Conversations Reconstruction in the Social Web

Juan Antonio Lossio Ventura
University of Lyon 1, France
juan.lossio@etu.univ-lyon1.fr

Hakim Hacid, Arnaud Ansiaux, Maria Laura Maag
Bell Labs, Nozay - France
fname.lname@alcatel-lucent.com

ABSTRACT

We propose a socio-semantic approach for building *conversations* from social interactions following three steps: (i) content linkage, (ii) participants (users) linkage, and (iii) temporal linkage. Preliminary evaluations on a Twitter dataset show promising and interesting results.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Information storage and retrieval—*On-line Information Services*

General Terms

Experimentation

Keywords

Social Networks, Social Conversations, Social Dynamics

1. INTRODUCTION

In this work, we focus on the management of scattered data generated by social interactions. We consider that the exchanged content in social networks contain hidden and fragmented knowledge. Also, separately, these fragments may have limited utility and may even constitute noise, especially if processed with automated tools. We propose to reconstitute those fragments by providing a socio-semantic linkage of content in social networks. The result of this approach is an aggregation of content according to different social aspects that can convey meaning to an end user or a third party application.

We define the problem we are intending to address as follows: *Having a broad set of interactions between users of a social network (like Twitter) with disparate messages and relationships without additional information (meta-data), how can these interactions be linked so that they are correlated consistently and significantly for either an end user or an automatic process?* This problem has never been tackled before under this form although some initiatives exist [1]. Our proposed solution is unique in that it combines the semantic, social, and temporal dimensions to generate the possible connections between short messages in social networks defined with the different constraints discussed beforehand. In this work, all the observations have been performed on a Twitter dataset.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. PROPOSED APPROACH

It is generally difficult for the user to be aware about the different threads of discussions or for an automatic process to get some useful insights from such disparate content. Our initial hypothesis is that a message can be interesting for a user if it is highly similar to content exchanged between users he may know (directly or indirectly, i.e. through other social relatives) or sent during a period of time. To perform this linkage, our approach is composed of four main levels: (i) content, (ii) participants, (iii) temporal, and finally (iv) an effective linkage.

Content Level: Messages in social networks are written with an informal language containing slang, shortcuts, hyper-links, emotions (expressed mainly by character sets), etc. This limits the possibility of directly exploiting this content for automatic understanding. After cleaning, normalization, and enrichment (for hyper-links content), we proceed to the keywords extraction, weighting, and their similarity.

- **Importance (I)**: Generally, the importance of a keyword is considered dependent on its usage frequency in a given corpus. We believe that this is insufficient in our case and is certainly not likely to capture the real importance. Thus, the importance of a keyword k is calculated as a function combining: (i) the strength of the keyword and (ii) a propagation estimation of the keyword. These features are recovered using a *users* vs. *keywords* matrix.

Strength: It is calculated as in Equation 1 where $a_{u_i k_j}$ is the number of times a user u_i used the keyword k_j, $\mid U_{k_j} \mid$ the number of users in the community of the keyword k_j and $\mid K_{k_j} \mid$ the total number of keywords used in that community.

$$S_{k_j} = \frac{1}{\mid U_{k_j} \mid} \times \sum_{i=1}^{|U_{k_j}|} \frac{a_{u_i k_j}}{\sum_{m=1}^{|K_{k_j}|} a_{u_i k_m}} \quad (1)$$

Likelihood of propagation(D): A high usage rate of a keyword does not necessarily mean that it is important. We use the social characteristics of social communities to estimate a propagation degree. These characteristics are: (i) activity (representing the number of sent messages vs. total of messages of a user denoted A_{u_i}), (ii) participation (computing the amount of messages fired by user u containing keyword k denoted $Part_{u_i}$), and (iii) density (recovering the user's network density and denoted Den_{u_i}). We compute then the propagation likelihood as follows: $D_{k_j} = \frac{1}{|U_{k_j}|} \sum_{i=1}^{|U_{k_j}|} (A_{u_i} \times Part_{u_i} \times Den_{u_i})$

After these two computations, we can combine them to estimate keyword importance in the system: $I_{k_j} = S_{k_j} \times D_{k_j}$. After this step we obtain a set of pairs (k_j, v_j), where v_i is the keyword weight corresponding to the value of I_{k_j}.

- **Keywords Similarity (Sim):** We use a combination of the Jaccard and Dice measures as follows. If two keywords have close importance values, it is likely that these two keywords have a higher probability of being similar. This is particularly true since the importance integrates different dimensions as discussed before. The proximity is then: $prox_{k_ik_j} = 2 \times Min(I_{k_i}, I_{k_j})/(I_{k_i} + I_{k_j})$.

This proximity promotes the keywords that have high importance and penalizes low values. In fact, for small values, this measure requires that the importance values are closer in order to the value of this proximity to exceed the threshold. On the other hand, it allows a greater difference between the importance of values. Let α and β be parameters between 0 and 1. Their values are chosen manually, $\beta = 1 - \alpha$. Let $c_{k_ik_j}$ be the community formed by the keywords k_i and k_j and let $\mid U_{k_ik_j} \mid$ be the number of users of community $c_{k_ik_j}$. The similarity measure is then:

$$\begin{aligned} Sim_{k_ik_j} =& \alpha \left(prox_{k_ik_j} \times \frac{\mid c_{k_i} \cap c_{k_j} \mid}{\mid c_{k_i} \cup c_{k_j} \mid} \right) \\ &+ \beta \sum_{r=1}^{t} \left(prox_{(k_i \cap k_j)k_r} \times \frac{2\left(\mid c_{k_ik_j} \cap c_{k_r} \mid\right)}{\mid U_{k_ik_j} \mid + \mid U_{k_r} \mid} \right) \end{aligned} \qquad (2)$$

Participants level (LP): In the context of social networks, there is generally no explicit and evident relation between the "answers" and the root message. Consider two messages p and q. Let u_p and u_q be users who send messages p and q respectively. Let U_p and U_q be the set of users who appear in p and q respectively, including u_p and u_q. Let $f_{u_pu_q}$ be 1 if user u_p follows[1] and 0 otherwise. Now, let's consider in_{q-u_p} to be the value that represents whether the user u_p is in the message content of q and in_{p-u_q} the value that represents whether the user u_q is in the message content p. Participants proximity is computed then as follows:

$$LP_{pq} = \frac{1}{3}\Big(\frac{f_{u_pu_q} + f_{u_qu_p}}{2} + \frac{in_{q-u_p} + in_{p-u_q}}{2} + \frac{2 \mid U_p \cap U_q \mid}{\mid U_p \mid + \mid U_q \mid}\Big) \qquad (3)$$

Temporal level: Our assumption is that two messages sent at large intervals of time would not tend to be linked.Although the assumption seems strong, it is justified by the high dynamics related to social networks. Indeed, information in this type of structure has value for a short period of time. For this problem, we exploit the reactivity of a person as an indicator for message correlation. We analyzed our dataset to find the average time of a user logs per day, per month and finally a connection general average time of all users. After an evaluation in the social interactions database that we have, we found that a user has an average of three connections per day. This gives a logging interval of 8 hours for each user. It has been also shown [3] that (i) the propagation of information has a behavior with a two pulse curve and (ii) users generally react on messages within 3 hours after the launch of the discussion. Thus, given these two observations we decided to use an average value that represents the reaction time of 5.5 hours. This value means that once this period has passed, the link is penalized. Let d_p and d_q be the dates of each message. The connection time between two messages would be: $LT_{pq} = 1 - \frac{|d_p - d_q|}{5.5}$

Effective messages linkage: After the previous computations, we reach the final calculation that aims to make the effective linkage between messages. In our case, we consider this connection as a linear combination of similarity measures of the content, participants, time: $Link_{pq} = w_{cont} \times Sim_{pq} + w_{part} \times LP_{pq} + w_{tmp} \times LT_{pq}$ where, w_{cont}, w_{part} and w_{tmp} represent the weights given to each measure. These weights are between 0 and 1, are selected manually, and their sum $w_{cont} + w_{part} + w_{tmp} = 1$. Example of the obtained results is shown hereafter.

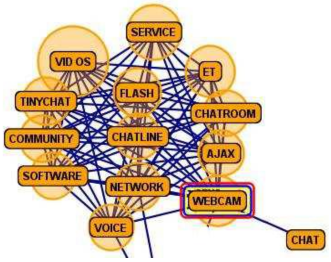

Figure 1: Subgraph of WEBCAM similarity

3. PRELIMINARY RESULTS

We operated two types of evaluations: **(i) evaluation with WordNet [2],** which is intended to check if the quality of the obtained links is similar to a human built structure. We used a collection of twitter messages containing the first $10K$ messages. The graph obtained is processed to keep only the extracted keywords from interactions which are also in Wordnet. We also keep the relationships between these keywords computed using our approach. From extracted keywords of the database of interactions, we keep only the first 471 (most important). From these results we compute an incompatibility of 97.4%, i.e. the relationship between keywords in social networks are not in WordNet. This confirms our initial hypothesis about the content of social interactions. However, we believe that consideration of a larger set of data could reduce this rate. **(ii) Manual evaluation:** In this step, we manually check the list of results if it is consistent and can be meaningful to the user. We use the same graph as before and we check manually all relationships. A relationship that is meaningful to a human has a positive note and that has no meaning is rated negative. The results obtained are very encouraging and show that the link quality is satisfactory with a precision of 0.78 and a recall of 0.97.

4. CONCLUSION AND FUTURE WORK

We discussed the problem of linking social interactions for building conversations. We have proposed an approach considering several levels and using the social network information: (i) content, (ii) users, and (iii) time. The innovation in this approach is also represented by the "massive" use of the social dimension at all levels of the process ensuring a contextual linkage. The preliminary results are encouraging and show the interest of the approach. As a next step, we intend to improve the approach and perform more evaluations.

5. REFERENCES

[1] S. Erera and D. Carmel. Conversation detection in email systems. In *ECIR*, pages 498–505, 2008.

[2] P. University. Wordnet, large lexical database of english, v2.1, 2008.

[3] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.

[1] Here "follows" is in the micro-blogging (Twitter) meaning.