Aligning Tweets with Events: Automation via Semantics

Editor(s): Alexandre Passant, DERI, NUI Galway, Ireland; Fabien Gandon, INRIA Sophia Antipolis, France; Harith Alani, KMI, The Open University, UK
Solicited review(s): Eraldo Fernandes, PUC-Rio, Brazil; David Laniado, Politecnico di Milano, Italy and Barcelona Media, Spain; one anonymous reviewer

Matthew Rowe [a,*], and Milan Stankovic [b]
[a] *Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*
E-mail: m.c.rowe@open.ac.uk
[b] *Hypios Research, 187 rue du Temple, 75003 Paris, France*
STIH, University Paris-Sorbonne, 28 rue de Serpente, 75006 Paris, France
E-mail: milan.stankovic@hypios.com

Abstract. Microblogging platforms, such as Twitter, now provide web users with an on-demand service to share and consume fragments of information. Such fragments often refer to real-world events (e.g., shows, conferences) and often refer to a particular event component (such as a particular talk), providing a bridge between the real and virtual worlds. The utility of tweets allows companies and organisations to quickly gauge feedback about their services, and provides event organisers with information describing how participants feel about their event. However, the scale of the Web, and the sheer number of Tweets which are published on an hourly basis, makes manually identifying event tweets difficult. In this paper we present an automated approach to align tweets with the events which they refer to. We aim to provide alignments on the sub-event level of granularity. We test two different machine learning-based techniques: proximity-based clustering and classification using Naive Bayes. We evaluate the performance of our approach using a dataset of tweets collected from the Extended Semantic Web Conference 2010. The best $F_{0.2}$ scores obtained in our experiments for proximity-based clustering and Naive Bayes were 0.544 and 0.728 respectively.

Keywords: Social Web, Semantic Web, Machine Learning, Twitter

1. Introduction

The microblogging service Twitter has become popular, among other purposes, as a general backchannel for commenting on events. Fashion shows, elections, conferences and professional events are very often the subject of discussions and commenting in the form of 140 character messages - tweets, exchanged in real-time during events. This practice becomes especially interesting for composite events (composed of several smaller events) such as conferences where participants use Twitter to comment on presentations, and discuss and ask questions during panels and question-answering sessions. Studies, such as the one presented in [6], show that tweets which are related to a conference mostly contain useful information about conference events and serve to exchange links related to support materials and mentioned systems and websites. Events or their sub-events cited in such microblogs are often done so implicitly, or with no formal declaration or link to the event - i.e., stating the title of a paper or commenting on a keynote. In parallel to the rise in usage of microblogs to share information, structured event data is also being increasingly published by services such as Semantic Web Dog Food[1] - which pub-

[*] Corresponding author. E-mail: m.c.rowe@open.ac.uk

[1] `http://data.semanticweb.org/`

lishes semantically rich structured data about Semantic Web events, and Live Matrix [2] - which crawls and organises many kinds of events found on the Web.

At present explicit links do not exist between tweets and the events which they refer to. Some big events have a Twitter hashtag associated with them that offers a way to associate a tweet with the corresponding main event, but often events are composed of several smaller events that remain unbound to their corresponding tweets. In the case of conferences, knowing that a tweet is about a particular conference, where several parallel presentations provoke independent reactions and discussions is insufficient, instead a more finely grained event description is required. By providing a clear and machine-readable relation between a given tweet and the event which it refers to, one would leverage implicit knowledge locked away within the post. In particular the mappings would give way to many advanced analytics, for instance taking the topics of events that a user has mentioned into account when deciding about his/her interests. In certain cases a user's tweets might contain insufficient topics to profile the user - given that the profile contains explicitly declared interests - but if a user's tweets point to events that have their own topics of interest, profiling could be enhanced. Similar utility can be found for providing feedback for composite events based on tweet streams in gaining access to tweet-to-event mappings, and identifying key points of discussion or popular events. Additionally tracking interest dynamics and topic popularity would also benefit from such mappings.

Motivated by the need to align tweets with the events which they refer to, in this paper we present an automated mapping solution. Our contributions in this paper are three-fold:

1. *Techniques to enrich tweets with metadata:* We present an approach to process a given corpus of tweets into a semantically rich and structured format using available ontologies, describing the publication of such metadata as linked data.
2. *An approach to automatically align tweets with events:* We present a machine learning-based approach to map tweets with events (particular events that are part of a larger composite event), testing two different techniques: proximity-based clustering inspired by K-Means and generative classification using Naive Bayes. We supply labelled data to our techniques using URIs from the Web of Linked Data, and test various feature sets from dereferencing those URIs.
3. *A benchmarking dataset for evaluation:* The evaluation of our approach uses a dataset of tweets collected from the Extended Semantic Web Conference 2010, a portion of which has been manually annotated for testing. We have placed this dataset online for the community to perform benchmarking experiments.

We have structured the paper as follows: section 2 presents an overview of our approach for enriching tweets with semantics and how they are passed on for alignment with events. Section 3 describes the tweet processing stage in more detail, describing the metadata enrichment techniques. Section 4 contains the central contribution of this paper detailing our approach for aligning tweets with events, the feature sets used and the different techniques tested. Section 5 presents the evaluation of our approach, describing the method for collecting the dataset, the evaluation measures used and the results and findings from our experiments. Section 6 describes applications of the aligned tweets and events in the context of the dataset used for experiments - i.e., the Extended Semantic Web Conference. Section 7 presents related work to our approach including existing work within the Semantic Web community using tweets and similar work within the field of reference reconciliation. Section 8 finishes the paper with the conclusions drawn from this work and our plans for future work.

2. Approach Overview

Twitter exposes data using a variety of formats including JSON and XML. Enriching tweets with semantics provides a common machine-readable format which can then be exposed as linked data, enabling the linkage of tweets with events in the Web of Data, and providing an enriched network of information in the process. To realise this end-goal we utilise the approach presented in Fig. 1 which is composed of four sequential stages: *first*, we collect microblog posts from Twitter and store these in a local repository, utilising the proprietary information representation format used by the platform. *Second*, we convert the collected tweets into triples by generating metadata to describe information using the Semantically Interlinked Online Communities (SIOC) [1] and Online Presence (OPO)

[2] http://livematrix.com/

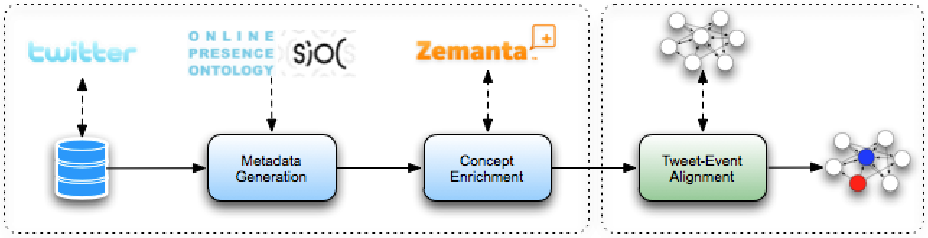

Fig. 1. Overview of the approach to a) process tweets into triples and enabling publication as Linked Data, and b) align of tweets with events in the Web of Data

```
@prefix opo:   <http://online-presence.net/opo/ns#>
@prefix dc:    <http://purl.org/dc/elements/1.1/>
@prefix sioct: <http://rdfs.org/sioc/types#>
<http://data.hypios.com/tweets/tweet-15162225891> rdf:type opo:OnlinePresence ;
  opo:customMessage <http://data.hypios.com/tweets/tweet-15162225891-cm> ;
  opo:declaredBy <http://data.hypios.com/tweets/user-ciro> ;
  opo:startTime "2010-06-01T09:16:46+0200" ;
  opo:publishedFrom <http://data.hypios.com/tweets/tweet-15162225891-source> .
<http://data.hypios.com/tweets/tweet-15162225891-cm> rdf:type sioct:MicroblogPost ;
  sioc:content "Noshir Contractor at #eswc2010 speaking of data-driven social network analysis of MMORPG.." ;
  sioc:id "15162225891" ;
  dcterms:language "en" ;
  foaf:maker <http://data.hypios.com/tweets/user-ciro> ;
  dcterms:date "2010-06-01T09:16:46+0200" ;
  dcterms:subject <http://dbpedia.org/resource/Social_network> .
<http://data.hypios.com/tweets/tweet-15162225891-source> rdf:type opo:SourceOfPublishing ;
  opo:sourceName "Twitter.com" .
```

Fig. 2. RDF/Turtle extract of a Tweet following metadata generation

[15] ontologies, thereby providing machine-readable metadata descriptions. *Third*, we enrich the tweets with DBPedia concepts by querying the web service Zemanta. This initial processing and enrichment enables the tweets to be published as linked data using dereferenceable URIs, and the provision of access to the tweets at a SPARQL end-point. *Fourth*, we align the tweets with the events which they refer to, where each event is represented as a URI on the Web of Data.

The final product of our approach is therefore a corpus of tweets which have been woven into the Web of Linked Data, allowing paths to be traversed from an event to those talking about it and perform analysis of what has been said. In a later section of this paper we describe several applications following this alignment and results from such applications using our alignment method. We now describe each section of our approach in greater detail, starting with the processing of tweets into triples before moving on to present our automated alignment technique.

3. Tweet Processing

Due to overload of tweets per second[3] and the scale of storing such information, messages on Twitter disappear from public searches following a week or less. Twitter archiving services (e.g., TweetBackup,[4] TwapperKeeper[5]) and desktop tools (e.g., Archivist,[6] Twinbox[7]) have emerged to resolve this problem by offering to save tweets for future searchers. In our processing of tweets we rely on archives created by the collaborative public archiving service TwapperKeeper. Data coming from this service is raw, is not connected with additional user metadata and is thus disconnected from potentially relevant topics. Therefore we transform it into a structured form to enable sophisticated and precise

[3] Recorded in February 2010 as being 600 per second according to http://blog.twitter.com/2010/02/measuring-tweets.html

[4] http://tweetbackup.com/

[5] http://twapperkeeper.com

[6] http://visitmix.com/labs/archivist-desktop/

[7] http://www.techhit.com/TwInbox/twitter_plugin_outlook.html

queries and analysis. We also link the tweet data with the metadata about the tweet author that is returned from querying the Twitter API.[8] Most importantly, we use Zemanta API to extract relevant topic concepts from the tweets. In this way we create a complete dataset about the tweets, that can be a solid ground for machine processing of different kinds.

3.1. Metadata Generation

We have built a Java-based parser that can process TwapperKeeper archives in comma separated values format. Once the tweets are imported, we return the user information and user account data (e.g., name, biography) from the Twitter API. At present the system is capable of using Jena and Talis triple stores. We have tried to make the most general representation of tweets possible in order to maximise the potential use of published data. Tweets are therefore represented as instances of *sioct:MicroblogPost* from the SIOC Types ontology,[9] and the maker/creator of a given tweet is expressed as an instance of *foaf:Person* - attributing the relevant person attributes to this instance (i.e., name, homepage, etc). We utilise the Online Presence Ontology (OPO) to define the act of publishing the tweet, and express this as an instance of *opo:OnlinePresence*, relating this instance to the source from which the information was published. General properties like titles are represented using the Dublin Core Terms vocabulary.[10] An example tweet, following metadata generation, is presented in Fig. 2. Apart from tweets, full descriptions of tweet authors and their twitter user accounts are also converted into triples, thereby providing additional contextual information. We mint URIs for the processed tweets, the authors, the content of the tweets and the source from which the information was published, enabling the posts to be looked up and the surrounding information graph traversed.

3.2. Concept Enrichment

Information contained within a microblog post often contains ambiguous or abbreviated terms which refer to distinct concepts. Furthermore, several tweets may contain distinct terms which refer to the same underlying concept. Querying over such data can render inconclusive results due to the lack of disambiguation performed over terms. We therefore perform lightweight concept enrichment by processing the text of tweets using the Zemanta[11] keyword extraction API. This approach returns DBPedia concepts related to a microblog post, which we then associated to a given tweet using the *dc:subject* property. This therefore enables all the tweets in a given semantic corpus to be returned which refer to a distinct topic.

4. Automatic Alignment of Tweets with Events

Following the processing of tweets into a machine-readable form we now wish to align the collection of microblog posts with the events which they refer to. By automatically aligning tweets we will enable the implicit semantics within such information snippets to be leveraged for reuse - we describe several applications of this leveraging within a later section of this paper. Our problem can be considered in an abstract sense as inferring a relation between an instance of a tweet (e.g., *sioct:MicroblogPost*) with the event (e.g., *swrc:InProceedings*[12]) that it cites. We use the Linking Open Descriptions of Events (LODE) ontology[13] to define this relation, given its ability to model descriptions of events, and we apply the *lode:illustrate* predicate to define the link between the tweet and the event, thus providing the necessary link into the Web of Data. Given the rise in both production and consumption of Linked Data, we now have a wide variety of machine-readable, usable data, containing richly typed semantics. Our task therefore, at a low-level, is one of reference reconciliation [3], where we must identify the link between a tweet and the URI of an event in the Web of Data. Fig. 3 presents an overview of our alignment approach pipeline.

Our solution is to consider this task as a machine-learning problem: consider the following scenario where we have a collection of tweets published during an event (e.g., the Extended Semantic Web Conference), we wish to identify sub-events (i.e., talks) which those tweets are referring to. Using the Semantic Dog Food service[14] we are able to derefer-

[8] http://apiwiki.twitter.com

[9] http://rdfs.org/sioc/types

[10] http://dublincore.org/documents/dcmi-terms/

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

[12] This namespace corresponds to the Semantic Web located at http://swrc.ontoware.org/ontology#

[13] http://linkedevents.org/ontology/

[14] http://data.semanticweb.org/

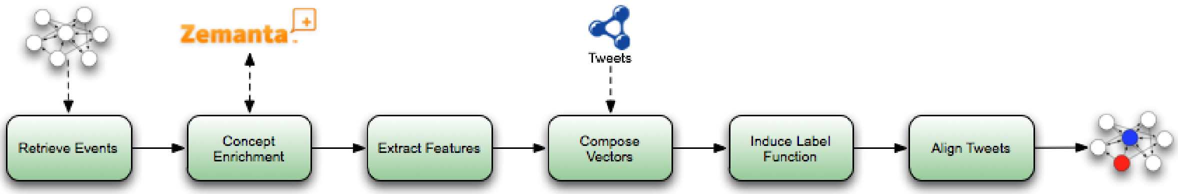

Fig. 3. Alignment Pipeline: the events are retrieved with which the tweets are to be aligned against, and the necessary features are extracted. Both the tweets and events are converted into feature vectors, where the latter are used to induce the labelling function for the tweets and the former are assigned the relevant URIs.

```
@prefix swrc: <http://swrc.ontoware.org/ontology#>
@prefix swc:  <http://data.semanticweb.org/ns/swc/ontology#>
@prefix dog:  <http://data.semanticweb.org>
@prefix dc:   <http://purl.org/dc/elements/1.1/>
<http://data.semanticweb.org/conference/eswc/2010/> rdf:type swrc:ConferenceEvent ;
  swrc:hasRelatedDocument <http://data.semanticweb.org/conference/eswc/2010/proceedings> ;
  swrc:isSuperEventOf <http://data.semanticweb.org/conference/eswc/2010/keynote/1> ;
  swrc:isSuperEventOf <http://data.semanticweb.org/workshop/apresw/2010> ;
  swrc:isSuperEventOf <http://data.semanticweb.org/workshop/irmles/2010> ;
  swrc:isSuperEventOf <http://data.semanticweb.org/workshop/nefors/2010> ;
  swrc:isSuperEventOf <http://data.semanticweb.org/workshop/ores/2010> ;
<http://data.semanticweb.org/conference/eswc/2010/proceedings> rdf:type src:Proceedings ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/inuse/13> ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/mobility/10 ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/phd_symposium/23> ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/onto/56> ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/social_web/1> ;
  swc:hasPart <http://data.semanticweb.org/conference/eswc/2010/paper/web_of_data/44> .
```

Fig. 4. RDF/Turtle extract of sub-event URIs surrounding the Extended Semantic Web Conference URI on Semantic Dog Food

ence the URI of the conference,[15] and use a *follow-your-nose* strategy to traverse paths within the Web of Data to the URIs of sub-events. Fig. 4 shows an extract from the surrounding graph of the Extended Semantic Web Conference's URI, from this data structure we are able to identify the sub-events which took place at the conference, where each sub-event is using a URI. The URIs of these sub-events together with the parent URI therefore provide the set of class labels Y, and the inherent features of these events form our *labelled* instances ($\{(x_i, y_i)\}_{i=1}^{L}$), while our *unlabelled* instances are composed from the collection of tweets: ($\{(x_i)\}_{i=1}^{U}$). Our goal is to induce some function which assigns the most appropriate event label to a tweet: $f : X \rightarrow Y$.

4.1. Feature Extraction

Various techniques exist for inducing the labelling function, within this paper we investigate two such techniques, comparing the performance levels of a sample clustering method based on proximity measures and the Naive Bayes classifier. Before moving on to present these techniques we must first decide on the features used by each method to induce its labelling function. To provide a range of features and explore their effects on the alignment process we define three distinct feature sets as follows:

F1: Immediate Resource Leaves Our natural intuition is to utilise what we know about each event to induce the labelling function, such that the features which are most indicative of an event can be detected within a tweet and the appropriate label assigned. Our first feature set covers this intuition by using only the literals and resources which exist within the instance description of the event URI as features. To extract such features, given an event resource (r) we extract the *Resource Leaves* surrounding the event from the Linked Data graph (G) using the following construct:

$$
\begin{aligned}
RLS_G(r) &= < r, p, o > \mid < r, p, o > \in G \\
&\wedge \nexists p', o' < o, p', o' > \in G \qquad (1)
\end{aligned}
$$

[15] e.g., Extended Semantic Web Conference uses the URI: `http://data.semanticweb.org/conference/eswc/2010/proceedings`

```
@prefix swrc: <http://swrc.ontoware.org/ontology#>
@prefix swc:  <http://data.semanticweb.org/ns/swc/ontology#>
@prefix dog:  <http://data.semanticweb.org>
@prefix dc:   <http://purl.org/dc/elements/1.1/>
<http://data.semanticweb.org/conference/eswc/2010/paper/phd_symposium/23> rdf:type swrc:InProceedings ;
  dc:subject "Knowledge Acquisition" ;
  dc:subject "Semantic Analysis" ;
  dc:subject "Social Web" ;
  dc:subject "Microblogs" ;
  dc:title "Exploring the Wisdom of the Tweets:  Knowledge Acquisition from Social Awareness Streams" ;
  swrc:abstract "Although one might argue that little wisdom can be conveyed in messages of 140 ..." ;
  swrc:author <http://data.semanticweb.org/person/claudia-wagner> .
<http://data.semanticweb.org/person/claudia-wagner> rdf:type foaf:Person ;
  foaf:name "Claudia Wagner" ;
  swrc:affiliation <http://data.semanticweb.org/organization/joanneum-research> ;
  foaf:based_near <http://dbpedia.org/resource/Austria>
```

Fig. 5. RDF/Turtle extract of a given sub-event (paper) and the associated author

This construct functions by extracting the surrounding triples of a given resource where the objects of those triples are not the subjects of other triples. Consider the example in Fig. 5 where the instance description of a paper is shown. Applying the above construct to the resource would extract all of the immediate attributes of the paper while ignoring the author details - given that this information is provided by traversing to another resource away from the event - producing the triples shown in Fig. 6.

F2: 1-Step Resource Leaves The second feature set utilises the graph structure of the Web of Data to gather information which is one step away from the event. The intuition behind this method is that information which is not directly stored within the event description provides a wider context of features - given that the feature scope will enlarge as links are traversed and the instance descriptions retrieved. For this we gather a collection of resources which are 1-step away from the event URI in the Web of Data and return this set of URIs. For each URI we use the same construct as from Equation (1), by dereferencing the URI and retrieving the literals and resources within the returned instance description.

To demonstrate the feature set provided using this method, consider the example shown in Fig. 5: the URI of the paper is *looked-up* and the URIs which are 1-step away in the data graph are retrieved - in this case it is the URI of the paper author. The returned URI is then dereferenced and the Resource Leaves are extracted from the description. This produces the triples shown in Fig. 7 containing the author's name as a literal. Using this method uses information which is further away in the Linked Data graph space from the resource.

```
http://dbpedia.org/resource/Social_consciousness
http://dbpedia.org/resource/Knowledge_acquisition
http://dbpedia.org/resource/Stream
http://dbpedia.org/resource/Doctor_of_Philosophy
```

Fig. 8. URIs of DBPedia concepts returned using the Resource Leaves as input to Zemanta

F3: DBPedia Concepts The third feature set used in our approach uses DBPedia concepts describing the events. To generate these concepts we employ the same approach as previous for the concept enrichment of tweets by querying the Zemanta API using the Resource Leaves of an event. This returns the concepts as dereferenceable URIs which are then used as features for that event. Our intuition is that this feature set will provide comparable features with the tweets following concept enrichment. Furthermore ambiguity of terms within the event descriptions and tweets could be avoided given the utility of the wider graph and the normalisation of the event to concepts, thus impacting upon accuracy levels and improving alignment. As an example Fig. 8 shows the URIs returned when querying Zemanta using the Resource Leaves from Fig. 6 returned when dereferencing the event URI.

4.2. Feature Vector Composition

Our alignment technique uses a bag-of-words approach to represent *unigram* features of both events and tweets. For tweets we compose this bag-of-words by taking each tweet's content and the enriched concepts and removing any stop words from the bag. We also normalise everything to lowercase. For our events we take the returned features and use the same process as above: removing stop words and returning a collection of unigrams. This method allows hashtags to be incorporated into the decision process as features,

```
<http://data.semanticweb.org/conference/eswc/2010/paper/phd_symposium/23> rdf:type swrc:InProceedings ;
  dc:subject "Knowledge Acquisition" ;
  dc:subject "Semantic Analysis" ;
  dc:subject "Social Web" ;
  dc:subject "Microblogs" ;
  dc:title "Exploring the Wisdom of the Tweets:  Knowledge Acquisition from Social Awareness Streams" ;
  swrc:abstract "Although one might argue that little wisdom can be conveyed in messages of 140 ..." .
```

Fig. 6. RDF/Turtle extract of a given sub-event (paper) using the Immediate Resource Leaves

```
<http://data.semanticweb.org/person/claudia-wagner> rdf:type foaf:Person ;
  foaf:name "Claudia Wagner" ;
```

Fig. 7. RDF/Turtle extract of a given sub-event (paper) using the 1-step resource leaves

however the utilisation of such unigrams is only possible if the training instances contain such features. In our later experiments we found that the resources on the Web of Linked Data that corresponded to workshops provided the acronym of the workshop, where such an abbreviation was used as the hashtag for citing the sub-event on Twitter.

It is worth noting that in certain cases we return several bags for a single class label depending on the event and feature set used. For instance in the case of a paper with many authors (e.g., 6 people) and the use of only F2 as the feature set we will have 6 bags for that single event. Conversely, if we use F1 as the feature set for the event we will have only one bag. Our evaluation not only tests the use of one single feature set, but several different feature sets combined. Our intuition is therefore that the use of a single feature set will result in poor performance, while a combination of feature sets will increase accuracy when performing alignment.

From our bag-of-words representations for both our events and tweets we must map this representation into a statistical model - this enables our labelling function to be induced. For this we compose feature vectors for each instance $x \in X$ such that x_i represents a unique feature from the input vector. Our mapping process involves using a binary indexer such that each input instance has its features compared against the indexer and the relevant indexes returned. If only F3 is used for features of the events then the dimensionality of the indexer will be low - given the limited number of DBPedia concepts that will be returned - in comparison with the combination of F1+F2+F3 which will result in higher dimension feature vectors.

4.3. Inducing the Labelling Function

At this stage in our approach we have a consistent feature vector form for both the tweets and events. From this form we wish to induce some labelling function to derive class labels for our tweets: $f : X \to Y$. To do this we tested 2 different approaches, each differing in their functionality and method of inducing f: a proximity-based clustering approach, similar to K-Means, and a classification-based approach using a Naive Bayes classifier. We now describe these methods in terms of our alignment task.

4.3.1. Proximity-based Clustering

The first approach to induce our labelling function f for labelling tweets is the use of the proximity-based clustering algorithm, similar to K-Means [5]. K-Means functions to divide a given set of input vectors (X) into (k) clusters or classes. A common problem when applying this method is approximating k should the number of distinct clusters not already be known. However in our context we already have a predefined number of events against which we wish to align tweets and can therefore apply this approach - where k corresponds to the number of event URIs coupled. K-Means is applied in an unsupervised setting using a two-step process of *assignment* and *update*. The algorithm is initialised with k random vectors (means), each data point is then assigned to the nearest mean by minimising its dissimilarity/distance from itself to a given mean. Once all data points have been assigned then each of the k means are recalculated (*updated*) and the process repeated until convergence.

In our case we are applying K-Means in a slightly different way. We build the feature space in the same manner, and construct the k means for each of the k events which act as class labels. We then take our collection of tweets (X) and return the class where the distance between the event mean and the tweet vector is minimised. We define this formally as:

$$y = \underset{y \in Y}{\operatorname{argmin}}\; d(x, \mu_y) \tag{2}$$

Where μ_y is the mean (centroid) of the event within the feature space and $d(x, \mu_y)$ is a distance function between the two vectors. In essence this distance function measures the dissimilarity between the two vectors, therefore the set which produces the minimal distance between its centroid and the tweet is chosen as the event label. The selection of different feature sets will impact on the utility of this method, where in cases where multiple instances are provided for a single event (i.e., in the case of F2) then the class distribution will be greater than for cases where only a single instance is supplied - given that in the latter instance μ will resolve to this vector.

Given our generic notion of the above distance measure we implement and contrast two distance metrics: Manhattan distance and Euclidean distance. The Manhattan distance, defined formally in Equation (3), is a non-euclidean geometric measure between two points within a feature space. It is often referred to as the "*taxicab distance*" as it measures the distance between two points, not along a straight line, but via single unit increments along axis steps in the feature space. In contrast the Euclidean distance, defined in Equation (4), measures the direct distance between two points within the feature space.

$$manhattan(x, \mu) = \sum_{i=1}^{n} |x_i - \mu_i| \tag{3}$$

$$euclidean(x, \mu) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2} \tag{4}$$

4.3.2. *Naive Bayes*

To contrast the performance of a proximity-based clustering method against a multiclass classification technique we implemented the Naive Bayes classifier. This classifier constructs a generative model for classification using a probabilistic model learnt from the event features. Consider an unlabelled instance $x \in X$, such as a tweet, which is to be classified. Given that the tweet is represented as a feature vector - $\{x_1, x_2, ..x_n\} \in x$ - Naive Bayes tries to assign the most probable class label (event URI: $y \in Y$) to the tweet based on these features. Therefore the class label y for x is the most likely event given the known features of x: $y = \underset{y \in Y}{\text{argmax}} P(y|x_1, x_2, ..., x_n)$. Using Bayes theorem we can write this as:

$$y = \underset{y \in Y}{\text{argmax}} \frac{P(x_1, x_2, ..., x_n|y)P(y)}{P(x_1, x_2, ..., x_n)} \tag{5}$$

$$y = \underset{y \in Y}{\text{argmax}} P(x_1, x_2, ..., x_n|v_j)P(y) \tag{6}$$

Naive Bayes uses the assumption of variable independence, where the probability of $P(x_1, x_2, ..., x_n|y)$ is derived from the product of the probability of each feature in the tweet x given the class (event) y as follows:

$$y = \underset{y \in Y}{\text{argmax}} P(y) \prod_{i} P(x_i|y) \tag{7}$$

The probabilities in Equation (7) are built from the event data by using frequency distributions over the observations: $P(y)$ is built from the number of times this event has been observed in the labelled data. $P(x_i|y)$ is built from the number of times that the given feature (unigram) x_i has been observed when associated with that class (event). For example this could be the probability that the term *skos* appears in a given keynote speech description. An ideal scenario would be to build these probabilities from tweets which have been identified as citing an event, however this is not realistic given the expensive process of labelling such data - we performed such labelling for evaluating our methods, a discussion of which follows this section - therefore we build these frequency distributions from the collection of events that we have retrieved from the Web of Data, constructing our conditional probabilities from those features appearing given the event. This is, in essence, one of the novel contributions of our work in that we rely on freely accessible information for labelled data, requiring no annotation or manual labelling.

5. Experiments

To evaluate the success of our approach for aligning tweets with events we performed a set of experiments using a corpus of tweets. In this section we describe this dataset, construction of a gold standard used for evaluating our method and the evaluation metrics used to measure the performance of our approach. We then present the results from our experiments.

5.1. Dataset

We collected a dataset of Tweets posted during the Extended Semantic Web Conference in May 2010. For collection we used the Twapper Keeper service for logging any Tweets tagged with "*#eswc*" and the accompanying provenance information: e.g., author details, date/time. Once collected, we then processed the corpus into triples, expressing metadata using the previously described ontologies, and enriched tweets with the DBPedia concepts they refer to. Due to the bespoke topics of the conference, many tweets which described distinct events shared the same concepts, therefore when using only F3 as the feature set we achieved poor performance due to the lack of differentiating features between events - we discuss this further in the following section. Our produced dataset contained 1082 tweets, and following concept enrichment 213 tweets were associated with at least one DBPedia concept.

5.2. Gold Standard Construction

In order to gauge the accuracy of our method we constructed a gold standard from a random sample of 200 tweets from the corpus. For constructing this gold standard 3 raters were set the task of manually labelling this sample with the events that the tweets referred to, if any. From these ratings the majority opinion would then be taken to produce the event labels for the tweets. We used the κ-statistic [4] to judge the level of agreement between raters, where a higher value ($0 \leq \kappa \leq 1$) indicates a greater level of agreement between the raters. The κ-statistic is calculated as follows, using the set definitions from Table 1:

$$\kappa = \frac{2(ad - bc)}{(a+c)(c+d) + (b+d)(d+b)} \tag{8}$$

Matching agreement between raters however does not render a binary outcome - i.e., 2 raters can each label a tweet with different event URIs, rather than one rater labelling the tweet with an event and the other rater not doing so - therefore we modify the set definitions in Table 1 so that set b includes the outcome where both raters label the tweet with a URI yet the URI differs. Following the initial rating we identified a low-level of agreement between the 3 raters, where $\kappa = 0.328$, such that using the agreed labels would have rendered evaluation inaccurate. Therefore we used the Delphi method [11] to perform a second round of ratings where each human rater was able to see the event labels assigned by other raters and allow their alignment to be rectified - i.e., identifying that they were incorrect previously and amend their decision accordingly. Following the second round of ratings the level of agreement increased significantly to $\kappa = 0.820$, indicating a large level of agreement between the raters - sufficient for gold standard construction. Therefore we used the labels from this round of ratings as the gold standard for evaluation of the sample 200 tweets.

Table 1

Confusion matrix for calculating interrater agreement

		Rater 1	
		Positive	Negative
Rater 2	Positive	a	b
	Negative	c	d

5.3. Evaluation Measures

To provide binary cases we evaluate our method at the event level, determining the performance levels achieved by each technique as an average of the evaluation measures for all events. For evaluation we use commonly applied metrics from information retrieval, thus requiring the definition of two document sets initially as follows: for each event evaluated, let A be the set of relevant tweets which refer to the event and B be the set of retrieved tweets, following these definitions precision is defined as in Equation 9 and recall is defined as in Equation 10. Precision measures the proportion of event tweets that were labelled correctly, and recall measures the proportion of relevant tweets that were successfully retrieved for an event.

$$P = \frac{|A \cap B|}{|B|} \tag{9}$$

$$R = \frac{|A \cap B|}{|A|} \tag{10}$$

To provide a unary measure of performance we also use f-measure as described in Equation (11) with a suitable setting for β. In our approach we evaluate our methods using three settings of f-measure: $F_{0.2}$, $F_{0.5}$ and F_1. The first weighs precision as being 5 times more important than recall, in the second we weight precision as being twice as important as recall and in the third permutation we weigh precision and recall as being of equal importance. Our preference for precision over recall is due to the context of application - which we discuss in detail in the following section - given that we must map tweets with events but ensure that we minimise false positives (incorrect tweets) when doing so. In cases where this occurs the affects of tweets labelled with the incorrect event will impact on applications which utilise such data. It is worth noting that F-measure values are calculated on the micro-evaluation level (for each event evaluated), therefore the following results that we present represent mean values of respective F-measures for all micro-evaluations - this enables statistical significance to be tested using the Sign Test [10].

$$F_\beta = \frac{(1+\beta^2) * P * R}{\beta^2 * P + R} \tag{11}$$

5.4. Experimental Setup

Our approach requires no tuning, simply the provision of the event URIs which act as class labels, thus demonstrating the utility of our methods and the lack of manually labelled data required for training - although we still regard our approach as a *supervised* method. For each of our alignment methods experiments were set up as follows: for proximity-based clustering no set up or tuning was required, the centroid vectors of each set were computed using the feature vectors for each event. We test the differences in performance between the two distance measures when clustering the tweets with the event centroids, and abbreviate these measures to PBC_{man} and PBC_{eucl} for Manhattan distance and Euclidean distance respectively. For Naive Bayes we trained the classifier for a multi-class problem using the provided event feature vectors and their associated labels. For each of the tested methods we ran each technique using each of the previously described feature sets, the differences in performance of which are now discussed.

5.5. Results

5.5.1. Proximity-based Clustering

The results from our experiments are shown in Table 2. The findings indicate that for all but one of the tested feature sets, the use of Euclidean distance outperforms Manhattan distance and yields higher levels of both precision and recall. In the case of feature set 1 (F1), Euclidean distance and Manhattan distance yield the same performance levels. This is due to the computation of the centroid vector being derived from only a single feature vector - i.e., the resource leaves surrounding the URI of the event. The results show that as the number of feature vectors for each event grows the ability of Manhattan distance to accurately cluster tweets with events reduces. Conversely Euclidean distance maintains higher levels of performance, particularly as the feature scope expands.

Feature Set 3 (F3) is omitted from the results in Table 2 due to its solitary use yielding poor, and in some cases no, results. The intuition behind our feature sets is such that as they are combined together accuracy levels increase. For proximity-based clustering this is evident in our findings, where $F_{0.2}$ levels are found to be significantly different - using the Sign Test ($p < 0.01$) - between F1 and F1+F2. An interesting characteristic however is the deterioration in performance when using all of the feature sets combined with this alignment method. For both of the tested distance measures we observe a significant reduction in performance levels when comparing F1+F3 with F1+F2+F3. This is due to the overlap of features between events which are created when using the second feature set. In this instance the feature vectors contain fewer discriminatory features, for instance a common feature vector produced using the 1-step Resource Leaves is composed from dereferencing the URI of the conference. This is consistent across all of the event URIs, given that each of them are associated with the conference in which the event occurs - using the *swrc:isPartOf* relation. As a result, this lack of discrimination between the events makes clustering limited.

5.5.2. Naive Bayes

The findings in Table 2 show that the Naive Bayes classifier consistently outperforms PBC for all of the tested feature sets. The generative model learnt by the classifier is also seen to improve in performance as the feature set is increased. Comparing all f-measure levels tested for F1 in comparison with all the feature sets shows a significant improvement as additional features are used - this increase is found to be statistically different using the Sign Test ($p < 0.01$). In particular, we observe that, like PBC, the use of F1+F3 outperforms F2+F3 for all tested evaluation measures. The inclusion of the immediate properties of the event - i.e., paper title - when combined with DBPedia concepts

Table 2

Accuracy levels of alignments using Proximity-based Clustering and Naive Bayes

		P	R	$F_{0.2}$	$F_{0.5}$	F_1
F1	PBC_{man}	0.389	0.311	0.365	0.317	0.284
	PBC_{eucl}	0.389	0.311	0.365	0.317	0.283
	Naive Bayes	**0.593**	**0.563**	**0.573**	**0.536**	**0.520**
F2	PBC_{man}	0.030	0.018	0.030	0.027	0.022
	PBC_{eucl}	0.088	0.106	0.088	0.090	0.093
	Naive Bayes	**0.288**	**0.239**	**0.278**	**0.261**	**0.249**
F1+F2	PBC_{man}	0.141	0.042	0.121	0.085	0.061
	PBC_{eucl}	0.539	0.437	0.507	0.457	0.429
	Naive Bayes	**0.732**	**0.722**	**0.721**	**0.701**	**0.689**
F2+F3	PBC_{man}	0.092	0.058	0.080	0.058	0.042
	PBC_{eucl}	0.092	0.101	0.087	0.076	0.070
	Naive Bayes	**0.323**	**0.269**	**0.313**	**0.293**	**0.279**
F1+F3	PBC_{man}	0.349	0.283	0.332	0.293	0.262
	PBC_{eucl}	0.562	0.556	0.544	0.513	0.501
	Naive Bayes	**0.591**	**0.566**	**0.574**	**0.540**	**0.522**
All	PBC_{man}	0.143	0.115	0.129	0.104	0.093
	PBC_{eucl}	0.421	0.416	0.415	0.399	0.387
	Naive Bayes	**0.738**	**0.732**	**0.728**	**0.709**	**0.698**

derived from this information achieves superior performance than the use of features derived from 1-step away in the Web of Data. However, the inclusion of F2 with F1+F3 shows improvement and achieves the optimum feature set combination for aligning tweets with events - the difference in performance is found to be statistically significant using a more liberal significance level ($p < 0.05$).

It is worth noting that the Naive Bayes classifier was not the solitary classifier that we tested in our experiments. We also evaluated the performance of Support Vector Machines in a multi-class classification setting by testing a one-vs-all and a one-vs-one with voting strategy. In each case, and when testing these permutations with different kernels, we yielded poor results, where no classifications were made or all were erroneous. We believe that this is due to the lack of discriminatory features between the events that we used to induce our labelling function ($f : X \rightarrow Y$) from. For instance, the majority of events will contain the unigrams *semantic* and *web* in their feature list, given the nature of the conference, thus rendering the observations of differences between classes (events) limited. The labelling function induced using Naives Bayes attempts to learn a complete, or rather general, model of features from which the most likely event label can be returned. The utility of such an approach is evident in our alignment task, given the high levels of $F_{0.2}$ that we achieve with respect to the alternative clustering mechanism and the failed discriminative models (i.e., SVMs).

6. Applications

In this section we explain how the alignment of tweets with events can be put to use. As mentioned before, these mappings provide connections into the existing Web of Data, and thus gain additional knowledge about current user activities. Tweets mapped to events provide access to richer information and allow concept extraction tools to produce lists of relevant topic concepts. This in turn allows user profiles to be extended with additional interests which are not explicitly defined by the user.

Apart from giving access to more information than the tweets contain, the mappings to events also introduce a dynamic layer over these events by providing insights into the level of activity taking place around a particular event. This can be useful for providing feedback to the conference organiser and for the presenters themselves. In addition to the quantitative dimension of this dynamic layer (e.g., the number of tweets that follow a talk), some qualitative dimensions, like sentiment, are also interesting to explore. Apart from studying the dynamics that arises around the events, such mappings would allow an explanation of the dynamics that arise around the topics of discussion, including the sentiment towards given topics, popularity tendencies and other similar metrics.

6.1. Improving User Interest Profiles

One of the potential use-cases where tweet-event mappings are useful is the creation of user interest profiles. Firstly we assume that a user who tweeted about some concept, or who attended a event defined by a concept, might have an interest in that subject. In particular we will focus on improving user profiles, based on the assumption that the tweet-event links give access to more information about users' attention and activities during a conference, since event descriptions might be richer in topics than tweets, which are well-known for being low in information content. We thus rely on our automatically generated mappings using the Naive Bayes classifier to propagate topics from the events that a given user tweeted about to the user as his/her interest.

For comparison we contrast the interest list constructed in such a way against the list of topics that can be found directly in the user's tweets (we refer to topic concepts extracted by Zemanta). For each user for whom we had generated mappings, we calculated the lists of interest concepts based on (1) concepts extracted from the content of tweets, and (2) concepts extracted from the event descriptions that the tweets were mapped to. We sent the custom generated concept lists to the 20 most active twitter users on the ESWC2010 conference, and obtained feedback from 6 of them. From this feedback we observed that the method of using mappings to identify interests always gave a larger number of interest concepts, and users found more of their true interests in those lists. For the purposes of example, Table 3 and Table 4 show the lists of interest concepts generated for the Twitter user Claudia Wagner (@clauwa on Twitter).

Table 3

Interest concepts obtained from the text of tweets

http://dbpedia.org/resource/Semantic_Web
http://dbpedia.org/resource/Resource_Description_Framework

Table 4

Interest concepts obtained from the event descriptions that @clauwa's tweets were mapped to

http://dbpedia.org/resource/Ontology_(information_science)
http://dbpedia.org/resource/Semantic_Web
http://dbpedia.org/resource/Data
http://dbpedia.org/resource/Unsupervised_learning
http://dbpedia.org/resource/Tag_cloud
http://dbpedia.org/resource/Semantics
http://dbpedia.org/resource/Rule_of_inference
http://dbpedia.org/resource/Resource_Description_Framework
http://dbpedia.org/resource/Research
http://dbpedia.org/resource/Proprietary_format
http://dbpedia.org/resource/Logical_schema
http://dbpedia.org/resource/Gold_standard
http://dbpedia.org/resource/Folksonomy
http://dbpedia.org/resource/Flickr
http://dbpedia.org/resource/Biology

6.2. *Providing Conference Feedback*

In a similar manner to improving user interest profiles, the mappings help in grasping popular topics and popular events. The intuition behind this use being the same: events provide more content and topics then merely the content of tweets. Based on event popularity detected in our dataset, we found the most popular paper, reflecting *vox populi*, was "*LESS - Template-based Syndication and Presentation of Linked Data for End-users*" by Soren Auer, Raphael Doehring, and Sebastian Dietzold. Similar analysis found the most popular workshop to be *Linking of User Profiles and Applications in the Social Semantic Web (LUPAS) 2010*[16] whose tweets outnumbered the first runner-up 6-fold, thus indicating the popularity amongst conference participants of this workshop.

Table 5

Top Interests that appear in papers (derived from mapped tweets)

Concept	Count
http://dbpedia.org/resource/Data	149
http://dbpedia.org/resource/Semantic_Web	127
http://dbpedia.org/resource/Semantics	113
http://dbpedia.org/resource/SPARQL	68
http://dbpedia.org/resource/Resource_Description_Framework	65
http://dbpedia.org/resource/Technology	54
http://dbpedia.org/resource/Ontology_(information_science)	52
http://dbpedia.org/resource/Application_software	49
http://dbpedia.org/resource/Web_2.0	45
http://dbpedia.org/resource/Text-based_(computing)	44

Table 6

TOP Interests that appear in Tweets (derived from mapped events)

Concept	Count
http://dbpedia.org/resource/Semantic_Web	30
http://dbpedia.org/resource/Linked_Data	24
http://dbpedia.org/resource/SPARQL	20
http://dbpedia.org/resource/Resource_Description_Framework	11
http://dbpedia.org/resource/Social_network	9
http://dbpedia.org/resource/Uniform_Resource_Identifier	8
http://dbpedia.org/resource/Radio-frequency_identification	7
http://dbpedia.org/resource/Hypertext_Transfer_Protocol	6
http://dbpedia.org/resource/Twitter	4
http://dbpedia.org/resource/Level_of_detail	4
http://dbpedia.org/resource/Business_model	4
http://dbpedia.org/resource/Business	4

In addition, differences between topics of popular talks and topics popular in tweets opens the prospect for various additional analyses. For instance we could say that topics that appear in tweets, but are not covered in talks, could indicate promising content for future conferences. *Social Networks* is one of such topics, as well as *Business* and *Business model*. Those are indeed topics that target the issues related to the Semantic Web as it matured and a business model for it was sought. Those topics might be of potential interest for inclusion in future conference agendas. On the other hand Radio Frequency Identification (RFID) is a also a topic that appears in the tweets but not in events. This is due to an RFID-based experiment that was conducted during the conference, however it could be misleading to think that its popularity indicates a demand for such a subject in the conference agenda.

[16] http://www.personal-reader.de/lupas/

One could imagine how topics that appear in tweets related to a particular event will give additional information about the event itself. Such topics, as well as links shared in the tweets, might provide valuable insight to the event hoster as well as to those observing the conference from the Social Web sphere. For instance, the tweets related to the paper presentation "*An Unsupervised Approach for Acquiring Ontologies and RDF Data from Online Life Science Databases*" given by Mohammad S. Mir, Steffen Staab, and Isabel Rojas, point to the topic `http://dbpedia.org/resource/ScienceDirect`, that does not appear in the paper abstract, but it might be something that was evoked during the actual presentation.

7. Related Work

The explicit linkage of media resources - i.e., videos, photos, microblog posts, etc - with the events which they refer to has been presented in work by [16], using an ontological model to formalise such relations. The approach in [16] extracts media resources from Social Web platforms using the previously labelled associations between media resources and events on the platforms. The *lode:illustrate* property from the Linking Data with Events ontology, described in [16], is used within our work to associate a given event (URI of a keynote/talk in our case) with the media resource (i.e., a tweet in our case) that refers to it.

As Twitter now provides an on-demand service for up-to-the-minute information about real-world events, work by [14] analysed the relationship between television debate performance in the run-up to the 2008 US presidential election and the usage of twitter during the events. The findings from this body of work demonstrate the correlation between peaks and troughs in debate performance - characterised by key points won and lost during the debate - and the comments which were published on Twitter at the same time. Attempting to correlate real-world events with information on the Web, and in particular Twitter, has been presented in work by [13], which proposes an automated approach to sense and predict earthquakes based on Tweets. The authors train a Support Vector Machine, using a linear kernel, with a collection of tweets labelled as being predictive of an earthquake or not. Similar to our approach, different feature sets are used including keywords and statistical features, however the approach presented in [13] differs in their use of a binary classification task: *does a given tweet refer to an earthquake or not?*.

The approach presented in this paper to convert tweets into linked data - via metadata processing and concept enrichment - is similar to the methodology presented in recent work by [9]. The key difference between our work and Mendes et al however is our processing of static datasets of tweets, whereas the latter's approach functions in real time. Similar to our approach, Mendes et al utilise an information extraction module to process tweet content and identify entities and concepts which are then turned into DBPedia concepts - enabling richer querying. However, our work goes one step further by aligning tweets with the events they are referring to, thereby exposing richer semantics when a given tweet is observed or queried. Additional recent work by [17] presents a formal model to associate users with their tweets and in turn topics, similar to our applications listed in the previous section. Wagner emphasises the need to extract semantic models from tweets, and how such models would enable the leveraging of implicit knowledge from such sources.

The automated technique to align tweets with events can be correlated with existing work within the field of reference reconciliation - where references are detected in disparate sources and linked together. For instance existing work within the field of object identification has explored the effects of training a *profiler* to detect object matches [2]. The profiler contains attributes and values which make up the description of the concept. This is then used to match other concepts which are deemed to refer to the same *thing* or object. The profiler runs using *If, Then* rules, such that objects are matched if they satisfy certain criteria. References are reconciled in [3] by using contextual information and past reconciliation decisions in the form of a dependency graph. By using information which is external to a given entity, reconciliation decisions can be performed, for instance when reconciling authors considering the individuals they have coauthored papers with. Techniques for the reconciliation of references are presented in [12] which combine both numerical and logical (based on the semantics of distinct data items) approaches.

Although we use Linked Data as the data source from which our classification functions are induced, there are similarities between this approach and existing Ontology-based Information Extraction methods. For instance, work by [7] uses an ontology and an annotated corpus as input, from which a Perceptron classifier is then learnt for each concept in the ontology. This is similar to our approach of using the URIs of the resources corresponding to the events that we wish to label tweets with. However, our experiments using the discriminative classifier SVM, which also learns a separating hyperplane in the feature space like Perceptron but using different a optimisation method, demonstrated the poor performance when applying such methods. This could be attributed to the standard way in which we trained the classifier - testing both a one-vs-one and one-vs-all voting strategy - while in [7], the learning algorithm is modified to account for the hierarchical relations between concepts - something that would be interesting to explore for future work within the domain of Linked Data. Likewise, earlier work by [8] described the use of evaluation measures when assessing the accuracy of OBIE, again using the same method as from [7] to train an individual Perceptron classifier for each concept from a given domain ontology. Their contribution is a series of evaluation measures, stating that if a label cannot be assigned to a piece information - thereby enabling its extraction - then the label with the lowest error should be assigned. The notion of error is derived from the semantic distance between concepts, the intuition being that a more general concept label is better than no label. We also subscribe to this thesis, and believe that the adaption of our approach could utilise the alternative class label - i.e., the URI of the conference event.

8. Conclusions

In this paper we have presented an approach to align tweets with the events which they refer to. Our approach tests two automated labelling techniques, one using a proximity-based clustering method and the other using a Naive Bayes classifier. Following evaluation of each of our tested methods we have empirically observed that the use of a generative model for alignment outperforms both tested discriminative models - i.e., SVMs - and the explored clustering method. For the clustering method we tested two distance measures for aligning tweets with events using a vector space model. Our future work will investigate the use of other distance metrics such as Mahalanobis distance which takes into consideration the shape of a given class distribution, unlike Euclidean distance which measures a spherical radius surrounding the mean of the event distribution. We are also currently implementing a graph-based method which utilises a graph-space composed from the features of both tweets and events. Our intuition is that common features between events and tweets will be harnessed when clustering using local similarity measures, such as centrality and betweenness.

The feature sets used by each of our alignment techniques demonstrates the effects of feature selection on performance. In the case of proximity-based clustering we found that the provision of additional features actually reduced the performance of the alignments, yielding lower levels of $F_{0.2}$ over feature sets which were smaller in size. Conversely Naive Bayes improved in performance as more features were introduced, resulting in 0.728 for $F_{0.2}$ for the combination of all feature sets.

The evaluation presented within this paper uses a dataset collected from the Extended Semantic Web Conference 2010 of all published tweets. We chose this conference as a representative sample of most Computer Science conferences - where Twitter is used in parallel to talks as an additional means of conversation and feedback. The utility of using Linked Data to train our method is evident given the levels of accuracy that we have obtained. One can imagine that future events, not necessarily tied to the Semantic Web community, will also produce metadata describing the talks and workshops at conferences in a machine-readable form. The growth of the Web of Data has seen large-scale production of information describing a range of subjects and topic areas from Encyclopaedic information through to publications. We therefore anticipate the production of Linked Data describing conferences throughout various disciplines, not just the Semantic Web, in doing so providing our approach with the necessary labelled data from which alignments can be made between tweets and events. For our future work we will be performing experiments using tweets collected from various events - not only Semantic Web-related - and exploring the contribution of the different feature sets when aligning tweets with the events, and sub-events, they refer to.

The large number of sub-events found in our experimental dataset requires an approach capable of handling multi-class classification. Through our evaluation we found Naive Bayes to offer such a solution, while the large number of classes led to problems in applying SVMs. Similarities exist between our presented methods and existing information retrieval approaches such as Term Frequency-Inverse Document Frequency (TF-IDF). For instance, our proximity-based measures attempt to minimise distance within the feature space between tweets and events, building vector representations of each using term weightings derived from a BoW model. TF-IDF gauges the importance of a term within a given document - in our case a tweet - based on its frequency in the document measured in proportion to its frequency in the entire corpus. Such a weighting heuristic is to be explored in our future work.

As discussed previously, the manual labelling of a portion of this dataset provided a gold standard against which we were able to test our approach. We believe there are many possible techniques and methods to perform this alignment task and have therefore published our dataset online[17] for the community to use. Our hope is that benchmarking evaluation will provide comparisons between differing methods and advance the state of the art in this area of work.

For future work we plan to extend the application discussed previously to enable automated, large-scale feedback to event organisers by aligning tweets with events and performing sentiment analysis over the published microblogs. At present our applications only expose lightweight usage of the mappings and therefore do not fully demonstrate the utility of such alignments. We are also exploring the use of additional features to improve alignment accuracy, such as the social network of a user on Twitter and how that information can be related to co-citation and authorship networks on the Web of Linked Data. Another interesting feature set to utilise would be the aspect of time - given that tweets are published during the same time slot as events - however, at present, such information is not provided as Linked data, requiring it to be obtained from an external source, the process of which could add noise into the approach and therefore reduce accuracy.

Within this paper we have used the third-party service Zemanta to provide DBPedia URIs for both tweets and events, where we see an improvement in performance when these features are incorporated into our experiments. That said, we do not know if this tool is the best for the job - given that it was not designed to provide concepts from content that is short in length. Therefore our future work will deploy several similar concept enrichment tools - e.g., Sem4Tag[18] - in the presented experimental setting and assess which service yields the best alignment performance in terms of our experimental results. At present no such comparison exists within the literature, nor have such tools been tested over a corpus of tweets.

9. Acknowledgements

The work of Matthew Rowe was supported in part by the EU-FP7 projects WeGov (grant no. 248512) and Robust (grant no. 257859). The work of Milan Stankovic has been partially funded by ANRT (French National Research Agency) under the grant number CIFRE N 789/2009. We are also grateful to kind people from Talis who have provided us with a free Linked Data triple store to put the data online. Zemanta was also kind to provide us with a generous API call limit that enabled us to do the topic enrichment of the tweets and events. We would also like to acknowledge the support of the Semantic Web community for their feedback from our experiments.

References

[1] John G. Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards semantically-interlinked online communities. In Asunción Gómez-Pérez and Jérôme Euzenat, editors, *Proceedings of the 2nd European Semantic Web Conference*. Springer, 2005. .

[2] Anhai Doan, Ying Lu, Yoonkyong Lee, and Jiawei Han. Object matching for information integration: A profiler-based approach. In Subbarao Kambhampati and Craig A. Knoblock, editors, *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web*, 2003.

[3] Xin Dong, Alon Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. In Fatma Özcan, editor, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005. .

[4] Joseph Fleiss. *Statistical Methods for Rates and Proportions*. Wiley-Interscience, 1981.

[5] John A Hartigan. *Clustering algorithms*. Wiley New York,, 1975. ISBN 047135645.

[6] Julie Letierce, Alexandre Passant, Stefan Decker, and JG Breslin. Understanding how Twitter is used to spread scientific messages. In Wendy Hall and Jim Hendler, editors, *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[7] Yaoyong Li and Kalina Bontcheva. Hierarchical, perceptron-like learning for ontology-based information extraction. In Juliana Freire Michael Rappa, Paul Jones and Soumen Chakrabarti, editors, *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007. ISBN 978-1-59593-654-7. .

[8] Diana Maynard, Wim Peters, and Yaoyong Li. Metrics for evaluation of ontology-based information. In *Proceedings of WWW2006 workshop on "Evaluation of Ontologies for the Web"*, 2006.

[17] `http://groups.google.com/group/tweet-event-mappings`

[18] `http://grafias.dia.fi.upm.es/Sem4Tags/`

[9] Pablo N. Mendes, Alexandre Passant, and Pavan Kapanipathi. Twarql: tapping into the wisdom of the crowd. In Adrian Paschke, Nicola Henze, and Tassilo Pellegrini, editors, *Proceedings of the 6th International Conference on Semantic Systems*. ACM, 2010. ISBN 978-1-4503-0014-8. .

[10] Deborah Nolan and Terry Speed. *Stat labs: mathematical statistics through applications*. Springer, 2000.

[11] Gene Rowe and George Wright. The delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, 15(4):353–375, October 1999.

[12] Fatiha Saïs, Nathalie Pernelle, and Marie-Christine Rousset. Combining a logical and a numerical method for data reconciliation. *Journal on Data Semantics*, 12:66–94, 2009. .

[13] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In Michael Rappa and Paul Jones, editors, *Proceedings of the 19th International World Wide Web Conference*. ACM, April 2010.

[14] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In Susanne Boll, Steven Hoi, and Jiebo Luo, editors, *Proceedings of the first SIGMM workshop on Social media*. ACM, 2009. ISBN 978-1-60558-759-2. .

[15] M. Stankovic. Modeling Online Presence. In John Breslin, Uldis Bojars, Alexandre Passant, and Sergio Fernández., editors, *Proceedings of the First Social Data on the Web Workshop*. CEUR Workshop Proceedings, 2008.

[16] Raphaël Troncy, Bartosz Malocha, and André T. S. Fialho. Linking events with media. In Adrian Paschke, Nicola Henze, and Tassilo Pellegrini, editors, *Proceedings of the 6th International Conference on Semantic Systems*. ACM, 2010. ISBN 978-1-4503-0014-8. .

[17] Claudia Wagner. Exploring the wisdom of the tweets: Knowledge acquisition from social awareness streams. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *Proceedings of the 7th Extended Semantic Web Conference*. Springer, 2010.