguation of the various categories. For example, note the difference between the following pairs of phrases: "Just watched video about awful days of Tory rule" vs "Ah good, the entertainment is here." In the first phrase, "awful" is an adjective and refers to the "days of Tory rule": this would be appropriate as a match for a negative word. In the second phrase, "good" is an adverb and should not be used as a match for a positive sentiment about the entertainment (it does not actually denote that the entertainment itself is good, only that the author is looking forward to the entertainment). Similarly, note the difference between the preposition and verb "like" in the following pair of phrases, which again express very different sentiments about the person in question: "People like her should be shot." vs "People like her."

5.2 Other clues

Hashtags can also be a source of information about the opinion of the author. Some are fairly explicit, for example #VoteSNP, #Labourfail, while others are more subtle, e.g., #torytombstone, #VoteFodderForTheTories. Currently, we list a number of frequently occurring hashtags of this sort in a gazetteer list, but future work will involve deconstructing some of these hashtags in order to deduce their meaning on the fly (since they are not correctly tokenised, they will not be recognised by our regular gazetteers and grammars). Some hashtags are easier to decipher the meaning of than others: for example, #torytombstone requires some implicit knowledge about the use of the word "tombstone" being used in a derogatory way.

5.3 Opinionated statements

This phase checks the tweets to see if they are opinionated, or whether they contain questions or doubtful statements. For example, it is hard to tell from the question: "Wont Unite's victory be beneficial to Labour?" whether the author is a supporter of Labour or not, so we posit simply a neutral opinion here. Initially, we match any statement containing an Affect annotation as being opinionated, unless it contains a question, but this could be extended to deal with other cases. We annotate any tweet that contains a question mark (or potentially other distinguishing question-related features) as a Question, and retain it for possible later use, but do not annotate it as an Opinion at this point.

5.4 Negatives

This phase checks the tweet to see if it contains some negative word (as found in a gazetteer list), such as "not", "couldn't", "never" and so on. In most cases, it will reverse the opinion already found: the existing feature value on the Sentiment annotation is changed from "pro" to "anti" or vice versa. More complex negatives include checking for words such as "against", "stop" and so on as part of a noun phrase involving a political party, or as part of a verb phrase followed by a mention of a political party.

5.5 Political Party

Finding the name of the Political Party in the tweet is generally straightforward as there are only a limited number of ways in which they are usually referred to. As mentioned above, however, there is some ambiguity possible. We therefore use other clues in the tweet, where possible, to help resolve these. For example, if "Tory" is part of a person's name (identified by ANNIE), we discard it as a possible political party.

5.6 Identifying the Opinion Holder

Where a Person is identified in the tweet as the holder of the opinion (via another set of grammar rules), we create a Person annotation. If the opinion holder in the pattern matched is a Person or Organization, we create a Person annotation with the text string as the value of an opinion_holder feature on the annotation. If the opinion holder in the pattern matched is a pronoun, we first find the value of the string of the matching proper noun antecedent and use this as the value of the opinion_holder feature. Currently, we only match opinion holders within the same sentence.

However, there may not always be an explicit opinion holder. In many cases, the author of the tweet is the opinion holder, e.g., "I'm also going to vote Tory. Hello new world." Here we can co-refer "I" with the person tweeting. In other cases, there is no opinion holder explicitly mentioned, e.g., "Vote for Labour. Harry Potter would." In this case, we can assume that the opinion is also held by the author. In both cases, therefore, we use "author" as the value of opinion_holder, and get the details of the tweet author from the xml metadata.

5.7 Creating triples

As described above, we first create temporary annotations for Person, Organization, Vote, Party, Negatives etc. based on gazetteer lookup, named entities and so on. We then use a set of rules to combine these into triples, for example:

<*Person*, *Vote*, *Party*>
"Tory Phip admits he voted LibDem" → <*Phip*, *pro*, *LibDem*>

<*Person*, *Party*, *Affect*>
"When they get a Tory government they'll be sorry." → <*author*, *anti*, *Tory*>

Finally, we create an annotation "Sentiment" which has the following features:

- kind = pro_Labour, anti_LibDem, etc.
- opinion_holder = Bob Smith, author, etc.

Currently, we restrict ourselves to rules which are very likely to be successful, thus achieving high Precision at the expense of Recall. These rules should be eventually expanded in order to get more hits, although Precision may suffer as a result.

6 Evaluation and Discussion

We evaluated the first stage of this work, i.e., the basic opinion finding application, on a sample set of 1000 tweets from the large political corpus (selected randomly by a Python script). We then ran the application over this test set and compared the results. Table 1 gives some examples of the different opinions recognised by the system: it shows the tweet (or the relevant portion of the tweet), the opinion generated by the system (labelled "System") and the opinion generated by the manual annotator (labelled "Key").

Out of 1000 tweets, the system identified 143 as being opinionated (about a political party), i.e., it created a Sentiment annotation for that tweet. We analysed these tweets manually and classified them into the following categories: ProCon, AntiCon, ProLab, AntiLab, ProLib, AntiLib, Unknown and Irrelevant. The first 6 of these categories match the system annotations. Unknown is marked when either a political opinion is not expressed or where it is unclear what the political opinion expressed is, e.g., *"Labour got less this time than John Major did in 1997."* Irrelevant is marked when the tweet is not relevant to politics or the election, e.g., *"i am sooooooooo bored, want to go into labour just for something to do for a couple of hours :)"*. The distinction between Irrelevant and Unknown is only important in that it tells us which tweets should ideally be excluded from the corpus before analysis: we want to include the Unknown ones in the corpus (even though the system should not annotate them), in order to ensure that the system does not annotate false positives as containing a political sentiment, but not the Irrelevant ones. While only 2 documents out of the 143 were classed as

Tweet	**System**	**Key**
I just constantly said "Vote Labour" in a tourettes kinda way	pro-Lab	pro-Lab
Daily Mail reveals PM's wife has ugly feet http://bit.ly/b6ZNlK ¡–Eww! Another reason not to vote Labour.	pro-Lab	pro-Lab
Still, can't bring myself to vote tactically for Labour	anti-Lab	anti-Lab
@WilliamJHague If you fancy Interest Rates at 15.8% Vote Tory .... they will throw you out of your house...back to the 80's	pro-Con	anti-Con
Vote Tory to stop them bleating! You know it's worth it.	pro-Con	pro-Con
George Osborne. Reason number 437 not to vote Tory.	anti-Con	anti-Con
Vote Tory or Labour, get Lib Dems. Might as well vote LibDem and have done with it	pro-Lib	pro-Lib
@Simon_Rayner sorry but laughing so much it hurts. Who in their right mind will vote for libdem savage cuts?	anti-Lib	anti-Lib

Table 1. Examples of tweets and the opinions generated

Key/System	ProCon	AntiCon	ProLab	AntiLab	ProLib	AntiLib	Total
ProCon	**5**	0	0	0	0	0	5
AntiCon	10	**5**	0	2	0	0	17
ProLab	0	0	**69**	2	0	0	70
AntiLab	0	0	4	**4**	0	0	8
ProLib	3	0	1	0	**6**	0	10
AntiLib	0	0	0	0	0	**1**	1
Unknown	10	1	11	5	2	0	29
Irrelevant	0	1	0	1	0	0	2
Total	28	7	85	14	8	1	143

Table 2. Confusion matrix for evaluation corpus

Irrelevant, 29 were classed as Unknown (roughly 20%). This means that roughly 80% of the documents that the system classified with a Sentiment, were in fact opinionated, though not all of them had the correct opinion.

Table 2 shows a confusion matrix for the different sentiments recognised by the system, compared with those in the Key (the manually generated sentiments). This table only depicts the results for those tweets where the system recognised a Sentiment as present: it does not show the Missing annotations (where the system failed to recognise a valid Sentiment). The figures in bold indicate a correct match between System and Key. Overall, the system achieved a Precision of 62.2%, which is promising for work at this early stage.

Unfortunately, it was not feasible in this preliminary evaluation to manually annotate 1000 tweets, so we cannot calculate system Recall easily. However, we can extrapolate some hypothesised Recall based on a smaller sample. We took 150 of the tweets which were not classified as opinionated by the system, and annotated them manually. Of these, 127 (85%) were correct, i.e., were classified as Unknown or Irrelevant by the human annotator. Assuming that this sample

is representative, we can predict the Recall. For the task of finding whether a political sentiment exists or not (regardless of its orientation), we get 78% Precision and predict 47% Recall. Where a document was found to contain a political sentiment, the polarity of this sentiment was correct in 79% of cases. Overall, for the task of both correctly identifying that a document contained a political sentiment, and correctly identifying its polarity, we get 62% Precision and predict 37% Recall.

While the Recall of our system is clearly less than ideal, this is unsurprising at this stage because it has been developed with Precision rather than Recall in mind, i.e., only to produce a result if it is reasonably certain. As we have discussed earlier, there is plenty of scope for improvements to the NLP, in order to improve the Recall of the system. The Precision could also be tightened up further by improving the negation aspect of the rules (most of the errors are related either to not correctly identiying a negative, or by missing out on language nuances such as sarcasm, which are hard for an automated system to deal with). Further evaluation will focus on a larger number of tweets.

It is important also to recognise in the context of evaluation, that performing NLP tasks on social media is in general a harder task than on news texts, for example, because of the style and lack of correct punctuation, grammar etc. Shorter posts such as tweets suffer even more in this respect, and therefore performance of NLP is likely to be lower than for other kinds of text. Also, tweets in particular assume a high level of contextual and world knowledge by the reader, and this information can be very difficult to acquire automatically. For example, one tweet in our dataset likened a politician to Voldemort, a fictional character from the Harry Potter series of books. This kind of world knowledge is unlikely to be readily available in a knowledge base for such an application, and we may have to just accept that this kind of comment cannot be readily understood by auomatic means (unless we have sufficient examples of it occurring). It is also worth experimenting with machine learning techniques, although this also requires a fairly substantial set of manually annotated data. Previous experiments with supervised machine learning techniques on classifying opinions about businesses and transactions using a data-driven approach rather than relying on *a priori* information proved relatively successful [11], and we shall look to expand on this work in the future.

7 Conclusions

Typically, opinion mining looks at social media content to analyse people's explicit opinions about an organisation, product or service. However, this backwards-looking approach often aims primarily at dealing with problems, e.g., unflattering comments, while a forwards-looking approach aims at looking ahead to understanding potential new needs from consumers. This is achieved by trying to understand people's needs and interests in a more general way, e.g., drawing conclusions from their opinions about other products, services and interests. It

is not sufficient, therefore, just to look at specific comments in isolation: non-specific sentiment is also an important part of the overall picture.

One of the difficulties of drawing conclusions from traditional opinion mining techniques is the sparse data issue. Opinions about products and services tend to be based on one very specific thing, such as a particular model of camera or brand of washing powder, but do not necessarily hold for every other model of that brand of camera, or for every other product sold by the company, so a set of very isolated viewpoints is typically identified. The same applies, in some sense, to political viewpoints: a person may not like a particular politician even if they support the party represented by that person, overall. Furthermore, political opinions are often more suject to variations along a timeline than products and brands. A person who prefers Coke to Pepsi is unlikely to change their point of view suddenly one day, but there are many people whose political leanings change frequently, depending on the particular government, the politicians involved and events which may occur (if this were not the case, then of course the party in power in the UK would never change). Similarly, people's interests and opinions in general may change over the course of time, so an opinion mining system which investigates such things (rather than just products, films and so on) needs to take this into consideration. In order to overcome such issues, we need to be able to figure out which statements can be generalised to other models/products/issues, and which are specific. Another solution is to leverage sentiment analysis from more generic expressions of motivation, behaviour, emotions and so on, e.g., what type of person buys what kind of camera, what kind of person is a Labour supporter, and so on. To do this, we need to combine the kind of approach to opinion mining which we have described here, with additional information about people's likes, dislikes, interests, social groups and so on. Such techniques will form part of our future work.

As discussed earlier, there are many improvements which can be made to the opinion mining application in terms of making use of further linguistic and contextual clues: this work reports the development of this application as a first stage towards a more complete system, and also contextualises the work within a wider framework of social media monitoring which can lead to interesting new perspectives when combined with relevant research in related areas such as trust, archiving and digital libraries, amongst other things. In particular, the exploitation of Web 2.0 and the wisdom of crowds can make web archiving a more selective and meaning-based process. Analysis of social media can help archivists select material for inclusion, providing content appraisal via the social web, while social media mining itself can enrich archives, moving towards structured preservation around semantic categories.

Acknowledgements

This research is conducted as part of the EU FP7 project ARCOMEM[9].

[9] `http://www.arcomem.eu/`

References

1. Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval **12**(5) (2009) 526–558
2. Scharl, A., Weichselbraun, A.: An automated approach to investigating the online media coverage of US presidential elections. Journal of Information Technology and Politics **5**(1) (2008) 121–132
3. Mullaly, A., Gagné, C., Spalding, T., Marchak, K.: Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. The Mental Lexicon **5**(1) (2010) 87–114
4. Weichselbraun, A., Gindl, S., Scharl, A.: A context-dependent supervised learning approach to sentiment detection in large textual databases. Journal of Information and Data Management **1**(3) (2010) 329–342
5. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics **35**(3) (2009) 399–433
6. Gindl, S., Weichselbraun, A., Scharl, A.: Cross-domain contextualisation of sentiment lexicons. In: Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010). (2010) 771–776
7. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). (2002)
8. Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., Wilks, Y.: Architectural Elements of Language Engineering Robustness. Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data **8**(2/3) (2002) 257–274
9. Cunningham, H., Maynard, D., Tablan, V.: JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield (November 2000)
10. Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Miller, G.A., Beckwith, R., Felbaum, C., Gross, D., Miller, C.Minsky, M.: Five papers on WordNetklines: A theory of memory. (1980)
11. Funk, A., Li, Y., Saggion, H., Bontcheva, K., Leibold, C.: Opinion analysis for business intelligence applications. In: First international workshop on Ontology-Supported Business Intelligence (at ISWC), Karlsruhe, ACM (October 2008)

A new ANEW: Evaluation of a word list for sentiment analysis in microblogs

Finn Årup Nielsen

DTU Informatics, Technical University of Denmark, Lyngby, Denmark.
fn@imm.dtu.dk, http://www.imm.dtu.dk/~fn/

Abstract. Sentiment analysis of microblogs such as Twitter has recently gained a fair amount of attention. One of the simplest sentiment analysis approaches compares the words of a posting against a labeled word list, where each word has been scored for valence, — a "sentiment lexicon" or "affective word lists". There exist several affective word lists, e.g., ANEW (Affective Norms for English Words) developed before the advent of microblogging and sentiment analysis. I wanted to examine how well ANEW and other word lists performs for the detection of sentiment strength in microblog posts in comparison with a new word list specifically constructed for microblogs. I used manually labeled postings from Twitter scored for sentiment. Using a simple word matching I show that the new word list may perform better than ANEW, though not as good as the more elaborate approach found in SentiStrength.

1 Introduction

Sentiment analysis has become popular in recent years. Web services, such as socialmention.com, may even score microblog postings on Identi.ca and Twitter for sentiment in real-time. One approach to sentiment analysis starts with labeled texts and uses supervised machine learning trained on the labeled text data to classify the polarity of new texts [1]. Another approach creates a sentiment lexicon and scores the text based on some function that describes how the words and phrases of the text matches the lexicon. This approach is, e.g., at the core of the *SentiStrength* algorithm [2].

It is unclear how the best way is to build a sentiment lexicon. There exist several word lists labeled with emotional valence, e.g., ANEW [3], General Inquirer, OpinionFinder [4], SentiWordNet and WordNet-Affect as well as the word list included in the SentiStrength software [2]. These word lists differ by the words they include, e.g., some do not include strong obscene words and Internet slang acronyms, such as "WTF" and "LOL". The inclusion of such terms could be important for reaching good performance when working with short informal text found in Internet fora and microblogs. Word lists may also differ in whether the words are scored with sentiment strength or just positive/negative polarity.

I have begun to construct a new word list with sentiment strength and the inclusion of Internet slang and obscene words. Although we have used it for sentiment analysis on Twitter data [5] we have not yet validated it. Data sets with

manually labeled texts can evaluate the performance of the different sentiment analysis methods. Researchers increasingly use Amazon Mechanical Turk (AMT) for creating labeled language data, see, e.g., [6]. Here I take advantage of this approach.

2 Construction of word list

My new word list was initially set up in 2009 for tweets downloaded for online sentiment analysis in relation to the United Nation Climate Conference (COP15). Since then it has been extended. The version termed AFINN-96 distributed on the Internet[1] has 1468 different words, including a few phrases. The newest version has 2477 unique words, including 15 phrases that were not used for this study. As SentiStrength[2] it uses a scoring range from −5 (very negative) to +5 (very positive). For ease of labeling I only scored for valence, leaving out, e.g., subjectivity/objectivity, arousal and dominance. The words were scored manually by the author.

The word list initiated from a set of obscene words [7, 8] as well as a few positive words. It was gradually extended by examining Twitter postings collected for COP15 particularly the postings which scored high on sentiment using the list as it grew. I included words from the public domain *Original Balanced Affective Word List*[3] by Greg Siegle. Later I added Internet slang by browsing the Urban Dictionary[4] including acronyms such as WTF, LOL and ROFL. The most recent additions come from the large word list by Steven J. DeRose, *The Compass DeRose Guide to Emotion Words*.[5] The words of DeRose are categorized but not scored for valence with numerical values. Together with the DeRose words I browsed Wiktionary and the synonyms it provided to further enhance the list. In some cases I used Twitter to determine in which contexts the word appeared. I also used the Microsoft Web n-gram similarity Web service ("Clustering words based on context similarity"[6]) to discover relevant words. I do not distinguish between word categories so to avoid ambiguities I excluded words such as patient, firm, mean, power and frank. Words such as "surprise"—with high arousal but with variable sentiment—were not included in the word list.

Most of the positive words were labeled with +2 and most of the negative words with –2, see the histogram in Figure 1. I typically rated strong obscene words, e.g., as listed in [7], with either –4 or –5. The word list have a bias towards negative words (1598, corresponding to 65%) compared to positive words (878). A single phrase was labeled with valence 0. The bias corresponds closely to the bias found in the OpinionFinder sentiment lexicon (4911 (64%) negative and 2718 positive words).

[1] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=59819
[2] http://sentistrength.wlv.ac.uk/
[3] http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html
[4] http://www.urbandictionary.com
[5] http://www.derose.net/steve/resources/emotionwords/ewords.html
[6] http://web-ngram.research.microsoft.com/similarity/

I compared the score of each word with mean valence of ANEW. Figure 2 shows a scatter plot for this comparison yielding a Spearman's rank correlation on 0.81 when words are directly matched and including words only in the intersection of the two word lists. I also tried to match entries in ANEW and my word list by applying Porter word stemming (on both word lists) and WordNet lemmatization (on my word list) as implemented in NLTK [9]. The results did not change significantly.

When splitting the ANEW at valence 5 and my list at valence 0 I find a few discrepancies: aggressive, mischief, ennui, hard, silly, alert, mischiefs, noisy. Word stemming generates a few further discrepancies, e.g., alien/alienation, affection/affected, profit/profiteer.

Apart from ANEW I also examined General Inquirer and the OpinionFinder word lists. As these word lists report polarity I associated words with positive sentiment with the valence +1 and negative with –1. I furthermore obtained the sentiment strength from SentiStrength via its Web service[7] and converted its positive and negative sentiments to one single value by selecting the one with the numerical largest value and zeroing the sentiment if the positive and negative sentiment magnitudes were equal.

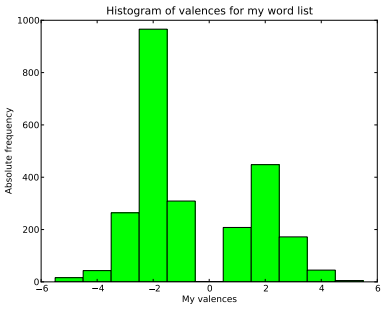


Fig. 1. Histogram of my valences.

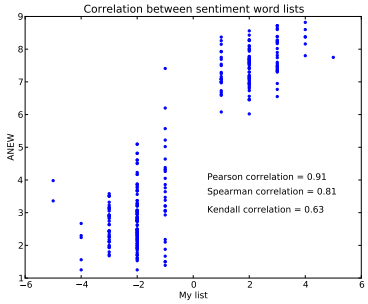


Fig. 2. Correlation between ANEW and my new word list.

3 Twitter data

For evaluating and comparing the word list with ANEW, General Inquirer, OpinionFinder and SentiStrength a data set of 1,000 tweets labeled with AMT was applied. These labeled tweets were collected by Alan Mislove for the *Twittermood*/"Pulse of a Nation"[8] study [10]. Each tweet was rated ten times to get a more reliable estimate of the human-perceived mood, and each rating was a sentiment strength with an integer between 1 (negative) and 9 (positive). The average over the ten values represented the canonical "ground truth" for this study. The tweets were not used during the construction of the word list.

To compute a sentiment score of a tweet I identified words and found the va-

[7] http://sentistrength.wlv.ac.uk/
[8] http://www.ccs.neu.edu/home/amislove/twittermood/

Table 1. Example tweet scoring. –5 has been subtracted from the original ANEW score. SentiStrength reported "positive strength 1 and negative strength –2".

Words:	ear	infection	making	it	impossible	2	sleep	headed	2	the	doctors	2	get	new	prescription	so	fucking	early	
My	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4	0	-4
ANEW	0	-3.34	0	0	0	0	2.2	0	0	0	0	0	0	0	0	0	0	0	-1.14
GI	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
OF	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
SS																			-2

lence for each word by lookup in the sentiment lexicons. The sum of the valences of the words divided by the number of words represented the combined sentiment strength for a tweet. I also tried a few other weighting schemes: The sum of valence without normalization of words, normalizing the sum with the number of words with non-zero valence, choosing the most extreme valence among the words and quantisizing the tweet valences to +1, 0 and –1. For ANEW I also applied a version with match using the NLTK WordNet lemmatizer.

4 Results

My word tokenization identified 15,768 words in total among the 1,000 tweets with 4,095 unique words. 422 of these 4,095 words hit my 2,477 word sized list, while the corresponding number for ANEW was 398 of its 1034 words. Of the 3392 words in General Inquirer I labeled with non-zero sentiment 358 were found in our Twitter corpus and for OpinionFinder this number was 562 from a total of 6442, see Table 1 for a scored example tweet.

I found my list to have a higher correlation (Pearson correlation: 0.564, Spearman's rank correlation: 0.596, see the scatter plot in Figure 3) with the labeling from the AMT than ANEW had (Pearson: 0.525, Spearman: 0.544). In my application of the General Inquirer word list it did not perform well having a considerable lower AMT correlation than my list and ANEW (Pearson: 0.374, Spearman: 0.422). OpinionFinder with its 90% larger lexicon performed better than General Inquirer but not as good as my list and

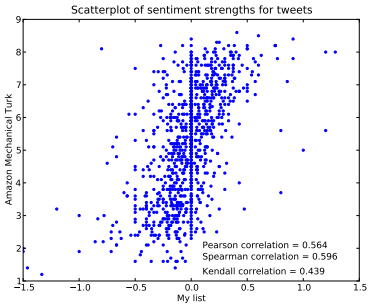


Fig. 3. Scatter plot of sentiment strengths for 1,000 tweets with AMT sentiment plotted against sentiment found by application or my word list.

	My	ANEW	GI	OF	SS
AMT	.564	.525	.374	.458	.610
My		.696	.525	.675	.604
ANEW			.592	.624	.546
GI				.705	.474
OF					.512

Table 2. Pearson correlations between sentiment strength detections methods on 1,000 tweets. AMT: Amazon Mechanical Turk, GI: General Inquirer, OF: OpinionFinder, SS: SentiStrength.

ANEW (Pearson: 0.458, Spearman: 0.491). The SentiStrength analyzer showed superior performance with a Pearson correlation on 0.610 and Spearman on 0.616, see Table 2.

I saw little effect of the different tweet sentiment scoring approaches: For ANEW 4 different Pearson correlations were in the range 0.522–0.526. For my list I observed correlations in the range 0.543–0.581 with the extreme scoring as the lowest and sum scoring without normalization the highest. With quantization of the tweet scores to +1, 0 and –1 the correlation only dropped to 0.548. For the Spearman correlation the sum scoring with normalization for the number of words appeared as the one with the highest value (0.596).

To examine whether the difference in performance between the application of ANEW and my list is due to a different lexicon or a different scoring I looked on the intersection between the two word lists. With a direct match this intersection consisted of 299 words. Building two new sentiment lexicons with these 299 words, one with the valences from my list, the other with valences from ANEW, and applying them on the Twitter data I found that the Pearson correlations were 0.49 and 0.52 to ANEW's advantage.

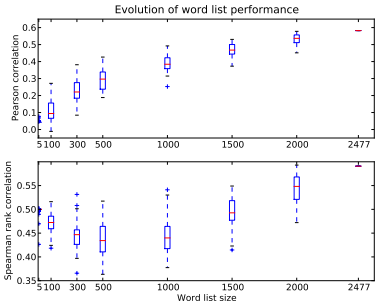


Fig. 4. Performance growth with word list extension from 5 words 2477 words. Upper panel: Pearson, lower: Spearman rank correlation, generated from 50 resamples among the 2477 words.

5 Discussion

On the simple word list approach for sentiment analysis I found my list performing slightly ahead of ANEW. However the more elaborate sentiment analysis in SentiStrength showed the overall best performance with a correlation to AMT labels on 0.610. This figure is close to the correlations reported in the evaluation of the SentiStrength algorithm on 1,041 MySpace comments (0.60 and 0.56) [2].

Even though General Inquirer and OpinionFinder have the largest word lists I found I could not make them perform as good as SentiStrength, my list and ANEW for sentiment strength detection in microblog posting. The two former lists both score words on polarity rather than strength and it could explain the difference in performance.

Is the difference between my list and ANEW due to better scoring or more words? The analysis of the intersection between the two word list indicated that the ANEW scoring is better. The slightly better performance of my list with the entire lexicon may be due to its inclusion of Internet slang and obscene words.

Newer methods, e.g., as implemented in SentiStrength, use a range of techniques: detection of negation, handling of emoticons and spelling variations [2]. The present application of my list used none of these approaches and might have

benefited. However, the SentiStrength evaluation showed that valence switching at negation and emoticon detection might not necessarily increase the performance of sentiment analyzers (Tables 4 and 5 in [2]).

The evolution of the performance (Figure 4) suggests that the addition of words to my list might still improve its performance slightly.

Although my list comes slightly ahead of ANEW in Twitter sentiment analysis, ANEW is still preferable for scientific psycholinguistic studies as the scoring has been validated across several persons. Also note that ANEW's standard deviation was not used in the scoring. It might have improved its performance.

Acknowledgment I am grateful to Alan Mislove and Sune Lehmann for providing the 1,000 tweets with the Amazon Mechanical Turk labels and to Steven J. DeRose and Greg Siegle for providing their word lists. Mislove, Lehmann and Daniela Balslev also provided input to the article. I thank the Danish Strategic Research Councils for generous support to the 'Responsible Business in the Blogosphere' project.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2) (2008) 1–135
2. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12) (2010) 2544–2558
3. Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999)
4. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics (2005)
5. Hansen, L.K., Arvidsson, A., Nielsen, F.Å., Colleoni, E., Etter, M.: Good friends, bad news — affect and virality in Twitter. Accepted for The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011) (2011)
6. Akkaya, C., Conrad, A., Wiebe, J., Mihalcea, R.: Amazon Mechanical Turk for subjectivity word sense disambiguation. In: Proceedings of the NAACL HLT 2010 Workshop on Creating, Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics (2010) 195–203
7. Baudhuin, E.S.: Obscene language and evaluative response: an empirical study. Psychological Reports **32** (1973)
8. Sapolsky, B.S., Shafer, D.M., Kaye, B.K.: Rating offensive words in three television program contexts. BEA 2008, Research Division (2008)
9. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly, Sebastopol, California (June 2009)
10. Biever, C.: Twitter mood maps reveal emotional states of America. The New Scientist **207**(2771) (July 2010) 14