On the Semantics of User Interaction in Social Media

(Extended Abstract*)

Folke Mitzlaff¹, Martin Atzmueller¹, Gerd Stumme¹ and Andreas Hotho²

¹Knowledge and Data Engineering Group, University of Kassel
²Data Mining and Information Retrieval Group, University of Würzburg

Abstract

In ubiquitous and social web applications, there are different user traces, for example, produced explicitly by “tweeting” via twitter or implicitly, when the corresponding activities are logged within the application’s internal databases and log files. Each set of user interactions can then be mapped to a network, with links between users according to their observed interactions. In this paper, we analyze correlations between different interaction networks. We collect for every user certain external properties which are independent of the given network structure. Based on these properties, we then calculate semantically grounded reference relations among users and present a framework for capturing semantics of user relations. The experiments are performed using different interaction networks from the twitter, flickr and Bibsonomy systems.

1 Introduction

By interacting with social and ubiquitous systems, the user is leaving traces within the different databases and log files, e.g., by updating the current status via twitter or chatting with social acquaintances via facebook. Ultimately, each type of such traces gives rise to a corresponding network of user relatedness, where users are connected if they interacted either explicitly (e.g., by establishing a “friendship” link within in an online social network) or implicitly (e.g., by visiting a user’s profile page). We consider a link within such a network as evidence for user relatedness and call it accordingly evidence network or interaction network. These interaction networks are of large interest for many applications, such as recommending contacts in online social networks or for identifying groups of related users [8]. Nevertheless, it is not clear, whether every such interaction network captures meaningful notions of relatedness and what the semantics of different aggregation levels really are. As multifaceted as humans are, as many reasons for individuals being related exists. Ultimately, it is therefore not possible to judge whether an interaction network is “meaningful” or not. Nevertheless, certain networks are more probable than others and give rise to more traceable notions of relatedness.

This extended abstract summarizes the paper [9]: Folke Mitzlaff, Martin Atzmueller, Gerd Stumme, and Andreas Hotho. Semantics of User Interaction in Social Media. In Gourab Ghosal, Julia Poncela-Casasnovas, and Robert Tolksdorf (Eds.), Complex Networks IV, Springer Verlag, Heidelberg, Germany, 2013.

2 Experiments and Results

This paper summarizes work presented in [9], focusing on an experimental methodology for assessing the semantics of evidence networks and similarity metrics therein. The methodology is applied to a broad range of evidence networks. The obtained results thus yield a semantic grounding of evidence networks and similarity metrics, which are merely based on structural properties of the networks. Furthermore, we consider both established reference sources such as tagging data, as well as geographical locational data as a proxy for semantic relatedness.

Evidence Networks in Bibsonomy Beside explicit relations among users, i.e., the “friends” in Bibsonomy, different relations are established implicitly by user interactions, e.g., when user $u$ looks at user $v$’s resources. In particular, we considered the directed Friend-Graph, containing an edge $(u, v)$ iff user $u$ has added user $v$ as a friend, the directed Copy-Graph which contains an edge $(u, v)$ with weight $w \in \mathbb{N}$, iff user $u$ has copied $v$’s resources, i.e., a publication reference from user $u$ and the directed Visit-Graph, containing an edge $(u, v)$ with label $c \in \mathbb{N}$, iff user $u$ has navigated $c$ times to the user page of user $v$.

Evidence Networks in twitter Each user publishes short text messages (“tweets”) which may contain freely chosen hashtags, i.e., distinguished words being used for marking keywords or topics. Furthermore, users may “cite” each other by “retweeting”. A user $u$ retweets user $v$’s content, if $u$ publishes a text message containing “RT @v:” followed by (an excerpt of) $v$’s corresponding tweet. Users may also explicitly follow other user’s tweets by establishing a corresponding friendship-like link. For analysis, we considered the directed Follow-Graph, containing an edge $(u, v)$ iff user $u$ follows the tweets of user $v$, and the ReTweet-Graph, containing an edge $(u, v)$ with label $c \in \mathbb{N}$ if user $u$ cited (or “retweeted”) exactly $c$ of user $v$’s tweets.

Evidence Networks in flickr In flickr, users mainly upload images and assign arbitrary tags but also interact, e.g., by establishing contacts or commenting on other users images. For our analysis we extracted the directed Contact-Graph, containing an edge $(u, v)$ iff user $u$ added user $v$ to its personal contact list, the directed Favorite-Graph, containing an edge $(u, v)$ with label $c \in \mathbb{N}$ if user $u$ added exactly $c$ of $v$’s images to its personal list of favorite images as well as the directed Comment-Graph, containing edge $(u, v)$ with label $c \in \mathbb{N}$ if user $u$ posted exactly $c$ comments on $v$’s images.
Table 1: High level statistics for all networks with density \( d \), the number of strongly connected components \( \# \text{ecc} \) and the size of the largest strongly connected component SCC.

|       | \(|V_i|\) | \(|E_i|\) | \( d \) | \( \# \text{ecc} \) | SCC |
|-------|---------|---------|--------|----------------|------|
| Copy  | 1,427   | 4,144   | 2 \( \cdot 10^{-4} \) | 1,108          | 309  |
| Visit | 3,381   | 8,214   | 10^{-3}  | 2,599          | 717  |
| Friend| 700     | 1,012   | 2 \( \cdot 10^{-3} \) | 515            | 17   |
| ReTweet| 826,104 | 2,286,416| 3.4 \( \cdot 10^{-6} \) | 699,067        | 123,055 |
| Follower| 1,486,403| 72,590,619| 3.3 \( \cdot 10^{-5} \) | 198,883        | 1,284,201 |
| Comment| 525,902 | 3,817,626| 1.4 \( \cdot 10^{-6} \) | 472,232        | 53,359 |
| Favorite| 1,381,812| 20,206,779| 1.1 \( \cdot 10^{-6} \) | 1,305,350      | 76,423 |
| Contact| 5,542,705| 119,061,843| 3.9 \( \cdot 10^{-6} \) | 4,820,219      | 722,327 |

3 Analysis of Network Semantics

In the following, we tackle the problem of assessing the “meaning” of relations among pairs of vertices within such a network. This analysis then gives insights into the question, whether and to which extent the networks give rise to a common notion of semantic relatedness among the contained vertices. For this, we apply an experimental methodology, which was previously used for assessing semantic relationships within co-occurrence networks [10]. The basic idea is simple: We consider well founded notions of relatedness, which are naturally induced by external properties of the corresponding vertex sets, as, e.g., similarity of the applied tag assignments in BibSonomy or geographical distance between users in twitter. We then compute for each pair of vertices within a network these “semantic” similarity metrics and correlate them with different measures of structural similarity in the considered network.

3.1 Vertex Similarities

Below, we apply two well-established similarity functions in corresponding unweighted variants, namely the cosine similarity \( \text{COS} \) and the Jaccard Index \( JC \) as well as the corresponding weighted variants \( \text{COS}^{w} \) and \( JC^{w} \), following the presentation in [2]. Additionally we apply a modification of the preferential PageRank which we adopted from our previous work on folksonomies [3]. For a column stochastic adjacency matrix \( A \) and damping factor \( \alpha \), the global PageRank vector \( \vec{\omega} \) with uniform preference vector \( \vec{\beta} \) is given as the fixpoint of \( \vec{\omega} = \alpha A \vec{\omega} + (1 - \alpha) \vec{\beta} \). In case of the preferential PageRank for a given node \( i \), only the corresponding component of the preference vector is set. For vertices \( x,y \) we set accordingly \( \text{PPR}(x,y) := \vec{\omega}(c)[y] \), that is, we compute the preferential PageRank vector \( \vec{\omega}(c) \) for node \( x \) and take its \( y \)’th component. We calculate the adopted preferential PageRank score \( \text{PPR} \) from the preferential PageRank score in order to reduce frequency effects and set

\[
\text{PPR}_{\text{t}}(x,y) := \text{PPR}(x,y) - \text{PR}(x,y).
\]

3.2 Semantic Reference Relations

For assessing the semantic similarity of two nodes within a network, we consider the similarity of users based on the applied tags or hashtags, respectively, and the geographical distance of users in twitter and flickr.

Tag Similarity In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness (see, e.g., [11]).

We compute the cosine similarity in the vector space \( \mathbb{R}^T \), where, for user \( u \), the entries of the vector \( (u_1, \ldots, u_T) \in \mathbb{R}^T \) are defined by \( u_t := w(u,t) \) for tags \( t \) is the number of times user \( u \) has used tag \( t \) to tag one of her resources (in case of BibSonomy and flickr) or the number of times user \( u \) has used hash tag \( t \) in one of her tweets.

Geographical Distance In twitter and flickr, users may provide an arbitrary text for describing his or her location. Accordingly, these location strings may either denote a place by its geographic coordinates, a semi-structured place name (e.g., “San Francisco, US”), a colloquial place name (e.g., “Motor City” for Detroit) or just a fantasy name. Also the inherent ambiguity of place names (consider, e.g., “Springfield, US”) renders the task of exactly determining the place of a user impossible. Nevertheless, by applying best matching approaches, we assume that geographic locations can be determined up to a given uncertainty and that significant tendencies can be observed by averaging over many observations.

We used Yahoo!’s PlacemakerTM API for matching user provided location strings to geographic locations with automatic place disambiguation. In case of flickr, we obtained geographic locations for 320,849 users and in case of twitter for 294,668 users. Geographical distance of users is then simply given by the distance of the centroids for the correspondingly matched places.

3.3 Grounding of Shortest Path Distance

For analyzing the interdependence of semantic and structural similarity between users, we firstly consider a very basic measure of structural relatedness between two nodes in a network, namely their respective shortest path distance. We ask, whether users which are direct neighbors in an evidence network tend to be more similar than distant users. That is, for every shortest path distance \( d \) and every pair of nodes \( u,v \) with a shortest path distance \( d \), we calculated the average corresponding similarity scores \( \text{COS}(u,v), \text{JC}(u,v), \text{PPR}(u,v) \) with variants and geographic distance. To rule out statistical effects, we repeated for each network \( G \) the same calculations on shuffled null model graphs.
Geographic Distance

For average geographic distances in less pronounced tendencies for very distant nodes. diameter, the number of observations is too small, resulting average similarity again. For distances around a network’s Comment-Graph, the Follower-Graph and the ReTweet average up to a distance of eight). For the Visit-Graph, case of the ReTweet graph, users are more similar than in average (in case of the Contact-Graph), they also share a common pattern: Direct neighbors are in average significantly more similar than distant pairs of users. And with a distance of two to three, users tend to be less similar than in average (in case of the ReTweet graph, users are more similar than in average up to a distance of eight). For the Visit-Graph, the Comment-Graph, the Follower-Graph and the ReTweet graph, the average similarity scores approach the global average similarity again. For distances around a network’s diameter, the number of observations is too small, resulting in less pronounced tendencies for very distant nodes.

Semantic Similarity

For average geographic distances of users in flickr and twitter, we repeated the same calculations, as depicted in Figure 2. Firstly, we note the overall tendency, that direct neighbors tend to be located more closely than distant pairs of users within a network. Additionally, the average geographic distance of users then approaches the global average, and increases again after a certain plateau. As for the ReTweet-Graph, the average geographic distance remains at the global average level, once reached at a shortest path distance of ten.

Discussion

It is worth emphasizing, that in all considered evidence networks, the relative position of users already gives rise to a semantically grounded notion of relatedness, even in case of implicit networks, which are merely aggregated from usage logs as, e.g., the Visit-Graph. But one has to keep in mind that all observed tendencies are the result of averaging over a very large number of observations (e.g., 34,282,803,978 pairs of nodes at distance four in the Follower-Graph). Therefore, we cannot deduce geographic proximity from topological proximity for a given pair of users, as even direct neighbors in the Follower-Graph are in average located 4,000 kilometers apart from each other. But the proposed analysis aims at revealing semantic tendencies within a network and for comparing different networks (e.g., the Retweet-Graph better captures geographic proximity of direct neighbors in the graph).

3.4 Grounding of Structural Similarity

We now turn our focus towards different measures of structural similarity for nodes within a given network. There is a broad literature on such similarity metrics for various applications, such as link prediction [7] and distributional semantics [4; 10]. We thus extend the question under consideration in Section 3.3, and ask, which measure of structural similarity best captures a given semantically grounded notion of relatedness among users. In the scope of the present work, we consider the cosine similarity and Jaccard index, which are based only on the direct neighborhood of a node as well as the (adjusted) preferential PageRank similarity which is based on the whole graph structure (refer to Section 3.1 for details).

Ultimately, we want to visualize correlations among structural similarity in a network and semantic similarity, based on external properties of nodes within it. We consider, again, semantical similarity based on users’ tag assignments in BibSonomy, flickr and hashtag usage in twitter as well as geographic distance of users in flickr and twitter. In detail: For a given network \( G = (V, E) \) and structural similarity metric \( S \), we calculate for every pair of vertices \( u, v \in V \) their structural similarity \( S(u, v) \) in \( G \) as well as their semantic similarity and geographic distance. For visualizing correlations, we create plots with structural similarity at the x-axis and semantic similarity at the y-axis. As plotting the raw data points is computationally infeasible (in case of the Contact-Graph 30, 721, 580, 000, 000 data points), we binned the x-axis and calculated average semantical similarity scores per bin. As the distribution of structural similarity scores is highly skewed towards lower similarity scores (most pairs of nodes have very low similarity scores), we applied logarithmic binning, that is, for a structural similarity score \( x \in [0, 1] \) we determined the corresponding bin via \( \lfloor \log(x \cdot b^N) \rfloor \) for given number of bins \( N \) and suitable base \( b \). Pragmatically, we determined the base relative to the machine’s floating point precision \( \epsilon \) resulting in \( b := \epsilon^\frac{1}{N} \).

Semantic Similarity

Figure 3 shows the obtained results for each considered network separately. We firstly note, that the cosine similarity metric and the Jaccard index are highly correlated. Secondly, the adjusted preferential PageRank similarity consistently outperforms the other similarity metrics with respect to magnitude and monotonicity (except for BibSonomy’s Friend-Graph and flickr’s Contact-Graph).

Geographic Distance

As for geographic distances, Figure 4 shows the observed correlations for structural similarity in the different evidence networks and the corresponding average pairwise distance. In all but flickr’s Favorite-Graph, for both local neighborhood based similarity metrics \( COS \) and \( JC \), the average distance first decreases, but then increases again. This behavior is most pronounced in twitter’s ReTweet-Graph. In the Favorite-Graph, both \( COS \) and \( JC \) monotonically decrease with increasing similarity score. On the other hand, the average distance decreases monotonically with increasing preferential PageRank score \( PPR \) consistently in all considered networks, ex-
Discussion  Again, the obtained results only point at tendencies of the considered similarity metrics in capturing geographic proximity by means of structural similarity. Nevertheless, the adjusted preferential PageRank similarity consistently outperforms the other considered metrics. We therefore conclude that from all considered similarity metrics, the adjusted preferential PageRank similarity best captures the notion of geographic proximity. This is especially of interest, as the geographic proximity is a prior for many properties users may have in common, such as, e.g., language, cultural background or habits. twitter’s ReTweet-Graph seems to encompass the strongest geographic binding, as indicated in the relative low average distance for direct neighbors (cf. Figure 2 and the overall low average distance for higher preferential PageRank similarity scores (cf. Figure 4). Of course, other established similarity metrics (e.g., [6; 5; 4]) can be applied as well and are the subject of future considerations.

4 Conclusion & Future Work

With the present work, we introduced an experimental framework for assessing the semantics of social networks. The proposed methodology has a broad range of applications, such as user recommendation or community mining tasks, as it allows semantically grounded pre-processing of given networks (e.g., merging different small networks, scaling edge weights, selecting certain groups of users or directedness of networks). The conducted experiments give insights into the semantics of evidence networks from flickr, twitter and BibSonomy and well known similarity metrics.

Ultimately, the proposed experimental setup allows to formulate the assessment of semantic user relatedness as a regression task, which will be subject to future work.

References


