

Time Aware Knowledge Extraction for Microblog Summarization on Twitter

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

^aDepartment 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 noise and redundant. In general, social media analytics services have caught increasing attention from both side 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 distinguish 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, Wikification

*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[1], media events[2], and other real world events[3]. 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 the 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 [4, 5, 6]. 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 with the following problem. Given a topic-focused timestamped tweet stream and a level of *shrinking* s , the task is aimed to filter and chronologically order tweets in order to produce a Microblog Summary MS_s 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 closure level of *shrinking* .

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. More specifically, the summarization process is achieved taking into account both semantics and timestamps of the tweets. Firstly, content will be annotated via sentence wikification that is the practice of representing a sentence with a set of Wikipedia concepts (i.e., entries)[7, 8]. Secondly, it is possible to identify temporal peaks of tweet frequency analyzing timestamps and exploiting the Offline Peak-Finding Algorithm (OPAD), proposed in [9]. Then, taking into account the meaning of the tweet content and time dependences among detected peaks, temporal extension of Fuzzy Formal Concept Analysis [10, 11] 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 structure. The algorithm extracts chronologically ordered tweets summarizing main concepts of the story according to their temporal evolution.

Experimental Results. The proposed framework has been performed on the same tweet streams exploited in [4] 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 level of *shrinking* s in $[0 - 1]$. For all of the used metrics the system produces good performances. Specifically, the algorithm outperforms the results in [4] in terms of *Novelty Measurement* and *Text-based Coverage of Wikipedia*. Furthermore, evaluating *Concept-based Coverage of Wikipedia* setting level of *shrinking* with values ~ 0.9 (that is a verbose summary), the algorithm outperforms the results shown in [4] in terms of F-Measure, with optimal Recall and comparable values of Precision.

Outlines. 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 4 introduces the overall framework detailing each phase in the sections 5, 6 and 7; finally, Section 8 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 [12]. Unlikely, 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 [13] summarizes what's happening on Twitter providing a list of relevant terms that should explain Twitter topics. [14] and [5] extract a succinct summary for each topic using a phrase reinforcement ranking approach. [15] explores tweets and linked web contents to discover relevant information about topics. Moreover, [16] generates summaries especially for sport topics. Furthermore, [17] 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 [18] or clustering to perform the summarization of Evolving Tweet Streams [19]. 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 [20] and TwitInfo [9] allow users to graphically browse a large collection of tweets. Specifically, [20] visualizes conversations in Twitter data

using topic streams that is visually represented as stacked graphs and TwitInfo [9] uses a timeline-based display that highlights peaks of high tweet activity.

Considering our proposal we find some similarities in [4] and in [6]. Specifically, [4] 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 [6] 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 [6] 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 [10] introducing time dependencies among objects, in order to provide a summary that follow 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) [21]. FCA is a theoretical framework which supplies a basis for conceptual data analysis, knowledge processing and extraction. Fuzzy FCA [10] combines fuzzy logic into FCA representing the uncertainty through membership values in the range $[0, 1]$.

Following, some definitions about Fuzzy FCA are given.

Definition 1: A **Fuzzy Formal Context** is a triple $K = (G, M, I = \varphi(G \times M))$, where G is a set of objects, M is a set of attributes, and I is a fuzzy set on domain $G \times M$. Each relation $(g, m) \in I$ has a membership value $\mu(g, m)$ in $[0, 1]$.

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 $\Phi(O) = \{A_1(\mu_1), A_2(\mu_2), \dots, A_m(\mu_m)\}$, where $\{A_1, A_2, \dots, A_m\}$ 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 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 1(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 1(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

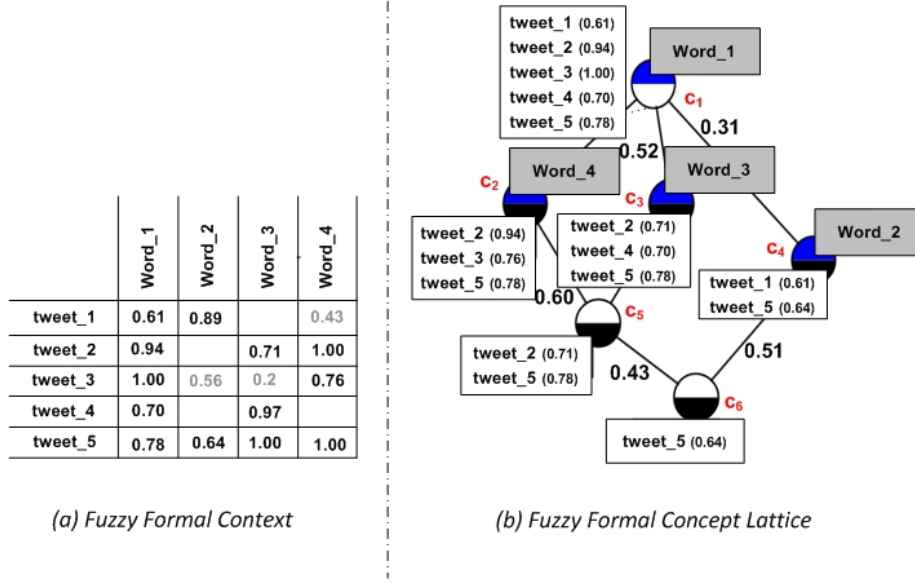


Figure 1: Portion of fuzzy formal context (a) and the relative concept lattice with threshold $T = 0.6$ (b)

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:

Definition 3: Fuzzy Formal Concept. Given a fuzzy formal context $K=(G, M, I= \varphi(G \times M))$ and a confidence threshold T , we define $A^* = \{m \in M \mid \forall g \in A: \mu(g, m) \geq T\}$ for $A \subseteq G$ and $B^* = \{g \in G \mid \forall m \in B: \mu(g, m) \geq T\}$ for $B \subseteq M$. A fuzzy formal concept (or fuzzy concept) of a fuzzy formal context K with a confidence threshold T is a pair $(A_f = \varphi(A), B)$, where $A \subseteq G$, $B \subseteq M$, $A^*=B$ and $B^*=A$. Each object $g \in \varphi(A)$ has a membership μ_g defined as

$$\mu_g = \min_{m \in B} \mu(g, m)$$

where $\mu(g, m)$ is the membership value between object g and attribute m , which is defined in I . Note that if $B=\{\}$ then $\mu_g = 1$ for every g . A and B are the extent and intent of the formal concept $(\varphi(A), B)$ respectively.

Definition 4: Let (A_1, B_1) and (A_2, B_2) be two fuzzy concepts of a fuzzy formal context (G, M, I) . $(\varphi(A_1), B_1)$ is the **subconcept** of $(\varphi(A_2), B_2)$,

denoted as $(\varphi(A_1), B_1) \leq (\varphi(A_2), B_2)$, if and only if $\varphi(A_1) \subseteq \varphi(A_2)$ ($\Leftrightarrow B_2 \subseteq B_1$). Equivalently, (A_2, B_2) is the **superconcept** of (A_1, B_1) .

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).

An example of Fuzzy Formal Concept is C_4 that is composed of objects $A_f = tweet_1, tweet_5$ and attributes $B = \text{"word}_1, word_2'', \dots$) with $\mu_{tweet_1} = 0.61$ and $\mu_{tweet_5} = 0.64$, as shown in 1(b). Furthermore, Fuzzy FCA carries out Fuzzy Concept Lattice, i.e. a hierarchical structure of the concepts according to the order relation (see Definition 4), as shown in Figure 1(b). For instance, let us observe in Figure 1(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 .

Now, it is possible to define Fuzzy Concept Lattice as follows:

Definition 5: *A Fuzzy Concept Lattice of a fuzzy formal context K with a confidence threshold T is a set $F(K)$ of all fuzzy concepts of K with the partial order \leq with the confidence threshold T .*

Figure 1(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 [21]. Furthermore, the Fuzzy FCA introduces the definition of Fuzzy Formal Concept Similarity.

Definition 6: *Fuzzy Formal Concept Similarity between concept $K_1 = (\mu(A_1), B_1)$ and its subconcept $K_2 = (\mu(A_2), B_2)$ is defined as*

$$E(K_1, K_2) = \frac{|\varphi(A_1) \cap \varphi(A_2)|}{|\varphi(A_1) \cup \varphi(A_2)|}$$

where \cap and \cup refer intersection and union operators¹ on fuzzy sets, respectively.

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 6.

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 2 sketches the whole process of the system that is composed of following main phases:

- *Microblog Content Analysis* (see Section 5). It takes as input a tweet stream and detects tweet frequency peaks, then performs tweet’s features

¹The 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 $\mu_B(x)$, $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$ and $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$.

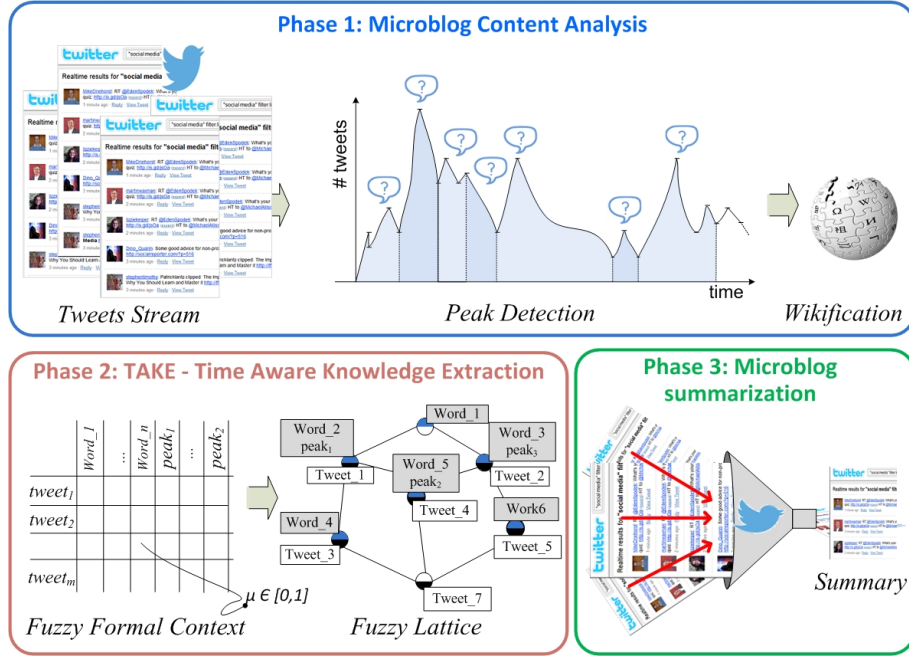


Figure 2: Overall Process of the framework

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 6). It takes as input term weighted tweets and their timestamps and performs Time Aware FFCA in order to arrange tweets into a hierarchy carrying out also time dependence relation among extracted concepts;
- *Microblog Summarization Algorithm* (see Section 7). 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 *shrinking* level specified as input (See later for the definition and discussion of *shrinking*).

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 *shrinking*. 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 [9]. 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 [9].

Given a time-sorted collection of tweets, the algorithm locates surges by tracing tweet volume changes. Let $T = (t_1, t_2, \dots, t_n)$ 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 $C = (C_1, C_2, \dots, C_t)$ where C_i is the number of tweets in bin i . The objective is identify each bin i such that C_i is large relative to the recent history $C_{i-1}, C_{i-2}, \dots, C_1$.

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., $\frac{|C_i - \text{mean}|}{\text{meandev}} > \tau$), 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 [9].

Listing 1: OPAD- Offline Peak Area Detection

```

1  windows = []
2  mean = C1
3  meandev = variance ( C1, ... , Cp )
4
5  for i = 2; i < len(C); i++ do
6      if  $\frac{|C_i - \text{mean}|}{\text{meandev}} > \tau$  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, Ci )

```

```

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 ++
19         end if
20     end while
21     if ( $C_i < C_{start}$ ) then
22         end = i --
23     windows.append(start, end)
24 else
25     (mean, meandev) = update(mean, meandev,  $C_i$ )
26 end if
27 end for
28 return windows
29
30 function update(oldmean, oldmeandev, updatevalue):
31     diff = |oldmean - updatevalue|
32     newmeandev =  $\alpha * diff + (1 - \alpha) * oldmeandev$ 
33     newmean =  $\alpha * updatevalue + (1 - \alpha) * oldmean$ 
34 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 [7]. 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:

$tweet_i = \text{"President Obama just designated the largest marine reserve in the world"}$.

The wikification process extracts from the above text a set of $\langle \text{topic}, \text{relevance} \rangle$ pairs. These pairs are features characterizing meaning of the input text. Taking into account the example above, the extracted topic (shown in Figure 3) are:

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

Then, at this point, considering the example defined in Section 5.1 about $tweet_i$, the content will be annotated via sentence wikification as:

$$tweet_i = \{ \langle \text{peak}_i \rangle \} \cup$$

$$\{ \langle \text{topic}_{i_1}, \text{relevance}_{i_1} \rangle, \langle \text{topic}_{i_2}, \text{relevance}_{i_2} \rangle, \dots, \langle \text{topic}_{i_m}, \text{relevance}_{i_m} \rangle \}$$

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 5.1) and the wikification of tweet content (see Section 5.2), *ITF* (i.e., In-

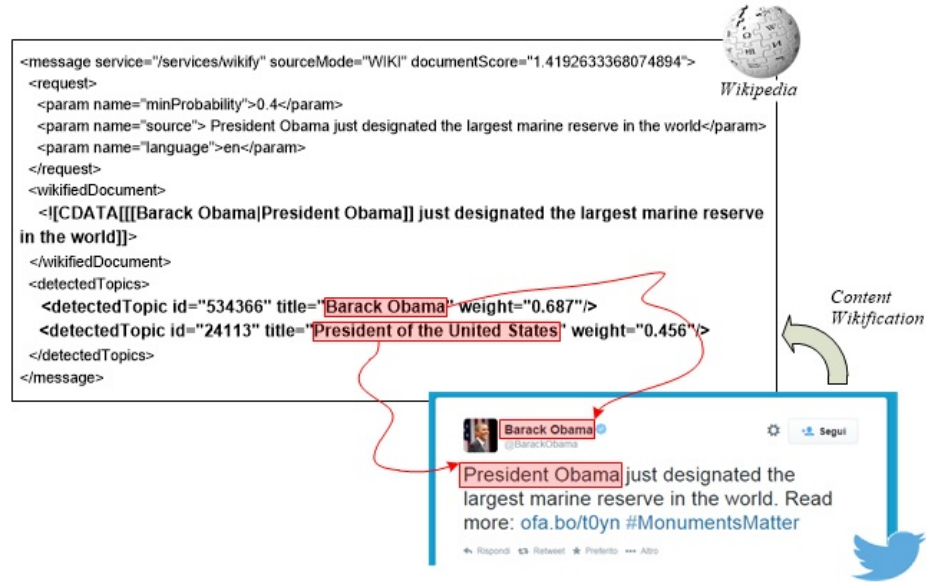


Figure 3: Example of tweet's content wikification.

verse *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 5.2) provides whether it is common or rare across all tweets. Specifically, let $W = \{w_1, w_2, \dots, w_n\}$ be the set of topics extracted by means of wikification process from set of tweets $T = \{t_1, t_2, \dots, t_m\}$. Let us compute the ITF for each one topic as:

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

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:

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

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

$$\text{tweet}_i = \{\langle \text{peak}_i \rangle\} \cup \{\langle \text{topic}_j, f_{rel_j} \rangle, \langle \text{topic}_{j+1}, f_{rel_{j+1}} \rangle, \dots, \langle \text{topic}_m, f_{rel_m} \rangle\}$$

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 [22]. Specifically, this temporal extension has been applied in [23] to search pedophiles

on the Internet analyzing chat conversation over the 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 $K_t = (G, M^+ = M \cup T, I_M = \varphi(G \times M), I_T)$, 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 $t \in T$.

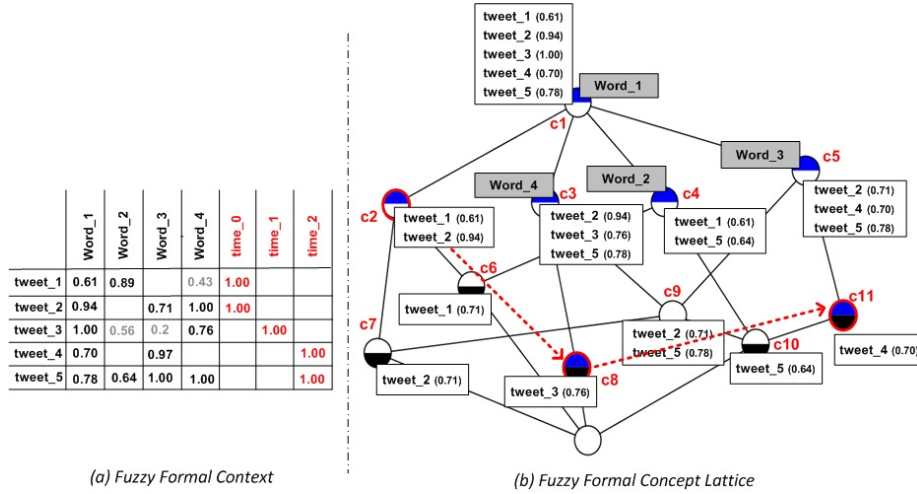


Figure 4: Time Aware Fuzzy FCA: portion of fuzzy temporal fuzzy formal context (a) and the relative temporal fuzzy concept lattice

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 4. Specifically, Figure 4(b) emphasizes that each node (i.e., a formal concept) includes the objects, attributes and time attributes. For example in the lattice in Figure 4(b), a concept is $(A_f = tweet_1, tweet_2, B = \text{"word}_1, time_0'')$ with $\mu_{tweet_1} = 0.61$ and $\mu_{tweet_2} = 0.94$.

The resulting timed fuzzy lattice emphasizes a temporal correlation among concepts and highlights how the concepts change over the timeline (Figure 4). To represent the concept development over the timeline in a timed fuzzy lattice have been introduced temporal edges (in Figure 4 red dashed arrows) among related concepts. The temporal edges allow the evolution of attributes to be followed over the time. A temporal precedence relation is defined over time points. The direction of the arrow indicates this precedence. In the lattice in Figure 4(b), the evolution of attributes is represented as: $c_2 \rightarrow c_8 \rightarrow c_{11}$, i.e., $\{Obama\} \rightarrow \{Obama, election\} \rightarrow \{Obama, President\}$.

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 $w(C)$ defined later.

More formally, given a fuzzy formal context $K=(G, M, I= \varphi(G \times M))$ and fuzzy formal concept $C = (\varphi(A), B)$ (see Section 3), the level of *shrinking* of

formal concept C is defined as:

$$s(C) = 1 - \text{support}(C) = 1 - \frac{|A|}{|G|}$$

where $|A|$ is the number of tweets in C and $|G|$ is the number of tweets in the overall stream. Then, the weight $w(C)$ of fuzzy formal concept C will be evaluated as follows:

$$w(c) = \frac{\sum_{m \in B} \mu_m}{|B|}$$

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

$$\mu_m = \max_{g \in \varphi(A)} \mu(g, m)$$

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 2. First of all, the sets of covered attributes (i.e., CA), covered concepts (i.e., CC) and summary (i.e., MS_s) are initialized as empty set (line 4-6). Then, the algorithm selects concepts of the timed fuzzy lattice whose *shrinking* is greater than a threshold s 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 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). At the end of each iteration, the algorithm includes t_{max} in the resulting summary (i.e., MS_s) (line 16) and updates the set of covered attributes CA (line 17) and the set of covered concepts CC (line 18).

Listing 2: Summarization by Time Aware FFCA

```

1 Input: timed fuzzy lattice  $L$ , peak areas  $P$ , and shrinking level  $s$ .
2 Output: a microblog summary  $MS_s$  of  $|T|$  tweets;
3
4  $CA = \emptyset$ ,
5  $CC = \emptyset$ ,
6  $MS_s = \emptyset$ 
7
8  $C^* = \{c_i \in L \mid s(c_i) = 1 - \text{support}(c_i) > s\}$ 
```

```

9  $P = (p_1, p_2, \dots, p_n) \quad \forall i, j: p_i > p_j \Leftrightarrow i < j$ 
10  $C_{p_i}^* = \{c_i \in C^* \mid c_i = (A, B), p_i \in B\}$ 
11
12 for  $i = 1; i < \text{len}(P); i++$  do
13   while  $C_{p_i}^* \setminus CC \neq \emptyset$ 
14      $c_{max} = \underset{c_i \in (C_{p_i}^* \setminus CC)}{\text{argmax}} \left( w(c_i) = \frac{\sum_{m \in (B \setminus CA)} \mu_m}{|B|} \right)$ 
15      $t_{max} = \underset{g \in c_{max}}{\text{argmax}} (\mu_g)$ 
16      $MS_s = MS_s \cup t_{max}$ 
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 *shrinking* $s = 70\%$. Figure 5 shows the concepts with level of *shrinking* greater than 70% resulting from the execution of line 8 in Listing 2. Table 7.1 lists the set of candidate concepts grouped by peak area to which they belong to.

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

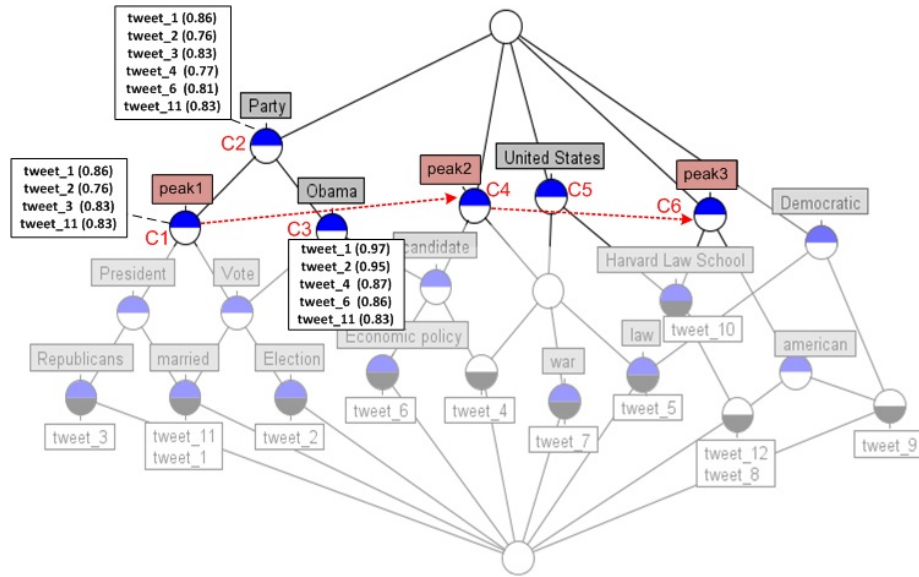


Figure 5: Example of timed Fuzzy Concept Lattice

Table 7.1: Concepts of the timed fuzzy lattice in Figure 5 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

and the concept with maximum value of w will be selected (see Listing 2, 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*.

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 2, 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. *peak2*. At the end of the execution the resulting summary will be composed of following tweets:

$$MS_s = \{tweet_1, tweet_7\}.$$

8. Framework Evaluation

This section details the experimental results obtained performing the proposed summarization algorithm on specific tweet streams. As said before, the summarization algorithm relies on timed fuzzy lattice of tweets resulting from Time Aware Knowledge Extraction methodology execution. Since the timed fuzzy lattice allows to perform the summarization algorithm with different lev-

els of *shrinking*, the results have been evaluated by varying these levels. In particular, the higher the level of *shrinking* the more detailed the resulting summary will be, that is the summary will include a greater amount of tweets.

The discussion will continue as follows: description of the dataset of tweets (i.e. tweet streams) on which the framework has been executed (Section 8.1), definition of evaluation measures (Section 8.2), and finally the experimental results will be discussed (Section 8.3).

8.1. Tweet Streams

The summarization framework has been applied on tweet streams focused on four real-world events²: Facebook IPO³, Obamacare⁴, Japan Earthquake⁵ and BP Oil Spill⁶. 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 8.1 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.

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

²Specifically, the data have been provided by authors of [4]

³http://en.wikipedia.org/wiki/Initial_public_offering_of_Facebook

⁴http://en.wikipedia.org/wiki/Patient_Protection_and_Affordable_Care_Act

⁵http://en.wikipedia.org/wiki/2011_T%C5%8Dhoku_earthquake_and_tsunami

⁶http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill

8.2. Measurements

The proposed framework has been evaluated considering the following metrics:

- *Novelty Measurements*, specifically *Sequence Novelty Measurement* introduced in [6] and *Historical Novelty Measurement*.

- *Sequence Novelty Measurement* measures average novelty among chronologically 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.

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

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

$$Novelty = \frac{1}{|D| - 1} \sum_{i>1} \left(I_{d_i} - \left[\left(\bigcup_{k<i} I_{d_k} \right) \cap I_{d_i} \right] \right) \quad (2)$$

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 $k = 1 \dots |D|$ correspond to all tweets in the summary.
- *Text-based Coverage of Wikipedia*, introduced in [4] 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 Wikipedia⁷. 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:

$$g_n = \frac{1}{|NG_n^{gold}|} \sum_{ng \in NG_n^{gold}} \min(|ng \in NG_n^{gold}|, |ng \in NG_n^{gen}|) \quad (3)$$

$$Sim(S^{gold}, S^{gen}) = 0,2 \cdot g_1 + 0,3 \cdot g_2 + 0,5 \cdot g_3 \quad (4)$$

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 sum-

⁷Indeed, gold summaries have been provided by authors of [4]. 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 [7, 8]. More formally, let $C_{gold} = \{c_1, c_2, \dots, c_m\}$ and $C'_{gen} = \{c'_1, c'_2, \dots, c'_n\}$ 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:

$$P = \frac{|C_{gold} \cap C'_{gen}|}{|C'_{gen}|} \quad R = \frac{|C_{gold} \cap C'_{gen}|}{|C_{gold}|} \quad (5)$$

Then, F-measure F is computed as follows:

$$F = 2 \times \frac{P \times R}{P + R} \quad (6)$$

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.

8.3. 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 [4]. Since TAKE produces different summaries with different level of *shrinking*, the results have been evaluated by varying the level of s in $[0 - 1]$ and for all of the used metrics the system reveals good performances.

8.3.1. Novelty Results

Figures 6 and Figure 7 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 level of *shrinking* s in $[0 - 1]$ and plotting the obtained minimum and maximum values for both novelty measures.

Figure 6 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 7 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 enough close 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 shorten summaries with acceptable levels of redundancy. Thus, the proposed method incrementally includes tweets in the resulting summary introducing significant amount of novel con-

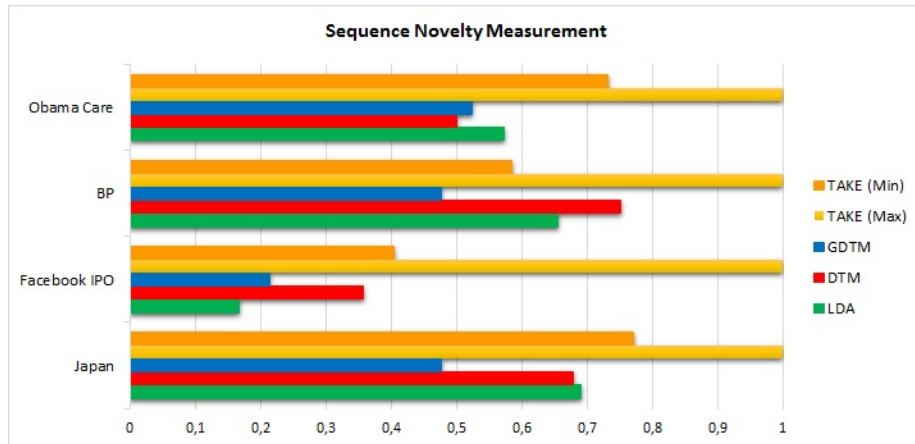


Figure 6: Sequence Novelty Measurement Results.

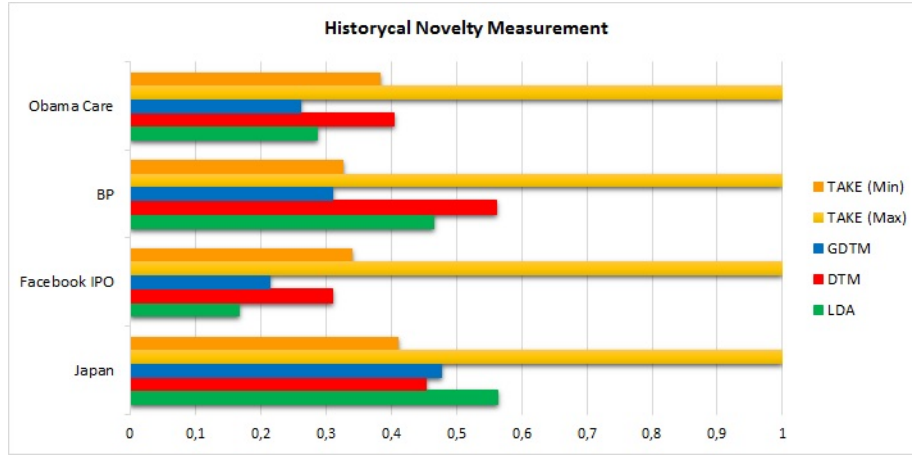


Figure 7: Historical Novelty Measurement Results.

cepts improving the description of the event according to its development over the timeline.

8.3.2. Text-based and Concept-based Coverage Results

Figure 8 and Figure 9 show the results obtained evaluating *Text-based Coverage of Wikipedia* and *Concept-based Coverage of Wikipedia*, respectively. These

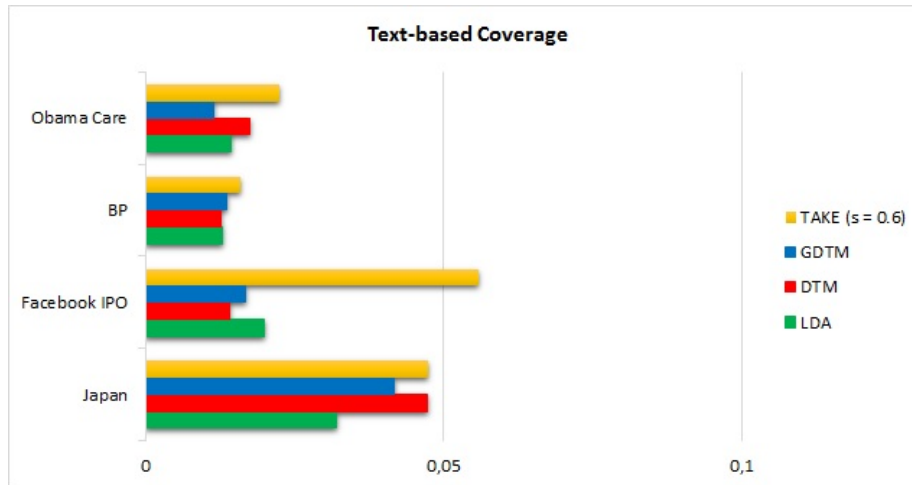


Figure 8: *Text-based Coverage of Wikipedia* Results.

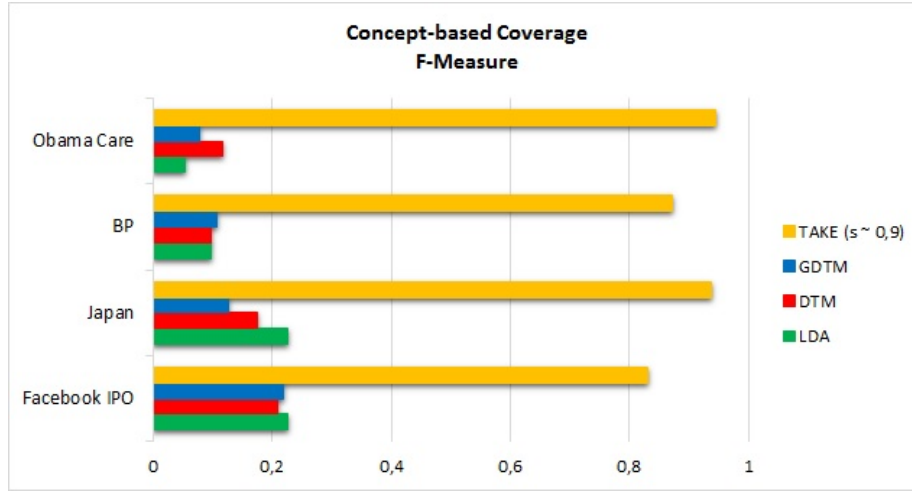


Figure 9: Concept-based Coverage of Wikipedia Results.

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 *shrinking*, in Figure 8 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 *shrinking* to 0.6. For levels of *shrinking* 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 9 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 10 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 *shrinking* s from 0.0 to 1.0. The figure highlights that TAKE reveals valuable performance

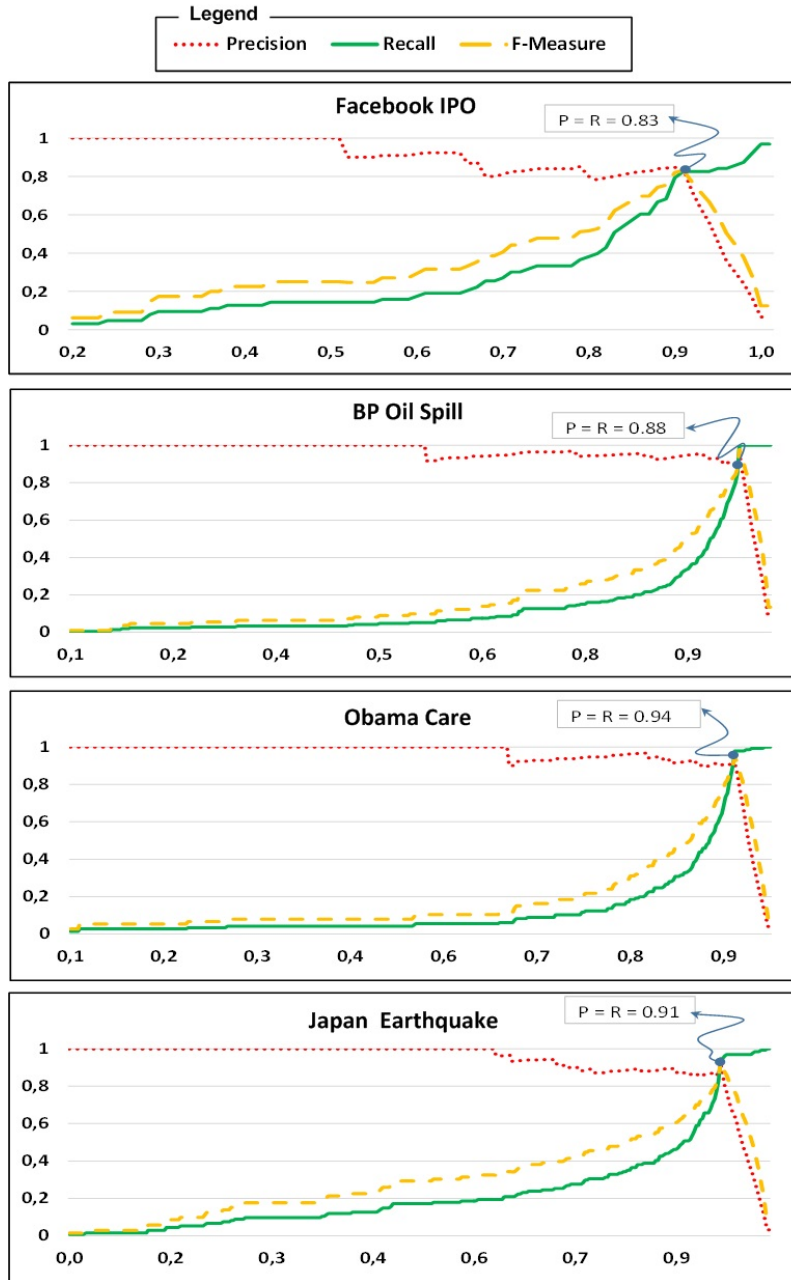


Figure 10: Precision, Recall and F-Measure curves of Concept-based Coverage of Wikipedia varying the level of *shrinking* in $[0 - 1]$.

in terms of Recall with acceptable values of Precision with level of *shrinking* between 0.7 and 0.9.

In general, the distinguishing feature introduced by TAKE approach is the possibility to have more or less shorten summary ensuring good trade off between quality and completeness both at syntactic and semantic level as shown by the experimental results.

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 service of wikification enabling semantic annotation of the text with wikipedia’s entities. Finally, a microblog summarization algorithm has been defined walking across concepts of the resulting timed fuzzy lattice in order to select right tweets covering main concepts of the story and their development over the timeline. Specifically, the distinguishing feature introduced with this work is the level of *shrinking* 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 *shrinking* 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 [4], 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.

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. Another interesting future direction is to apply the verification techniques described in [24] for hierarchical structures to the FCA lattice to verify properties of the concepts.

Acknowledgement

The authors are indebted to the authors of [4] 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] 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.
- [8] Y. Miao, C. Li, Enhancing query-oriented summarization based on sentence wikification, in: Workshop of the 33 rd Annual International, 2010, p. 32.
- [9] 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>
- [10] C. D. 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>

- [11] K. E. Wolff, States, transitions, and life tracks in temporal concept analysis, in: *Formal Concept Analysis*, Springer, 2005, pp. 127–148.
- [12] 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.
- [13] M. Krieger, D. Ahn, Tweetmotif: exploratory search and topic summarization for twitter, in: *In Proc. of AAAI Conference on Weblogs and Social*, 2010.
- [14] B. Sharifi, M.-A. Hutton, J. Kalita, Automatic summarization of twitter topics, in: *National Workshop on Design and Analysis of Algorithm*, Tezpur, India, 2010.
- [15] 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.
- [16] 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>
- [17] 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>
- [18] 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>

- [19] 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>

- [20] 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>.

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

- [22] 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

- [23] 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.

- [24] 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>