Recommending Venues Using Continuous Predictive Social Media Analytics

Marco Balduini
Politecnico di Milano

Alessandro Bozzon
Delft University of Technology

Emanuele Della Valle
Politecnico di Milano

Yi Huang
Siemens Corporate Research and Technology

Geert-Jan Houben
Delft University of Technology

The authors' Continuous Predictive Social Media Analytics system operates in real time on social media streams and graphs to recommend venues to visitors of geo- and temporally bounded city-scale events. By combining deductive and inductive stream reasoning techniques with visitor-modeling functionalities, this system semantically analyzes and links visitors' social network activities to produce high-quality link predictions when information about preferences is sparse. The authors demonstrate their system's quality with experiments on real-world data.

The diffused use of location-based social networks via smartphones scales up geographically and temporally bounded events to a broader audience. This phenomenon offers a window on people's interests, habits, and preferences, conveying an increasingly accurate and up-to-date representation of reality.[1,2] This graph of interests, places, and people is constantly reshaping. Many applications would benefit from being able to track and predict its changes. For instance, by observing social media for conversations about an event, an application could predict new activities and offer personalized recommendations for event-related content, offers, or advertisements.

Deductive and inductive stream reasoners[3] can address such applications' needs by tracking how social media phenomena evolve in a graph of webpages,[3] places,[4] and users, and recommending new links. The key to success is adopting techniques that can resolve semantic ambiguity in the processed data and having a sufficiently large time window of training data for accurate real-time predictions. For instance, previous approaches required at least three months of social conversation to train a system to deliver accurate predictions about microblog posts mentioning physical locations.[3,4] This raises questions about what to do if such training data isn't available: How can a system predict the appearance of links in a graph when the graph is sparse – that is, when little or no training information about existing links is available?

To make this problem concrete, consider city-scale events such as the 2013

Milano Design Week (MDW13). Geographically limited to the Milan area, MDW13 featured 681 venues (normally employed as public spaces, parking lots, bars, and so on) that served as temporary exhibition centers for hosting 1,127 events attended by 500,000 visitors in one week. MDW13 is a scenario in which exploiting social media conversation for visitor-venue prediction is difficult. First, past microposts (such as tweets) are useless. Second, historical data about the relationship between visitors, venues, and events isn't available because these elements change across MDW editions. Third, background information about venues is irrelevant, because their functions change during the event. Given these limitations, the central question is how a system can predict which venues visitors will mention in social media. As a running example, let's use a visitor named Alice who performs a FourSquare check-in at the Apple Temporary Showroom the first day of the event: With only this information available, how can the system recommend a new venue for her to visit?

We present a novel Continuous Predictive Social Media Analytics (CP-SMA) system and provide experimental evidence of its quality using a real-world and real-time evolving graph collected during MDW13. Choosing a robust machine-learning-based predictive model that isn't sensitive to its model-specific parameters is relevant because a lack of historical data makes it nearly impossible to identify the optimal configuration upfront. The most-talked-about venues play a key role in predicting new visitor-venue links. We empirically demonstrate that by including visitor-modeling strategies — which semantically analyze and link visitors' social network activities to produce historical and event-related topical profiles — our system can effectively gather relevant information for link prediction.

Continuous Predictive Social Media Analytics

Social media analytics (SMA) identifies a set of methods and technologies for collecting, monitoring, analyzing, summarizing, and visualizing media generated and disseminated in a conversational and distributed mode among online communities.[2] In our previous work, we specified a generic and flexible stream reasoning architecture for SMA.[3] This system listens to streaming APIs in microblogging services and continuously processes the input stream with a deductive stream reasoner to enrich it with implicit facts derived from explicit ones. The system then instructs an inductive stream reasoner to find potential facts that haven't been observed but are likely to become true in the future. We used best practices from linked data to ensure interoperability among the components, using RDF streams for data exchange and SPARQL (extended to cope with streaming and probabilistic data) as the query language.

Here, we build on this successful foundation and propose a novel CP-SMA system that includes semantic user profiling capabilities as offered by the U-Sem framework,[5] a user modeling infrastructure for extracting semantically enriched knowledge about social-Web users.

Figure 1 depicts CP-SMA's architecture using a city-scale event example. The *social listener* (SL), which specializes the deductive stream reasoner, focuses on identifying phenomena on location-based microblogging streams. The SL continuously processes microposts to link them to venues and thus identify the most-talked-about ones and the most active users.

The *visitor modeler* (VM) analyzes relevant information on the social Web to create semantically enriched user models. For each social network user identified as an event visitor, the VM builds both an historical and an event profile. Historical profiles semantically describe the visitors' interests as expressed before their presence at the monitored event. The VM builds event profiles by analyzing visitors' social activities during the event. It supplies both profiles to the *visitor-venue recommender* (VVR) to enrich and personalize predictions. The VM also supplies analytical information about visitors and their online conversations, such as their demographics, trends, online presence, and influence.

The inductive stream reasoner is specialized in the VVR to predict visitor-venue links that are likely to be observed in the future. The VVR is trained with visitor-venue links and visitor profiles from the SL and VM, respectively. It then applies the resulting model to predict, in real time, links that the system can use to recommend venues to visitors. Because the training step is time consuming, it's carried out in batch mode using the *statistical unit node set* (SUNS),[6]

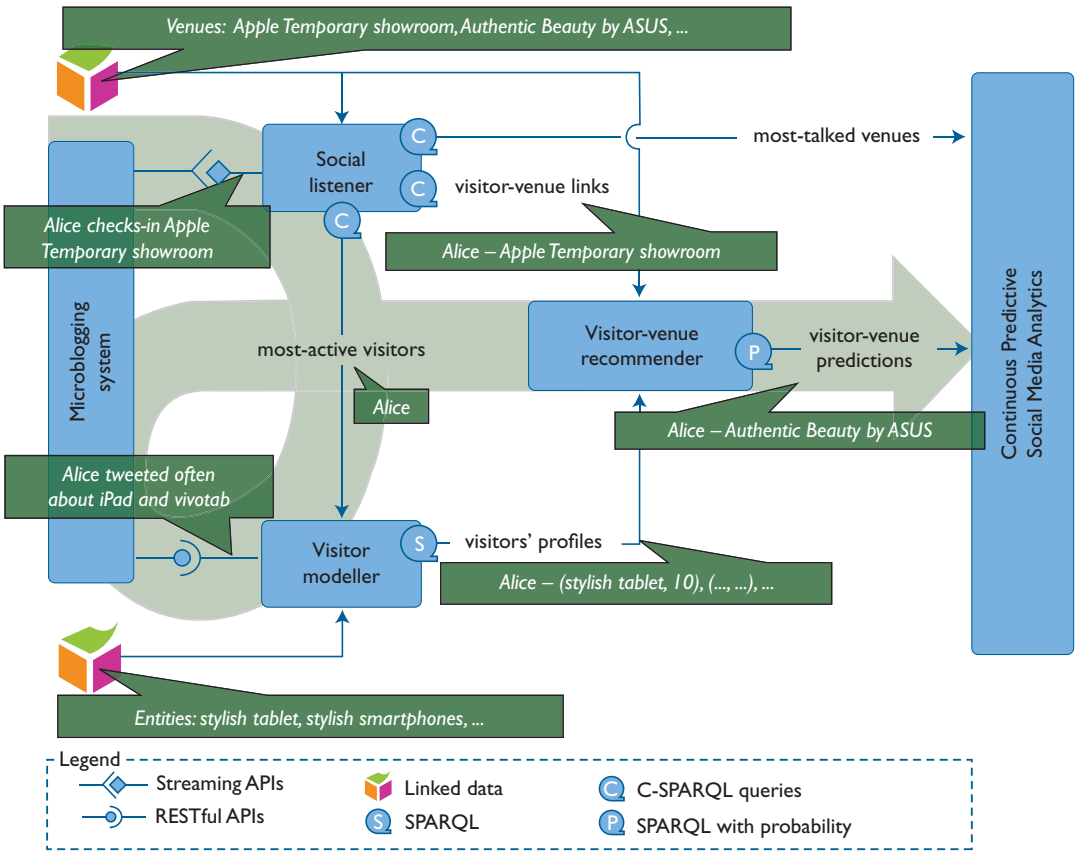


Figure 1. Architecture of the Continuous Predictive Social Media Analytics (CP-SMA) system. Microblogging systems push microposts to the social listener that are linked to venues for the monitored city-scale event. Those who post them are identified as visitors, and the visitor modeler builds their profiles by linking their past tweets to semantic entities. The visitor-venue recommender uses the venue-visitor links and the visitor profiles to predict unseen visitor-venue links and, thus, recommend venues to visitors.

a scalable link-prediction approach. The SUNS predictive model can self-advance through continuous updating — for example, in a predefined time interval — to incorporate changes in the graph due to new visitor-venue links or new or updated profiles. These features enable the VVR to cope with social media stream dynamics.

Considering the running example, the CP-SMA system will perform as follows. When Alice checks in at the `Apple Temporary Showroom`, the SL receives the tweet this check-in generates and extracts the link `Alice – Apple temporary showroom`. The VM builds Alice's historical profile based on her past tweets, indicating that she's interested in `stylish tablets` and `smartphones`. The combination of her historical profile and the known links associated with venues lets the VVR recommend that `Alice` visit `Authentic Beauty`, the venue for the Asus computer and tablet manufacturer.

Recommending Venues for City-Scale Events

We tested our CP-SMA system on the Twitter stream recorded for MDW13 (9–14 April).[1] To simulate a real-world deployment, the system at startup has access to general linked data (that is, DBpedia; http://dbpedia.org) and MDW13-specific data: an address, geocoordinates, and a set of events describe each MDW13 venue. Each event has a name and a set of categories (such as hotel architecture, lighting design, or consumer electronics). No information about visitors is initially available.

The SL continuously processes tweets geolocated in Milan to link them to one or more MDW13 venues. It listens to the result stream of a spatial-bounding-box query registered in Twitter streaming APIs and creates time-boxed buckets of tweets using the underlying stream

Table 1. Evaluation of content-linking strategies for venue-tweet linking.

Method	Input A	Input B	Precision (%)	Recall (%)	F1 (%)
Automatically generated regex	Tweet content	Venue name	100.00	65.56	79.20
Dice distance, top 15%	Tweet content	Venue name	73.45	46.68	57.08
Dice distance, top 15%	Tweet content	Event name	62.62	29.72	40.31
Jaccard distance, top 15%	Tweet content	Venue name	80.32	26.05	39.34
Levenshtein distance < 3, top 30%	Tweet hashtag	Event name	51.95	18.62	27.41
Levenshtein distance < 5, top 30%	Tweet hashtag	Venue name	48.21	16.52	24.61
Dice distance, top 15%	Tweet hashtag	Event name	66.87	9.70	16.95
Dice distance, top 15%	Tweet hashtag	Venue name	84.62	5.77	10.80
Jaccard distance, top 15%	Tweet hashtag	Venue name	82.76	4.20	7.99

reasoner's time window feature. The SL performs linking by considering both spatial aspects (the post occurs in the venue's surroundings) and content aspects (the post talks about an event hosted in a venue). For each aspect, the SL can apply different strategies.

Table 1 reports various content-linking strategies' performance, where both precision and recall[5] are important in providing accurate and abundant training data. We compute precision and recall on a set of links picked from the 57,154 recorded tweets. With the exception of the Levenshtein edit distance, similarity metrics based on the distance between a tweet's content and the name of venues or events provide good precision but low recall, whereas using hashtags as a basis for comparison generally leads to poor performance. Note that before computing the distances, we cleanse the tweet content by removing special characters, stop words, and words that identify the event or the district.

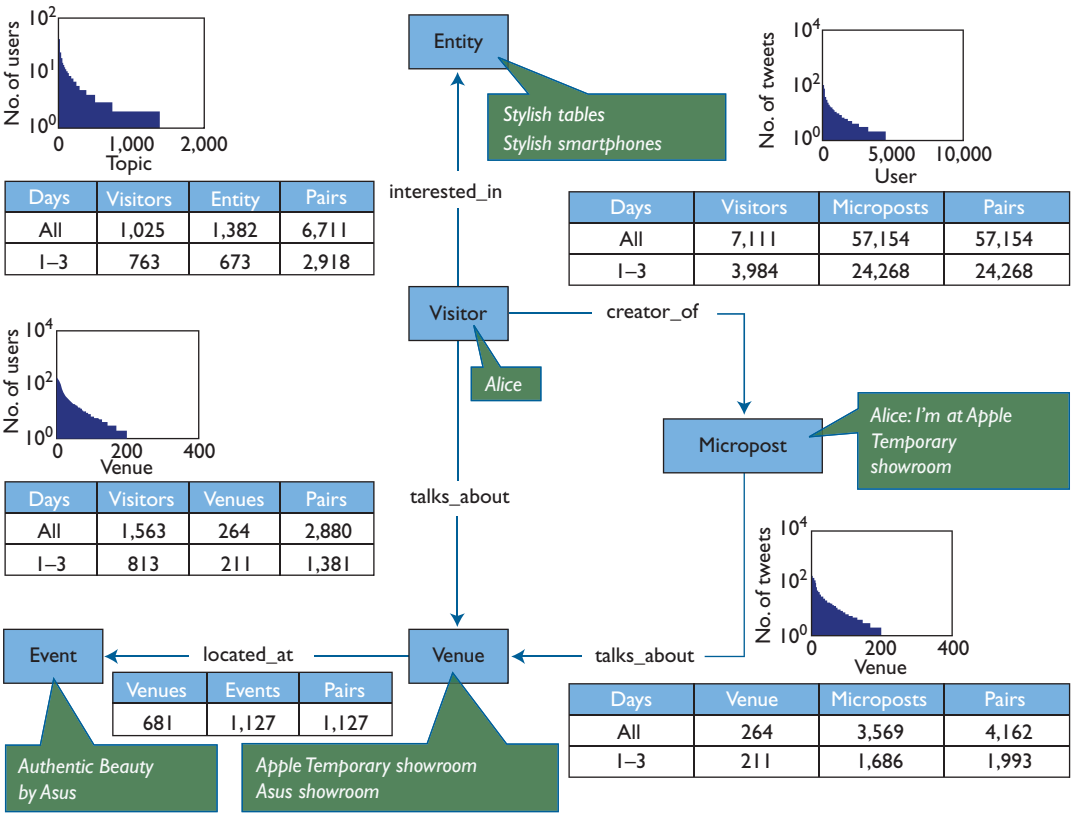
Best performances are achieved through regular expressions that are automatically generated from MDW13-specific data and manually tuned for maximum accuracy. (For instance, the venue "Paolo Curti and Annamaria Gambuzzi Gallery" generates the expression `(Curti|Gambuzzi)`, where "gallery," "Paolo," and "Annamaria" are ignored as domain- and language-dependent stop words.)

The SL notifies the VM about the creator u of a new tweet: if the system knows u, it updates his or her event profile. Otherwise, the VM starts a crawling activity on the social graph to build the historical visitor profile.

Profiles are created by distilling from tweets topics and entities that visitors are concerned with. The VM relies on DBpedia Spotlight to extract semantic entities (identified with a DBpedia URI), such as people and organizations. The connection between semantically enriched microposts lets the VM construct a profile, defined as a set of pairs $(e,w(u,e))$, where e is an entity. The weight $w(u,e)$ models the importance in u's profile of the entity e, and we can calculate it according to different strategies. In our experiments, the VM counts the occurrences of the visitor's social activities.

Figure 2 illustrates the graph that CP-SMA manages, annotated with data from the running example and enriched with some statistics about the MDW13 dataset. The dataset contains 57,154 tweets from 7,111 visitors, half of which were observed during the first three days. The SL linked a subset of 3,569 tweets created by 1,563 visitors (813 in the first three days) to 264 venues (211 in the first three days). Notably, between the third and the last day, the number of visitors doubled. The VM processed more than 2.7 million tweets to create historical (94,806 entity instances) and event (1,383 entity instances) profiles.

The input data for the VVR consists of two parts. The first covers the visitor profiles that the VM generates, whereas the other is visitor-venue links the SL detects. The VVR transforms the incoming data into a matrix that's typically high-dimensional and sparse. In this situation, low-rank matrix-factorization methods have been successfully applied to model interactions in the graphs.[7] In particular, the winning entry to the Netflix competition (which determined the next title each user would rent) used low-rank matrix-factorization approaches.[8] All venues are jointly recommended such that statistical strength can be shared between them. The most important model-specific parameter is the number of latent variables — that is,



Days	Visitors	Entity	Pairs
All	1,025	1,382	6,711
1–3	763	673	2,918

Days	Visitors	Microposts	Pairs
All	7,111	57,154	57,154
1–3	3,984	24,268	24,268

Days	Visitors	Venues	Pairs
All	1,563	264	2,880
1–3	813	211	1,381

Venues	Events	Pairs
681	1,127	1,127

Days	Venue	Microposts	Pairs
All	264	3,569	4,162
1–3	211	1,686	1,993

Figure 2. Data model and data statistics. The Continuous Predictive Social Media Analytics (CP-SMA) system knows who talked about given topics in the past and about venues for the mo`nitored city-scale event in that person's current microposts. Information about events hosted in the venues is also available.

the data matrix's rank. As mentioned, we train the VVR model using SUNS, which is a regularized matrix-factorization approach. Regularization can reduce the model's sensitivity to the number of latent variables. The VVR applies the trained model to suggest the most interesting venues to visitors, based on both their profiles and those venues they've talked about during the event.

Evaluation

To evaluate our system's performance, we considered two settings. The first, Day6, provides predictions based on data collected at the event's end, when both visitor profiles and visitor-venue links are completely available. The second scenario, Day3, is the actual target of our research, and considers as training data only the graph available three days from the event's beginning. On Day3, approximately 50 percent of the data — that is, visitor profiles and visitor-venue links — wasn't available, and some venues had yet to appear in the social stream. In each setting, we randomly withheld a visitor-venue link for each visitor who talked about at least two venues. Consequently, we selected roughly 300 links as ground truth in the test phase. We trained the inductive stream reasoner using all remaining links and then recommended venues for all visitors. We repeated this data split five times.

We evaluated prediction performance using normalized discounted cumulative gain (nDCG)[9] to account for the predicted venue rankings. Figures 3 and 4 depict the system's performance in both settings with different prediction techniques. Using only visitor-venue links, we compared SUNS' prediction capabilities with

- a *Random* ranking for unknown venues for each visitor, where we assigned a random

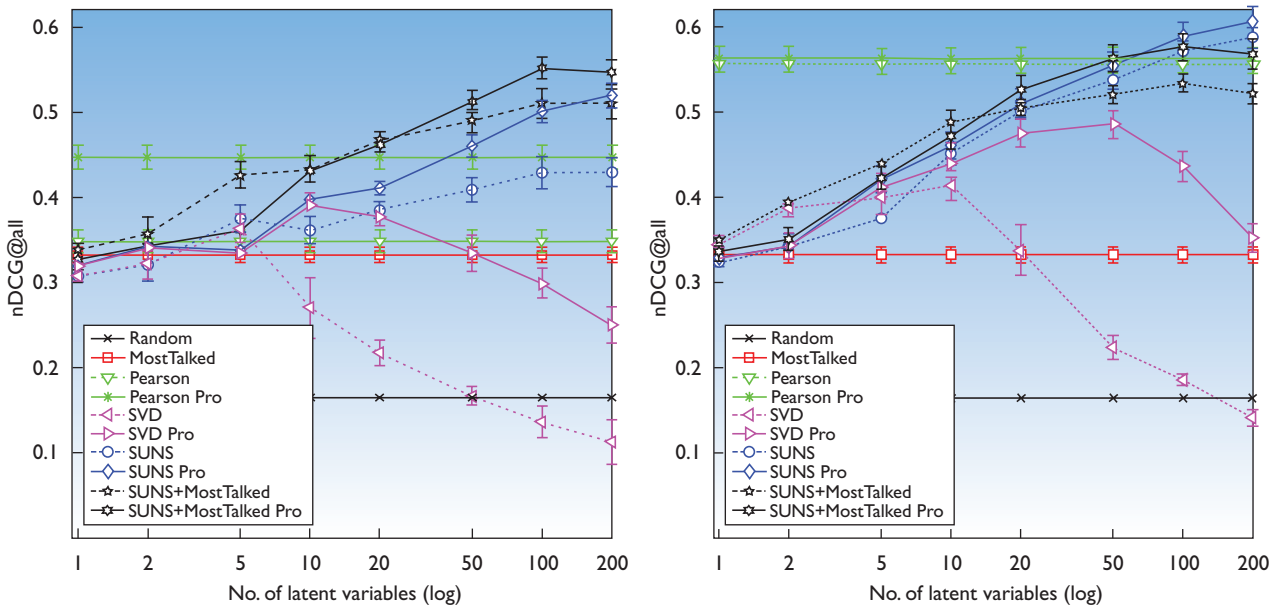


Figure 3. The nDCG@all values. For (a) Day3 and (b) Day6, we measured these values over the number of latent variables for Continuous Predictive Social Media Analytics methods (SUNS, SUNS Pro, SUNS+MostTalked, and SUNS+MostTalked Pro) *as well as several baselines.*

likelihood to every venue not mentioned by a particular visitor;
- the *Pearson* correlation coefficient, where predictions are based on the similarity of visitor-venue links; and
- a list of venues ranked by popularity (*MostTalked*).

In addition, we compared a machine learning method based on *singular value decomposition* (SVD) to a method combining *SUNS* with *MostTalked* venues (*SUNS+MostTalked*), which models both venue popularity and visitor preferences. To study how visitor profiles affect prediction quality, we incorporated them in *Pearson*, *SVD*, and *SUNS* (*Pearson Pro*, *SVD Pro*, and *SUNS Pro*, respectively) and compared the resulting performances.

Figure 3 shows the advantages of using a predictive model that doesn't require identifying the optimal configuration upfront. The figure plots nDCG@all values against the number of latent variables we use to configure a method. Note that the baselines *Random*, *Pearson*, and *MostTalked* are independent of the number of latent variables. Intuitively, a lower sensibility to the number of latent variables makes a method more appropriate to answer our research question, because no historical data is available in advance.

As expected, on Day3 (Figure 3a), *Random* had the worst performance, while *Pearson* achieved a good prediction quality. *SVD* achieved its best performance when using 10 latent variables, but dramatically decreased thereafter. In contrast to *SVD*, *SUNS*' performance kept increasing as the number of latent variables rose. *SUNS* reached its optimum at 200 latent variables, with no later observable descent. Interestingly, *SUNS+MostTalked* achieved the overall best nDCG performance using 100 latent variables. Including visitor profiles significantly increased nDCG values for all considered models, clearly demonstrating the advantage gained by including visitor profiles of the VM. We obtained the best results by combining CP-SMA's three subsystems.

The Day6 results (Figure 3b) confirmed the same relative order of performance for the tested prediction models. Notably, *SUNS+MostTalked* was no longer the best, but was outperformed by *SUNS Pro* (which incorporates visitor profiles).

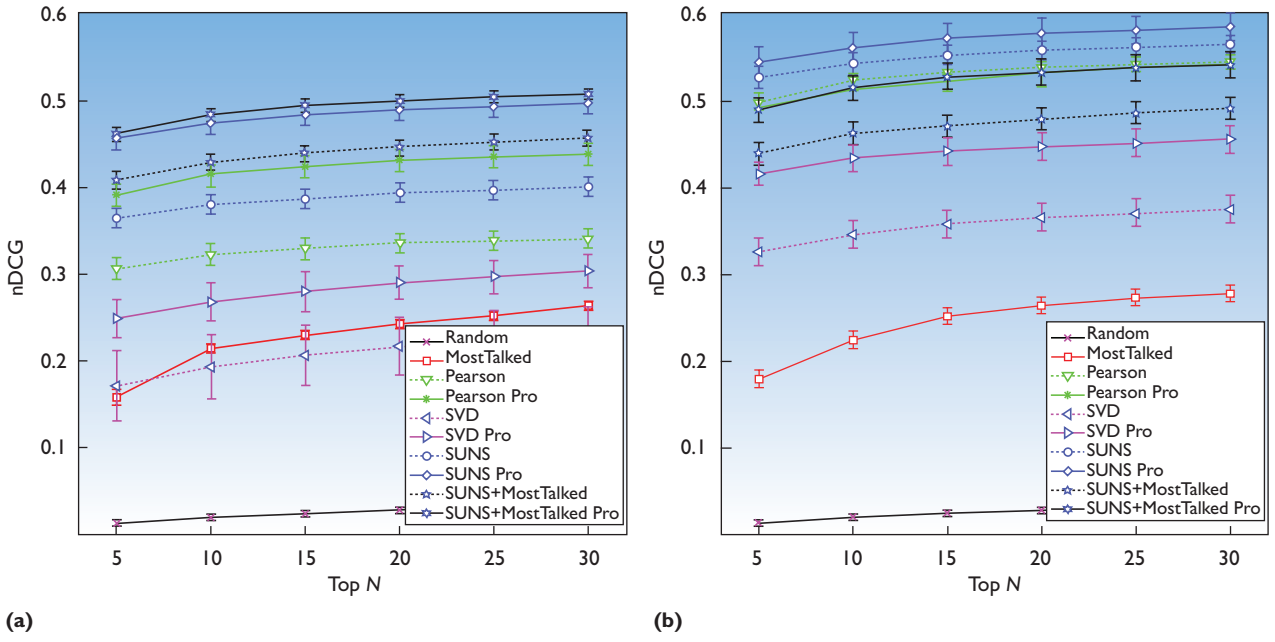


Figure 4. The nDCG values of top N *recommendations. For (a) Day3 and (b) Day6, we measured these values for Continuous Predictive Social Media Analytics methods* (SUNS, SUNS Pro, SUNS+MostTalked, and SUNS+MostTalked Pro)*, as well as several baselines.*

Another important performance metric for recommendation systems is the quality of the top *N* recommendations. We consider the number of latent variables yielding the best results (for example, *SUNS* with 200 latent variables). Figure 4 better highlights the advantage the *SUNS Pro* methods have, confirming the Figure 3 results: *SUNS+MostTalked Pro* was the best method for Day3, whereas *SUNS Pro* won by Day6.

Our evaluation uncovered several interesting findings. Regarding our research question, we can conclude that we can achieve accurate predictions even when little information is available in an event's earlier days. The predictive model *SUNS* achieves outstanding performance when using a reasonably large number of latent variables. It's more robust than *SVD* because it has a stronger generalization capability thanks to regularization. The most-talked-about venues that the SL computed as well as the visitor profiles compensate for the lack of information on Day3. On Day6, when visitor profiles and visitor-venue links were considerably enriched, combining *SUNS* with *MostTalked* is no longer the best option, whereas incorporating the visitor profiles remains important. In the future, this can guide us to choose an appropriate strategy for the VVR in real-world application scenarios.

These promising results provide new insights and pave the way to new and interesting research directions. In future work, we plan to investigate recommendation strategies that exploit event features — for example, visitors with similar profiles might have common interests in events of the same category. Finally, we aim to better understand the role visitor profiles play for recommendation purposes by exploring the use of weighted-matrix representations.

References

1. M. Balduini et al., "Social Listening of City Scale Events Using the Streaming Linked Data Framework," *Proc. Int'l Semantic Web Conf.*, LNCS 8219, Springer, 2013, pp. 1–16.
2. D. Zeng et al., "Social Media Analytics and Intelligence," *IEEE Intelligent Systems*, vol. 25, no. 6, 2010, pp. 13–16.
3. D.F. Barbieri et al., "Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics," *IEEE Intelligent Systems*, vol. 25, no. 6, 2010, pp. 32–41.

4. M. Balduini et al., "BOTTARI: An Augmented Reality Mobile Application to Deliver Personalized and Location-Based Recommendations by Continuous Analysis of Social Media Streams," *J. Web Semantics*, vol. 16, Nov. 2012, pp. 33–41.
5. P. De Meo et al., "Analyzing User Behavior across Social Sharing Environments," *ACM Trans. Intelligent Systems and Technology*, vol. 5, no. 1, 2013, pp. 14–45.
6. Y. Huang, V. Tresp, and H.-P. Kriegel, "Multivariate Prediction for Learning in Relational Graphs," *Proc. NIPS 2009 Workshop: Analyzing Networks and Learning with Graphs*, 2009, pp. 92–104.
7. E.J. Cands and B. Recht, "Exact Matrix Completion via Convex Optimization," Computing Research Repository, 2008.
8. R.M. Bell, Y. Koren, and C. Volinsky, *All Together Now: A Perspective on the Netflix Prize,* Chance, 2010.
9. K. Jarvelin and J. Kekalainen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," *Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2000, pp. 41–48.

Marco Balduini is a research assistant in the Department of Electronics, Information, and Bioengineering at Politecnico di Milano. His research focuses on data stream processing and the Semantic Web, in particular, the management of heterogeneous data streams generated by social and sensor networks. Balduini received a master's degree in computer science from Politecnico di Milano. He's a member of the W3C. Contact him at marco.balduini@polimi.it.

Alessandro Bozzon is an assistant professor with the Web Information Systems group at Delft University of Technology. His research interests are in Web data management; human computation, crowdsourcing, and games with a purpose; and information retrieval. Bozzon received a PhD in computer science from Politecnico di Milano, with a thesis focused on model-driven approaches for the design, development, and automatic code generation of search-based applications. Contact him at a.bozzon@tudelft.nl.

Emanuele Della Valle is an assistant professor of software project management in the Department of Electronics, Information, and Bioengineering at Politecnico di Milano. His research interests include the Semantic Web, service-oriented architectures, search engines, stream management systems, and rank-aware databases. Della Valle has a master's degree in computer science from Politecnico di Milano. He's a member of the W3C. Contact him at emanuele.dellavalle@polimi.it.

Yi Huang is a staff scientist at Siemens Corporate Research and Technology and is finishing his PhD at Ludwig Maximilian University of Munich, Germany. His research interests focus on statistical machine learning, text mining, information retrieval, and the Semantic Web. Huang received a Diploma in computer science from Ludwig Maximilian University of Munich. Contact him at yihuang@siemens.com.

Geert-Jan Houben is a full professor of Web information systems in the Software Technology Department at Delft University of Technology. His main research interests are in Web engineering, in particular the engineering of Web information systems that involve Web and Semantic Web technology, and user modeling, adaptation, and personalization. Contact him at g.j.p.m.houben@tudelft.nl.