A Signal-Based Approach to News Recommendation

Sirian Caldarelli, Davide Feltoni Gurini, Alessandro Micarelli and Giuseppe Sansonetti
Department of Engineering
Roma Tre University
Via della Vasca Navale 79
Rome, 00146 Italy
sirian.caldarelli@gmail.com,{feltoni,micarel,gsansone}@dia.uniroma3.it

ABSTRACT

In this paper, we describe our research activity on an approach to personalized news recommendation, which captures the temporal dynamics of the active user's interests. In such recommender, the user profile explicitly involves the time dimension in representing her interests and preferences. Each user's interest is represented as a signal, thus characterizing its evolution over time. To this aim, a signal processing technique (i.e., the discrete wavelet transform) is adopted to represent and analyze such signals. Furthermore, we report the experimental results of a very preliminary comparative evaluation on an online available dataset. Such results seem encouraging, thus spurring us to continue developing our approach.

Keywords

News recommendation; user profiling; temporal dynamics

1. INTRODUCTION

With the development in electronics and Internet technologies, online information available has been constantly increasing. In such scenario, users are confused and more and more feel the need to be guided in the selection of the information to pay attention to. News recommenders are a possible solution, since help users find the information of possible interest to them. In order to provide personalized suggestions, such systems rely on a representation of the target user's interests and preferences. A vast amount of user profiling techniques have been proposed and deeply evaluated [7]. However, representing how users' interests evolve over time remains a difficult challenge. In this paper, we apply an approach to user profiling, called *bag-of-signals* [2], whose aim is to represent the diversity and time-dependent evolving nature of users' interests. Based on such approach, we realized a recommender system of news articles. In order to assess its performance, we performed a very preliminary off-line evaluation as follows. Starting from a public database, we built users profiles extracting their interests from news articles linked to contents generated by them on social media. More specifically, we examined users' timelines on Twitter [1] considering all the tweets and the related news articles in the entire observation period. Then, we extracted users' interests as concepts (e.g., topics) from those news and represented their evolution over time as signals. For analyzing and comparing such signals, we made use of a signal processing tool that characterizes the frequency content of any signal, along with its accurate location in the time domain. A comparative evaluation with a classic approach that completely ignores the time-dependence of users' interests revealed the benefits of the proposed news recommender.

2. BAG-OF-SIGNALS MODEL

The representation of users' interests as signals requires some definitions. We define *pseudo-document* related to a user $u \in U$ (with U set of all the users) and an observation period ΔT, the set of all the news articles mentioned by u in the period ΔT:

$$PseudoDoc(u, \Delta T) = \{news | \, user(news) = u, \, date(news) \in \Delta T\}$$

The notation $user(news) = u$ means that the user u has mentioned that particular $news$, while $date(news) \in \Delta T$ means that u has mentioned that $news$ in the period ΔT. An extension of the *bag-of-words* representation, well-known in Information Retrieval, is the *bag-of-concepts* model, where *concepts* instead of keywords are extracted from pseudo-documents. Concepts are entities more semantically significant than simple keywords. We define *bag-of-concepts* user model the following set of weighted concepts:

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

where the function $w(u, c)$ gives the weight of the concept $c \in C$ for the user $u \in U$ (with C and U set of concepts and users, respectively). Then, we define *pseudo-fragment* related to a user $u \in U$ in an interval $\Delta t \in \Delta T$, the set of all the news mentioned by u in the interval Δt:

$$PseudoFrag(u, \Delta t) = \{news | \, user(news) = u, \, date(news) \in \Delta t\}$$

By analyzing a single pseudo-fragment related to an interval Δt, it is possible to determine the *signal components* for the concepts in the text fragment. A signal component $f_{u,c,\Delta t}$ related to a user $u \in U$, a concept $c \in C$, and an interval $\Delta t \in \Delta T$, is determined by the number of times the concept

[1] https://twitter.com

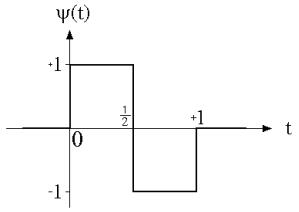

Figure 1: Haar wavelet.

c occurs in the pseudo-fragment $PseudoFrag(u, \Delta t)$, based on the weighting function $\omega(u, c, \Delta t)$

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

This function is used to reduce the impact of typical problems of Information Retrieval, which may affect the proposed model too. More specifically, $\omega(u, c, \Delta t)$ takes into account (i) the discriminating power of the concept c within the time interval Δt, and (ii) the relevance of the same concept within the user u's profile. We define *signal* $S_{u,c}$ related to a user u and a concept c the ordered set of signal components $f_{u,c,\Delta t_i}$ with $\Delta t_i \in \Delta T$

$$S_{u,c} = [f_{u,c,\Delta t_1}, f_{u,c,\Delta t_2}, \cdots, f_{u,c,\Delta t_n}]$$

where ΔT consists of n consecutive and same length intervals Δt_i (with $i = 1, 2, ...n$). As seen in the bag-of-concepts 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. Now, we define the *bag-of-signals* model of user $u \in U$ as the set of the signals related to the user u, where the components $f_{u,c,\Delta t}$ are determined by the weighting function $\omega(u, c, \Delta t)$:

$$P_{BoS}(u) = \{S_{u,c} = [f_{u,c,\Delta t_1}, f_{u,c,\Delta t_2}, \cdots, f_{u,c,\Delta t_n}] \mid c \in C\}$$

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. These signals show strong discontinuities and sharp spikes. Signal processing provides an ideal tool for representing and analyzing such kind of signals: the wavelet transform [5]. Wavelets are mathematical functions that may be located both in time (space), as well as in scale (frequency), thus providing an accurate *time-scale map* of the signal. The wavelet-based analysis relies on the use of a prototype function, so-called *mother wavelet*, whose translated and scaled versions constitute the basis functions for the series expansion that ensures the representation of the original signal through coefficients. Operations involving signals can, therefore, be developed - in a more streamlined and efficient way - directly on corresponding wavelet coefficients. If the mother wavelet is properly selected (in our approach we choose the Haar wavelet for its compact support, as can be seen from Figure 1), the wavelet transform allows for best capturing signal dynamics. Computation of the wavelet transform can be performed in a fast way (with computational cost $O(n)$, if n is the number of signal samples) by means of the *fast discrete wavelet transform (DWT)* [10]. Preliminary attempts of leveraging the wavelet theory for music and movies recommendation tasks have been proposed [4, 3]. Once defined the bag-of-signals model for representing user profiles, we also need to define a method for evaluating the similarity between users. Concretely, we considered two different similarity functions $f1$ and $f2$.

Given two users u_1, u_2 and their corresponding profiles $P_{BoS}(u_1)$, $P_{BoS}(u_2)$ based on the bag-of-signals representation, the *similarity function f1* between those users is defined as follows:

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

where $s_{u_1,c} \in P_{BoS}(u_1)$ and $s_{u_2,c} \in P_{BoS}(u_2)$, C_1 and C_2 are the sets of the concepts related to the signals belonging to $P_{BoS}(u_1)$ and $P_{BoS}(u_2)$, the function $\xi(s)$ expresses the energy of the signal s and $temp_{level}(s_1, s_2)$ is a function that analyzes whether the signals s_1 and s_2 show similar time use patterns. The importance of a signal within the profile is given by its energy. Given a discrete-time signal s, limited and with real components, its *energy* $\xi(s)$ 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 intervals will be greater than the contribution of the concepts used in different intervals. 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 $A_{level}(s_2) = [a_{s_2}, ..., a_{s_2}]$, the function $temp_{level}(s_1, s_2)$ is defined as follows:

$$temp_{level}(s_1, s_2) = \frac{C(s_1, s_2)}{\sqrt{C(s_1, s_1)C(s_2, s_2)}}$$

where

$$C(s1, s2) = \sum_{i=0}^{|2^l|} A_{level}(s_1)[i] A_{level}(s_2)[i]$$

Given two users u_1, u_2 and their respective user profiles $P_{BoS}(u_1)$ and $P_{BoS}(u_2)$ based on the bag-of-signals representation, the *similarity function f2* between those users is defined as follows:

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

where $s_{u_1,c} \in P_{BoS}(u_1)$ and $s_{u_2,c} \in P_{BoS}(u_2)$, C_1 and C_2 are the sets of the concepts related to the signals belonging to $P_{BoS}(u_1)$ and $P_{BoS}(u_2)$.

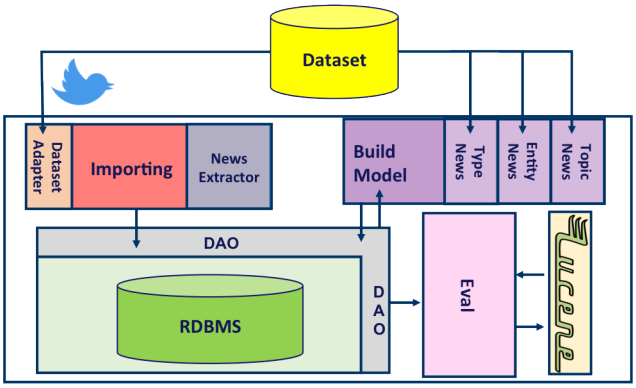

Figure 2: Schema of the experimental evaluation.

3. EXPERIMENTAL EVALUATION

In order to perform our experimental tests, we resorted to the dataset presented and employed in [1]. Such dataset was obtained by monitoring a sample of 20,000 English speaking users' timelines on Twitter for a given time period ΔT. From the original sample, the authors selected only those 1619 users that posted at least ten tweets at month and at least 20 tweets in the whole observation period, thus gathering more than two million tweets. From the news articles mentioned in such tweets, concepts (i.e., entities, types, and topics) were extracted through the web service *OpenCalais* [2]. We associated such concepts to the creation time of the corresponding tweet, in order to temporally localize them. The whole observation period ΔT was about three months, so we considered the tweets (and the linked news) of the first two months as training dataset, the remaining tweets as testing dataset. After that, the evaluation procedure was as follows (see Figure 2).

Training phase

- the news linked to the tweets belonging to the training dataset were retrieved;
- the concepts extracted from such news were considered;
- a bag-of-signals profile was built for each user, using the concepts obtained in the previous step;
- for each user a list of users more similar to her was returned.

Testing phase

- the news linked to the tweets belonging to the testing dataset were retrieved;
- a pseudo-document for each user was generated from those news;
- all the pseudo-documents were indexed using the open source Lucene platform [3], as proposed in [6];
- for each pseudo-document a list of pseudo-documents more similar to it was returned.

[2]http://www.opencalais.com/
[3]https://lucene.apache.org/

The performance of the recommender system was assessed in terms of the normalized version of *Discounted Cumulative Gain (nDCG)* [8, 9]. $nDCG$ is usually truncated at a particular rank level to emphasize the importance of the first retrieved documents. The measure is defined as follows:

$$nDCG@n = \frac{DCG@n}{IDCG@n} \quad (1)$$

and the Discounted Cumulative Gain (DCG) is defined as follows:

$$DCG@n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i} \quad (2)$$

where rel_i is the graded relevance of the $i-$th result (i.e., 0 = *non-significant*, 1 = *significant*, and 2 = *very significant*), and the Ideal DCG ($IDCG$) for a query corresponds to the DCG measure where scores are resorted monotonically decreasing, that is, the maximum possible DCG value over that query. $nDCG$ is often used to evaluate search engine algorithms and other techniques whose goal is to order a subset of items in such a way that highly relevant documents are placed on the top of the list, while less important ones are moved lower. Basically, higher values of $nDCG$ mean that the system output gets closer to the ideal ranked output. Figure 3 shows the experimental results obtained considering the two similarity functions $f1$ and $f2$ introduced above, and the *function S1* proposed in [6], which was obtained by indexing the contents of all the news articles using Lucene. It is possible to notice that the first two approaches, which consider the evolution of interests over time, outperform the last one that, instead, ignores the temporal dimension.

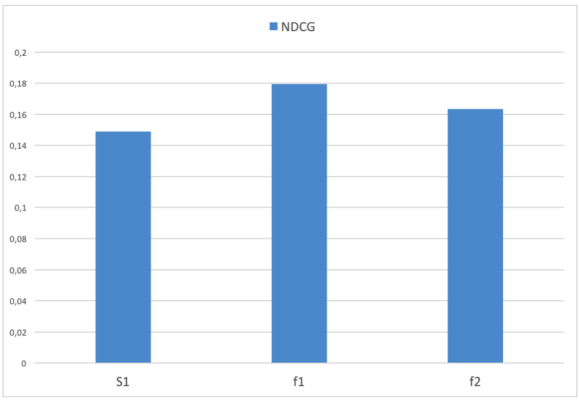

Figure 3: Comparative analysis between the two proposed similarity functions $f1$ and $f2$ and the function $S1$ proposed in [6], in terms of $nDCG$ values.

Figure 4 reports the best results (i.e., those obtained through the $f1$ similarity function) when varying the nature of the concepts represented as signals in the user profile. As we could expect, bag-of-signals user profiles representing entities as signals allow the news recommender to obtain the best performance. In fact, the maximum number for topics and types extracted by OpenCalais is 18 and 39, respectively. On the contrary, there is no limit for the number of entities extracted from news articles. In the used dataset, a bag-of-signals user profile with entities as signals can have more than 3500 represented concepts. Hence, the smaller amount of information in case of topics and types brought about worse results than those obtaining using entities.

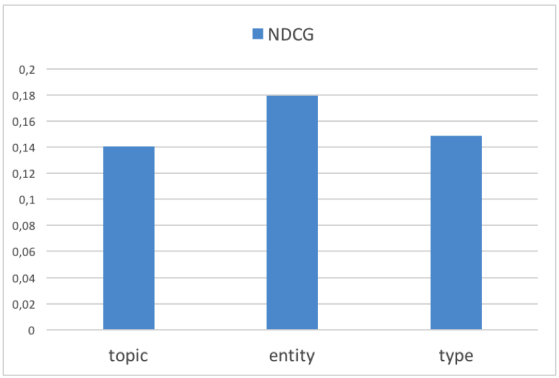

Figure 4: Best results when varying the nature of the concepts extracted from the news articles and represented as signals in the proposed user profiling.

4. CONCLUSIONS

In this paper, we have presented a news recommender system based on the *bag-of-signals* user model, which leverages 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-signals user model involves modeling the user interests through a set of signals and the adoption of similarity functions suitably defined. More specifically, for the signal analysis and representation we employ the wavelet mathematical tool for its main characteristic of time-frequency localization. Practically, the discrete wavelet transform allows us to effectively analyze the sampled signals with a different time window.

Although the experimental results on an online available dataset are positive, this work is still in a preliminary stage and leaves much space for future developments. For instance, the similarity function is an open issue that should be further investigated. Starting from the bag-of-signals model, we could explore new functions considering the same data but in a different way, developing new aspects, and using other tools from the signal processing domain. Moreover, we intend to test our news recommender on real news datasets. Finally, another interesting development could involve sentiment analysis. Concretely, we propose to add a further module to the described news recommender, whereby extract the positive, negative, or neutral opinion expressed by the user about a given concept. In this way, the profile may take into account not only the level and the temporal localization of users' interests, but also their nature.

5. 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 the 3rd International Web Science Conference*, WebSci '11. ACM, 2011.

[2] G. Arru, D. Feltoni Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. Signal-based user recommendation on twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 941–944, New York, NY, USA, 2013. ACM.

[3] 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.

[4] 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.

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

[6] 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. ACM, 2010.

[7] M. Harandi and J. A. Gulla. Survey of user profiling in news recommender systems. In J. A. Gulla, B. Yu, Ö. Özgöbek, and N. Shabib, editors, *Proceedings of the 3rd International Workshop on News Recommendation and Analytics (INRA 2015) co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 20, 2015.*, volume 1542 of *CEUR Workshop Proceedings*, pages 20–26. CEUR-WS.org, 2015.

[8] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48, New York, NY, USA, 2000. ACM.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[10] 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.