HAL
archives-ouvertes.fr

How to Extract Relevant Knowledge from Tweets?

Flavien Bouillot, Phan Nhat Hai, Nicolas Béchet, Sandra Bringay, Dino Ienco, Stan Matwin, Pascal Poncelet, Mathieu Roche, Maguelonne Teisseire

▶ **To cite this version:**

Flavien Bouillot, Phan Nhat Hai, Nicolas Béchet, Sandra Bringay, Dino Ienco, et al.. How to Extract Relevant Knowledge from Tweets?. Communications in Computer and Information Science, 2013. <lirmm-00798662>

HAL Id: lirmm-00798662
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00798662

Submitted on 29 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

How to Extract Relevant Knowledge from Tweets?

F. Bouillot[1], P. Nhat Hai[1], N. Béchet[2], S. Bringay[1,3], D. Ienco[1,4], S. Matwin[5], P. Poncelet[1], M. Roche[1], and M. Teisseire[1,4]

[1] LIRMM – CNRS, 161 rue Ada, Montpellier, France
{bringay,bouillot,poncelet,mroche}@lirmm.fr
[2] Univ. Caen Basse-Normandie, Caen, France nicolas.bechet@unicaen.fr
[3] Univ. Montpellier 3, Montpellier, France
[4] IRSTEA - UMR TETIS, Montpellier France
{nhat-hai.phan,dino.ienco,maguelonne.teisseire}@teledetection.fr
[5] University of Ottawa, Ontario, Canada stan@site.uottawa.ca

Abstract. Tweets exchanged over the Internet are an important source of information even if their characteristics make them difficult to analyze (e.g., a maximum of 140 characters; noisy data). In this paper, we investigate two different problems. The first one is related to the extraction of representative terms from a set of tweets. More precisely we address the following question: *are traditional information retrieval measures appropriate when dealing with tweets?*. The second problem is related to the evolution of tweets over time for a set of users. With the development of data mining approaches, lots of very efficient methods have been defined to extract patterns hidden in the huge amount of data available. More recently new spatio-temporal data mining approaches have specifically been defined for dealing with the huge amount of moving object data that can be obtained from the improvement in positioning technology. Due to particularity of tweets, the second question we investigate is the following: *are spatio-temporal mining algorithms appropriate for better understanding the behavior of communities over time?* These two problems are illustrated through real applications concerning both health and political tweets.

1 Introduction

In recent years, the development of social and collaborative Web 2.0 underlines the central and active role of users in collaborative networks. Blogs to spread diaries, RSS news to track last information on a specific topic, tweets to publish social actions, are now extremely widespread. Easy to create and manage these tools are used by Internet users, businesses or other organizations to communicate about themselves and the growing use of this technology starts to influence many aspect of the real life. Furthermore, this data represents an important source of information that can be exploited in the decision making process.

Since its introduction in 2006, the Twitter website[6] has become so popular that it is currently ranked as the 10^{th} most visited site over the world[7]. In January 2012, Twitter has been visited 2.5 billion times and in October 2011, more than 250 million tweets are posted every day with a user base of about 300 million people. Basically, Twitter is a platform for microblogging. It means that it is a system for sharing information where users can either follow other users who post short messages (i.e. 140 characters), or can be followed. When a user follows a person, the user receives all messages from this person, and in turn, when that user tweets, all his followers will receive the messages. 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. For example the tag *@username* means that you are sending a message to a particular user, the *# topic* assigns a specific topic, *RT* means that the message was re-tweeted, i.e. send to all the followers.
Actually, by taking into account all this meta-information, we can observe that tweets can be represented in a multidimensional way with, for instance, one dimension for the location, one dimension for the time, one dimension for the set of words used, etc. In this context, different systems were proposed to analyze this flow of information [1–3]. For instance, they can focus on event detection [3], Name Entity recognition [4], can combine different types of information such as timeline and sentiment features [5] or even analyze propagation from specific features such as hashtags [6].

Nevertheless all these approaches do not really exploit their multidimensional characteristics. In [7], we propose to address these multidimensional characteristics by focusing on datawarehouses [8] since they provide very efficient tool for the storage and analysis of multidimensional and historized data. Our main goal was to using the facilities provided by these tool to manipulate a set of indicators (measures) according to the different dimensions obtained from tweets and for which we can be provide some hierarchies. Furthermore associated operators (e.g., Roll-up, Drill-down, etc.) allow an intuitive navigation on different levels of the hierarchy. In this paper we focus on the problems associated with the definition of such measures when dealing with hierarchies and propose some extension that are well adapted to our concern. In order to illustrate how such a tool can be useful we report some results on experiments conducted on the health domain.

By using a multidimensional tool we are able to really help the end user to analyze the data over time. Nevertheless one problem remains. As we are provided with a huge amount of data, applying data mining techniques seems relevant to highlight knowledge hidden in the data. Among the different techniques we investigate if pattern mining approaches are appropriate to understand users' behaviors. More precisely we focus on trajectory mining algorithms to analyze behavior of communities of user. Basically they are defined for dealing with spatio-temporal data. We show that measures defined for tweet datawarehouse

[6] http://twitter.com
[7] http://www.alexa.com/topsites

can also be used to define new kinds of trajectories related to tweet terms rather than spatial information. We illustrate some trajectories that can be extracted from the analysis of the evolution of French political communities[8] over Twitter during 2012 which was particularly important for French political communities dues the two main elections: Presidential and Legislative.

The remainder of this paper is organized as follows. Section 2 investigates how multidimensional characteristics can be handled and focus on different kinds of useful measures. It also illustrates some results by using tweets related to diseases. We address the problem of pattern mining in Section 3. After a brief presentation of trajectory approaches we illustrate how these patterns can be useful to understand users' behaviors over time. Finally Section 4 concludes the paper and presents future work.

2 Towards a Datawarehouse for Analyzing Tweets

In this section we propose a minimal model for dealing with multidimensional characteristics of tweets. We mainly focus on three dimensions: the dimension Word corresponding to the set of words (or terms) that can be extracted from tweets, the dimension location that can be extracted from metadata and the time dimension. Furthermore we illustrate the model through tweets expressed in the health domain.

2.1 The Model

According to [9], a fact table F is defined on the schema $D = \{T, \ldots, T_n, M\}$ where T_i $(i = 1, .., n)$ are the dimensions and M stands for a measure. Each dimension T_i is defined over a domain D partitioned into a set of categories C_j. We thus have $D = \cup_j C_j$. D is also provided with a partial order $\sqsubseteq_D$ to compare the values of the domain D. Each category represents the values associated with a level of granularity. We note $e \in D$ to specify that e is a value of the dimension D if there is a category $C_j \subseteq D$ such that $e \in \cup_j C_j$. Note that two special categories are distinguished and are present on all dimensions: $\bot_D$ et $\top_D \in C_D$ corresponding respectively to the level of finer and higher granularity. In our approach, the partial order defined on the domains of the dimensions stands for the inclusion of keywords associated to the values of the dimensions. Thus, let $e_1, e_2 \in \cup_j C_j$ be two values, we have $e_1 \sqsubseteq_D e_2$ if e_1 is logically contained in e_2.

Example 1 *Figure 1 illustrates the location dimension having a hierarchy such that* $\bot_{location} = City \leq State \leq Country \leq \top_{location}$. *The values of the dimension are* $dom(Localisation) = \{$*New York, Albany, Los Angeles, Northeastern, California, United States, ...*$\}$ *divided into these categories (levels of granularity) as follows:* $City = \{NewYork,$ *Albany,* $Los\ Angeles\}$, $State = \{$*Northeastern,*

[8] This work is a part of the POLOP Project (*Political Opinion Mining*) which aims to cope with the analysis of the evolution of French political communities over Twitter during 2012 both in terms of relevant terms, opinions, behaviors.

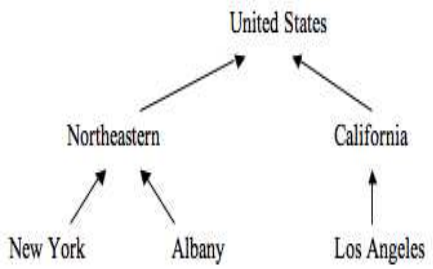


Fig. 1. A part of the hierarchy associated to the location dimension.

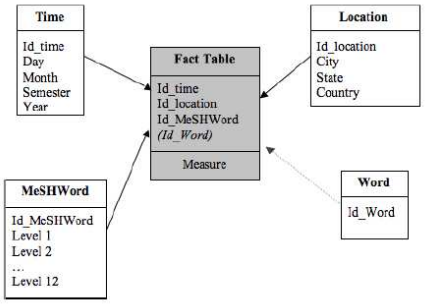


Fig. 2. An example of a schema for dealing with tweets in the health domain.

California}, $Country = \{United\ States...\}$. *The partial order* $\sqsubseteq_D$ *over the values of dimensions can be generalized to categories: for* $C_1, C_2 \in C_D$, *we thus have* $C_1 \leq_D C_2$ *if* $\exists e_1 \in C_1, e_2 \in C_2$ *such as* $e_1 \sqsubseteq_D e_2$. *For example, we have Los Angeles* $\sqsubseteq_D$ *California* $\sqsubseteq_D$ *United States* $\sqsubseteq_D \top$. *The taking into account of the dynamic hierarchy is such that all categories of this dimension must respect the defined partial order.*

Figure 2 illustrates the associated schema. We find the dimension *location* and the dimension *time* as $\perp_{temps} = day \leq month \leq semester \leq year \leq \top_{temps}$. For the hierarchy, we use the MeSH (Medical Subject Headings)[9] National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a twelve-level hierarchy that permits searching at various levels of specificity. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels, such as "Ankle" and "Conduct Disorder". In 2011, 26,142 descriptors are available in MeSH. There are also over 177,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid".

2.2 Some Proposed Measures

Traditionally, the $TF\text{-}IDF$ measure (Term Frequency - Inverse Document Frequency), introduced by [10], is a very useful measure that giving greater weight to the discriminant terms and can thus be well adapted to our concern. As a first step, it is necessary to compute the frequency of a term (*Term Frequency*) corresponding to the number of occurrences of the term in the document[10]. Thus,

[9] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

[10] Here *document* is used to be compliant with the original definition of the $TF\text{-}IDF$ measure and refers to a tweet in our context.

for the document d_j and the term t_i, the frequency of the term in the document is given by the following equation:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ stands for the number of occurrences of the term t_i in d_j. The denominator is the number of occurrences of all terms in the document d_j.

The IDF (*Inverse Document Frequency*) measures the importance of the term in the corpus. It is defined as follows:

$$IDF_i = log_2 \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

where $|D|$ stands for the total number of documents in the corpus and $|\{d_j : t_i \in d_j\}|$ is the number of documents having the term t_i.

Finally, the TD-IDF is obtained as follows:

$$TF\text{-}IDF_{i,j} = TF_{i,j} \times IDF_i$$

Nevertheless relying only on knowledge of the hierarchy in a cube does not always allow a good aggregation (i.e., corresponding to a real situation). For instance, the characteristics of the words in tweets are not necessarily the same in a State and in a City.

In [7], in a very different context, we proposed a new measure called $TF\text{-}IDF_{adaptative}$. This measure has been defined in order not to focus on the number of documents but rather to the number of documents for a specific class and take into account the level in the hierarchy. So in our case, this measure is well adapted for handling available hierarchies as it does not calculate the representative terms from the number of documents but rather from the desired class at a specific level. It is defined as follows:

$$TF_{i,j} - IDF_i^k = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log_2 \frac{|E^k|}{|\{e_j^k : t_i \in e_j^k\}|}$$

where $|E^k|$ stands for the total number of elements of type k (in our example, $k = \{$*City*, *State*, *Country*$\}$) which corresponds to the level of the hierarchy that the decision maker wants to aggregate. $|\{e_j : t_i \in e_j\}|$ is relative to the number of elements of type k where the term t_i appears. Thus, we define $IDF_{adaptive}$ as follows:

$$IDF_i^{C_l} = log_2 \frac{m}{|\{C_l : t_i \in C_l\}|}$$

where m stands for the total number of communities. $|\{C_l : t_i \in C_l\}|$ is the number of communities C_l where the term t_i appears.

Actually, this adaptive measure can be easily generalized for taking into account the different level of the hierarchies as well as the information the user

is interested in. Then it is possible to define a context $C = \langle n, l, type \rangle$ where n stands for one node in the hierarchy, l the level of the hierarchy the user would like to extract information and *type* corresponds to the type of information, i.e. element of a specific level (e.g. city) or tweets for this level. For a context C and a term t_i, the *Generalized IDF* is defined as follows:

$$Generalized\ IDF_{C,i,j \in type} = log_2 \frac{|type^l|}{f(type)}$$

where $|type^l|$ stands for the total number of elements of type *type* occurring in the scope of n at the level l of the hierarchy. $f(type)$ is defined as:

$$f(type) = \begin{cases} if\ type\ = Extension_D & |\{d_j : t_i \in Extension_D(n,l)\}| \\ else & |\{n_k : \exists t_i \in docs(n_{k_l})\ and\ n_{k_l} \in Ext(n,l)\}| \end{cases}$$

Depending on the value of type, the $f(type)$ function returns either the number of documents specified in the extension having the term t_i or the number of nodes having at least one document with the term t_i.

Finally, for a context C and a term t_i, the *Generalized TF-IDF* is defined as:

$$Generalized\ TF\text{-}IDF_{C,i,j \in type} = TF_{i,j} \times Generalized\ IDF_{i,j}$$

2.3 Illustration

In this section we illustrate how such a model can be useful for the decision maker. To extract the tweets related to the vocabulary used in MeSH, we focus on the tweets related to "Disease" and request Twitter by using all the terms of the corresponding hierarchy. We thus collected 1,801,310 tweets from Janurary 2010 to February 2011 having at least one term of the Disease hierarchy. Experiments were performed by using PostgreSQL 8.4 with the Pentaho Mondrian 3.20 environment. Visualization are done by using the visualization set of tools provided by Google are

Figure 3 shows the distribution of words *hepatitis*, *leukomia* and *pneunomia* over the period (excluding US). We can note that the word *pneunomia* appears a lot in the set of tweets in January 2009 and in February 2010. By using visualization tools it is possible to visualize the worldwide coverage of this disease (Figure 4). This coverage is obtained by fixing the location dimension and by examining the countries for which this word has been selected as more relevant for a country with our measures.

In Figure 4 we can notice that *pneunomia* has been extensively reported in tweets from Russia, India, Ecuador or Australia. News available at that period might be useful to better understanding this behavior. For instance, in the second week of January 2001, the British singer Trish Keenan died of pneumonia after contracting swine flu in Australia[11] and, in Ecuador, an alarm over cases of severe

[11] http://www.dailymail.co.uk/news/article-1347160/.

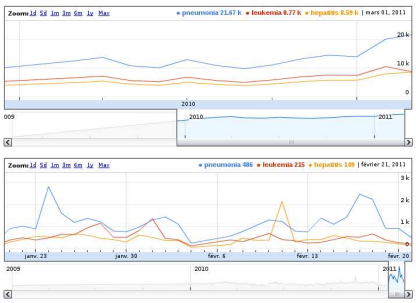


Fig. 3. Distribution of the use of words pneumonia, hepatitis and leukemia over the period (top) and for January and February (bottom).

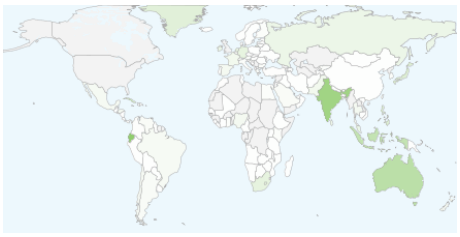


Fig. 4. Distribution of the use of the word pneunomia.

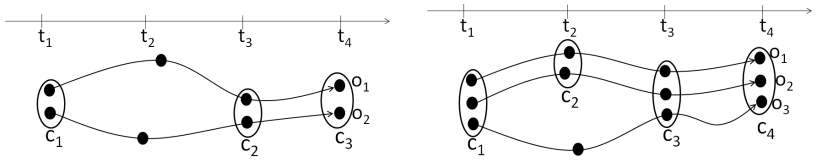
pneumonia similar to the H1N1 virus was triggered after that the Merced hospital has received 35 cases, 18 occurred in the first week of January, and reported that two people died. and just as serious[12]. In Russia also, In December of 2010, in Yurga /Kemerovo region, over 200 soldiers were taken to hospitals with a bad cold and several people were in critical condition: severe pneumonia[13]. All these events were extensively tweeted or re-tweeted in these countries.

3 Pattern Mining for Tweets

Usually pattern mining approaches focus on extracting different kinds of patterns (i.e. itemsets, sequences, trees, graphs, etc) hidden in the database. Using these patterns for analyzing tweets is one of the topic addressed by some research work. In this paper we investigate another kind of patterns coming from a very different context: spatio-temporal patterns. Our main objective is to highlight that knowledge extracting can be very useful for the decision maker. Basically, spatio-temporal patterns are defined in a totally different context (i.e spatio-temporal data) and aim to identify groups of moving objects for which a strong relationship and interaction exist within a defined spatial region during a given time duration. Recently, many patterns have been defined such as flocks [11], convoys [12], swarms, closed swarms [13], moving clusters [14], group pattern [15], etc.

Let us assume that we have a group of moving objects $O_{DB} = \{o_1, o_2, \ldots, o_z\}$, a set of timestamps $T_{DB} = \{t_1, t_2, \ldots, t_n\}$ and at each timestamp $t_i \in T_{DB}$, spatial information[14] x, y for each object. . Usually, in spatio-temporal mining, we are interested in extracting a group of objects staying together during a period of time. Therefore, from now, $O = \{o_{i_1}, o_{i_2}, \ldots, o_{i_p}\}(O \subseteq O_{DB})$ stands for a group

[12] http://www.flutrackers.com/forum/showthread.php?t=158136
[13] http://www.flutrackers.com/forum/showthread.php?t=156585
[14] Spatial information can be, for instance, GPS location.



(a) Swarm (b) Convoy

Fig. 5. An example of swarm and convoy where c_1, c_2, c_3, c_4 are clusters.

of objects, $T = \{t_{a_1}, t_{a_2}, \ldots, t_{a_m}\}$ ($T \subseteq T_{DB}$) is the set of timestamps within which objects stay together. Let ε be a user-defined threshold standing for a minimum number of objects and min_t a minimum number of timestamps. Thus $|O|$ (resp. $|T|$) must be greater than or equal to ε (resp. min_t). Here we focus on two particular patterns to illustrate. Informally, a *swarm* is a group of moving objects O containing at least ε individuals which are closed each other for at least min_t timestamps. For example, as shown in Figure 5a, if we set $\varepsilon = 2$ and $min_t = 2$, we can find the following swarms $(\{o_1, o_2\}, \{t_1, t_3\})$, $(\{o_1, o_2\}, \{t_1, t_4\})$, $(\{o_1, o_2\}, \{t_3, t_4\})$, $(\{o_1, o_2\}, \{t_1, t_3, t_4\})$. We can also note that these swarms are in fact redundant since they can be grouped together in the following closed swarm $(\{o_1, o_2\}, \{t_1, t_3, t_4\})$. A *convoy* is also a group of objects such that these objects are closed each other during at least min_t *consecutive* time points. For instance, on Figure 5, with $\varepsilon = 2, min_t = 2$ we have two convoys $(\{o_1, o_2\}, \{t_1, t_2, t_3, t_4\})$ and $(\{o_1, o_2, o_3\}, \{t_3, t_4\})$. Recently in [16] we proposed GET_MOVE an unifying approach for extracting all these kinds of patterns. Furthermore in [17], we proposed new kinds of patterns, called gradual patterns. For instance, they can be useful to extract gradual trajectories such as: "From October to December the more time passes, the more Eagle are moving from Canada to Mexico" or "From June to July, the more the time goes by, the more people are going to Miami". All these approaches share the same pre-processing: a clustering algorithm is applied (e.g. DBSCAN) for extracting clusters grouping together objects closed to the same location. During our research we considered that the clustering can be applied in other dimensions. Recently, we defined a new project called POLOP[15] (*Political Opinion Mining*) which aims to cope with the analysis of the evolution of French political communities over Twitter during 2012 both in terms of relevant terms, opinions, behaviors. 2012 is particularly important for French political communities dues the two main elections: Presidential and Legislative. From the 12th December 2011 to the 19th June 2012, we thus obtained 2,122,012 tweets from 213,005 users. For 130,618 tweets, 232 users can unambiguously be assigned to a political party (i.e. user is a politician or an official political community account). By using our defined measures (C.f. Section 2.2), we can select for different political parties the set of relevant words at different periods (see Figure 6). In order to extract interesting trajectories for different parties we applied a clustering technique, for each party, on the top-k set of relevant terms over time. Thus, we group together users from the same party sharing the same words. Figure 7 illustrates a kind of trajectory that might be extracted from the political tweets. In February 2012, in one of

[15] http://www.lirmm.fr/~bouillot/polop/polop.html

(a) Left Party (b) Right Party

Fig. 6. An example of the most relevant words of parties: (a) Left Party and (b) Right Party.

the first debate, the focus of the French election campaign suddenly shifts from the economy to racism and national identity. For the right party, as we can notice, during the debate, most of used words are such that lots of party members shared the same words. Just after the debate, some right party users move to another cluster where words such as "national identity", "France" while other ones shift to other topic like "elections". Actually, during the debate the main trend was to motivate some electors from the extreme right. During this time, as we can notice, in the left party, as it is illustrated through the gradual pattern, the left party uses completely different kinds of words but just when right users behavior split.

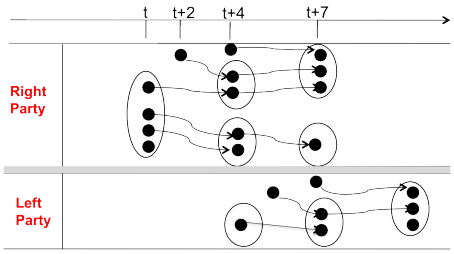


Fig. 7. An example of a trajectory that can be extracted from tweets.

4 Conclusions

In this paper we first proposed a new approach to analyze tweets from their multidimensional characteristics. The originality of our proposal is to define and manipulate cubes of tweets. We have shown that the analysis of tweets requires the definition of new measures and that a contextualization step is relevant. Then we investigated how pattern mining and more precisely trajectories can be useful for analyzing tweet user behaviors. Future work involves several issues. First we want to extend the proposed approach to take into account opinions or feelings expressed in the tweets. Furthermore, as too many trajectories can be extracted from tweets, we would like to investigate how the minimal description length can be used to report only the most relevant patterns, i.e. top-k patterns.

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proceedings of WWW. (2010) 851–860
2. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream. In: Proceedings of SIGMOD 2010. (2010) 1155–1158
3. Li, C., Sun, A., Datta, A.: Twevent: Segment-based event detection from tweets. In: Proceedings of CIKM. (2012)
4. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., , Lee, B.S.: Twiner: Named entity recognition in targeted twitter stream. In: Proceedings of SIGIR. (2012)
5. Tsolmon, B., Kwon, A., Lee, K.S.: Extracting social events based on timeline and sentiment analysis in twitter corpus. In: Proceedings of NLDB. (2012)
6. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING. (2010)
7. Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., Teisseire, M.: Towards an on-line analysis of tweets processing. In: Proceedings of DEXA. (2011)
8. Codd, E., Codd, S., Salley, C.: Providing olap (on-line analytical processing) to user-analysts: An it mandate. In: White Paper. (1993) 3–5
9. Pérez-Martínez, J.M., Llavori, R.B., Cabo, M.J.A., Pedersen, T.B.: Contextualizing data warehouses with documents. Decision Support Systems **45**(1) (2008) 77–94
10. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11) (1975) 613–620
11. Vieira, M., Bakalov, P., Tsotras, V.: On-line discovery of flock patterns in spatio-temporal data. In: Proceedings of SIGSPATIAL. (2009)
12. Jeung, H., Yiu, M., Zhou, X., CS, C.J., Shen, H.: Discovery of convoys in trajectory databases. PVLDB **1** (2008)
13. Li, Z., Ji, M., Lee, J.G., Tang, L., Yu, Y., Han, J., Kays, R.: Movemine: Mining moving object databases. In: Proceedings of SIGMOD. (2010)
14. Jensen, C., Lin, D., Ooi, B.: Continuous clustering of moving objects. IEEE TKDE (2007)
15. Wang, Y., Lim, E.P., Hwang, S.Y.: Efficient mining of group patterns from user movement data. DKE (2006)
16. Nhan, H.P., Poncelet, P., Teisseire, M.: Get_move: An efficient and unifying spatio-temporal pattern mining algorithm for moving objects. In: Proceedings of IDA. (2012)
17. Nhat, H.P., Ienco, D., Poncelet, P., Teisseire, M.: Mining time relaxed gradual moving object clusters. In: Proceedings of SIGSPATIAL. (2012)
18. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. Discourse Processes **25** (1998)
19. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of ECML. (2001)
20. Tang, J., Jin, R., Zhang, J.: A topic modeling approach and its integration into the random walk framework for academic search. In: Proceedings of ICDM. (2008)