Time Aware Knowledge Extraction for Microblog Summarization on Twitter

Carmen De Maio, Giuseppe Fenza, Vincenzo Loia*, Mimmo Parente

Department of Computer Science, University of Salerno, Fisciano (SA), Italy

Abstract

Microblogging services like Twitter and Facebook collect millions of user generated content every moment about trending news, occurring events, and so on. Nevertheless, it is really a nightmare to find information of interest through the huge amount of available posts that are often noisy and redundant. In the era of Big Data, social media analytics services have caught increasing attention from both research and industry. Specifically, the dynamic context of microblogging requires to manage not only meaning of information but also the evolution of knowledge over the timeline. This work defines Time Aware Knowledge Extraction (briefly TAKE) methodology that relies on temporal extension of Fuzzy Formal Concept Analysis. In particular, a microblog summarization algorithm has been defined filtering the concepts organized by TAKE in a time-dependent hierarchy. The algorithm addresses topic-based summarization on Twitter. Besides considering the timing of the concepts, another distinguishing feature of the proposed microblog summarization framework is the possibility to have more or less detailed summary, according to the user's needs, with good levels of quality and completeness as highlighted in the experimental results.

Keywords: Microblog Summarization, Time Awareness, Fuzzy Formal Concept Analysis, Big Data, Social Media Analytics

*Corresponding author

Email addresses: cdemaio@unisa.it (Carmen De Maio), gfenza@unisa.it (Giuseppe Fenza), loia@unisa.it (Vincenzo Loia), parente@unisa.it (Mimmo Parente)

1. Introduction

Context. Nowadays, microblogging streams are useful to detect and track political events[?], media events[?], and other real world events[?]. Nevertheless, it is really difficult to understand the main aspects of the news or events inquiring these microblogging services. In fact, given a specific topic on Twitter a huge amount of relevant tweets that are redundant or not relevant due to the ambiguity and noise of the social media exists. Furthermore, the dynamic context of microblogging requires to manage not only meaning of information but also the evolution of knowledge over time. To face with this side effect many applications have been realized on Twitter, like Tweetchup (tweetchup.com), Twitalyzer (twitalyzer.com), which provide social media analytics services to detect and track trending topics. Moreover, automatic microblog summarization algorithms that extend Latent Diriclet Allocation (LDA) exist that consider both chronological order of the tweets and their information content, but at two distinct stages [? ? ?]. In the light of the described scenario, this work defines topic-based microblog summarization framework extending Fuzzy Formal Concept Analysis to manage time relations among tweets and introducing two main distinguishing features, that are specifically: first considering both time and meaning of the tweet at the same time to analyze knowledge evolution over the timeline; and second providing summaries with different level of detail according to the user's needs exploiting the peculiar properties of timed fuzzy lattice.

Problem. Formally, this work tries to face the following problem. Given a topic-focused timestamped tweet stream and a level of *detail* d, the task is aimed to filter and chronologically order tweets in order to produce a Microblog Summary MS_d that provides a complete description of the story covering main concepts describing topic development over the timeline. The proposed framework is able to retrieve more or less detailed summary according to the user's demand in terms of the level of *detail*.

Proposed Solution. This work defines a Time Aware Knowledge Extraction (briefly, TAKE) as a new methodology to solve the problem of topic-based microblog summarization focusing here on Twitter to give experimental evidence. In particular, a temporal extension of Fuzzy FCA in order to arrange resources (i.e., tweets) into a hierarchy of time dependent concepts, that is a timed fuzzy lattice, is defined. This enhancement makes Fuzzy FCA theory more suitable to deal with microblog considering temporal dimension. The final summarization process is achieved taking into account both timestamps and semantics of the tweets.

Firstly, temporal peaks of tweet frequency analyzing timestamps and exploiting the Offline Peak-Finding Algorithm (OPAD), proposed in [**?**], are identified. Secondly, content will be annotated via sentence wikification that is the practice of representing a sentence with a set of Wikipedia concepts (i.e., entries)[**? ?**]. Wikify service[1], provided by University of Waikato, enable us to face with the short length nature of microblogs, specifically we didn't rely on co-occurrences of words in the tweets to identify a topic. it is the practice of representing a sentence with a set of Wikipedia concepts. Wikification enables us to recognize sense of main concepts and named entity mentioned in the tweet associating a Wikipedia link and a corresponding weight representing confidence degree of the disambiguation result. Sentence wikification has been already exploited for text summarization [**?**], [**?**], [**?**] and it reveals to be not compromised by the short nature of the sentence. Wikification is an alternative approach with respect to co-occurring words analysis usually exploited in the literature and less suitable, in our opinion, to address the short nature of the tweet.

Then, taking into account the meaning of the tweet content and time dependences among detected peaks, temporal extension of Fuzzy Formal Concept Analysis [**? ?**] will be performed in order to arrange tweets into a hierarchy of time dependent concepts, that is a *timed fuzzy lattice*. Finally, a summarization algorithm has been defined exploring resulting timed fuzzy lattice knowledge

[1] Publically available at `http://wikipedia-miner.cms.waikato.ac.nz/` Let us note that we have exploited a local installation of the Wikipediaminer installation.

structure. The algorithm extracts chronologically ordered tweets summarizing main concepts of the story according to their temporal evolution.

Motivation. Analogously to other unsupervised methods, like Topic Modeling and clustering, the FCA theory groups together resources that share their meaning (or topics) that is essentially represented by set of strictly related words (e.g., Bill Clinton, Election, Politics, and so on). Nowadays, Topic Modeling has been widely applied to ontology extraction [**?**], text categorization [**?**], topic detection and tracking [**?**], microblog summarization [**?**], and so forth. Topic modeling has been described as "a recurring pattern of co-occurring words". A topic modeling tool looks through a corpus for these clusters of words and groups them together by a process of similarity. Nevertheless, Topic models exploit the co-occurrences of words between documents to find relations between words. Given that documents d_1 and d_2 share the common words $\{w1, w2, w3, w4\}$, we can infer that this densely connected set of words forms a topic and has semantically similar meanings. However, in the case of tweets due to the smaller number of words, there is less likelihood for words to co-occur with one another across different tweets. The words which could be inferred as belonging to the same topics have a weaker co-occurrence relationship with other words. To overcome this drawback, the authors in [**?**] have assumed that the tweets written around the same time are similar in content and so they can "share" words from other tweets to compensate for their short length. While, in our case we do not make this assumption and each tweet is treated separately considering its time occurrence if more tweets share the meaning that is extracted disambiguating the words with Wikipedia.

Furthermore, LDA (*Latent Dirichlet Allocation*), that is the technique usually exploited to extract topic model, doesn't extract relations among the identified topics. FCA (and Fuzzy FCA), instead, arranges the tweets into a hierarchical conceptual structure of topics, where the topic is represented by Formal Concept that is composed of words (the intent) shared among different tweets (the extent of the formal concept, Definition 3).

Considering these aspects and our expertise in the area of Fuzzy FCA the proposed approach is in our opinion the right starting way to address social media research challenges, and specifically microblog summarization, proposing an alternative approach compared to the state of the art.

Experimental Results. The proposed framework has been applied on the same tweet streams used in [**?**] that are focused on some real-world events, such as: Obamacare, Japan Earthquake, and so on. The results have been evaluated considering the following metrics: *Novelty Measurements*, *Text-based Coverage of Wikipedia*, and *Concept-based Coverage of Wikipedia.* The evaluation has been performed by varying the level of *detail* d in $[0 - 1]$. For all of the used metrics the system produces good performances. Specifically, the algorithm outperforms the results in [**?**] in terms of *Novelty Measurement* and *Text-based Coverage of Wikipedia.* Furthermore, evaluating *Concept-based Coverage of Wikipedia* setting level of *detail* with values ~ 0.9 (that is a verbose summary), the algorithm outperforms the results shown in [**?**] in terms of F-Measure, with optimal Recall and comparable values of Precision.

Outline. The manuscript is organized as follows: Section 2 provides an overview of the literature describing some related works; Section 3 introduces the theoretical background, i.e. Fuzzy Formal Concept Analysis; Section **??** introduces the overall framework detailing each phase in the sections **??**, **??** and **??**; finally, Section **??** shows the obtained results and argues the comparison with other existing approaches.

2. Related Works

Nowadays, automatic microblog summarization has caught increasing attention from worldwide researchers.

From the time-dependent document summarization point of view, some existing approaches are aimed to address update summarization task defined in TAC (www.nist.gov/tac). Specifically, they emphasize the novelty of the subsequent summary [**?**]. The proposed approach focuses more on the temporal development of the story (i.e. topic or event) that is stressed by the multitude of the messages posted through microblogging service, i.e. Twitter.

From the microblog summarization point of view, some pioneering approaches working on Twitter exist that are essentially aimed to describe topic extracting list of relevant words or sentences. Specifically, TweetMotif [?] summarizes what's happening on Twitter providing a list of relevant terms that should explain Twitter topics. [?] and [?] extract a succinct summary for each topic using a phrase reinforcement ranking approach. [?] explores tweets and linked web contents to discover relevant information about topics. Moreover, [?] generates summaries especially for sport topics. Furthermore, [?] defines frequency and graph based method to select multiple tweets that conveyed information about a given topic without being redundant. Other approaches are based on integer linear programming [?] or clustering to perform the summarization of Evolving Tweet Streams [?]. Other approaches consist of aggregating tweets about specific topic into a visual summaries. These visualizations must be interpreted by users and do not include sentence-level textual summaries. For instance, Visual Backchannel [?] and TwitInfo [?] allow users to graphically browse a large collection of tweets. Specifically, [?] visualizes conversations in Twitter data using topic streams that is visually represented as stacked graphs and TwitInfo [?] uses a timeline-based display that highlights peaks of high tweet activity.

Considering our proposal we find some similarities in [?] and in [?]. Specifically, [?] describes a framework for summarizing events from tweet stream. The authors define two topic models, Decay Topic Model (DTM) and Gaussian DTM, to extract summaries from microblog, and they finally argue that these models outperforms LDA (Latent Dirichlet Allocation) baseline that doesn't consider temporal relation among tweets. Instead, the approach used in [?] introduces a sequential summarization for Twitter trending topics exploiting two approaches: a stream based approach that is aimed to extract important subtopic concerning with specific category (e.g., News, Sport, etc.) identifying peak areas according to the timestamps of the tweets; and a semantic based approach leveraging on Dynamic Topic Modeling, that extends LDA in order to consider timeline, to identify topic from a semantic prospective in the time interval. In [?] the authors argue that hybrid approach that considers stream and semantic of the tweets outperforms other ones.

In general, these research works highlight that to achieve microblog summarization, due to the dynamic nature of its content, it is crucial to consider both the chronological order of the posts and their information content. Unlike these microblog summarization approaches that consider the time and meaning of the tweets at two different stages, our solution considers both timestamps and meaning of the tweets at the same time. This work presents the Time Aware Knowledge Extraction (briefly TAKE) methodology, as a new approach to perform conceptual and temporal data analysis of tweets' content for microblog summarization. TAKE extends Fuzzy Formal Concept Analysis [?] introducing time dependencies among objects, in order to provide a summary that follows the evolution of the story over the timeline. Furthermore, the proposed framework reveals good performances in terms of F-Measure, with optimal Recall and comparable values of Precision with respect to the compared approaches. Specifically, the timed fuzzy lattice extracted by TAKE enable us to support user requests providing less or more succinct summary according to the specific needs.

3. Theoretical Background: Fuzzy Formal Concept Analysis

The formal model behind the proposed methodology for microblog summarization is the fuzzy extension of Formal Concept Analysis (briefly, Fuzzy FCA or FFCA) [?]. FCA is a theoretical framework which supplies a basis for conceptual data analysis, knowledge processing and extraction. Fuzzy FCA [?] combines fuzzy logic into FCA representing the uncertainty through membership values in the range [0, 1].

The fuzzy set theory first introduced by [?] has been widely applied to Text Analysis and Knowledge Extraction in approaches aimed to deal with imprecise knowledge. This theory has provided mathematical representation of vague concepts expressed through linguistic terms such as *high temperature*, *low cost* or *cold weather* [**?**] enabling interpretation of inaccurate statements typically used in the natural language. Considering the advantages of using fuzzy sets theory for knowledge representation, fuzzy based methods have been extensively exploited also for multi-label text categorization in which a document can partially belong to one or more than one category.

Specifically, the goal of this paper is to propose a microblog summarization algorithm leveraging on fuzzy hierarchical classification of the tweets by means of the extraction of fuzzy ontology or taxonomy. A discussion on the automatic generation of taxonomies is presented in [**?**], it provides an overview of various approaches to the development of taxonomies dealing with imprecision and uncertainty in textual information, including the fuzzy taxonomy. [**?**] proposes a method based on fuzzy hierarchical clustering. The fuzzy concept hierarchy is also used by [**?**] to propose a new fuzzy membership rule by which categories of attributes are generalized. Other approaches to fuzzy clustering and taxonomy can be found in [**?**], [**?**]. In general, the fuzzy ontologies extraction from unstructured resources has received much attention from text mining researchers also to define methodology capable to retrieve texts relevant for a specific query [**?**], [**?**], [**?**].

In particular, we have defined in [**?**] a fuzzy extension of well assessed Formal Concept Analysis (FCA) theory to address fuzzy ontology extraction taking into account uncertainty and imprecision embedded in unstructured data. Fuzzy FCA deals with fuzzy relations between objects (e.g., documents, tweets, etc.) and their features (e.g., topics, named entities, and so on) considering membership varying in [0,1], instead of binary relation of traditional FCA, and so it enables us to specify more or less relevant features to represent resources enabling granular representation of them and it enables us to carry out similarity among resources varying in [0,1].

Following, some definitions about Fuzzy FCA are given.

Definition 1: *A* **Fuzzy Formal Context** *is a triple $K = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I = ((G \times M), \mu)$ is a fuzzy set.*

Recall that, being I a fuzzy set, each pair $(g,m) \in I$ has a membership value $\mu(g,m)$ in [0,1]. In the following the fuzzy set function μ will be denoted by μ_I.

Definition 2: Fuzzy Representation of Object. *Each object O in a fuzzy formal context K can be represented by a fuzzy set $\Phi(O)$ as $\Phi(O)$=$\{A_1(\mu_1)$, $A_2(\mu_2)$,..., $A_m(\mu_m)\}$, where $\{A_1$, A_2,..., $A_m\}$ is the set of attributes in K and μ_i is the membership of O with attribute A_i in K. $\Phi(O)$ is called the fuzzy representation of O.*

Unlike FCA that use binary relation to represent formal context, Fuzzy Formal Context enables the representation of the fuzzy relation between objects and attributes in a given domain. So, fuzziness enables to model a relation among object and attribute in a more smoothed way ensuring more precise representation and uncertainty management. Fuzzy Formal Context (see Definition 1) is often represented as a cross-table as shown in Figure **??**(a), where the rows represent the objects, while the columns, the attributes. Let us note that each cell of the table contains a membership value in [0, 1]. Specifically, Fuzzy Formal Context shown in Figure **??**(a) has a confidence threshold T=0.6, that means all the relationship with membership values less than 0.6 are not shown.

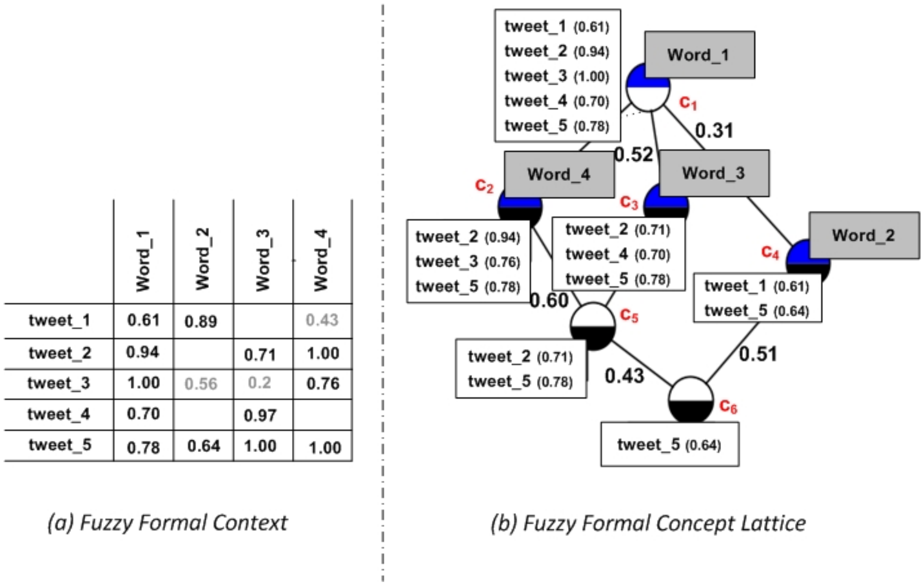
Taking into account Fuzzy Formal Context, Fuzzy FCA algorithm is able to identify Fuzzy Formal Concepts and subsumption relations among them. More formally, the definition of Fuzzy Formal Concept and order relation among them are given as follows:

Given a fuzzy formal context $K = (G, M, I)$ and a confidence threshold T, for $G' \subseteq G$ and $M' \subseteq M$, we define $G^* = \{m \in M \mid \forall g \in G',\ \mu_I(g,m) \geq \chi\}$ and $M^* = \{g \in G \mid \forall m \in M',\ \mu_I(g,m) \geq \chi\}$.

Definition 3: Fuzzy Formal Concept. *A fuzzy formal concept (or fuzzy concept) C of a fuzzy formal context K with a confidence threshold χ, is $C = (I_{G'}, M')$, where, for $G' \subseteq G$, $I_{G'} = (G', \mu), M' \subseteq M, G^* = M'$ and $M^* = G'$. Each object g has a membership $\mu_{I_{G'}}$ defined as*

$$\mu_{I_{G'}}(g) = min_{m \in M'} \left(\mu_I(g,m)\right) \tag{1}$$

where μ_I is the fuzzy function of I.



	Word_1	Word_2	Word_3	Word_4
tweet_1	0.61	0.89		0.43
tweet_2	0.94		0.71	1.00
tweet_3	1.00	0.56	0.2	0.76
tweet_4	0.70		0.97	
tweet_5	0.78	0.64	1.00	1.00

(a) Fuzzy Formal Context

(b) Fuzzy Formal Concept Lattice

Figure 1: Portion of fuzzy formal context (a) and the relative concept lattice with threshold T = 0.6 (b)

Note that if $M' = \emptyset$ then $\mu_I(g) = 1$ for every g. G' and M' are the extent and intent of the formal concept $(I_{G'}, M')$ respectively.

An example of Fuzzy Formal Concept is c_4 that is composed of objects $A_f = tweet_1, tweet_5$ and attributes $B =$ "$word_1, word_2'', \ldots$) with μ_{tweet_1}= 0.61 and μ_{tweet_5}= 0.64 , as shown in **??**(b).

Definition 4: *Let $(I_{G'}, M')$ and $(I_{G''}, M'')$ be two fuzzy concepts of a Fuzzy Formal Context (G, M, I). $(I_{G'}, M')$ is the **subconcept** of $(I_{G''}, M'')$, denoted as $(I_{G'}, M') \leq (I_{G''}, M'')$, if and only if $I_{G'} \sqsubseteq I_{G''} (\Leftrightarrow M'' \subseteq M')$. Equivalently, $(I_{G''}, M'')$ is the **superconcept** of $(I_{G'}, M')$.*

For instance, let us observe in Figure **??**(b), the concept c_5 is *subconcept* of the concepts c_2 and c_3. Equivalently the concepts c_2 and c_3 are *superconcepts* of the concept c_5.

Let us note that each node (i.e. a formal concept) is composed by the objects and the associated set of attributes, emphasizing by means fuzzy membership the object that are better represented by a set of attributes. In the figure, each node can be colored in different way, according to its characteristics: a half-blue colored node represents a concept with *own* attributes; a half-black colored node instead, outlines the presence of *own* objects in the concept; finally, a half-white colored node can represent a concept with no *own* objects (if the white colored portion is the half below of the circle) or attributes (if the white half is up on the circle).

Furthermore, given a Fuzzy Formal Concepts of Fuzzy Formal Context, it is easy to see that the subconcept relation $\leq$ induces a *Fuzzy Lattice* of Fuzzy Formal Concepts. As a matter of fact the lowest concept contains all attributes (Wikipedia entities) and the uppermost concept contains all object (tweets) of Fuzzy Formal Context.

Figure **??**(b) shows an example of lattice coming from the related table, with threshold $T = 0.6$. In fact, FCA provides also an alternative graphical representation of tabular data that is somewhat natural to navigate and use [**?**]. Furthermore, FFCA introduces also the definition of Fuzzy Formal Concept Similarity and Fuzzy Formal Concept Support. The former provides a degree of truth corresponding to each subsumption relation(i.e., an approximate subsumption). The definitions are given following.

Definition 5: ***Fuzzy Formal Concept Similarity*** *between concept* $C' = (I_{G'}, M')$ *and its subconcept* $C'' = (I_{G''}, M'')$ *is defined as*

$$E(C', C'') = \frac{|I_{G'} \sqcap I_{G''}|}{|I_{G'} \sqcup I_{G''}|} \tag{2}$$

where $\sqcap$ *and* $\sqcup$ *refer to intersection and union operators*[2] *on fuzzy sets, respectively.*

The notion of Fuzzy Formal Concept Support is based on the definition of *frequent concept intent* and closure systems introduced in [**?**]. Specifically:

[2] The fuzzy intersection and union are calculated using t-norm and t-conorm, respectively. The most commonly adopted t-norm is the minimum, while the most common t-conorm is the maximum. That is, given two fuzzy sets A and B with membership functions $\mu_A(x) and \mu_B(x), \mu_{A \bigcap B}(x) = min(\mu_A(x), \mu_B(x)) and \mu_{A \bigcup B}(x) = max(\mu_A(x), \mu_B(x))$.

Definition 6: ***Fuzzy Formal Concept Support***. *Let $K = (G, M, I)$ be a fuzzy formal context, the support of a Fuzzy Formal Concept $C' = (I_{G'}, M')$ is given by*

$$Supp(C') = \frac{|G'|}{|G|} \tag{3}$$

Let *minsupp* be a threshold $\in [0 - 1]$, then C' is said to be a frequent concept if $Supp(C') \geq minsupp$.

On one hand, the FCA provides a taxonomic arrangement of concepts and extracts the subsumption relationships (often known as a "hyponym-hypernym or is-a relationship") among them. On the other hand, Fuzzy FCA enables to considers these relations with a certain degree of truth (i.e., an approximate subsumption). In other words, the resulting fuzzy lattice elicits data-driven knowledge-based, hierarchical dependences, refining the taxonomic nature of this structure weighting interrelation among concepts introducing Fuzzy Formal Concept Similarity as stated in Definition 5.

4. Framework Overview

The proposed framework is aimed to address microblog summarization service on twitter. Specifically, this work defines a novel Time Aware Knowledge Extraction (briefly TAKE) methodology aimed to perform temporal and conceptual data analysis to foster dynamic nature of social media introducing intelligent analytics services. In particular we show how tweet stream will be analyzed to extract meaning of the tweets and to detect temporal correlation among them.

Specifically, Figure **??** sketches the whole process of the system that is composed of following main phases:

- *Microblog Content Analysis* (see Section **??**). It takes as input a tweet stream and detects tweet frequency peaks, then performs tweet's features extraction exploiting text analysis services, such as wikification, determining the meaning of the tweet and performing ad-hoc term weighting;
- *TAKE - Time Aware Knowledge Extraction* (see Section **??**). It takes as input term weighted tweets and their timestamps and performs Time

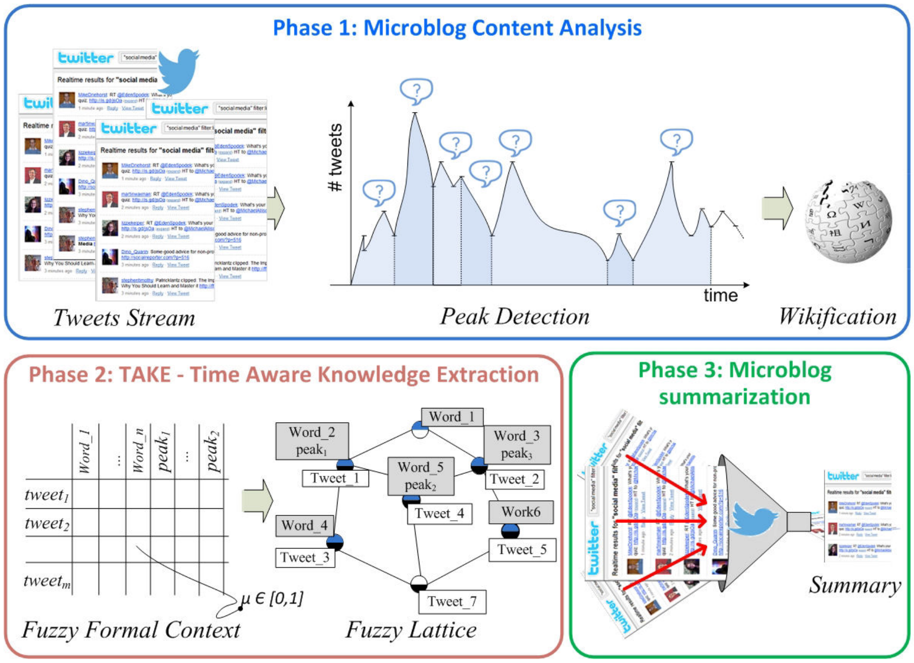


Figure 2: Overall Process of the framework

Aware FFCA in order to arrange tweets into a hierarchy carrying out also time dependence relation among extracted concepts;

- *Microblog Summarization Algorithm* (see Section **??**). It is a summarization algorithm that given the timed fuzzy lattice resulting by TAKE extracts a filtered set of tweets that covers the key concepts of the story considering the timeline and the *detail* level specified as input (See later for the definition and discussion of *detail*).

It is possible to distinguish phases performed online and offline. Specifically, *Microblog Content Analysis* and *Time Aware Knowledge Extraction* will be periodically performed offline also because they are time consuming activities. Instead, *Microblog Summarization Algorithm* is performed at execution time according to the user request in terms of topic and level of *detail*. Additional technical and formal details about each macro-phases are given in the next sections.

5. Microblog Content Analysis

This phase is aimed to characterize tweets extracting representing features considering both the timestamp and the meaning of the tweet. Specifically, this activity is preliminary to map the domain data (e.g., tweets content) into a fuzzy formal context, enabling FFCA execution (details are given in the following sections).

This phase is composed of these steps:

- *Peak Detection*, to detect temporal peaks from Twitter streams;
- *Content Wikification*, to identify and extract relevant features that characterize meaning of the input tweet;
- *Inverse Tweet Frequency*, to measure how important a concept is.

The mathematical modeling of the fuzzy formal context needs the representative features capable to represent both meaning and time dependencies among the tweets. The goal is to exploit a vector-based representation of each tweet and then build the matrix which represents the fuzzy formal context. This matrix will show the relationships (in terms of degree values) between the extracted features (i.e., peaks and wikipedia entities) and tweets in the application domain. Further details are given in the following subsections.

5.1. Peak Detection

This step identifies temporal peaks in tweet frequency exploiting the *Offline Peak-Finding Algorithm* (OPAD) (listing 1), proposed in [**?**]. The algorithm is based on the idea of TCP congestion control, which uses a weighted moving mean and variance to determine if there is a new peak area [**?**].

Given a time-sorted collection of tweets, the algorithm locates surges by tracing tweet volume changes. Let $T = (t_1, t_2, ..., t_n)$ a time-sorted collection of tweets, we group tweets that are posted within the same 1440 minute (i.e., 1 day) time window. At this point we have a list of tweet counts $C = (C_1, C_2, ...C_t)$

where C_i is the number of tweets in bin i. The objective is to identify each bin i such that C_i is large relative to the recent history $C_{i-1}, C_{i-2}, \ldots C_1$. Initializing the mean and variance with the first time interval (line 2-3), the algorithm loops through the whole tweet stream (line 5). If the number of tweets in the current bin (i.e., C_i) is greater than τ (we use $\tau = 2$) mean deviations from the current mean (i.e., $\frac{|C_i - mean|}{meandev} > \tau$), and the tweet number in current bin is increasing (i.e., $C_i > C_{i-1}$, line 6), then a new peak window starts (line 7). Then, the algorithm will loop until the condition $C_i > C_{i-1}$ is verified and updates the mean and variance (line 8-11). So, the peak search stops when the tweet number in the bin is less than the number of the previous one. After that, in the loop of lines 12-20 the bottom of peak interval is searched, which occurs either when the tweet number in the current bin is smaller than the tweet number at starting of the peak window (line 12) or another significant increase is found (line 13). At line 23, new peak window is included in the set of found peak areas. Every time we iterate over a new bin count, we update the mean and mean deviation (lines 9, 17 and 24) by means of *Update* function (line 30-34). In the function *Update* α is set to 0.125 as in [?].

Listing 1: OPAD- Offline Peak Area Detection

```
1  windows = []
2  mean = $C_1$
3  meandev = variance ( $C_1$, ... , $C_p$ )
4
5  for i = 2; i < len(C); i++ do
6      if $\frac{|C_i - mean|}{meandev} > \tau$ and $C_i > C_{i-1}$ then
7          start = i-1
8          while i < len(C) and $C_i > C_{i-1}$ do
9              (mean, meandev) = update(mean, meandev, $C_i$ )
10              i ++
11         end while
12         while i < len(C) and $C_i >$ $C_{start}$ do
13             if $\frac{|C_i - mean|}{meandev} > \tau$ and $C_i > C_{i-1}$ then
14                 end = -- i
15                 break
16             else
17                  (mean, meandev) = update(mean, meandev, $C_i$ )
18                  end = i ++
            end if
        end while
        if ( $C_i < C_{start}$ ) then
            end = i − −
        windows.append(start, end)
    else
          (mean, meandev) = update(mean, meandev, $C_i$ )
    end if
end for
return windows

function update(oldmean, oldmeandev, updatevalue):
    diff = |oldmean − updatevalue|
    newmeandev = $\alpha$* diff+(1−$\alpha$)*oldmeandev
    newmean = $\alpha$*updatevalue+(1−$\alpha$)*oldmean
return (newmean, newmeandev)
```

Then the i-th tweet will be annotated temporally, such as follows:

- $tweet_i = \{\langle peak_i \rangle\}$.

5.2. Content Wikification

The previous step of Microblog's content Analysis process involves the extraction of concepts from an unstructured text in the tweet content. To achieve this aim this work exploits common-sense knowledge available in Wikipedia. In order to do this, the tweet content is wikified to extract a set of $\langle topic, relevance \rangle$ pairs corresponding to Wikipedia articles that are related to the tweet content itself with a specific relevance degree [**?**]. In particular, topics returned by applying the wikification upon a tweet content helped us to characterize the given text.

Let us report an example by considering the following tweet:
$tweet_i$ = *"President Obama just designated the largest marine reserve in the world"*.

The wikification process extracts from the above text a set of $\langle topic, relevance \rangle$ pairs. These pairs are features characterizing meaning of the input text. Taking into account the example above, the extracted topic (shown in Figure **??**) are:

$\langle Barack\ Obama, 0.678 \rangle$, $\langle President\ of\ the\ United\ States, 0.456 \rangle$

Then, at this point, considering the example defined in Section **??** about $tweet_i$, the content will be annotated via sentence wikification as:

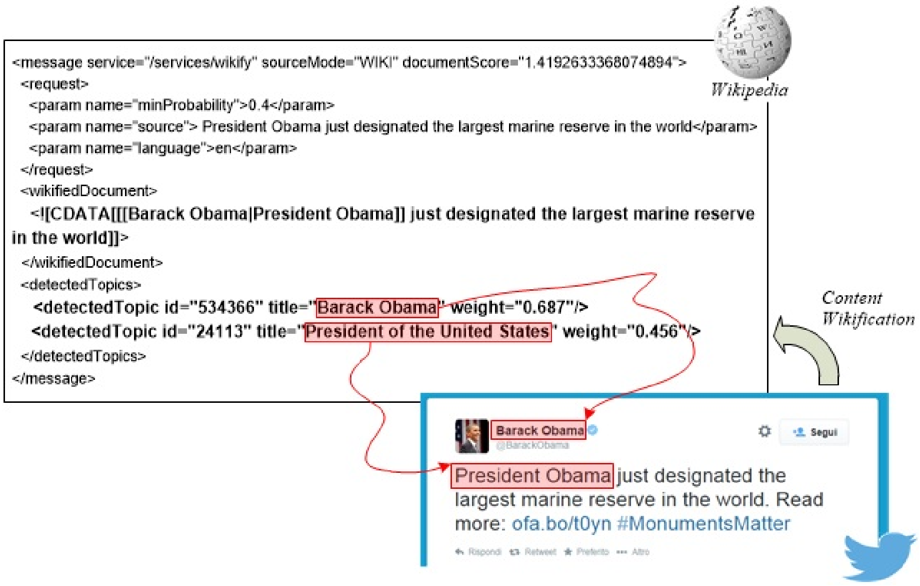
$$tweet_i = \{\langle peak_i \rangle\} \bigcup$$
$$\{\langle topic_{i_1}, relevance_{i_1} \rangle, \langle topic_{i_2}, relevance_{i_2} \rangle, \ldots, \langle topic_{i_m}, relevance_{i_m} \rangle\}$$

where m is the number of topics detected by sentence wikification of the $tweet_i$.

5.3. Inverse Tweet Frequency

After having analyzed the peak area which the tweets belong to (see Section **??**) and the wikification of tweet content (see Section **??**), *ITF (i.e., Inverse Tweet Frequency)* is exploited to refine membership of relevance degree of the topic found inside the tweet. It intuitively evaluates the measure of how much information each extracted topic (see Section **??**) provides whether it is common or rare across all tweets. Specifically, let $W = \{w_1, w_2, ..., w_n\}$ be the set of topics extracted by means of wikification process from set of tweets $T = \{t_1, t_2, ..., t_m\}$. Let us compute the ITF for each one topic as:

$$\mathrm{itf}(w_i, T) = \log \frac{N}{|\{t_j \in T : w_i \in t_j\}|}$$



```
<message service="/services/wikify" sourceMode="WIKI" documentScore="1.41926333680748944">
  <request>
    <param name="minProbability">0.4</param>
    <param name="source"> President Obama just designated the largest marine reserve in the world</param>
    <param name="language">en</param>
  </request>
  <wikifiedDocument>
    <![CDATA[[[Barack Obama|President Obama]] just designated the largest marine reserve
in the world]]>
  </wikifiedDocument>
  <detectedTopics>
    <detectedTopic id="534366" title="Barack Obama" weight="0.687"/>
    <detectedTopic id="24113" title="President of the United States" weight="0.456"/>
  </detectedTopics>
</message>
```

Barack Obama
@BarackObama

President Obama just designated the largest marine reserve in the world. Read more: ofa.bo/t0yn #MonumentsMatter

Figure 3: Example of tweet's content wikification.

where:

- N: total number of tweets analyzed;
- $|\{t_j \in T : w_i \in t_j\}|$: number of tweet from which the topic w_i has been extracted.

This value is exploited to compute the final value that characterizes the frequency associated to each topic extracted for a tweet. In particular, the final relevance f_{rel} associated to the topic w_i with respect to the tweet t_j is defined as:

$$f_{rel}(w_i, t_j) = relevance(w_i, t_j) \times itf(w_i, T);$$

Then, at this point, considering the example defined in previous sections about $tweet_i$, the content will be annotated as:

$$tweet_i = \{\langle peak_i \rangle\} \bigcup \{\left\langle topic_j, f_{rel_j} \right\rangle, \left\langle topic_{j+1}, f_{rel_{j+1}} \right\rangle, \ldots, \langle topic_m, f_{rel_m} \rangle\}$$

6. TAKE - Time Aware Knowledge Extraction

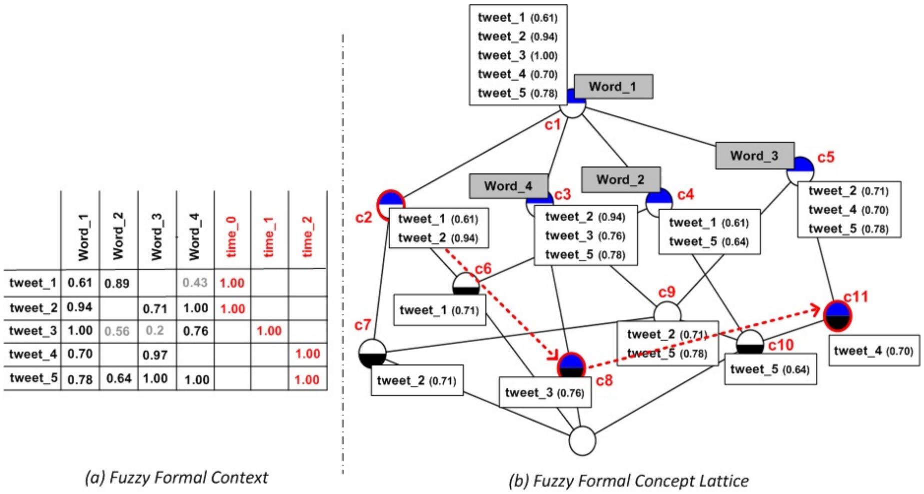
Time Aware Knowledge Extraction is an important feature to perform conceptual data analysis taking into account temporal relation among resources and to consequently carry out temporal correlation among concepts in order to represent their development over the timeline. The proposed approach to address this aim relies on Fuzzy Formal Concept Analysis, but as stated in Section 3, FFCA does not cover time dependences in the data.

In literature, some approaches that extend formal concept analysis to handle temporal properties and represent temporally evolving attributes exist [**?**]. Specifically, this temporal extension has been applied in [**?**] to search pedophiles on the Internet analyzing chat conversation over time. Here we adopt a distinct approach by extending FCA introducing fuzziness and temporal correlation among objects, in order to extract temporal dependencies among attributes in the concepts.

This work defines a time extension of *FFCA* to extract hierarchically and temporal related concepts. Indeed, besides classical contexts, timed *FFCA* extracts chronological relations among formal concepts inferred by analyzing time dependences among formal objects.

From a theoretical viewpoint, this work extends FFCA to consider timeline defining special attributes for representing time relations among formal objects. Formally, a time aware fuzzy formal context is defined as follow:

Definition 7: *A **Time Aware Fuzzy Formal Context** is a fuzzy formal contexts* $K_t = (G, M^+ = M \bigcup T, I_M = \varphi(G \times M), I_T)$*, where* T *is the set of time attributes and* I_T *is a binary time relation* $I_T \subseteq G \times T$ *representing the relation between formal object* $g \in G$ *and time attributes* $t \in T$*.*



	Word_1	Word_2	Word_3	Word_4	time_0	time_1	time_2
tweet_1	0.61	0.89		0.43	1.00		
tweet_2	0.94		0.71	1.00	1.00		
tweet_3	1.00	0.56	0.2	0.76		1.00	
tweet_4	0.70		0.97				1.00
tweet_5	0.78	0.64	1.00	1.00			1.00

(a) Fuzzy Formal Context

(b) Fuzzy Formal Concept Lattice

Figure 4: Time Aware Fuzzy FCA: portion of fuzzy temporal fuzzy formal context (a) and the relative temporal fuzzy concept lattice

For instance, if $g \in G$ and $t \in T$ are in relation I_T means that g happens at time $t \in T$.

Time extension of Fuzzy FCA allows to organize tweets in a weighted hierarchical knowledge structure, that is a timed fuzzy lattice. In particular, a straight mapping defines a correspondence between the set of attributes M and linguistic terms extracted from tweets content, as well as the set G of objects and the tweets collection.

Let us consider timed fuzzy formal context and correspondent timed fuzzy lattice in Figure **??**. Specifically, Figure **??**(b) emphasizes that each node (i.e., a formal concept) includes the objects, attributes and time attributes. For example in the lattice in Figure **??**(b), a concept is $(A_f = tweet_1, tweet_2,\ B =$ "$word_1, time''_0)$ with μ_{tweet_1}= 0.61 and μ_{tweet_2}= 0.94.

The resulting timed fuzzy lattice emphasizes a temporal correlation among concepts and highlights how the concepts change over the timeline (Figure **??**). To represent the concept development over the timeline in a timed fuzzy lattice temporal edges have been introduced (in Figure **??** red dashed arrows) among related concepts. The temporal edges allow the evolution of attributes to be followed over time. A temporal precedence relation is defined over time points. The direction of the arrow indicates this precedence. In the lattice in Figure **??**(b), the evolution of attributes is represented as: $c_2 \to c_8 \to c_{11}$, i.e., $\{Obama\} \to \{Obama, election\} \to \{Obama, President\}$.

7. Microblog Summarization Algorithm

The microblog summarization algorithm has been defined walking across concepts of the timed fuzzy lattice structure resulting from Time Aware Knowledge Extraction. The general idea behind is to explore fuzzy formal concepts according to the chronological order of the peak areas. The algorithm incrementally selects the *best tweet*, that is the tweet with highest degree of membership belonging to the most representative concept C, at each exploration stage. The most representative concept is one that has highest weight $w(C)$. Formally, the weight $w(C)$ of fuzzy formal concept C will be evaluated as follows:

$$w(C) = \frac{\sum_{m \in M'} \mu_m}{|M'|} \tag{4}$$

where $|M'|$ is the number of attributes in C and the membership μ_m is defined as follows:

$$\mu_m = max_{g \in I_{G'}} \mu(g, m) \tag{5}$$

where $\mu(g, m)$ is the membership value between object g and attribute m (see Section 3).

The microblog summarization algorithm is detailed in the Listing **??**. First of all, the sets of covered attributes (i.e., CA), covered concepts (i.e., CC) and microblog summary corresponding to a *detail* level d (i.e., MS_d) are all initialized as empty set (line 4-6). Then, the algorithm selects concepts (C^*) of the timed fuzzy lattice whose support is greater than of *detail* d specified as input (line 8). After that, the algorithm sorts peak areas in a descending order, that is the most recent peak area will appear first (line 9). Finally, the algorithm loops across each selected concepts that have been grouped by peak area $p_i \in P$ (line 10). At each iteration, the algorithm selects the most representative concept c_{max} (line 14) and the *best tweet* t_{max} with highest degree of membership belonging to c_{max} (line 15). t_{max} is included in the resulting summary (MS_d) (line 16) and both the set of covered attributes CA (line 17) and the set of covered concepts CC (line 18) are updated.

Listing 2: Summarization by Time Aware FFCA

```
1  Input: timed fuzzy lattice $L$, peak areas $P$, and $detail$ level $d$.
2  Output: a microblog summary $MS_d$ of $T^*$ tweets;
3
4   $CA = \emptyset$,
5   $CC = \emptyset$,
6  $MS_d = \emptyset$
7
8   $C^* = \{c_i \in L \mid 1 - Supp(c_i) > d\}$
9    $P = (p_1, p_2, \ldots, p_n)\ \forall\ i, j: p_i > p_j \Leftrightarrow i < j$
10 $C^*_{p_i} = \{c_i \in C^* \mid c_i = (I_{G'}, M'),\ p_i \in M'\}$
11
12 for i = 1; i < len($P$); i++ do
13   while $C^*_{p_i} \setminus CC \neq \emptyset$
14       $c_{max} = argmax_{c_i \in (C^*_{p_i} \setminus CC)} \left( w(c_i) = \frac{\sum_{m \in (M' \setminus CA)} \mu_m}{|M'|} \right)$
15       $t_{max} = argmax_{g \in c_{max}} (\mu_g)$
16       $MS_d = MS_d \cup t_{max}$
17       $CA = CA \cup \{m \mid \mu m \in c_{max}\}$
18       $CC = CC \cup c_{max}$
19    end while
20  end for
```

Just to give an example, let us suppose an input level of *detail* $d = 70\%$. Figure **??** shows the concepts with level of *detail* greater than 70% resulting from the execution of line 8 in Listing **??**. Table **??** lists the set of candidate concepts grouped by peak area to which they belong to.

Table 7.1: Concepts of the timed fuzzy lattice in Figure **??** grouped by peak areas

peak	**Concepts**
1	c_1, c_2, c_3
2	c_2, c_3, c_4, c_5
3	c_5, c_6

According to the algorithm, the set of concepts is analyzed starting from the most recent peak area, that is $peak_1$. For each concept the weight w is calculated and the concept with maximum value of w will be selected (see Listing **??**, line 14). Let us consider the following example:

$$w(c_1) = 0.89; \quad w(c_2) = 0.74; \quad w(c_3) = 0.94;$$

The first selected concept will be $c_3 = \{tweet\_1, tweet\_2, tweet\_11, tweet\_4, tweet\_6\}$ with maximum weight $w = 0.94$. The attributes covered by the concept c_3 are: *Obama, Party*.

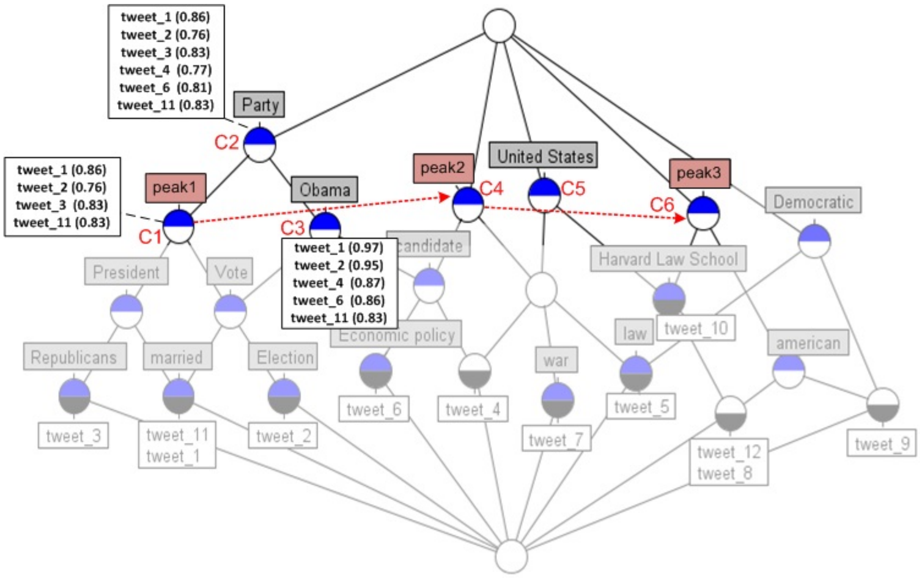


Figure 5: Example of timed Fuzzy Concept Lattice

The algorithm exploits fuzzy membership corresponding to the tweets in the selected concept (i.e., c_3) in order to look for the tweets with maximum membership degree. Thus, the tweet that will be introduced in the summary is $tweet\_1$ with highest degree of membership (i.e., 0.97) belonging to the concept c_3 (see Listing **??**, line 15). After updating summary including this tweet, the weight of remaining concepts will be updated removing attributes already covered by selecting c_3. In this case, the weight of remaining concepts is 0 for both c_1 and c_2. So, there are no more concepts to select in the peak area $peak1$. So, the algorithm proceeds with next peak area, i.e. $peak_2$. At the end of the execution the resulting summary will be composed of following tweets:

$$MS_d = \{tweet\_1, tweet\_7\}.$$

8. Framework Evaluation: Quantitative and Qualitative

This section details the Quantitative and Qualitative results obtained performing the proposed summarization algorithm on some tweet streams. As said before, the summarization algorithm relies on timed fuzzy lattice of tweets resulting from Time Aware Knowledge Extraction execution. Since the timed fuzzy lattice allows to perform the summarization algorithm with different levels of *detail*, the results have been evaluated by varying these levels. In particular, the higher the level of *detail* the more exhaustive the resulting summary will be, that is the summary will include a greater amount of tweets.

The next sections detail quantitative (Section **??**) and qualitative (Section **??**) results using different datasets and measures.

8.1. Quantitative Evaluation

In this section quantitative evaluation will be discussed describing the tweet streams on which the framework has been executed (Section **??**), defining measures (Section **??**), and showing the obtained results (Section **??**).

8.1.1. Tweet Streams

The summarization framework has been applied on tweet streams focused on four real-world events[3]: Facebook IPO[4], Obamacare[5], Japan Earthquake[6] and BP Oil Spill[7]. The number of tweets for these events ranges from 9.570 tweets for Facebook IPO to 251.802 tweets for the Japan Earthquake. Specifically, Table **??** synthesizes how many tweets are included in each tweets stream. Let us note that the multitude of tweets related to each event highlights that nowadays microblogging summarization as well as other social media analytics services are welcomed to foster social media usage.

As will be describe in next sections the difference among distinct evaluation results are due to the quality of content of the corresponding tweet stream: if it is composed of more or less significant or redundant tweets, number of keywords for tweets, and so on. For example, considering the "'Facebook IPO" and "Obama care" dataset, main difference among them it is essentially due to the fact that both tweet streams and corresponding gold summaries have very different size in terms of number of tweets and number of sentences, respectively. In particular, "Facebook IPO" tweet stream is composed of a number of tweets less or equal to 10k, and "Obama care" is composed of a number of tweets less or equal 137k (see Table **??**). Analogously, the gold summary size is 59 sentences for "Facebook IPO" and 259 sentences for "Obama care".

8.1.2. Measurements

The proposed framework has been evaluated considering the following metrics:

- *Novelty Measurements*, specifically *Sequence Novelty Measurement* introduced in [**?**] and *Historical Novelty Measurement.*

  – *Sequence Novelty Measurement* measures average novelty among chro-

[3] Specifically, the data have been provided by authors of [**?**]

[4] `http://en.wikipedia.org/wiki/Initial_public_offering_of_Facebook`

[5] `http://en.wikipedia.org/wiki/Patient_Protection_and_Affordable_Care_Act`

[6] `http://en.wikipedia.org/wiki/2011_T%C5%8Dhoku_earthquake_and_tsunami`

[7] `http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill`

Table 8.1: Number of tweets for each dataset

Name	# Tweets
Facebook IPO	9.570
Obamacare	136.761
Japan Earthquake	251.802
BP Oil Spill	79.676

nologically adjacent tweets included in the resulting summary. Information content I has been used to measure the novelty of update summaries. In particular, it is defined as the average of I increments of two adjacent new tweets added to summary.

$$Novelty = \frac{1}{|D|-1}\sum_{i>1}\left(I_{d_i} - I_{d_i,d_{i-1}}\right) \tag{6}$$

where:

- $|D|$ is the number of the tweets in the generated summary;
- I_{d_i} number of concepts in d_i;
- $I_{d_i,d_{i-1}}$ cardinality of intersection of d_i, d_{i-1}.

– *Historical Novelty Measurement* evaluates average novelty among each tweet and all previous ones included in the resulting summary. This measure has been defined in this work to represent the update summary ratio considering history of chronologically previous tweets included in the generated summary. Analogously to the *Sequence Novelty Measurement*, information content I has been used to measure the novelty of update summaries. In particular, it is defined as

$$Novelty = \frac{1}{|D|-1}\sum_{i>1}\left(I_{d_i} - \left[\left(\bigcup_{k<i} I_{d_k}\right)\bigcap I_{d_i}\right]\right) \tag{7}$$

where:

- $|D|$ is the number of new tweets added in the summary;
- I_{d_i} number of concepts in d_i;

- $I_{d_i,d_{i-1}}$ cardinality of intersection of d_i, d_{i-1};
- I_{d_k} with $k = 1...|D|$ correspond to all tweets in the summary.

- *Text-based Coverage of Wikipedia*, introduced in [?] where is called *Quantitative Comparison with Wikipedia*, evaluates how much generated summary covers the gold one at text-level (i.e., considering n-grams). Specifically, gold summaries are extracted from Wikipedia[8]. Specifically, the metric counts the total number of *n-grams* (excluding stop-words) in the generated summary S^{gen} that are also included in the gold summary S^{gold}. Let us define NG_n^{gold} the set of n-grams in the gold summary and NG_n^{gen} the set of n-grams in the generated summaries, this metric has been evaluated as follows:

$$g_n = \frac{1}{\left|NG_n^{gold}\right|} \sum_{ng \in NG_n^{gold}} min\left(\left|ng \in NG_n^{gold}\right|, \left|ng \in NG_n^{gen}\right|\right) \quad (8)$$

$$Sim\left(S^{gold}, S^{gen}\right) = 0,2 \cdot g_1 + 0,3 \cdot g_2 + 0,5 \cdot g_3 \quad (9)$$

First equation calculates the number of n-grams common to both S^{gold} and S^{gen}. In order to not let few frequent n-gram to dominate the counts, each n-gram is limited to the minimum number of counts between the gold summary and the generated summary. The other equation calculates the final similarity score between the summaries by aggregating the number of matched 1, 2 and 3-grams. The weights allocated are meant to give a higher importance to 3-grams and lower importance to 1-grams.

- *Concept-based Coverage of Wikipedia*, this metric has been defined in this work to evaluate how much the generated summary covers the gold

[8] Indeed, gold summaries have been provided by authors of [?]. They are extracted considering the references of the relevant news articles cited in Wikipedia article corresponding to the topic/event of the tweet stream. For each of the Wikipedia references for the selected events, we extract the headline text which gives a one line summary of the corresponding news article.

summary at concept level. Indeed, each sentence of gold and generated summaries will be annotated via wikification that is the practice of representing a sentence with a set of Wikipedia concepts [**?** **?**]. More formally, let $C_{gold} = \{c_1, c_2, ..., c_m\}$ and $C'_{gen} = \{c'_1, c'_2, ..., c'_n\}$ be, respectively, the set of concept extracted from the sentences included in the gold summary and the set of concepts extracted from the generated summary. Then *Concept-based Coverage of Wikipedia* will be evaluated in terms of well-known F-Measure that is obtained by combining measures of Precision and Recall. Specifically, Precision and Recall will be evaluated as follows:

$$P = \frac{\left|C_{gold} \bigcap C'_{gen}\right|}{\left|C'_{gen}\right|} \qquad R = \frac{\left|C_{gold} \bigcap C'_{gen}\right|}{\left|C_{gold}\right|} \tag{10}$$

Then, F-measure F is computed as follows:

$$F = 2 \times \frac{P \times R}{P + R} \tag{11}$$

So, this measure provides qualitative (i.e., Precision) and quantitative (i.e., Recall) information about how much generated summary covers the gold summary, and so, it evaluates semantically performances of the proposed microblog summarization approach.

- *Hashtag Coverage*, this metric has been defined to measure the coverage of the top fifty hashtags more frequently used in the considered tweet streams. More formally let $Ht_{dataset} = \{ht_1, ht_2, \ldots, ht_n\}$ and $Ht_{summary} = \{ht_1, ht_2, \ldots, ht_m\}$, with $m \leq n$, the set of different hashtag extracted from the sentences included in the tweet stream and the set of different hashtag extracted from the generated summary. Then *Hashtag Coverage* will be evaluated as:

$$Ht_{coverage} = \left|Ht_{dataset} \bigcap Ht_{summary}\right| \tag{12}$$

8.1.3. Experimental Results

The selected tweet streams and measures have been used to evaluate both the proposed approach (i.e., referred as TAKE) and methods defined in [**?**].

Since TAKE produces different summaries with different level of *detail*, the results have been evaluated by varying the level of d in $[0-1]$ and for all of the used metrics the system reveals good performances.

8.1.3.1 Novelty Results

Figures **??** and Figure **??** show the results about novelty, respectively *Sequence Novelty Measurement* and *Historical Novelty Measurement*. The results have been grouped by tweet streams (i.e., real world events of Facebook IPO, Japan Earthquake, and so on) and for each evaluated approach they are shown with different colors. In particular, TAKE has been evaluated by varying the level of *detail* d in $[0-1]$ and plotting the obtained minimum and maximum values for both novelty measures.

Figure **??** illustrates the results of novelty among adjacent tweets, that is *Sequence Novelty Measurement*. On the one hand, it points out that the proposed approach produces summaries with maximum values of novelty highest than other approaches for each tweet stream. On the other hand, TAKE produces summaries with minimum values of novelty lower than other approaches only for the tweet stream of *BP Oil Spill*.

Analogously, the results of *Historical Novelty Measurement* shown in Figure

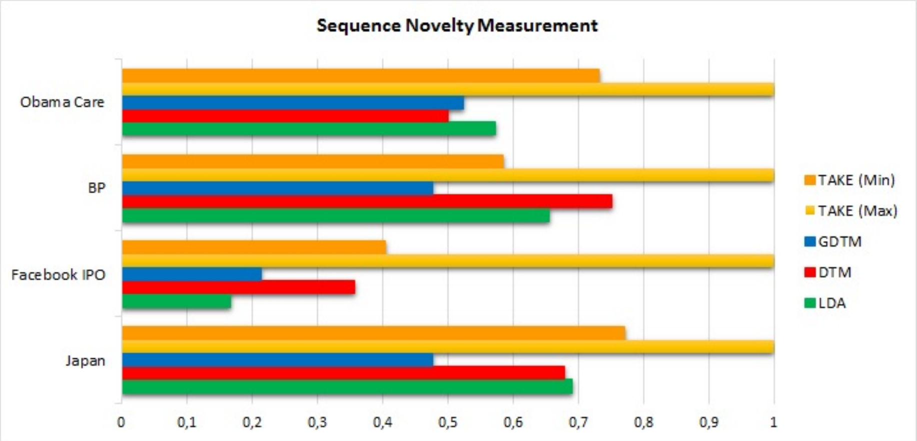


Figure 6: Sequence Novelty Measurement Results.

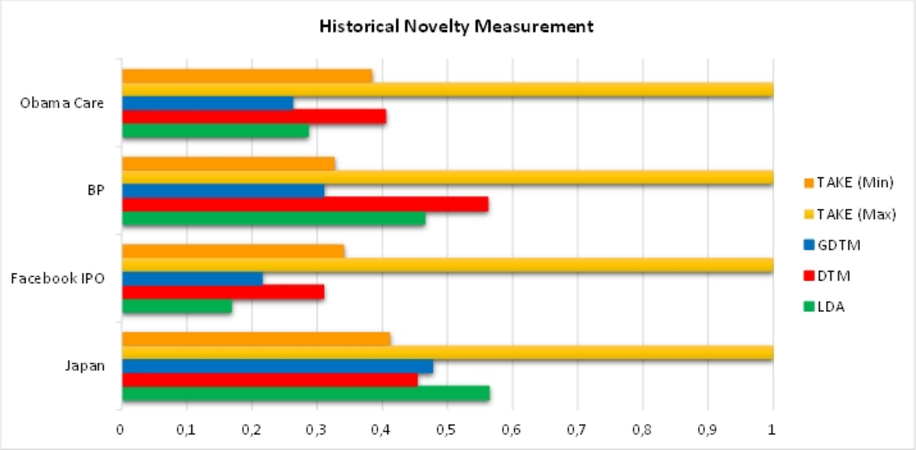


Figure 7: Historical Novelty Measurement Results.

?? highlight that TAKE produces summaries with maximum values of novelty highest than other approaches for each tweet stream. Minimum values of *Historical Novelty Measurement* produced by TAKE are close enough to the results obtained with other approaches, and so they are acceptable results.

Since, the proposed microblog summarization returns chronological ordered tweets starting from the most recent ones, the results of *Novelty Measurement* points out that TAKE generates more or less shortened summaries with acceptable levels of redundancy. Thus, the proposed method incrementally includes tweets in the resulting summary introducing significant amount of novel concepts improving the description of the event according to its development over the timeline.

8.1.3.2 Text-based and Concept-based Coverage Results

Figure **??** and Figure **??** show the results obtained evaluating *Text-based Coverage of Wikipedia* and *Concept-based Coverage of Wikipedia*, respectively. These outcomes are useful to measure quality and completeness of the generated summaries with respect to gold summaries. The results have been grouped by tweet streams and for each evaluated approach they are shown with different colors.

Since *Text-based Coverage of Wikipedia* grows by increasing the level of *detail*, in Figure **??** the minimum value produced by TAKE that is higher than the values produced by other approaches has been plotted. Specifically, it has been obtained setting the level of *detail* to 0.6. For levels of *detail* greater than 0.6, TAKE significantly outperforms other approaches revealing good performances in terms of complete description of summarized event at merely syntactically level.

Furthermore, Figure **??** shows that TAKE outperforms other techniques in terms of *Concept-based Coverage of Wikipedia*, and so it is possible to conclude that the proposed method reveals good performances in terms of quality and completeness also at the concept level.

In order to provide more details, Figure **??** shows the curves corresponding to Precision, Recall and F-measure of *Concept-based Coverage of Wikipedia* for each tweet stream. It has been evaluated by varying the level of *detail* d from 0.0 to 1.0. The figure highlights that TAKE reveals valuable performance in terms of Recall with acceptable values of Precision with level of *detail* between 0.7 and 0.9.

In general, the distinguishing feature introduced by TAKE approach is the

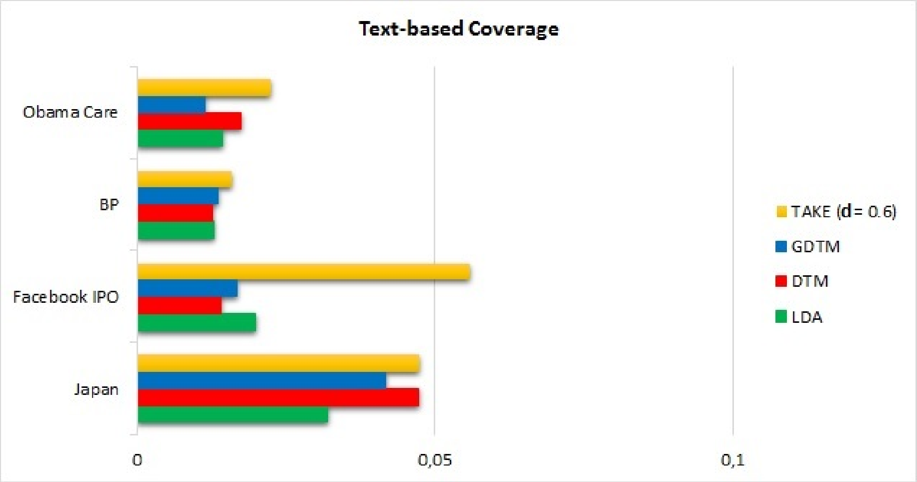


Figure 8: *Text-based Coverage of Wikipedia* Results.

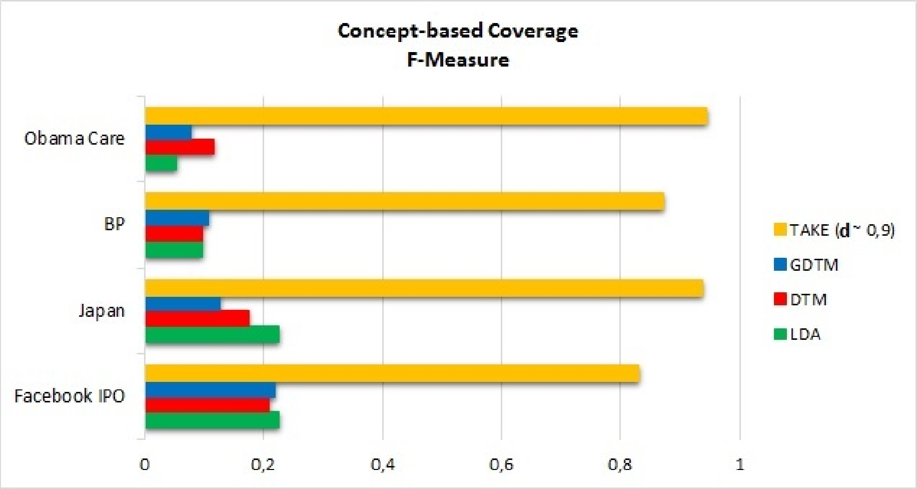


Figure 9: Concept-based Coverage of Wikipedia Results.

possibility to have more or less shortened summary ensuring a good trade off between quality and completeness both at syntactic and semantic level as shown by the experimental results.

8.1.3.3 Hashtag Coverage Results

Figure **??** shows the results obtained evaluating the *Hashtag Coverage* for the generated summaries with respect to the top fifty hashtags more frequently used in the tweet stream.

The results have been grouped by tweet streams and for each evaluated approach they are shown with different colors. As you can see in the figure, TAKE produces summaries with minimum hashtag coverage lower than other approaches only for the tweet stream of *BP Oil Spill.* While, it points out that the proposed approach outperform hashtag coverage of other approaches for the summaries generated with high level of *detail d*.

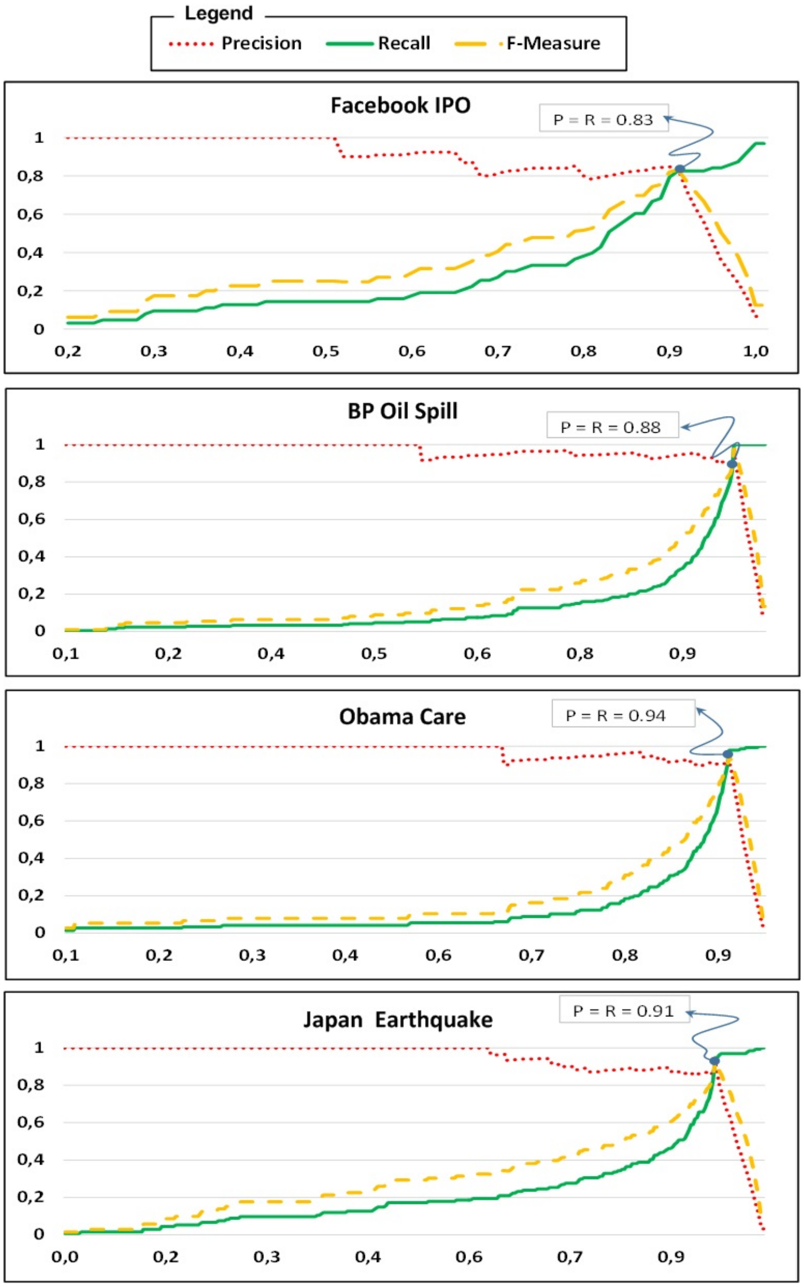


Figure 10: Precision, Recall and F-Measure curves of Concept-based Coverage of Wikipedia varying the level of *detail* in $[0 - 1]$.

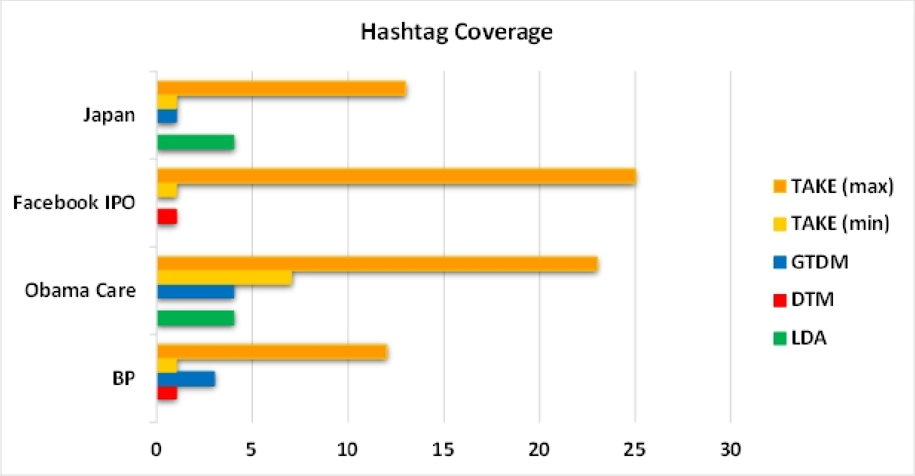


Figure 11: Hashtag Coverage Results.

8.2. Qualitative Evaluation

For a qualitative evaluation of resulting summaries, new tweet streams have been acquired via Twitter Stream API[9]. Totally 1,631,328 European tweets have been downloaded from April 7th to 14th, 2015 and we have selected subsets of tweets related to relevant world events occurred in that period. Specifically, in Table **??** we have shown the number of tweets selected for each of the selected three events (i.e., *"Hillary Clinton presidential campaign"*, *"7th Fast and the Furious movie"*, *"ISIS War"*).

Table 8.2: Number of tweets for each recently dataset

Name	**#Tweets**
Hillary Clinton	353
Fast and Furious 7	687
ISIS	764

The qualitative evaluation is assessed from human summarizers. In particular ten humans have been invited to read selected tweet sets (chronologically

[9] `https://dev.twitter.com/rest/public`

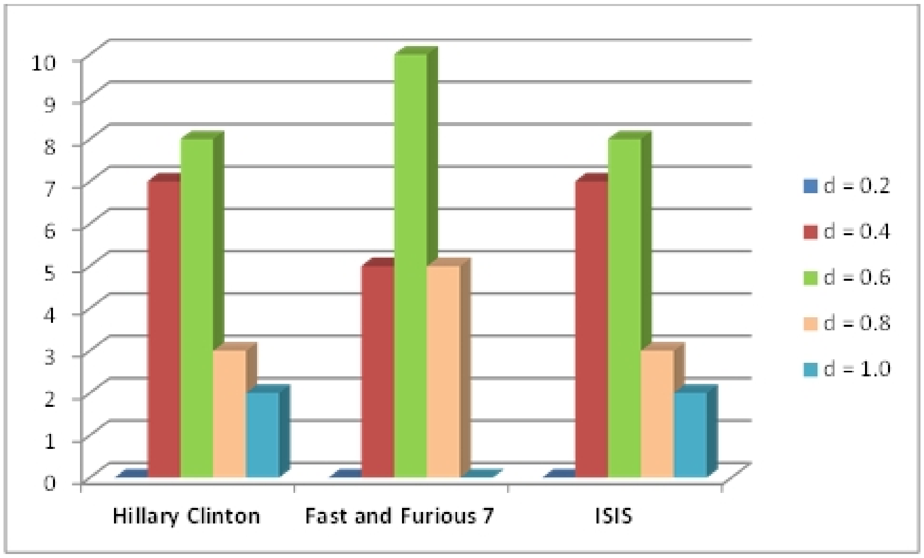


Figure 12: Qualitative assessment from human readers

ordered), along with the summary generated, for each event, at each of the five levels of *detail* ranging from the minimal level (d=0.2) to the maximum one (d=1.0) with step 0.2.

Summarizers were instructed to choose 1 or 2 out of the 5 summaries as the best representations of the event. Figure **??** shows the results of the human summarizers, with the x-axis differentiating among the three events and the y-axis showing the number of votes for each level of detail. By aggregating the total votes for all the events, the summaries with level of detail in [0.4, 0.6] have had the maximum number of votes.

Furthermore, we have performed summarization of tweets regarding Hillary Clinton (she has announced 2016 presidential run in 12th April) and the resulting summary for a level of *detail* d= 0.6 is shown in Listing **??**.

Listing 3: summarization of tweets regarding Hillary Clinton

```
Hillary Will Be The WOMEN'S Candidate! (Uh...Not) #HillarysBigAnnouncement #MyBirthRight #WarOnWomen #TCOT #CCOT
@NBCNews Tomorrow. Wait and see: NYC Mayor De Blasio will not endorse Hillary Clinton yet http://t.co/pjl1DHPlaC
```

```
3 #Democracy, On the Road Again To 2016? Hillary R. Clinton set to
    announce her 2016 presidential bid via social  and  accompanied by
    a video.
4 Long way to go yet. Don't think the Dems have much of an alternative to
    Hillary but feel Jeb B is the Rudy Giuliani of 2016
5 @GinoRaidy I'm not going to resort to unbacked arguments because Hillary
     isn't the current secretary of state, so why discuss Yemen?
6 @thehill It looks like John Kerry is making campaign for Hillary.. (?)
7 @TIMEPolitics: Hillary isn't the first. Here are five other women who
    ran for president http://t.co/N5cNAsQQ7H via @heroinebook #straffem
8 I understand that Hillary Clinton isn't too popular with Republican God
    fearing folk. Can't be sure if true it's just something I read
9 Hillary Clinton to declare for 2016 Democratic nomination
10 Hillary Clinton Set To Confirm Presidency Bid http://t.co/BfEV4U4qEv
11 Stop, think for a moment,,,,, 2016,,,,, Miliband in No−10   Hillary in
    the White House,, Pass the Gin bottle.
12 As Hillary Clinton approaches the 2016 race, she must decide how closely
     to align herself with President Obama
13 @KatrinaNation I'm inclined to believe Hillary does agree with McCain's
    position on Russia. She's the defence industry's ideal candidate.
14 Hillary Clinton or Jeb Bush (who at this moment is addressing the
    National Rifle Association)? http://t.co/hITx57oztN
15 Hillary had lead role in creating Adoption and Safe Families Act
    @Rickmayhem @TeaTraitors @lolalolita0 @p8triat @Centinel
    @TovarRasputin
16 Hillary helped start SCHIP (Children's Health Insurance) @Rickmayhem
    @TeaTraitors @lolalolita0 @p8triat @Centinel1787 @TovarRasputin
17 Hillary Clinton 4 President ? You could rendition me 2 your torture
    prisons; water board me or put me in Guantanamo but I'd never vote
    4 her
18 Pls sign/RT: Tell Secretary Hillary Clinton: Speak out against Fast
    Track and the TPP http://t.co/IW6kE8aukH http://t.co/ZwcoCOPWMi
```

9. Conclusion

This work defines Time Aware Knowledge Extraction methodology to support microblog summarization algorithm that has been applied on Twitter. The overall framework relies on Fuzzy Formal Concept Analysis introducing temporal correlation among tweets. Firstly, chronological ordered tweets have been analyzed to detect peaks of microblog activities around a specific topic. Secondly, tweet's content analysis exploits the service of wikification enabling semantic annotation of the text with wikipedia's entities. This enable us to treat time along with the meaning of the tweets at the same time to analyze knowledge evolution over the timeline during the extraction of microblog summary. Finally, a microblog summarization algorithm has been defined. It iteratively browses the timed fuzzy lattice of the concepts categorizing the tweet stream. At each iteration it incrementally selects the tweets corresponding to the concepts not yet included in the summary in order to cover the main arguments of the story developed over the time.

Specifically, the distinguishing feature introduced with this work is the level of *detail* that allows to filter the multitude of the concepts in the timed fuzzy lattice in order to zoom (in or out) the description of specific real world event. The *detail* enables the users to have more or less verbose update summary according to time constraints.

The framework has been validated comparing the obtained results with other existing methodologies, that are LDA (Latent Dirichlet Allocation), GDTM (Gaussian Decay Topic Model), and DTM (Decay Topic Model). As highlighted in [**?**], these methodologies outperform the LDA baseline by exploiting temporal correlation between tweets and their semantics at two different stages. The proposed framework outperforms the compared approaches considering at the same time temporal correlation among tweets and semantic of their content by means of Time Aware Knowledge Extraction ensuring good trade off between quality, completeness and redundancy.

Let us note that the wikify service that is exploited in the proposed approach to extract topics representing the meaning of the tweet content could be not able to identify meaning of emerging topics or named entities mentioned in the tweet if they are not available on the Wikipedia knowledge base snapshot used to train wikify service itself. So, one of the main weakness of the proposed approach is that it is limited to the Wikipedia knowledge base update.

Future works can exploit the Time Aware Knowledge Extraction methodology to address challenging research topics in the area of social media analytics, such as topic detection and monitoring, context-aware ad placement, and so on.

Furthermore, let us note that in the proposed work we didn't considered the following information available in the tweet metadata that could be exploited in the future works to better address microblog summarization as well as other analytics services on social media:

- *IRT*, it stands for In Reply To and it could be useful to analyse development of a discussion over the timeline;
- *Hashtags*, array of hashtags extracted from the Tweet text. It is a community-driven convention for adding additional context and metadata to your tweets. It makes it easier for users to find messages with a specific theme or content;
- *urls*, array of URLs extracted from the Tweet text. It can refers to image, video, web page etc;
- *user_mentions*, array of Twitter screen names extracted from the Tweet text;
- *media*, array of media file attached to the Tweet with the Twitter Photo Upload feature.

Another interesting future direction is to apply the verification techniques described in [**?**] for hierarchical structures to the FCA lattice to verify properties of the concepts. Moreover we plan to exploit the technique and tools used in [**?**] (specifically Timed Automata and non-repudiation protocol) to enforce a temporal and fairness criteria among the tweets received in the stream.

Acknowledgement

The authors are indebted to the authors of [**?**] for the provisioning of tweet streams and gold summary in order to enable the comparison evaluation of the proposed framework.

References

References

[1] N. A. Diakopoulos, D. A. Shamma, Characterizing debate performance via aggregated twitter sentiment, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, ACM, New York, NY, USA, 2010, pp. 1195–1198. doi:10.1145/1753326.1753504.
URL http://doi.acm.org/10.1145/1753326.1753504

[2] S. D., K. L., C. E., Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?, Horizon, In CSCW 2010.
URL http://www.research.yahoo.net/files/horizon4s-shamma.pdf

[3] K. Watanabe, M. Ochi, M. Okabe, R. Onai, Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, ACM, New York, NY, USA, 2011, pp. 2541–2544. doi:10.1145/2063576.2064014.
URL http://doi.acm.org/10.1145/2063576.2064014

[4] F. C. T. Chua, S. Asur, Automatic summarization of events from social media, in: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013., 2013.
URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6057

[5] B. Sharifi, M.-A. Hutton, J. K. Kalita, Experiments in microblog summarization, in: Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 49–56. doi:10.1109/SocialCom.2010.17.
URL http://dx.doi.org/10.1109/SocialCom.2010.17

[6] D. Gao, W. Li, X. Cai, R. Zhang, Y. Ouyang, Sequential summarization: A full view of twitter trending topics, Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22 (2) (2014) 293–302. `doi:10.1109/TASL.2013.2282191`.

[7] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, R. C. Miller, Twitinfo: Aggregating and visualizing microblogs for event exploration, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM, New York, NY, USA, 2011, pp. 227–236. `doi:10.1145/1978942.1978975`.
URL `http://doi.acm.org/10.1145/1978942.1978975`

[8] R. Mihalcea, A. Csomai, Wikify!: linking documents to encyclopedic knowledge, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, 2007, pp. 233–242.

[9] Y. Miao, C. Li, Enhancing query-oriented summarization based on sentence wikification, in: Workshop of the 33 rd Annual International, 2010, p. 32.

[10] S. Gong, Y. Qu, S. Tian, Summarization using wikipedia, TAC 2010 Proceedings.

[11] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 389–396.

[12] C. De Maio, G. Fenza, V. Loia, S. Senatore, Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis, Inf. Process. Manage. 48 (3) (2012) 399–418. `doi:10.1016/j.ipm.2011.04.003`.
URL `http://dx.doi.org/10.1016/j.ipm.2011.04.003`

[13] K. E. Wolff, States, transitions, and life tracks in temporal concept analysis, in: Formal Concept Analysis, Springer, 2005, pp. 127–148.

[14] Z. Lin, R. Lu, Y. Xiong, Y. Zhu, Learning ontology automatically using topic model, in: Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on, IEEE, 2012, pp. 360–363.

[15] P. Wang, H. Zhang, Y.-F. Wu, B. Xu, H.-W. Hao, A robust framework for short text categorization based on topic model and integrated classifier, in: Neural Networks (IJCNN), 2014 International Joint Conference on, IEEE, 2014, pp. 3534–3539.

[16] L. AlSumait, D. Barbará, C. Domeniconi, On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking, in: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE, 2008, pp. 3–12.

[17] X. Li, W. B. Croft, Improving novelty detection for general topics using sentence level information patterns, in: Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006, pp. 238–247.

[18] M. Krieger, D. Ahn, Tweetmotif: exploratory search and topic summarization for twitter, in: In Proc. of AAAI Conference on Weblogs and Social, 2010.

[19] B. Sharifi, M.-A. Hutton, J. Kalita, Automatic summarization of twitter topics, in: National Workshop on Design and Analysis of Algorithm, Tezpur, India, 2010.

[20] F. Liu, Y. Liu, F. Weng, Why is sxsw trending?: exploring multiple text sources for twitter topic summarization, in: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, 2011, pp. 66–75.

[21] D. Chakrabarti, K. Punera, Event summarization using tweets, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, 2011.

URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2885

[22] B. Sharifi, D. Inouye, J. K. Kalita, Summarization of twitter microblogs, Comput. J. 57 (3) (2014) 378–402. doi:10.1093/comjnl/bxt109.
URL http://dx.doi.org/10.1093/comjnl/bxt109

[23] F. Liu, Y. Liu, F. Weng, Why is ”sxsw” trending?: Exploring multiple text sources for twitter topic summarization, in: Proceedings of the Workshop on Languages in Social Media, LSM ’11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 66–75.
URL http://dl.acm.org/citation.cfm?id=2021109.2021118

[24] L. Shou, Z. Wang, K. Chen, G. Chen, Sumblr: Continuous summarization of evolving tweet streams, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’13, ACM, New York, NY, USA, 2013, pp. 533–542. doi:10.1145/2484028.2484045.
URL http://doi.acm.org/10.1145/2484028.2484045

[25] M. Dork, D. Gruen, C. Williamson, S. Carpendale, A visual backchannel for large-scale events, IEEE Transactions on Visualization and Computer Graphics 16 (6) (2010) 1129–1138. doi:http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.129.

[26] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, 1st Edition, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[27] L. A. Zadeh, Fuzzy sets, Information and control 8 (3) (1965) 338–353.

[28] K. GEORGE J, Y. Bo, Fuzzy sets and fuzzy logic, theory and applications, -.

[29] R. Krishnapuram, K. Kummamuru, Automatic taxonomy generation: Issues and possibilities, in: Fuzzy Sets and Systems—IFSA 2003, Springer, 2003, pp. 52–63.

[30] V. Torra, Fuzzy c-means for fuzzy hierarchical clustering., in: FUZZ-IEEE, 2005, pp. 646–651.

[31] K.-M. Lee, Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies, in: IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, IEEE, 2001, pp. 2977–2982.

[32] Y.-J. Horng, S.-M. Chen, Y.-C. Chang, C.-H. Lee, A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques, Fuzzy Systems, IEEE Transactions on 13 (2) (2005) 216–228.

[33] G. Bordogna, M. Pagani, G. Pasi, Soft computing for information retrieval on the web (2006).

[34] J. Zhai, Y. Liang, Y. Yu, J. Jiang, Semantic information retrieval based on fuzzy ontology for electronic commerce, Journal of Software 3 (9) (2008) 20–27.

[35] R. Y. Lau, D. Song, Y. Li, T. C. Cheung, J.-X. Hao, Toward a fuzzy domain ontology extraction method for adaptive e-learning, Knowledge and Data Engineering, IEEE Transactions on 21 (6) (2009) 800–813.

[36] C. De Maio, G. Fenza, M. Gaeta, V. Loia, F. Orciuoli, S. Senatore, Rss-based e-learning recommendations exploiting fuzzy fca for knowledge modeling, Applied Soft Computing 12 (1) (2012) 113–124.

[37] G. Stumme, Efficient data mining based on formal concept analysis, in: Database and Expert Systems Applications, Springer, 2002, pp. 534–546.

[38] R. Neouchi, A. Tawfik, R. Frost, Towards a temporal extension of formal concept analysis, in: E. Stroulia, S. Matwin (Eds.), Advances in Artificial Intelligence, Vol. 2056 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2001, pp. 335–344. `doi:10.1007/3-540-45153-6_33`. URL `http://dx.doi.org/10.1007/3-540-45153-6_33`

[39] P. Elzinga, K. Wolff, J. Poelmans, Analyzing chat conversations of pedophiles with temporal relational semantic systems, in: Intelligence and Security Informatics Conference (EISIC), 2012 European, 2012, pp. 242–249. doi:10.1109/EISIC.2012.12.

[40] S. La Torre, M. Napoli, M. Parente, G. Parlato, Verification of scope-dependent hierarchical state machines, Inf. Comput. 206 (9-10) (2008) 1161–1177. doi:10.1016/j.ic.2008.03.017.
URL http://dx.doi.org/10.1016/j.ic.2008.03.017

[41] M. Napoli, M. Parente, A. Peron, Specification and verification of protocols with time constraints, Electr. Notes Theor. Comput. Sci. 99 (2004) 205–227.