# Towards OLAP Analysis of Multidimensional Tweet Streams

Alfredo Cuzzocrea
DIA Department, University of Trieste and ICAR-CNR
Italy
alfredo.cuzzocrea@dia.units.it

Carmen De Maio
DI Department, University of Salerno
Italy
cdemaio@unisa.it

Giuseppe Fenza
DI Department, University of Salerno
Italy
gfenza@unisa.it

Vincenzo Loia
DI Department, University of Salerno
Italy
loia@unisa.it

Mimmo Parente
DI Department, University of Salerno
Italy
parente@unisa.it

## ABSTRACT

Social media and networks are used by millions of people to share with their friends across the world: tastes, opinions, ideas, etc. The volume and the speed at which these data are produced make it a challenging task to discover meaningful patterns in the data. Nevertheless, very interesting business goals could be achieved collecting these data and performing analytics on social media data streams, such as: addressing marketing strategies, targeting advertisements, and so forth. We emphasize that there is a need to investigate and define suitable knowledge mining approaches to go beyond explicitly available metadata by analyzing unstructured data to provide intelligent analytics services. Specifically, in this paper we provide first results on applying OLAP analysis to *multidimensional Tweet streams*.

## Categories and Subject Descriptors

H.2.7 [**DATABASE MANAGEMENT**]: Database Administration—*Data warehouse and repository*

## General Terms

Algorithms, Design, Management, Performance, Theory

## Keywords

Social Media Analytics, Knowledge Discovery, OLAP Analysis, Advanced Analytics

## 1. INTRODUCTION

Nowadays, people extensively use social networks (e.g., facebook, twitter, etc.) and multimedia aggregators (e.g., youtube, wikipedia, etc.) to share with their friends across the world: tastes, opinions, ideas, and so on. Millions of tweets about brands, news and so forth, hundreds of thousands of Facebook "likes" and check-ins on Foursquare, happen every day. So, we are heading into a social media data explosion. Collecting these data and performing analytics on social media data streams is one of the main challenge of big data because very interesting business goals could be achieved, such as: addressing marketing strategies, profiling people tastes, targeting advertisements, and so forth.

Specifically, this work is focused on the metadata available in Twitter and we point out that there is a need to investigate and define suitable knowledge mining approaches to go beyond explicitly available metadata analysing unstructured data in order to provide intelligent analytics services to support decision making.

Twitter is the most popular microblogging service with more than half billion of users. Recent statistics pointed out that the average number of tweets per day is more than 100 million, and thousands of them happen every minute. So, tweets exchanged over the Internet represent an important source of information. Tweets are associated with meta-information that cannot be included in messages (e.g., date, location, etc.) or included in the message in the form of tags having a special meaning. Tweets can be represented in a multidimensional way by taking into account all this meta-information as well as associated temporal relations. In Fig.1 there is a map of the overall metadata that accompanies every tweet. Twitter's API enable us to acquire public tweet metadata (e.g., Twitter's Search API[1] and Streaming API[2]). In general, we can distinguish two main groups of information explicitly available in the tweet metadata: (*i*) *User's Profile*: information about the user in terms of appearance, location, followers and following; (*ii*) *Tweet's Status*: information about the tweet in terms of content, timestamp, if the tweet is In Reply To (IRT) another one, location of the tweet, media object contained in the tweet (audio, video, image), and so on.

In this paper, we focus on the definition of multidimensional data model for the storage of tweet data stream en-


*DOLAP'15,* October 23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2811222.2811233.

[1] `https://dev.twitter.com/rest/public`
[2] `https://dev.twitter.com/streaming/overview`

abling OLAP analysis. Furthermore, we exploit the implicit information that could be derived or discovered in the tweets by investigating beyond explicit metadata.

Datawa rehouse model enables us to manipulate a set of indicators (measures) according to different dimensions which may be provided with one or more hierarchies. Associated operators (e.g., Roll-up, Drilldown, etc.) allow an intuitive navigation on different levels of the hierarchy. Indeed, the collected tweets will be represented by means of fuzzy mathematical model in order to extract timed fuzzy lattice through fuzzy extension of FCA[9].

## 2. BEYOND TWEET METADATA

In order to allow analytics services on the metadata of the tweets, a mapping of semi-structured data to multidimensional cubes has to be defined. An example of this mapping has been introduced in [17] and it takes into account explicitly available data in the tweet metadata. Specifically, in [17] the authors distinguish three main kinds of information that could be used to build analytics on tweet streams: (*i*) *Static*: information explicitly available in the tweet metadata; (*ii*) *Derived*: information inferred from the existing static ones – for instance, the *hashtags* can be used as a derived tweet dimension to group instances and to allow aggregation paths in OLAP queries (e.g., #worldcup, etc.); (*iii*) *Discovered*: information obtained performing knowledge mining and classification tasks.

Considering the maturity of data warehousing systems, there will be more demand for advanced and intelligent support to derive useful information integrating also unstructured data in Data Warehousing, as highlighted in [15] and [13]. In particular, some information can be discovered exploiting third-party text analysis APIs, such as: (*i*) *Language Detection*: this information adds the tweet's language as a dimension of the tweet record enabling cross-language analysis and aggregation – this task could be easily accomplished invoking external API, such as: MyMemory[3], Google API[4], Alchemy API[5], and so forth; (*ii*) *Topic, Entity and Event detection*: a tweet is enriched by storing along with it other information as: detected significant topics, entities (e.g., persons, locations, dates, products, etc.) and events (e.g., natural disasters, sports competitions, etc.). This information provides a multidimensional perspective and refines the grain of the data from a tweet record down to single terms (see Section 3.3); (*iii*) *Sentiment Detection*: it assesses the overall emotion of the content (i.e., positive, negative or neutral). AlchemyAPI and OpenCalais[6] are examples of platforms enabling this type of analysis.

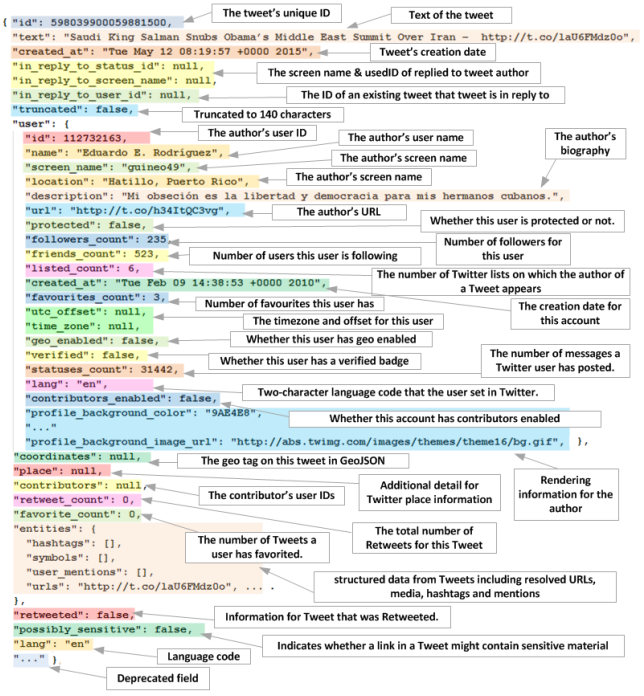


**Figure 1: A Descriptive Map of Tweet Metadata (http://www.slaw.ca/2011/11/17/the-anatomy-of-a-tweet-metadata-on-twitter/)**

Knowledge Mining algorithms may be helpful for discovering less obvious or hidden relationships and patterns in the dataset. The underlying dataset can be mined for a variety of descriptive and predictive tasks to build respective classification models. For example, users or tweets in the underlying dataset can be clustered into various groups based on their popularity, reputation, tweeting activity, topics discussed, etc. These discovered groups are said *analytics* as they offer new perspectives for multidimensional analysis and can be used as grouping criteria just like statically defined dimension categories.

Considering unstructured data, in the last decade several methods, aimed to support ontology learning from domain data, have been defined. Some of them use text-mining and machine learning techniques in order to conceptualize unstructured content [10], [2]. Others methodologies exploit Conceptual Data Analysis theory for knowledge structuring and ontology building [3]. Most of them are language dependent and doesn't exploit Commonsense Knowledge base (e.g., Wikipedia etc.) to enrich meaning of the analyzed content. In [9] the Fuzzy extension of Formal Concept Analysis (FFCA) theory has been exploited to build hierarchical classification of the collected resources. Furthermore, wikification service has been integrated in the text-analysis in order to enrich the meaning of the content by linking wikipedia entities (i.e., articles) [19]. Other approaches exploit probabilistic models and Latent Dirichlet Allocation (LDA) for ontology learning [24], [23]. The extraction of an unsupervised ontology could be a hierarchical dimension of OLAP cube.

In some cases Text Analysis API and Knowledge Mining methods cited above could be used to introduce hierarchical or flat dimensions. For instance topics dimension could be hierarchically structured according to a domain ontology as proposed in [1] where MeSH[7] (Medical Subject Headings) ontology has been used to index tweets. In the following Section, a case study about nesting knowledge mining technique (i.e., Fuzzy Formal Concept Analysis) in OLAP data model for multidimensional tweet stream will be proposed to address *summarization of microblogs*.

[3] http://mymemory.translated.net/doc/spec.php
[4] https://cloud.google.com/translate/v2/using_rest\#detect-language
[5] http://www.alchemyapi.com/
[6] http://www.opencalais.com/
[7] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

# 3. MULTIDIMENSIONAL ANALYSIS OF TWEET STREAMS

This Section describes the cube built taking into account the tweet stream.

## 3.1 Tweet Cube

A data cube can model an analysis subject called Fact $F$ defined on the schema $D = \{T_1, T_2, \ldots, T_n, M\}$ where $T_i$ $(i = 1, \ldots, n)$ are several dimensions and $M$ stands for a measure. Every dimension defined on a domain $D$ includes attributes $A = \langle a_1, a_2, \ldots, a_m \rangle$ that can be organized in several hierarchical levels $\langle l_1, l_2, \ldots, l_m \rangle$. In the proposed cube of tweets, the tweet represents the fact and some of its metadata explicit or implicit will be selected as hierarchical or flat dimensions.

## 3.2 Dimensions

In our tweet cube model, the dimensions can be classified in two types: $(i)$ *Semantic Dimension*: In our model, this dimension is extracted from Wikipedia knowledge base and it makes use of the titles of Wikipedia articles and of the Wikipedia category graph. Wikipedia provides a knowledge base for computing word relatedness in a more structured fashion than a search engine and with more coverage than WordNet. Obviously, semantic dimension could be extracted from other domain ontologies or vocabularies related to the dimension area as external knowledge source (e.g. GeoNames, WordNet, etc.). Each level $l_i = \langle c_1, c_2, \ldots, c_n \rangle$ includes a set of concepts $c_j$ $(j \in [1; n])$ extracted from Wikipedia category graph. $(ii)$ *Metadata Dimension*: Metadata refers to the information about the tweet that can be derived from its metadata, such as: timestamp, user, hashtag, location, etc. So, we have to create a metadata dimension to represent each metadata information that we want consider. In this case we focus on time and location that are both hierarchical dimensions.

## 3.3 Measures

Tweets are significant source of evidence when extracting information related to the reputation of a particular entity (e.g., a particular politician, singer or company) or, more general, a topic. In order to analyze the textual content of tweets, there is a need of an automated methods to disambiguate tweets with respect to entity or topic names in their content. To address this issue we propose a measure which exploit wikification service, wikify[8], that is the practice of representing a sentence with a set of Wikipedia concepts [19], [18]. 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 uncertainty degree of the disambiguation results.

Specifically, the tweet content is wikified to extract a set of $\langle topic, rd \rangle$ pairs corresponding to Wikipedia articles that are related to the tweet content itself with a specific relevance degree $(rd)$ [19]. Let us report an example by considering the following i-th tweet in the tweet stream:
*$tweet_i$ = "Hillary Clinton is running for president to be a champion for everyday Americans. Join Hillary for America today.".*

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

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 are:

$\langle Hillary\ Rodham\ Clinton, 0.94 \rangle$, $\langle Bill\ Clinton, 0.24 \rangle$

Hence, let $S$ be the vector space defined by the set of *topics*:

$$S \langle topic_1, topic_2, \ldots, topic_n \rangle.$$

A $tweet_i$ in a multidimensional space is represented by a weights vector as follows:

$tweet_i = \langle (topic_{i_1}, rd_{i_1}), (topic_{i_2}, rd_{i_2}), (topic_{i_m}, rd_{i_n}) \rangle$

where $topic_{i_k}$ is one of the topic associated to $tweet_i$, $rd_{i_k}$ is the *relevance degree* associated to $topic_{i_k}$, $n$ is the number of topics detected by sentence wikification of $tweet_i$.

## 3.4 Aggregating Multidimensional Tweet Streams via FFCA: An Overview

Our innovative aggregation model for Tweet data, which lies along classical aggregation mechanisms [14] with the novelty of being target to such kind of data that are inherently textual in nature, hence achieving the proposed methodology for microblog summarization. This methodology essentially makes use of a fuzzy extension of Formal Concept Analysis (briefly, Fuzzy FCA or FFCA) [12] as fundamental theoretical tool.

Based on the lattice theory, the FCA deals with concepts (objects and their attributes) and their hierarchical relationships. It supplies a basis for conceptual data analysis, knowledge processing and extraction. Fuzzy FCA [9] combines fuzzy logic into FCA representing the uncertainty through membership values in the range [0, 1]. We will provide an overview of our microblog summarization method in Section 4 through a comprehensive case study.

# 4. CASE STUDY: SUMMARIZING TWEET STREAMS DURING THE 2015 ITALIAN ELECTION CAMPAIGN

Nowadays, Twitter plays an important role in the political agenda at national and international level. So, the definition of advanced analytics services on Twitter is assuming much interest. Therefore, a motivating example of application of the methodology proposed in this work is about political elections. For the sake of clarity, this Section is aimed to illustrate a case study of OLAP and timed FFCA integration applied to a tweet stream collected during Italian regional election campaign held in May 2015, thus showing an overview of our microblog summarization method. A large round of regional elections were held in Italy in seven of the twenty regions composing the country, including four of the ten largest ones: Campania, Veneto, Apulia and Tuscany. The other three regions holding elections were Liguria, Marche, Umbria, along with more than 700 of Italy's municipalities went to the polls. An estimated 23 million Italians are eligible to vote.

The target case study clearly demonstrates how OLAP analysis methodologies are perfectly suitable of supporting advanced analytics over Tweet data, similarly to other recent and correlated research experience (e.g., [21, 4]). In fact, summarizing the information content of massive amounts of data like streaming Tweet data plays a leading role with respect to the goal of extracting useful knowledge from so abundant data. Indeed, even semantics issues must

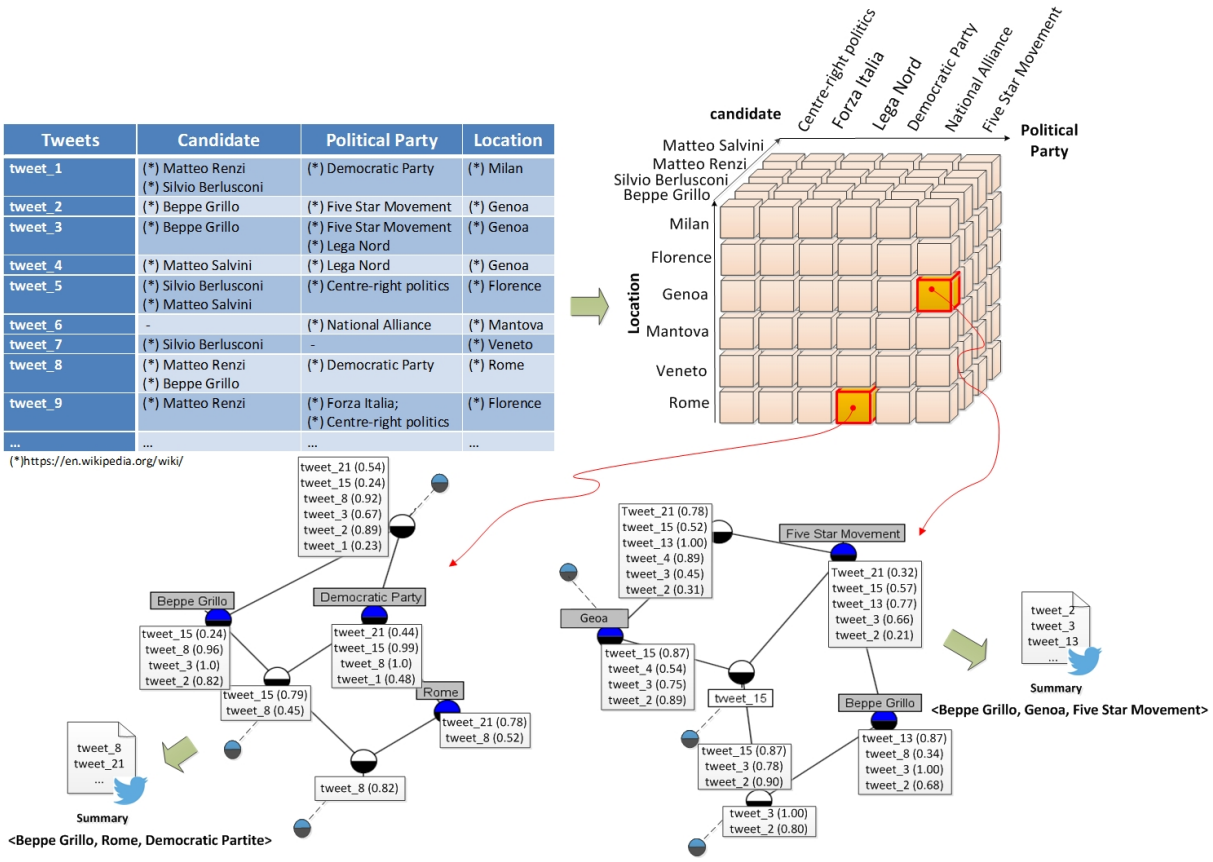


**Figure 2: A Case Study about the 2015 Italian Election Campaign**

be considered, as summarization is not only data reduction but, better, *knowledge synthesis* (e.g., [16]), which is now becoming more and more relevant in the emerging context of *big data analytics* (e.g., [8, 7, 11]).

Let us consider three-dimension data cube defined on a set of collected tweets (see Figure 2), having as measure the set of tweets belonging to the dimension *candidate* (e.g., Beppe Grillo), *location* (e.g., Rome) and *Political Party* (e.g., Democratic Party). The table on the top-left of Figure 2 shows a characterization of collected tweets. In particular, the table details the set of tweets shown in Figure 3 (Figure 4 shows English translation of them).

Let us suppose that a user wants to know about what are the main topics of twitting activity during election campaign for a specific city and political party, for example, as empathized in Figure 2, $\langle BeppeGrillo, Genova, FiveStarMovement \rangle$. The plethora of tweets that could satisfy the user request compromises the user understanding of the tweets flow. Let us note that according to recent statistics the average number of tweets per day is more than 100 million, and thousands of them happen every minute. In this sense, integrating the proposed microblog summarization process with OLAP enable us to provide the subset of tweets that covers the main topics discussed considering time dependence, and so updating user providing not redundant information. In the considered example the original set of tweets ($\simeq 500$ ) corresponding to the user request is reduced according to the *detail* level selected from the user (e.g., $\simeq 50$ for *detail* $d = 0.7$ ).

**Tweets**

- **tweet_1:** Stefano Fassina: "Oggi non mi ricandiderei col Partito Democratico. Matteo Renzi è peggio di Silvio Berlusconi". #sonomoltoperplesso
- **tweet_2**: Elezioni Liguria: Beppe Grillo a De Ferrari, tre ministri per Paita... #Grillo #M5S http://t.co/mn5TknT1ku
- **tweet_3:** Beppe Grillo La Lega Nord canta Bruciare il tricolore http://t.co/lOc3uxUdRp
- **tweet_4:** Matteo Salvini Lega Nord è PD che sono Forza Italia
- **tweet_5:** Nel centro destra il vero vincitore e' Matteo Salvini. Senza di lui Berlusconi e' soltanto doppio zero.
- **tweet_6:** #domenica #elezioni #Mantova vota Fratelli d'Italia Alleanza Nazionale!
- **tweet_7:** Berlusconi con #Ancellotti fa solo campagna elettorale: cerca l'8% in Liguria e Veneto @radiosportiva
- **tweet_8:** Matteo Renzi? Senza unità del Pd consegnerà il Paese a Beppe Grillo http://t.co/6D9VsfSTa4
- **tweet_9:** Amministrative: Berlusconi e crollo Fi, con centrodestra diviso vince Renzi: Forza Italia crolla sotto il 5% a... http://t.co/lAo32OHJNo
- ...

**Figure 3: Example of Tweet Streams Collected During the 2015 Italian Election Campaign**

**Tweets**

- **tweet_1:** Stefano Fassina: "Today I not nominate me as a candidate with the Democratic Party again. Matteo Renzi is worse than Silvio Berlusconi". #sonomoltoperplesso
- **tweet_2**: Elections Liguria : Beppe Grillo to De Ferrari , three ministers for Paita ... #Grillo # M5S http://t.co/mn5TknT1ku
- **tweet_3:** Beppe Grillo, the Lega Nord sings Burn tricolor http://t.co/lOc3uxUdRp
- **tweet_4:** Matteo Salvini, Lega Nord is PD that Forza Italia
- **tweet_5:** In the centerRight the real winner 'is Matteo Salvini '. Without him, Berlusconi ' is only double zero ..
- **tweet_6:** #sunday #election #Mantova rate Brothers Italian National Alliance!
- **tweet_7:** Berlusconi with #Ancellotti is just electoral campaign: He search the 8% in Liguria and Veneto @radiosportiva
- **tweet_8:** Matteo Renzi ? Without Democratic Party Unity, he will deliver the country to Beppe Grillo http://t.co/6D9VsfSTa4
- **tweet_9:** Administrative: Berlusconi e collapse Fi, with center- divided wins Renzi: Forza Italia falls below the 5% to... http://t.co/IAo32OHJNo
- ... ...

**Figure 4: English Translation of Tweets Shown in Figure 3**

# 5. CONCLUSIONS AND FUTURE WORK

This work focuses on the integration of knowledge mining approaches (i.e., FFCA) and OLAP technology analyzing unstructured data exchanged in social media and enabling advanced analytics services. Considering Twitter we emphasize the role of implicit information that could be derived or discovered in the tweets beyond explicit available metadata. A case study is proposed to address microblog summarization exploiting timed fuzzy lattice resulting from the execution of time-aware FFCA on the unstructured tweets' content. A fundamental contribution of our proposed framework is represented by the idea of performing the aggregation task by considering the temporal dimension mainly (via FFCA), and computing the target cube consequently, by also considering the remaining dimensions of the current multidimensional model. It should be noted that this approach is particularly targeted to the special case of data considered (i.e., multidimensional Tweet streams) as such kind of data are inherently time-dependent.

Finally, the paper represents a preliminary step toward the definition of advanced analytic services taking into account multidimensional social media stream considering also the latent semantics embedded in the information exchanged. On the other hand, the integration of knowledge representation and management methodologies with multidimensional analysis clearly represents a critical milestone to be considered in future, as highlighted by recent studies (e.g., [20]).

As future work, we plan to exploit the technique and tools used in [22] (specifically Timed Automata and non-repudiation protocol) to enforce a temporal and fairness criteria among the tweets received in the stream. On a system-oriented side, we are instead focused to embed in our framework specific *privacy-preserving OLAP features* (e.g., [5, 6]), as the latter are becoming relevant in emerging *Big Data* contexts.

# 6. REFERENCES


[1] S. Bringay, N. Béchet, F. Bouillot, P. Poncelet, M. Roche, M. Teisseire, Towards an on-line analysis of tweets processing, in: Database and Expert Systems Applications, Springer, 2011, pp. 154–161.

[2] P. Buitelaar, D. Olejnik, M. Sintek. A Protg plug-in for ontology extraction from text based on linguistic analysis. The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2004, 31-44.

[3] W. C. Cho and D. Richards. 2007. Ontology construction and concept reuse with formal concept analysis for improved web document retrieval. Web Intelli. and Agent Sys. 5, 1 (January 2007), 109-126.

[4] A. Cuzzocrea, "Analytics over Big Data: Exploring the Convergence of Data Warehousing, OLAP and Data-Intensive Cloud Infrastructures", in: *Proceedings of COMPSAC 2013*, pp. 481-483, 2013

[5] A. Cuzzocrea, V. Russo, Domenico Saccà, "A Robust Sampling-Based Framework for Privacy Preserving OLAP", in: *Proceedings of DaWaK 2008*, pp. 97-114, 2008

[6] A. Cuzzocrea, Domenico Saccà, "Balancing Accuracy and Privacy of OLAP Aggregations on Data Cubes", in: *Proceedings of DOLAP 2010*, pp. 93-98, 2010

[7] A. Cuzzocrea, and I.-Y. Song, "Big Graph Analytics: The State of the Art and Future Research Agenda", in: *Proceedings of DOLAP 2014*, pp. 99-101, 2014

[8] A. Cuzzocrea, I.-Y. Song, and K.C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!", in: *Proceedings of DOLAP 2011*, pp. 101-104, 2011

[9] 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.

[10] B. Fortuna, M. Grobelnik, D. Mladenic, OntoGen: semi-automatic ontology editor (pp. 309-318). Springer Berlin Heidelberg, 2007.

[11] M. Franklin, "Making Sense of Big Data with the Berkeley Data Analytics Stack", in: *Proceedings of WSDM 2014*, pp. 1-2, 2014

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

[13] S. M. González, T. d. R. L. Berbel, Considering unstructured data for olap: a feasibility study using a systematic review, Revista de Sistemas de Informação da FSMA (14) (2014) 26–35.

[14] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals", *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 29–53, 1997

[15] V. Gupta, N. Rathore, Deriving business intelligence from unstructured data, International Journal of Information and Computation Technology 3 (9) (2013) 971–976.

[16] L.V.S. Lakshmanan, J. Pei, and Y. Zhao, "QC-Trees: An Efficient Summary Structure for Semantic OLAP", in *Proceedings of SIGMOD Conference 2003*, pp. 64-75, 2003

[17] S. Mansmann, N. U. Rehman, A. Weiler, M. H. Scholl, Discovering olap dimensions in semi-structured data, Information Systems 44 (2014) 120–133.

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

[19] 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.

[20] K. Morfonios, and G. Koutrika, "OLAP Cubes for Social Searches: Standing on the Shoulders of Giants?", in: Proceedings of the 11th International Workshop on the Web and Databases, ACM, 2008

[21] T. MÃijhlbauer, W. RÃűdiger, A. Reiser, A. Kemper, and T. Neumann, ScyPer: "A Hybrid OLTPOLAP Distributed Main Memory Database System for Scalable Real-Time Analytics", in: *Proceedings of BTW 2013*, pp. 499-502, 2013

[22] M. Napoli, M. Parente, and A. Peron, "Specification and Verification of Protocols With Time Constraints," *Electr. Notes Theor. Comput. Sci.*, vol. 99, pp. 205–227, 2004.

[23] W. Wang, P. Barnaghi, A. Bargiela, Probabilistic topic models for learning terminological ontologies, Knowledge and Data Engineering, IEEE Transactions on 22 (7) (2010) 1028–1040. doi:10.1109/TKDE.2009.122.

[24] J.-h. Yeh, N. Yang, Ontology construction based on latent topic extraction in a digital library, in: Digital Libraries: Universal and Ubiquitous Access to Information, Springer, 2008, pp. 93–103.