Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles

Fedelucio Narducci[1], Cataldo Musto[2], Giovanni Semeraro[2], Pasquale Lops[2], and Marco de Gemmis[2]

[1] University of Milano-Bicocca, Italy
narducci@disco.unimib.it
[2] University of Bari Aldo Moro, Italy
name.surname@uniba.it

Abstract. The main contribution of this work[1] is the comparison of different techniques for representing user preferences extracted by analyzing data gathered from social networks, with the aim of constructing more transparent (human-readable) and serendipitous user profiles. We compared two different user models representations: one based on keywords and one exploiting encyclopedic knowledge extracted from Wikipedia. A preliminary evaluation involving 51 Facebook and Twitter users has shown that the use of an encyclopedic-based representation better reflects user preferences, and helps to introduce new interesting topics.

1 Motivation and Research Problem

In this work we investigate whether a Wikipedia-based representation of user interests allows to produce more transparent and serendipitous user profiles. Transparency is related to the readability of profiles, while serendipity to the ability to suggest surprisingly interesting items that users might not have otherwise discovered. The main motivation for this reseach is that a system that adopts a more understandable representation can lead towards a more transparent personalization process. For example, a recommender system that uses a human-understandable profile could easily explain the reason for a suggestion. Furthermore, serendipitous profiles help to overcome the overspecialization problem, that leads to accurate but obvious (and so unuseful) suggestions. We address the following two research questions:

- Are Wikipedia concepts (i.e., titles of Wikipedia articles) more representative than keywords for modeling user interests?
- Is it possible to leverage encyclopedic knowledge to enrich user profiles with novel topics of interest?

[1] This work fullfils the research objectives of the projects PON 02_00563_3470993 VINCENTE (A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems) and PON 01_00850 ASK-Health (Advanced system for the interpretations and sharing of knowledge in health care) funded by the Italian Ministry of University and Research (MIUR).

S. Carberry et al. (Eds.): UMAP 2013, LNCS 7899, pp. 350–352, 2013.

We indexed textual information extracted from Facebook and Twitter using different strategies, in order to obtain different user profiles whose effectiveness in terms of transparency and serendipity is evaluated in a user study.

2 User Profiling Strategies

For each Facebook user, we processed the title and description of liked groups and pages, attended events, personal status, title and summary of shared links, while for Twitter we used tweets of the user and her followings. We call *social items* the aforementioned pieces of information, that were indexed using three strategies and lead to the following representations of user profiles:

Social Profile: This is the simplest profile representation, based on keywords occurring in the social items collected for a user. The text of social items is tokenized and stopwords removed. Keywords are weighed using the TF-IDF score. This represents our baseline.

Tag.me Profile: This strategy exploits the *anchor disambiguation* algorithm implemented in TAG.ME [1] to identify Wikipedia concepts occurring in the text of social items. The titles of those concepts are included in the TAG.ME profile of the user, and weighed using TF-IDF.

Explicit Semantic Analysis (ESA) Profile: ESA represents a term or a text as a vector whose dimensions are Wikipedia pages. For example, the meaning of the term *Christmas* is described by a list of concepts it refers to (e.g. *Santa Claus, December 25*). Formally, given the space of Wikipedia concepts (articles) $C = \{c_1, c_2, ..., c_n\}$, a term t_i is represented by its *semantic interpretation vector* $v_i =< w_{i1}, w_{i2}, ..., w_{in} >$, where a weight w_{ij} represents the strength of the association between t_i and c_j, i.e. the TF-IDF of t_i in c_j [2]. The semantics of a text fragment f (i.e. a sentence, a tweet, a Facebook post) is the centroid of vectors associated with terms occurring in f. The ESA Profile of a user is built by including the 10 most relevant Wikipedia concepts in the semantic interpretation vector associated with her social items.

The difference between the last two approaches is that ESA implements a *feature generation* process, since it could generates *new* features related to the text to be indexed, while TAG.ME simply performs *feature selection*. As an example, let's consider the Facebook posts: *I'm in trepidation for my first riding lesson!*, *I'm anxious for the soccer match :(* , *I will flight by Ryanair to London!*, *Ryanair really cheapest company!*, *Ryanair lost my luggage :(*, *This summer holidays are amazing!*, *Relax during these holidays!*. Figure 1 depicts a sketch of the profiles produced by the different processing of social items. *Social* profile contains many non-relevant keywords, such as those referring to user moods (anxious, trepidation, etc.). The TAG.ME profile contains terms occurring in the *social* profile (soccer, London, etc.), whose weights are higher than those in social profile since the noise coming from non-relevant keyword has been filtered out. Finally, the ESA profile contains *new* topics somehow related to the other profiles (Vienna - which hosts a famous riding school, UK, low-cost), but not *explicitly* mentioned in the social profile.

Fig. 1. a) Social, b) Tag.me, c) ESA representations of user profiles shown as tag clouds

Table 1. Results of Transparency and Serendipity

	Transparency				Serendipity			
Profile	Avg. Rating	Min Rating	Max Rating	Std. dev.	Avg. Rating	Min Rating	Max Rating	Std. dev.
Social	1.33	0	3	0.65	0.42	0	2	0.57
Tag.me	**3.88**	2	5	0.82	0.54	0	2	0.61
ESA	1.16	0	4	1.00	**3.24**	0	5	1.24

3 Experimental Evaluation

The goal of the experiment was to identify which kind of user profile best represents user interests. For each kind of profile, we evaluated *transparency* as the overlap between actual user interests and keywords shown in the profile, and *serendipity* as the presence of *unexpected* and *interesting* features in the profile. 51 users were involved in the study, 36 of them gave us the consent to extract social items only from Facebook, 4 only from Twitter, 11 from both social networks. The experiment has been carried out for two weeks. For each user, the SOCIAL, ESA, and TAG.ME profiles were built and shown to her as tag clouds. The user provided feedback, on a 6-points discrete rating scale, on the extent to which 1) keywords in the profile reflect personal interests, and 2) the profile contains unexpected interesting topics. Results are reported in Table 1. Two main outcomes are observed: the best description of user interests is provided by the TAG.ME profile, while ESA shows the highest heterogeneity of results (highest standard deviation); highest serendipity is shown by ESA profiles ($p < 0.05$), since ESA is the only technique which enriches profiles with *new* features. In conclusion, we observe that a representation of user interests based on encyclopedic concepts might lead to more transparent and serendipitous profiles. As future work we will investigate possible ways to merge TAG.ME and ESA representations.

References

1. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management, pp. 1625–1628. ACM, New York (2010)
2. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal Artif. Intell. Res (JAIR) 34, 443–498 (2009)