Signal-Based User Recommendation on Twitter

Giuliano Arru, Davide Feltoni Gurini, Fabio Gasparetti,
Alessandro Micarelli and Giuseppe Sansonetti
Roma Tre University
Via della Vasca Navale 79
Rome, 00146 Italy
{arru,feltoni,gaspare,micarel,gsansone}@dia.uniroma3.it

ABSTRACT

In recent years, social networks have become one of the best ways to access information. The ease with which users connect to each other and the opportunity provided by *Twitter* and other social tools in order to follow person activities are increasing the use of such platforms for gathering information. The amount of available digital data is the core of the new challenges we now face. Social recommender systems can suggest both relevant content and users with common social interests. Our approach relies on a signal-based model, which explicitly includes a time dimension in the representation of the user interests. Specifically, this model takes advantage of a signal processing technique, namely, the wavelet transform, for defining an efficient pattern-based similarity function among users. Experimental comparisons with other approaches show the benefits of the proposed approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Information Filtering]

General Terms

Algorithms, Experimentation

Keywords

User recommendation, social network, twitter, signal processing, wavelet

1. INTRODUCTION

As most of the current social platforms, Twitter [1] allows users to build networks of relationships. One user has the chance to start following someone, obtaining the new tweets of the followed person that will appear in the personal homepage of the website (i.e., timeline). However, the diversity and time-dependent evolving nature of user interests are not represented by the traditional social graphs of friendships and subscribers. In this paper, we propose a new approach of user recommendation on Twitter. It relies on a novel user model, called *bag-of-signal*, that allows us to represent how user interests change over time, so adding the time dimension to user modeling in the Social Web. In order to fully exploit the potential of the new representation, we resort to mathematical tools used in the signal processing field, such as the discrete wavelet transform.

[1]https://twitter.com

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. RELATED WORK

Guy et al. [8, 3] propose a people recommendation engine within an enterprise social network site scenario. They aggregate several different sources to derive factors that might influence the similarity measure (e.g., co-authorships of publications, patents or project-wikis). An extended analysis [3] proves the effectiveness of content-based approaches as opposed to relationship-based algorithms, especially if histories of usage data in the social network are available. Further studies [9, 16, 15] show the benefits of tag-based profiles in the people recommendation task. Also Freyne et al. in [4] and Geyer et al. in [6] explore different recommendations strategies for improving the discovery of new users in social networks and social media. Twittomender [10, 12] lets users find pertinent profiles on Twitter exploiting different strategies, both content-based and collaborative ones, once the user submitted an initial query of interest. Hannon et al. in [11] advance a faceted profile structure that makes different types of interest more explicit. None of the above approaches takes explicitly into account the time as relevant factor to include during the recommendation process. A first preliminary attempt of using the wavelet theory for the recommendation task has been proposed in [2, 5]. The authors suggest a comparison among time habits in order to improve traditional collaborative approaches for music recommendation.

3. BAG-OF-SIGNAL MODEL

The idea behind the approach we propose is to bring the problem of the user representation to the problem of the document representation, so allowing us to take advantage of the Information Retrieval (IR) techniques. Particularly, we drew inspiration from the work presented in [14]. In the context of the user recommendation on Twitter, some definitions are needed. We define *pseudo-document* related to a user $u \in U$ the set of all the tweets $t \in T$ posted by u in a given observation period:

$$PD(u) = \{t \in T \,|\, user(t) = u\}$$

where U is the set of all the users and T the set of all the tweets. A natural extension of the bag-of-word representation is the *bag-of-concept* model, where *concepts* instead

of keywords are extracted from pseudo-documents. Concepts are entities more semantically significant than simple keywords. In this work, we consider the following types of concepts: (i) *hashtags*, namely, words or sentences prefixed with the symbol #, and (ii) *named entities*, namely, atomic elements in text classifiable into predefined categories (e.g., names of nations, people, etc.), which can be located by means of information extraction techniques. Specifically, we employed the *OpenCalais* [2] tool as named entity extractor. Therefore, we define *bag-of-concept* user model the set of weighted concepts:

$$P(u) = \{(c, w(u,c) \,|\, c \in C, u \in U\}$$

where $w(u,c)$ is the function that gives the weight of the concept $c \in C$ for the user $u \in U$, and C and U are the set of concepts and users, respectively. Now we have all the elements to define a new representation, which we called *bag-of-signal* to emphasize that the user model is made up of a set of signals, each of which is related to a different concept. The bag-of-signal representation is directly generated from the user activity. The context from which the profiling process starts is the user single post. In fact, the user activity consists of all the messages (tweets) posted during the observation period. In order to ensure the construction of the signals, it is necessary to extend the definition of pseudo-document to a more general one. We define *pseudo-fragment* related to a user $u \in U$ in a period $p \in PO$ the set of all the tweets $t \in T$ posted by u in the period p:

$$PF(u,p) = \{t \in T \,|\, user(t) = u,\, date(t) \in p\}$$

where U is the set of users, T the set of tweets, and PO the whole observation period. By analyzing a single pseudo-fragment related to a period p, it is possible to determine the *signal components* for the concepts in the text fragment. A signal component related to a user $u \in U$, a concept $c \in C$, and a period $p \in P$, is determined by the number of times the concept c occurs in the *pseudo-fragment* $PF(u,p)$, based on the weighting function $\omega(u,c,p)$

$$f_{u,c,p} = \omega(u,c,p)$$

where U is the set of users, C the set of concepts, and P the set of consecutive and same length periods. We define *signal* $S_{u,c}$ related to a user $u \in U$ and a concept c the ordered set of signal components $f_{u,c,p}$ with $p_i \in P$

$$S_{u,c} = [f_{u,c,p1}, f_{u,c,p2}, \cdots, f_{u,c,pn}]$$

where U is the set of users, C the set of concepts, and P a set of consecutive and same length periods. The value of the signal component $f_{u,c,p}$ is determined by a weighting function $\omega(u,c,p)$ where u is the user, c a concept, and p a period. This function is used to reduce the impact of typical problems of Information Retrieval, which may affect the proposed model too. The *concept-frequency* function takes into account all the pseudo-fragments related to the user u. This function is defined as a ratio whose numerator is the frequency whereby the concept c has been used inside the pseudo-fragment $PF_{u,p}$, and the denominator is the frequency of the most frequent concept within all the pseudo-fragments related to the user u:

$$CF_{u,p,c} = \frac{f_c}{max_{i \epsilon C}\{f_i\}}$$

[2]http://www.opencalais.com/

The *inverse-period-frequency* function gives the importance of a concept based on the number of pseudo-fragments in which it appears within the period. We define the inverse-period-frequency function for the concept c in the period p as follows:

$$IPF_{c,p} = log\left(\frac{|pf_{u,p}|}{|pf_{u,p} : c \epsilon pf_{u,p}|}\right)$$

where $|pf_{u,p} \in P|$ is the number of pseudo-fragments related to the period p, while $|pf_{u,p} : c \epsilon pf_{u,p}|$ is the number of pseudo-fragments where concept c appears. The weighting function assigning a value to the signal component $f_{u,c,p}$ (where u represents a user, c the concept which the signal component refers to, and p the period) is thus defined as follows:

$$\omega(u,c,p) = IPF_{c,p} * CF_{u,p,c}$$

As seen in the bag-of-concept model, a user is represented through a set of concepts weighted according to their occurrences within the pseudo-document. In the proposed model, a user is represented by a set of signals related to several concepts that appear in the pseudo-fragments concerning the user. Furthermore, each signal is made up of an ordered set of signal components weighted according to the weighting function. Below a definition of user model according to the proposed representation is given. We define the *bag-of-signal* model of user $u \in U$ as the set of the signals related to the user u, where the components $f_{u,c,p}$ are determined by a weighting function $\omega(u,p,c)$:

$$P_u = \{S_{u,c} = [f_{u,c,p0}, f_{u,c,p2}, \cdots, f_{u,c,pn}] \,|\, c \in C\}$$

where U and C are the set of users and concepts, respectively. Each signal contains two different information related to the concept: temporal and quantitative. Hence, the elementary units of bag-of-signal representation are signals and therefore they are the starting point for assessing the similarity between users. In order to analyze and represent such signals, the discrete wavelet transform (DWT) has been used. Wavelets are mathematical functions that decompose data into different frequency components and then analyze each component with a resolution depending on its scale [7]. Compared to the traditional Fourier analysis, wavelets are more suitable for analyzing not stationary signals containing discontinuities and sharp spikes. The wavelet analysis process makes use of a wavelet prototype function, named *mother wavelet*, for representing the signal in the wavelet domain with multiple levels of detail. This operation is called *multiresolution analysis*. The mother wavelet is compressed and expanded to analyze high and low frequencies, respectively. For signal processing purposes, the wavelet transform can be obtained through a bank of low-pass and high-pass filters [13]. The Mallat's intuition is fundamental for the practical use of wavelets, as it provides an efficient manner of implementing the discrete wavelet transform (with computational complexity $O(n)$, where n is the signal length). The low-pass filter has the effect of approximating the signal, while the high-pass filter has the effect of filtering the signal details. There are several wavelets; in this work we chose the Haar wavelet for its ease of implementation and compact support, which means that it vanishes outside a finite interval. In this context, we define two similarity functions $f1$ and $f2$. Given two users u_1, u_2 and their profiles P_{u_1}, P_{u_2} based on the bag-of-signal representation, the *sim-*

ilarity function f1 between those users is defined as follows:

$$f1(u_1,u_2)=\frac{\sum_{c\in C_1\cup C_2}\xi(s_{u_1,c})\cdot\xi(s_{u_2,c})\cdot temp_{level}(s_{u_1,c},s_{u_2,c})}{\sqrt{\sum_{c\in C_1}\xi^2(s_{u_1,c})}\cdot\sqrt{\sum_{c\in C_2}\xi^2(s_{u_2,c})}}$$

where $s_{u_1,c}\in P_{u_1}$ and $s_{u_2,c}\in P_{u_2}$, C_1 and C_2 are the concepts related to the signals belonging to P_{u1} and P_{u2}, the function ξ determines the energy of the signal and $temp_{level}$ is a function that analyzes whether the signals have similar time use patterns. The proposed similarity function is therefore very similar to cosine similarity. The importance of a signal within the profile is given by its energy. Given a discrete-time signal s, limited and with real components, the *energy* $xi(s)$ of the signal s is defined as follows:

$$\xi(s)=\sum_{i=0}^{|s|}s[i]^2$$

The function $temp_{level}$ returns a value between 0 and 1, providing a measure of how much the concepts belonging to the two profiles have been used with similar time patterns. In this way, the contribution of two concepts used in the same periods will be greater than the contribution of the concepts used in different periods. The approximation $A_l(s)$ of the signal s at level *l-th* is defined by the set of approximation coefficients of the DWT limited to the level *l-th*:

$$A_l(s)=\{a_{l,j}\ \ j=1,...,2^l\}$$

Given two signals s_1 and s_2 and their respective approximations at level $A_{level}(s_1)=[a_{s_1},...,a_{s_1}]$ and $A_{level}(s_2)=[a_{s_2},...,a_{s_2}]$, we have:

$$C(s_1,s_2)=\sum_{i=0}^{|2^l|}A_{level}(s_1)[i]A_{level}(s_2)[i]$$

A normalized value between 0 and 1 can be obtained as follows:

$$C_{normalized}(s_1,s_2)=\frac{C(s_1,s_2)}{\sqrt{C(s_1,s_1)C(s_2,s_2)}}$$

As $temp_{level}$ function we take $C_{normalized}$, which provides a similarity index between the two different signals. This function used in the similarity measure allows us to differently "weigh" the concepts occurring in the same period from the concepts occurring in different periods. Given two users u_1, u_2 and their respective profiles P_{u_1}, P_{u_2} based on the bag-of-signal representation, the *similarity function f2* between those users is defined as follows:

$$f2(u_1,u_2)=\frac{\sum\limits_{c\in C_1\cup C_2}\sum s_{u_1,c}[i]\cdot s_{u_2,c}[i]}{\sqrt{\sum\limits_{c\in C_1}\sum s_{u_1,c}[i]^2}\cdot\sqrt{\sum\limits_{c\in C_2}\sum s_{u_2,c}[i]^2}}$$

where $s_{u_1,c}\epsilon P_{u_1}$ and $s_{u_2,c}\in P_{u_2}$, C_1 and C_2 are the concepts related to the signals belonging to P_{u1} and P_{u2}.

4. EXPERIMENTAL EVALUATION

Testing a user recommender on a social network like Twitter raises many challenges that it is not possible to discuss here for reasons of space. In short, the basic idea we followed has been to exploit social relationships between users, in particular the *following relationship*, in order to evaluate the system performance. In the experimental tests we have used the *Success at Rank K (S@K)* and *Mean Reciprocal Rank (MRR)* as evaluation measures. The first one estimates the mean probability of finding a relevant item among the top K recommended items. In user recommendation the items are users, therefore *S@K* provides the mean probability of finding a relevant user in the top K positions of ranking. The ranking is defined according to the similarity function to calculate. The evaluation function exploits the social relationships between users in order to establish if a user is really relevant for another one. Therefore, it is needed to understand when a user u_1 is relevant for another user u_2. For this purpose, it is necessary to make a strong hypothesis: a user u_1 is relevant for a user u_2 if exists a *following relationship* between them. In other terms, if u_1, u_2, or both of them, have added the other one among his *followings*. This hypothesis is supported by the phenomenon of *homophily* according to which two *similar* users have more probability to follow each other than two *not similar* users. *S@K* provides the mean probability that a relevant user is located in the top K positions of suggested users, while *Mean Reciprocal Rank (MRR)* indicates the average position of a user in the recommended list. The dataset used for the experimental efforts has been obtained starting from the one proposed and employed in [1]. This corpus was built by following 20.000 English users from October 2010 to January 2011. Starting from these 20.000 users, we selected only the 1619 users that posted at least one tweet at month and at least 20 tweets in the whole observation period. We performed several tests in order to evaluate the effectiveness of the proposed approach. We also analyzed the performance of the user recommender while changing the parameters of the bag-of-signal model and the size of the recommendation list. Particularly, an important parameter of the model is the length of the observation period, namely, the number of days for each sample whereby the signals have been generated. In the following, we denote the signal representation used in a given similarity function by adding the number of days per period of the sample; for example, the notation $f1\_8$ indicates the similarity function $f1$ with a signal representation having a sampling period of 8 days. In this paper, for reasons of space we report only the results for the comparative analysis between the approach based on bag-of-signal model and two traditional approaches that do not consider the time dimension: (i) cosine similarity in a *Vector Space Model (VSM)* where vectors are weighted concepts, and (ii) the *function S1* proposed in [10], which is based on a vector user representation. Figure 1 shows the obtained results. Firstly, it can be noted that the functions $f1$ and $f2$ based on hashtags performed significantly better than the same functions based on named entities. This might seem an unexpected result, because in principle named entities should appear in users activities more often than hashtags. Indeed, while some users do not use hashtags, most of them report named entities in their posts, referring to names of celebrities, places, companies, etc. By analyzing the models constructed from the two different types of concepts, we found out that named entity based profiles were made up of 63 signals on average, while hashtag based profiles were composed of 223 signals on average. Hence, our theory is that the smaller amount of information in case of named entities resulted in similarities functions producing worse results than those obtained by extracting hashtags. This result also shows how the Twitter hashtag mechanism

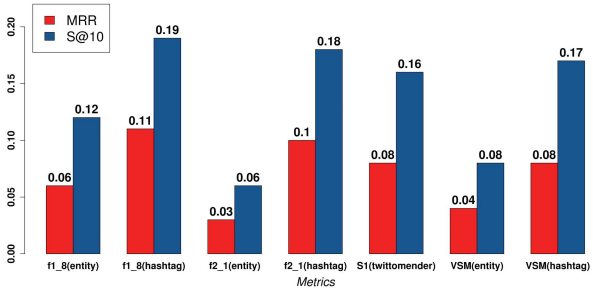

Figure 1: Comparative analysis among the two proposed similarity functions and two classical approaches advanced in literature, that is, *Vector Space Model (VSM)* and *function S1* (see [10]).

is well-established and widespread, and then it can be usefully exploited for user profiling purposes. Furthermore, it can be observed that the functions relying on signals built using hashtags (i.e., *f1(hashtag)* and *f2(hashtag)*) perform definitely better than the functions *VSM(hashtag)* and *S1*. These findings confirm that harnessing the time dimension guarantees better results in user profiling.

5. CONCLUSIONS

In this paper we have described a user recommender system on Twitter. Such a system is based on a novel user model, termed *bag-of-signal*, which makes use of signal processing techniques to represent not only the number of occurrences of the informative entities (concepts), but also the related time use patterns. The bag-of-signal user model involves modeling the user interests through a set of signals and the adoption of similarity functions suitably defined.

6. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web. In *Proceedings of ACM WebSci '11, 3rd International Conference on Web Science, Koblenz, Germany*. ACM, June 2011.

[2] C. Biancalana, F. Gasparetti, A. Micarelli, A. Miola, and G. Sansonetti. Context-aware movie recommendation based on signal processing and machine learning. In *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, CAMRa '11, pages 5–10, New York, NY, USA, 2011. ACM.

[3] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 201–210, New York, NY, USA, 2009. ACM.

[4] J. Freyne, M. Jacovi, I. Guy, and W. Geyer. Increasing engagement through early recommender intervention. In *Proceedings of the third ACM Conference on Recommender Systems*, RecSys '09, pages 85–92, New York, NY, USA, 2009. ACM.

[5] F. Gasparetti, C. Biancalana, A. Micarelli, A. Miola, and G. Sansonetti. Wavelet-based music recommendation. In K.-H. Krempels and J. Cordeiro, editors, *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, 18 - 21 April, 2012*, pages 399–402. SciTePress, 2012.

[6] W. Geyer, C. Dugan, D. R. Millen, M. Muller, and J. Freyne. Recommending topics for self-descriptions in online user profiles. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 59–66, New York, NY, USA, 2008. ACM.

[7] A. Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2), 1995.

[8] I. Guy, I. Ronen, and E. Wilcox. Do you know?: recommending people to invite into your social network. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 77–86, New York, NY, USA, 2009. ACM.

[9] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 194–201, New York, NY, USA, 2010. ACM.

[10] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM Conference on Recommender Systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM.

[11] J. Hannon, K. McCarthy, M. P. O'Mahony, and B. Smyth. A multi-faceted user model for twitter. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, UMAP'12, pages 303–309, Berlin, Heidelberg, 2012. Springer-Verlag.

[12] J. Hannon, K. McCarthy, and B. Smyth. Finding useful users on twitter: twittomender the followee recommender. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 784–787, Berlin, Heidelberg, 2011. Springer-Verlag.

[13] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans, on Pattern Analysis and Machine Intelligence*, PAMI-11(7):674–693, July 1989.

[14] L. A. F. Park, K. Ramamohanarao, and M. Palaniswami. A novel document retrieval method using the discrete wavelet transform. *ACM Trans. Inf. Syst.*, 23(3):267–298, July 2005.

[15] P. Symeonidis. User recommendations based on tensor dimensionality reduction. In Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, editors, *Artificial Intelligence Applications and Innovations III*, volume 296 of *IFIP International Federation for Information Processing*, pages 331–340. Springer US, 2009.

[16] Z. Yan and J. Zhou. User recommendation with tensor factorization in social networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3853 –3856, March 2012.