Real-Time Social Media Indexing and Search

Conference Paper · September 2014

DOI: 10.1049/ib.2014.0014

CITATION
1

READS
893

13 authors, including:

Ioannis (Yiannis) Kompatsiaris
The Centre for Research and Technology, Hellas
1,023 PUBLICATIONS **14,035** CITATIONS
SEE PROFILE

David P. A. Corney
Full Fact
39 PUBLICATIONS **1,925** CITATIONS
SEE PROFILE

Matthias Klusch
Deutsches Forschungszentrum für Künstliche Intelligenz
264 PUBLICATIONS **6,326** CITATIONS
SEE PROFILE

REAL-TIME SOCIAL MEDIA INDEXING AND SEARCH

I. Kompatsiaris[1], S. Diplaris[1], D. Corney[2], N. Heise[3], M. Klusch[4], E. Jaho[5], J. Geurts[6], Y. Liu[6], G. Petkos[1], S. Papadopoulos[1], N. Sarris[5], A. Goker[2], J. Spangenberg[3]

[1]Information Technologies Institute, Greece; [2]Robert Gordon University, UK; [3]Deutsche Welle, Germany; [4]German Research Center for Artificial Intelligence, Germany; [5]Athens Technology Center, Greece; [6]JCP-Consult SAS, France

ABSTRACT

This paper presents a real-time system that incorporates emerging knowledge from social media using crawling and mining techniques, developed within the SocialSensor FP7 project, that are designed for addressing the particularities of social web . The surfaced information is presented using an interface that is optimized and adapted to the context of the user, taking into account efficient content delivery techniques to optimize quality of experience and enhance user interaction. Research approaches are discussed, particularly focusing in social media analytics, trending topic detection, multimedia indexing and search, and efficient social content delivery and service coordination in mobile settings. We also discuss how these new methodologies are incorporated to build novel tools for the News and Infotainment domains.

INTRODUCTION

Social networks combined with smart mobile technologies have become an integral part of modern life driving more and faster communication than ever before. Sharing of media content at the time and place of an event creates an “online reflection” of what happens in the real world. However, the fast pace, the huge volume and the unpredictable nature of user-contributed content make it extremely challenging to obtain informative views of evolving news stories and events in real time.

Especially regarding the news industry the challenge is to embrace new media content authoring and provision methods to their fullest advantage, for both information gathering and distribution. A key pertinent challenge is to develop appropriate tools for quickly surfacing trends, sentiments and discussions in social media, in relevant and useful ways.

Real-time social media analytics are valuable also for the organisers of large events, such as festivals and expos. New methods providing context-aware information are becoming necessary to enhance user experience and help attendants organise their visits. Event organisers can also benefit from novel tools that capture the pulse of their events and help them make sense of the large amounts of online shared content.

This paper presents an algorithmic framework, developed in terms of the FP7 EU project SocialSensor, enabling real-time multimedia indexing and search across multiple social media sources. It places particular emphasis on the real-time, social and contextual nature

of content and information consumption and integrates topic and event detection, mining, search and retrieval, based on aggregation and indexing of shared user-generated content. The framework introduces the concept of Dynamic Social COntainers (DySCOs) as a layer of multimedia content organisation, i.e. an aggregate representation of data, content, metadata and inferred data around a topic, story or event. Through the proposed DySCOs-centered media search, SocialSensor integrates content mining, search and intelligent presentation in a personalised, context and network-aware way, based on aggregation and indexing of both UGC and multimedia Web content.

On top of the analysis results, SocialSensor delivers practical tools that incorporate novel user-centric media visualisation and browsing methods in the domains of News, with the goal of facilitating the discovery of newsworthy social multimedia, and in Infotainment, helping event organisers extract insights from large events by mining large amounts of online messages shared through social media.

The following sections present algorithmic approaches for topic detection, social multimedia indexing, social media verification, and peer-to-peer coordination of media services that utilize the above framework in order to build applications for the two use cases of news and infotainment that are later presented in detail.

TOPIC MINING

Of the millions of tweets posted each day, only a small number are related to significant news events. We therefore only collect messages from a small, group of active, news-focussed accounts, such as journalists, commentators and bloggers. As we collect the tweets, we also analyse them to discover topics in real-time.

Most topic mining techniques belong to one of three categories. First, f*eature-pivot* approaches mine topics in a corpus of documents by grouping together textual features according to their co-occurrence patterns. Second, d*ocument-pivot* approaches identify topics by clustering together documents - instead of features – according to their textual similarity. Finally, *probabilistic topic models* discover topics by performing probabilistic inference on the joint probability distribution of topics and terms.

We explored and evaluated several topic mining techniques during the development of the SocialSensor system. For instance, two varieties of a document-pivot approach were tested: first, a variation of a classic method that uses the cosine similarity of the standard *tf-idf* representations as the similarity metric between documents [1]; and second, a novel approach using joint complexity as the similarity metric [2]. However, several experiments with different datasets showed that a third approach performs best: a novel feature-pivot approach named BNgram (for bursty *n*-gram) [3].

Having developed several topic detection techniques, we needed a way to compare and evaluate them. Recently, as part of this process, we organized an international data challenge as part of the SNOW 2014 workshop [4]. This challenge produced a benchmark and evaluation resource for the problem of news-focused topic detection from social media streams. Participants collected over one million news-related tweets over a 24-hour period. We then defined a “ground truth" consisting of 59 stories selected from mainstream media (newspapers, TV news sites etc.).

Of the eleven teams that completed the challenge, the teams finishing first and second both used variants of the BNgram approach outlined above. The winning team used a strong set of filters to remove less useful tweets, including those with multiple hashtags or which are very short. This system found two-thirds of the target stories along with many

others extending into the long-tail of topics. SocialSensor's own BNgram method finished a close second; our joint complexity method finished third, and our *tf-idf* document-pivot method finished fifth.

One outcome of the evaluation was that no single team dominated all aspects of the evaluation. This suggests that combining different aspects of different systems may lead to even better topic mining systems in the future. All the data is now available for others wishing to evaluate their own work against a fixed benchmark.

SOCIAL MULTIMEDIA INDEXING

The platform supports the targeted collection, indexing and browsing of shared media content through a hybrid crawling strategy, comprising both a stream-based and a query-driven approach. In addition, it integrates very efficient and scalable image indexing and clustering implementations.

The crawler is responsible for the collection of data and content from online sources in the form of Items (posts made in a social platform, e.g. tweets), WebPages (URLs embedded in collected Items) and MediaItems (images/videos embedded in Items or WebPages), given a set of crawling specifications (arguments specifying what to crawl, e.g. a hashtag

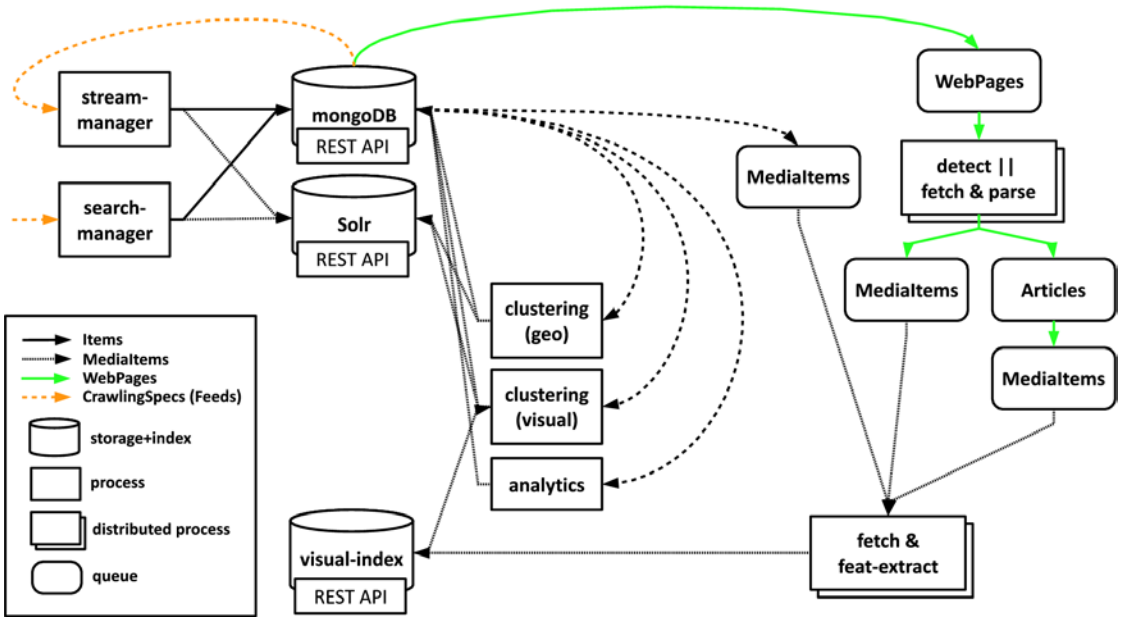

Figure 1 – Social multimedia crawling and search.

on Twitter) as input. The proposed crawling and indexing framework, depicted in Figure 1, elaborates on the generic SocialSensor architecture, comprising the following components: (a) Item collection, comprising stream-based (stream-manager) and query-driven (search-manager) components, (b) structured data repositories based on mongoDB and Solr, (c) fetching and indexing components, including WebPage fetching, article extraction, MediaItem detection and extraction, feature extraction and visual indexing [5], (d) aggregation components, including geo, visual clustering and analytics. Several of the system components are available as open-source projects on GitHub, intended for use in the professional journalism domain [6].

SOCIAL MEDIA VERIFICATION

The verification of information in social media is important for dealing with negative aspects of information dissemination, such as the spreading of rumours and fake news. Within SocialSensor, the problem is approached from a journalistic aspect, from three different angles: *Contributor*, *Content* and *Context* validity analysis. Analysis of the validity of *Contributor* means analysis of any information relating to the source of information, such as trust, reputation and influence of this source. The *Content* angle relates to the analysis of data such as the language used, the reliability of references to web data and the discovery of possible manipulations in linked multimedia content. Finally, *Context* analysis examines whether the 'what', 'when' and 'where' concur with each other. This involves the analysis of metadata revealing information about the geospatial and temporal dimensions of content, its provenance and relation to other similar information. Joint analysis of the validity of *Contributor*, *Content* and *Context* provides a more thorough approach for revealing trustworthiness compared to previous works [7, 8].

Our approach is to collect information from social media and analyse it with respect to the above properties in near-real-time to derive quantified parameters, called modalities. The modalities are then combined together to provide a single metric that gives a sense of trustworthiness to the reader. At the same time, all evidence that leads to this result is provided, so that the user can also subjectively decide on how credible every piece of information is. This framework is embedded in the 'Truthmeter', a web tool integrated into the SocialSensor system. The tool is currently under development, and will undergo a series of evaluations, aiming to examine the correctness of the produced rankings and weightings. A more extended description of the verification framework and the rating of modality parameters is available in [9].

PEER-TO-PEER COORDINATION OF MEDIA SERVICES

Semantic P2P Media Search and Network Adaptive Live Streaming

SocialSensor also adds a new dimension to social interactions by enabling users to jointly search, share and experience relevant media and live recordings in a mobile P2P network. For this purpose, the semantic middleware of the SocialSensor framework offers two innovative key techniques: High-performance dynamic semantic search in unstructured P2P networks (S2P2P) [10] and dynamic adaptive P2P video streaming over HTTP (pDASH) [14] based on the ISO/IEC standard MPEG-DASH [16]. Both techniques are used, for example, in the mobile P2P application MyMedia for Android devices [15] which is part of the infotainment use case of SocialSensor.

Semantic P2P media search

Each peer uses its observations of the semantics of all traversing queries and items to generate a local view of the semantic overlay of the unstructured P2P network it is part of. Media items like film trailers or live recordings are locally maintained and offered by peers through media services with semantics described in OWL-S. Service descriptions are generated by peers based on respective user input, a local tag cloud and domain ontology in OWL2. The highly precise semantic selection of relevant media services is performed with the hybrid semantic service matchmaker iSeM [12, 13]. Any query q is routed by peers with their S2P2P component [10] through the network on a shortest path with a maximal amount of semantic expertise of peers for q which are reachable within the given time-to-live (TTL). Once some relevant video content is found the question is how a peer group of users can synchronously watch it at the best quality possible?

Network adaptive P2P live streaming

Each peer maintains and encodes its local videos on Android devices according to the MPEG-DASH standard. The videos are adaptively segmented by the peer for different network bandwidths per streaming period in advance and the respective XML-based media presentation description files are created. It then uses its pDASH component to perform a network adaptive and pure P2P live streaming [14]. Each peer of the mobile P2P network dynamically keeps track of and exploits its gradually evolving local knowledge on other peers which downloaded segments of relevant DASH-encoded media. Unlike other current P2P streaming applications there is no central server-based streaming or tracking involved while the content is optimally distributed among peers based on preferred quality of service and available network bandwidths.

The preliminary experimental evaluation of the performance of P2P live streaming with pDASH was based on simulations in OMNeT++ using the INET framework with 40 peer clients with different bandwidths for each of them, an HTTP server with bandwidth of 1Gbit/s and 10 representations of video content with different bitrates and resolutions. The results showed that the bandwidth savings by peer-assisted streaming are lower for a representation bitrate limit of 6 Mbit/s than for 1.4 Mbit/s since more different representations were requested by peers in this case [14]. Besides, the DASH encoding of MPEG video live streams consumes most energy (e.g. for 1080p @ 1Mbit/s: 54.7mJ) in contrast to the remaining peer functionalities of the MyMedia app (5.8mJ). Furthermore, the averaged search precision and latency of P2P live streaming between peers of the MyMedia app with S2P2P and pDASH appear acceptable for users in practice. For example, the latency time of jointly watching a video among four peer networked Android devices is about 4 seconds. Finally, the experimental evaluation of semantic search with S2P2P [10] in a random power law graph P2P network with one million peers, random distribution of items and ten thousand queries (TTL=20, k=3) to peers with different large ontologies in OWL2 generated from dbpedia revealed that S2P2P is robust and offers an average precision of 0.68 (over 0.57 for k-random search) and 0.82 if combined with semantic replication [11].

Disconnection Prediction

Loosely controlled WIFI networks, such as those deployed at festivals and events, typically do not provide full coverage of the terrain. As a result, the experience of a mobile user watching a video stream is rather erratic, because of frequent disconnections. However, these periods of disconnections are generally short (10-50 secs for coverage of around 70%). To ensure continuous experience in this case one needs to graciously handle these “connection” gaps until a client re-connects to the subsequent access point. SocialSensor developed a disconnection prediction (DP) and a pre-fetcher (PF) to successfully initiate a transparent handover ensuring continuous video streaming experience. The DP predicts disconnections and reconnections based on a collaboratively generated connectivity map. Once a disconnection is imminent, the DP also has an understanding when the user is expected to reconnect. This information is used to activate the PF, which starts to download video material to the user’s device, which will be needed during the disconnected period. When the mobile user physically disconnects from the network, the video will continue to play as the requested data has been buffered on the mobile device. When she re-enters a connected area, the wireless network connection is restored and operation is resumed as normal.

The technology has been evaluated in various scenarios using network simulators (NS-3)

and geo-spatial user mobility generators (BonnMotion). Promising results include 85% prediction accuracy for an area of about 2 square km consisting of 7 hotspots and 50 mobile users. We are currently integrating the technology in the SocialSensor mobile App and planning field studies.

USE CASES

The SocialSensor News Use Case

Content residing in Social Networks plays an important role when it comes to information gathering as it can represent a valuable additional source. It thus allows for speedier reporting, the use of previously untapped sources, or access to information that would otherwise be unavailable. However, the use of Social Media in news reporting also includes a number of challenges. Many of these were tackled in the SocialSensor news use case: we analysed the needs of news professionals, and casual newsreaders.

With regard to news professionals, the SocialSensor team identified three target groups that guided the elicitation of user requirements and the development of the SocialSensor prototype as well as of individual components: a) *Editors and managers* interested in spotting stories or angles that rivals may have missed and understanding the level of audience interest in a particular story at any point in time. b) *Reporters and producers* directly engaged in the process of gathering and communicating news in text, video or audio. c) *Social media specialists* that identify trends for the wider organisation or particular sections and inject findings into the editorial process. Based on the needs of the above target groups, the SocialSensor professional news use case application focuses on the following three aspects: a) Identifying and visualising events and trends across social media sources in real time. b) Identifying key influencers and opinion formers around any event in real time. c) Creating a simple way to verify or authenticate user generated content (text, images, video and audio) from social media sources.

With regard to casual newsreaders, SocialSensor has analysed the needs of the heterogeneous group of non-professionals who frequently consume social media content in many different ways and devices. Although the specific needs very much depend on usage context and circumstances, personal interests and educational background, the SocialSensor casual newsreaders use case has identified the following key aspects: a) Alerting about breaking news in specific areas of personal interest in near real time. b) Identifying relevant multimedia content that is related to a specific area of personal interest. c) Providing links to related articles from different sources.

To make sure that development work and subsequent research results meet the formulated user needs, several evaluation rounds took place in the course of the project. Some are still ongoing. The first version of the SocialSensor professional news application was validated in a number of in-depth usability sessions and by a focus group of journalists. The casual news application was tested in a series of qualitative surveys among frequent consumers of online news. Overall, the testing strongly confirmed the need for a system that makes it easier to surface and analyse content residing in social networks, based on personalised criteria. The fine grained control across time and across networks was of great interest to professional journalists, whilst the range and breadth of multimedia content attracted general users. The simplicity of the mobile interface showed much promise along with new approaches to verification and credibility assessments.

Based on the evaluation findings, the final efforts of SocialSensor R&D work will be spent on improving key features and components rather than continuing to build an overarching multipurpose system. The main foci are “trending” and the detection of relevant multimedia items, as well as the identification and credibility of key influencers.

The SocialSensor Infotainment Use Case

To satisfy the requirements of different types of users in Infotainment events, namely event organisers and attendants, the SocialSensor Infotainment use case builds upon two research directions, the EventSense and the EventLive frameworks respectively: a) EventSense and b) Event Live.

EventSense is a social media sensing framework that can help event organisers and event enthusiasts capture the pulse of large events and gain valuable insights into their impact on visitors. Online messages about the event are organised around entities of interest (e.g. films) and topics, and sentiment scores are extracted, by aggregating the sentiment expressed by individual messages. This kind of aggregation enables the ranking of entities, topics and online users based on social interest and disposition, and thus conveys a succinct and informative view of the event highlights. The aggregated information is communicated to the organisers in the form of an online dashboard. Through a real-world evaluation on the 53rd Thessaloniki International Film Festival, it became evident that real-world event variables, such as film ratings, are correlated with aggregate statistics mined from the stream of online messages.

EventLive pertains to an intelligent real-time app that incorporates advanced social media search and analysis features, intuitive visualisations and contextual recommendations, to deliver relevant and timely content to event attendants, and to ultimately leverage the users’ event experience. To showcase its EventLive prototype, SocialSensor supports with mobile apps the Thessaloniki International Film and Documentary Festivals, and the Fête de la Musique Berlin event, a yearly music event that takes place in Berlin. A first version of the EventLive prototype was developed, featuring social recommendations, sentiment analysis, film-tweet matching, video play, network-adaptive live streaming and recording, P2P media search and share, and 2D-3D visualisation modules.

An initial evaluation of the system was carried out including immersive user testing in actual festival environments, and validations of the prototype by selected focus groups via questionnaires, in an effort to assess the performance, innovation, added value and USP of the research and project outcomes aggregated in the prototypes. This first set of evaluations highlighted the potential of the Infotainment application to enhance the overall event experience (for event attendants) and provide valuable insights and analytics to event organizers via the technologies coming from SocialSensor.

CONCLUSIONS

In this paper we outlined and highlighted an algorithmic framework for real-time social media content indexing, search and delivery. SocialSensor aspires to provide such tools for professional journalists, casual newsreaders, organisers and attendees of large events by innovative analysis techniques of social media, assisted by effective indexing of real-time social media streams. Evaluation studies demonstrate the effectiveness of such techniques in collecting diverse content from social networks, in analysing and searching it in multiple ways, and in perceiving the pulse of large-scale events.

REFERENCES

1. Petkos, G. et al., I. Two-level message clustering for topic detection in Twitter. In Proceedings of the SNOW 2014 Data Challenge, 2014.
2. Burnside, G., et al. One Day in Twitter: Topic detection via joint complexity. In Proceedings of the SNOW 2014 Data Challenge, 2014.
3. Martin, C. and Goker, A. Real-time topic detection with bursty n-grams. In Proceedings of the SNOW 2014 Data Challenge, 2014.
4. Papadopoulos, S., et al. SNOW 2014 data challenge: Assessing the performance of news topic detection methods in social media. In Proceedings of the SNOW 2014 Data Challenge, 2014.
5. Spyromitros-Xioufis, E., et al. An empirical study on the combination of surf features with VLAD vectors for image search. In Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on. pp. 1 to 4.
6. Papadopoulos, S., et al. Social Multimedia Crawling and Search. In IEEE Computer Society Special Technical Community on Social Networking E-Letter, vol. 1, no. 3, October, 2013.
7. Canini, K. R. et al. Finding credible information sources in social networks based on content and social structure. In IEEE SocialCom/PASSAT, pp. 1 to 8., 2011.
8. Cha, M., et al. Measuring user influence in twitter: The million follower fallacy. In Proceedings of international AAAI Conference on Weblogs and Social, 2010.
9. Jaho, E., et al. Alethiometer: a Framework for Assessing Trustworthiness. International World Wide Web Conference Committee (IW3C2), April, 2014, Seoul, Korea.
10. Cao, X., Klusch, M., 2013. S2P2P: Semantic Search in Unstructured Peer-to-Peer Networks. Proceedings of 15th IEEE International Conference on High-Performance Computing and Communications (HPCC)
11. Cao, X., Klusch, M., 2012. Dynamic Semantic Data Replication for K-Random Search in Peer-to-Peer Networks. Proceedings of 11th International IEEE Symposium on Network Computing and Applications
12. Klusch, M., Kapahnke, P., 2012. The iSeM Matchmaker: A Flexible Approach For Adaptive Hybrid Semantic Service Selection. International Journal of Web Semantics, vol. 15, Elsevier.
13. Klusch, M., 2012. The S3 Contest: Performance Evaluation of Semantic Service Matchmakers. Semantic Web Services – Advancement through Evaluation, Springer.
14. Lederer, S., et al. (2012): Towards peer-assisted dynamic adaptive streaming over HTTP. Proceedings of 19th International Packet Video Workshop (PV)
15. Rainer, B. et al. 2014. Real-time Multimedia Streaming in Unstructured Peer-to-Peer Networks. Proceedings of 11th IEEE Consumer Communications and Networking Conference (CCNC)
16. ISO/IEC 23009-1. 2012. Information Technology - Dynamic Adaptive Streaming over HTTP (DASH) - Part 1: Media Presentation Description and Segment Formats.

ACKNOWLEDGEMENTS

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.