# Extracting Emergent Semantics from Large-Scale User-Generated Content

Ioannis Kompatsiaris[1], Sotiris Diplaris[1], Symeon Papadopoulos[1]

[1] Informatics & Telematics Institute,
6th Km Charilaou-Thermi road, 57001 Thessaloniki, Greece
{ikom, diplaris, papadop}@iti.gr

**Abstract.** This paper presents a survey of novel technologies for uncovering implicit knowledge through the analysis of user-contributed content in Web2.0 applications. The special features of emergent semantics are herein described, along with the various dimensions that the techniques should be able to handle. Consequently a series of application domains is given where the extracted information can be consumed. The relevant techniques are reviewed and categorised according to their capability for scaling, multi-modal analysis, social networks analysis, semantic representation, real-time and spatio-temporal processing. A showcase of such an emergent semantics extraction application, namely ClustTour, is also presented, and open issues and future challenges in this new field are discussed.

**Keywords:** emergent semantics, social media analysis

## 1 Introduction

Social media sharing properties, such as Flickr, Facebook and Picasa host billions of images and videos that have been annotated and shared among friends, or published in groups covering a specific topic of interest. The fact that users annotate and comment on the content in the form of tags, ratings, preferences, etc. and that these are applied on a daily basis, gives this data source an extremely dynamic nature that reflects events and the evolution of community focus. Although current Web2.0 applications encourage annotations and feedback by the users, these are not sufficient for extracting such "hidden" knowledge, because they lack clear semantics and it is the combination of visual, textual and social context, which provides the ingredients for a more thorough understanding of social content. Therefore, there is a need for scalable and distributed approaches capable of handling the massive amounts of available data and capturing the emergent semantics, a procedure also called Collective Intelligence, that would enable the exploitation of the knowledge hidden in the user contributed content (UGC).

Recent advances of Web technologies have effectively turned ordinary people into active members of the Web: casual users act as co-developers and their actions and collaborations with one another have added a new social dimension on Web data. For example, Wikipedia motivates thousands of volunteers around the world to create the world's largest encyclopedia. An image in Flickr is annotated with descriptive tags, associated with the users that seem to like it and mark it as favorite, described by the visual features encoding its visual content, and often spatial and temporal information denoting its context. Even though all these facets of information are not combined naturally with each other, still they carry knowledge about the resource, with each facet providing a representation of the particular resource in a different feature space.

There already exist a number of approaches that are based on user-contributed content in order to provide useful information for various applications. For example, mobile location information and uploaded content is used to monitor online traffic and generate traffic patterns in **Error! Reference source not found.**, connect citizens in Boston **Error! Reference source not found.**, share nature experience **Error! Reference source not found.**, discover travel patterns and provide travel advice **Error! Reference source not found.Error! Reference source not found.**, communicate problems in a city **Error! Reference source not found.** and deal with climatic changes as in the Climate Collaboratorium [7] project of the MIT Center for Collective Intelligence[1]. The Collective Prediction effort, tries to make accurate predictions about future events such as product sales, political events, and outcomes of medical treatments [8]. However, the main characteristic of such applications is that they are mostly based on collecting well-structured contributions through specific applications, on shallow statistical processing of the contributions and their visualisation. Very few focus on analysis of implicit relations in UGC and feedback and on dealing with unstructured large-scale data, where an important source of knowledge is hidden.

This paper focuses on approaches that capture and analyse the emergent semantics lying under the large-scale multimedia content, which are extracted by using media resource and user relations, actions and interactions in social networking and sharing applications. A state-of-the-art survey is given by first defining the different features of existing applications dealing with emergent semantics extraction in terms of analysis dimensions and application domains (Section 2). Then the applications are reviewed and categorised according to the previously defined criteria (Section 3). We also present an example application, namely ClustTour, which encompasses many of the different aspects/dimensions discussed herein, and showcasing the capability of emergent semantics extraction to exploit different input data sources with combined analysis techniques for the delivery of enhanced services to the end users (Section 4). Finally, the paper concludes with a discussion on the identified issues and future challenges in developing emergent semantics extraction applications in the large scale (Section 5).

## 2 Social media analysis dimensions and applications

Understanding large-scale social media content involves the consideration of a multitude of features in the development of suitable analysis techniques.

[1] http://cci.mit.edu

A key issue to be addressed is the presence of *noisy and ambiguous data*. User contributed content is very noisy containing many non-relevant contributions either intentionally (spamming) or unintentionally. The lack of constraints with respect to tagging is the source of numerous annotation quality problems, such as spam, misspellings, and ambiguity of semantics. Moreover, the lack of (hierarchical) structure in the contributed information results in tag ambiguity, tag synonymy and granularity variation.

UGC can be viewed as a *rich multi-modal source of information* including attributes such as time, favorites and social connections. Social media analysis systems can exploit different modalities of input, ranging from single visual, textual, audio or user information to fused sets of such media sources, i.e. annotated images, audio or audiovisual content, or even the combination of such content with the graph structure of social networks. In such approaches, also called feature-level approaches, the need to obtain a joint, unique representation for the multimedia object demands techniques that manage to handle the very different characteristics exhibited by the different types of data. This is true both in terms of the nature of raw features (e.g., sparse, high-dimensional word co-occurrence vectors extracted from text descriptions, compared to usually dense and low-dimensional descriptors extracted from visual content), as well as in terms of their semantic capacity (e.g., while abstract concepts like “freedom” are more easily described with text, concrete concepts like “sun” are more easily grounded using visual information).

*Scalability* is an important issue since the discovery of implicit information is based on massive amounts of data. Such huge volumes of user generated data raise scalability issues that significantly compromise the performance (in terms of accuracy) of algorithms operating on such data. The situation gets worse in cases where either the employed algorithms aim at extracting knowledge patterns that only become stable after a specific usage period or processing needs to address *(near) real-time data*. These cases pose very demanding requirements in terms of algorithmic design, computational power, memory and storage. This includes parallelisation and distributed techniques, speed, storage, memory and relevant considerations, which are indispensable in order to advance to real-time systems offering emergent semantic discovery capabilities.

Another important aspect characterising how analysis techniques can unveil the hidden semantics is the presence or absence of *spatial processing* in the data sources. Following current web socialising and multimedia handling user trends, most users nowadays upload, geo-tag and localise their personal photos. Moreover, most Web2.0 systems make heavy use of geotagged resources. Therefore, many recommendation, presentation or prediction techniques are used for enhancing location-based systems and services.

A different dimension, orthogonal to the location aspect is time. The possibility of processing *temporal* features endows such systems with event processing capabilities. When time aspects of the UGC are considered, it is possible to derive information about situations or events. Based on massive user contributions, the emergent semantic extraction results could range from the single representation of events, and expand to past/current events detection or even future events prediction.

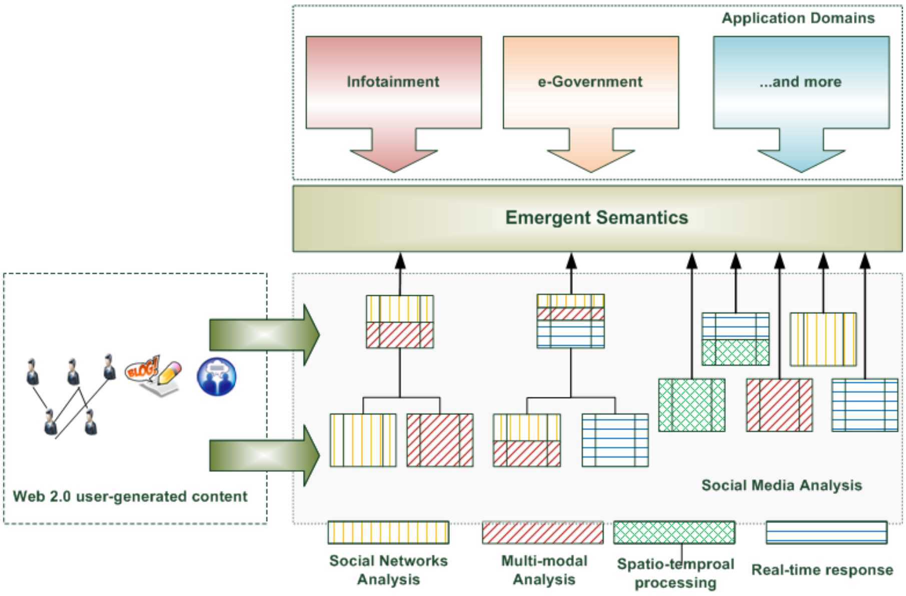


**Fig. 1.** Extracting emergent semantics from Web2.0 applications.

More elaborate applications are able to capture knowledge in the combined *spatio-temporal* dimension. Considering both space and time information, it is possible to depart from mere presentations of geotagged collections and discover knowledge about routes or map areas with particular interest, as well as to detect events in the combined spatio-temporal dimension.

To summarise, analysis techniques that deal with Web2.0-based and social network content and structure can be categorised with respect to the different dimensions they embrace (multi-modality), their ability to process the underlying social network structure information, the different features they consider (visual, audio, textual or their fusion), the capability of scaling, as well as their potential for running in real-time.

Exploiting the information carried by social media can then help tackle a variety of issues in different disciplines, such as information retrieval (e.g., poor recall and precision), machine learning, data mining, and multimedia understanding, where for example, social media sites can be used as a rich source of weakly labelled data for solving large-scale artificial intelligence problems. They can also be applied to domains such as tourism, culture, social sciences, politics, economics, and marketing. In tourism and culture, uploaded media can reveal “off-the-beaten-path” points of interest and events, otherwise difficult to discover through usual Web sources. In economics, marketing and brand monitoring, the number of related media uploaded online can reflect the number and locations of products sold in the market. Fig. 1 depicts the procedure by which the emergent semantics are analyzed and consumed.

# 3 Approaches for emergent semantics extraction

In this section we will review some of the social media analysis applications that take into account large-scale semi-structured UGC and apply computational methods in discovering implicit relations and therefore "hidden" knowledge. We also attempt to classify them according to the categorisation and the features these techniques consider, as discussed in Section 2.

**Multi-modal analysis approaches**

In [9] geo-location and tag information are used in order to generate representative city maps. In [10, 11] tags and visual information together with geo-location are used for object (e.g. monuments) and event extraction. The description of city cores can be derived automatically, by exploiting tag and location information [12]. The approach is able of distinguishing between administrative and vernacular uses of place names, thus avoiding the potential for confusion in the dispatch of emergency services. Tags from Flickr images and timestamp information are used in [13] to form a chronologically ordered set of geographically referenced photos and distinguish locals from tourist travelling. Another work in the line of multi-modality–aided localisation is VIRaL [14], a web-based tool which is used to identify and localise similar multimedia content under different viewpoints, applicable to any case that involves still image search and retrieval.

**Community, trend and event detection**

Although still at the level of research, there are several applications that exploit the knowledge extracted from massive user contributions. For instance, it is common to derive community-based views of networks, i.e. networks of which the nodes correspond to the identified communities of the original networks and the edges to the relations between the communities. Such views are more succinct and informative than the original networks. It is for this reason that community detection has found applications in the field of recommendation systems [15-18], as well as for representing user profiles [19, 20]. Other applications that make use of the knowledge extracted from tag communities include sense disambiguation [21] and ontology evolution/population [16].

The patterns emerging often show deep interconnections with various world events [22, 23] in a way that the evolving world reality is captured at each instant.

A leap in the exploitation of UGC research is the work in [24] which explores global trends and sentiments that can be drawn by analyzing the sharing patterns of uploaded and downloaded social multimedia. Taking into account both spatial and temporal aspects of content item views and uploads in social media sites and aggregating them, the authors are trying to forecast future events impacting politics, economics and marketing.

In [25] the authors are using human sensors to detect real world events, and generate situation awareness. They describe how spatio-temporal-thematic data in various social media can be aggregated into 'social pixels'. They derive image-like representations which enable sophisticated data processing, offering to users a query algebra tool for posing sophisticated situational queries. The technique is applied in business analysis, seasonal characteristics analysis and political event analytics.

Further, other work shows that the actions of individual Web users, when properly pooled, can indicate macro trends. There are studies using Search Engine queries for influenza surveillance over the Web [26], such as Google Trends [27], search advertisement click through [28], Yahoo search queries [29] and health website access logs [30]. Specifically in [27], Google search engine queries and data from the Centers for Disease Control (CDC) are used to monitor influenza rates 1-2 weeks ahead of the CDC reports.

**Real-time applications**

Finally, a separate class of applications involves the real-time aspect in the analysis. An early, but not quite scalable, tool [31] deals with the analysis of user profiles and query logs for the extraction of personalised touristic information (places and events suggestions) using a hierarchical semantic geospatial model as well as an event notification system. SCENT [32] is a framework for monitoring the evolution of multi-faceted social network data resulting from users' continuous online interactions. It enables very large scale data management and analysis by modeling social data in the form of tensor (multi-dimensional array) streams, tracking changes in the spectral properties of the tensor over time. It has mainly been used in recommendation and monitoring use cases.

Another tool [33] is developed to improve search quality (by reranking) and recommend supplementary information in query time (i.e., search-related tags and canonical images) analyzing visual content, high-level concept scores, time and location metadata. In [34] a real-time approach combining the contextual information from time duration, number of views, and thumbnail images with content analysis derived from color and local points is used in order to achieve near-duplicate elimination in video search in Web 2.0 applications. Finally, the authors in [35] exploit georeferenced toponyms found in community websites to model vernacular places in cities. The system is capable of answering queries in real time.

**Table 1.** Categorisation of emergent semantic extraction applications.

| Analysis Methods/ Dimensions | Scalability | Spatio-temporal | SNA | Multi-modal | Semantic representation | Real-time |
|---|---|---|---|---|---|---|
| Recommendation | [32-34] | [31, 32] | [15-20, 31, 32, 34] | [32-34] | [19, 20] | [31-34] |
| Clustering | [36] | [13, 36] | [13, 21, 36] | [10, 11, 36] | [13] | |
| Localisation | [12, 14, 35] | [14] | [14, 35] | [9, 14] | [9, 12] | [35] |
| Trend/event detection | [26-30, 36] | [22-25, 36] | [16, 22-24, 30, 36] | [9, 10, 22, 23, 36] | [9, 22, 23, 25] | |

In the following section another example of emergent semantics extraction, namely ClustTour, is described in more detail. Table 1 presents a categorisation of the above techniques and applications with respect to the kind of analysis they employ (i.e. recommendation, clustering, localisation, monitoring, event detection, etc.) and the parameters discussed in Section 2.

It can be observed that most applications do not employ the spatial or spatio-temporal dimension; instead they mostly make use of textual annotations in order to provide recommendations, or represent events and situations. However, more advanced techniques and applications [9,10,13,22,23,24,25] have also been presented, capable of processing more kinds of input modalities enabling the spatio-temporal and situational dimension. Scalability is addressed by a wide range of applications; however the amount of works enabling the real-time aspect is still very limited. It is notable that very few real-time applications are also able to handle and analyze multi-modal information.

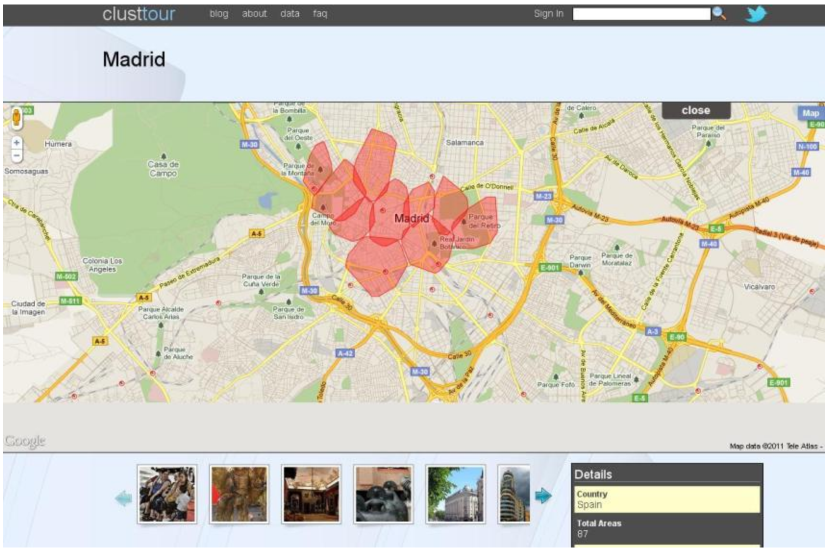
## 4 ClustTour: An example application

ClustTour brings to surface emergent semantics by using photo clusters corresponding to landmarks and events to assist the online exploration of a city. It employs an efficient graph-based scheme that clusters massive amounts of photos into clusters by use of a community detection method that is applied on a hybrid similarity graph encoding both visual and textual similarities between photos. Subsequently, it classifies these clusters into landmarks and events [11]. ClustTour leverages the obtained clusters through a map-based interface that enables users to navigate through a city. The application offers two modes of exploration: a city view depicting a high-level view of the most important areas in the city (Fig. 2) and an area view centered on the selected area and showing landmarks and events in its vicinity. ClustTour also provides a temporal content organisation layer on top of the detected areas and photo clusters. In that way, it endows users with enhanced content exploration and browsing capabilities, and at the same time, it improves upon the quality of the presented clusters. In the end, users obtain informative views over the interesting spots and areas in a city depending on the temporal context that is of interest to them.

With its ability to analyse massive visual and textual information from UGC (Flickr) and to present both landmarks and events in a spatio-temporal dimension, ClustTour is a typical case of implicit knowledge extraction application enabling the analysis of multi-modal input in the large scale.

## 5 Conclusions and open issues

The herein reviewed emergent semantics extraction techniques are trying to harness one or more forms of online user contributions in order to benefit end-users and organisations by employing large-scale recommendation, prediction, detection, representation or summarisation analysis techniques. However, such advanced efforts are still very limited, compared to the widespread usage of social media. Several aspects are currently covered by these methodologies, addressing issues such as scalability, efficiency, fusion and integration of multi-modal sources, as well as real-time analysis. Additional issues and dimensions relevant to social media analysis and data mining include:



Country
Spain

Total Areas
87

**Fig. 2.** Areas detected by spatial clustering of Flickr photos in Madrid.

- **Aggregation of multiple input sources.** Analysis of multiple Web2.0 sources by semantics extraction techniques is a step beyond the existing approaches that mostly consider single sources (e.g. Flickr photos).
- **Linking with Open Data**. Exposing, sharing, and connecting pieces of information extracted from social media, using URIs and RDF, with existing Open Data is a crucial step towards a more semantic description and representation of the extracted information.
- **User Interfaces and Visualisation** must be developed in order to support user interaction and understanding of the results of social media analysis.
- **Trust, security, privacy** is always a concern for users when they contribute content and especially when this is further analyzed by the application. The users should be guaranteed that their contribution remains anonymous and/or is used for an objective they consent to.
- **Social data sampling, modeling and representation.** In order to deal with the enormity of real-world social media datasets, appropriate sampling and data reduction techniques are needed, which manage to reduce the size of the data but at the same time they preserve their structure, topology and social context. These techniques are closely related to the representation approach used (for example, in graph-based representations they lead to sub-graph sampling) and there are still many open questions when dealing with social media-related sampling strategies.

## 6 References

1. Work, D., Blandin, S., Piccoli, B., Bayen, A.: A traffic model for velocity data assimilation. Applied Mathematics Research eXpress (2010)

2. Torres, L.H.: Citizen sourcing in the public interest. KM4D Journal, 3(1), 134--145 (2007)
3. iSpot, your place to share nature. URL: http://ispot.org.uk/.
4. Mobnotes. URL: http://www.mobnotes.com/
5. Dopplr. URL: http://www.dopplr.com/
6. FixMyStreet. URL: http://www.fixmystreet.com/
7. Malone, T.W., Klein, M.: Harnessing Collective Intelligence to Address Global Climate Change. Innovations, 2(3), 15--26 (2007)
8. Kemp, C., Shafto, P., Berke, A., Tenenbaum, J.B.: Combining causal and similarity-based reasoning. Advances in Neural Information Processing Systems 19 , in press.
9. Kennedy, L.S., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How Flickr helps us make sense of the world: context and content in community-contributed media collections. ACM Multimedia, pp. 631--640 (2007).
10. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In Proc. Int. Conf. on Content-based image and video retrieval, pp. 47-56 (2008)
11. S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, A. Vakali: Cluster-based Landmark and Event Detection on Tagged Photo Collections. IEEE Multimedia 18(1), pp. 52-63, (2011)
12. Hollenstein, L., Purves, R.S.: Exploring place through user- generated content: using Flickr to describe city cores. Journal of Spatial Information Science (2009)
13. Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., and Blat, J.: Digital footprinting: Uncovering tourists with user-generated content. IEEE Pervasive Computing, 7(4), 36--43 (2008)
14. Kalantidis, Y., Tolias, G., Avrithis, Y., Phinikettos, M., Spyrou, E., Mylonas, P., Kollias, S.: VIRaL: Visual Image Retrieval and Localization. Multimedia Tools Appl. 51(2), 555--592 (2011)
15. Nanopoulos, A., Gabriel, H.H., Spiliopoulou, M.: Spectral clustering in social-tagging systems. In: WISE ’09: Proceedings of the 10th International Conference on Web Information Systems Engineering, pp. 87–100. Springer-Verlag, Berlin, Heidelberg (2009).
16. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: ESWC ’07: Proceedings of the 4th European conference on The SemanticWeb, pp. 624–639. Springer-Verlag, Berlin, Heidelberg (2007).
17. Diederich, J., Iofciu, T.: Finding communities of practice from user profiles based on folksonomies. In: Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs’06) (2006).
18. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., Menczer, F.: Folks in folksonomies: social link prediction from shared metadata. In: WSDM ’10: Proc. 3$^{rd}$ ACM int. conference on Web search and data mining, pp. 271–280. ACM, New York, NY, USA (2010).
19. Au Yeung, C.M., Gibbins, N., Shadbolt, N.: A study of user profile generation from folksonomies. In: SWKM (2008).
20. Gemmell, J., Shepitsen, A., Mobasher, B., Burke, R.: Personalizing navigation in folksonomies using hierarchical tag clustering. In: DaWaK ’08: Proc. 10th int. conf. on Data Warehousing and Knowledge Discovery, pp. 196–205. Springer-Verlag, Heidelberg (2008).
21. Au Yeung, C.M., Gibbins, N., Shadbolt, N.: Contextualising tags in collaborative tagging systems. In: HT ’09: Proceedings of the 20th ACM conference on Hypertext and hypermedia, pp. 251–260. ACM, New York, NY, USA (2009)

22. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In World Wide Web Conference, (2010)
23. Signorini, A.: Swine Flu monitoring using twitter. In http : //compepi.cs.uiowa.edu/ alessio/twitter − monitor − swine − flu/.
24. Jin, X., Gallagher, A., Cao, L., Luo, J., Han, J.: The Wisdom of Social Multimedia: Using Flickr For Prediction and Forecast, MM '10 Proceedings of the international conference on Multimedia, Firenze, Italy (2010)
25. Singh, V.K., Gao, M., Jain, R.:. Social Pixels: Genesis and Evaluation. MM '10 Proceedings of the international conference on Multimedia, Firenze, Italy (2010)
26. Corley, C.D., Mikler, A.R.: A computational framework to study public health epidemiology. In International Joint Conferences on System Biology, Bioinformatics and Intelligent Computing (IJCBS09), Shanghai, China, (2009)
27. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature, 457, 1012--1014 (2009)
28. Eysenbach, G.: Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In AMIA 2006 Symposium Proceedings, pp 244--248 (2006)
29. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D.: Using internet searches for influenza surveillance. Clinical Infectious Diseases (Supplement), pp. 1443--1448 (2008)
30. Johnson, H.A., Wagner, M.M., Hogan, W.R., Chapman, W., Olszewski, R.T., Dowling, J., Barnas, G.: Analysis of web access logs for surveillance of influenza. Stud Health Technol Inform., 107(Pt 2), 1202´lC6 (2004)
31. Hinze, A., Voisard, A.: Location and time-based information delivery in tourism, Advances in spatial and temporal databases. Lect Notes Comput Sci 2750, 489–507 (2003)
32. Lin, Y.R., Candan, K.S., Sundaram, H., Xie, L.: "Scent: Scalable compressed monitoring of evolving multi-relational social networks," in review at ACM Trans. On Multimedia Computing, Communications and Applications (2010)
33. Yang, Y.H., Wu, P.T., Lee, C.W., Lin, K.H., Hsu, W.H., Chen, H.H.: ContextSeer: context search and recommendation at query time for shared consumer photos. In Proc. 16th ACM int. conf. on Multimedia (MM '08). ACM, New York, NY, USA, pp. 199--208 (2008)
34. Wu, X., Ngo, C.W., Hauptmann, A.G., Tan, H.K.: Real-Time Near-Duplicate Elimination for Web Video Search With Content and Context, Multimedia, IEEE Transactions on , 11(2), 196--207, (2009)
35. Henrich, A., Lüdecke, V.: Determining geographic representations for arbitrary concepts at query time. In Proceedings of the first international workshop on Location and the web (LOCWEB '08). ACM, New York, NY, USA, pp. 17---24 (2008)
36. Papadopoulos, S., Zigkolis, C., Kapiris, S., Kompatsiaris, Y., Vakali, A.: City exploration by use of spatio-temporal analysis and clustering of user contributed photos. Demo paper accepted in ACM International Conference on Multimedia Retrieval (ICMR) (2011)