Distinguishing Knowledge vs Social Capital in Social Media with Roles and Context

Vladimir Barash*, Marc Smith, Lise Getoor[%], Howard T. Welser^**

*Cornell University **Telligent Systems [%]University of MD ^Ohio University
*Ithaca, NY, USA **Texas, USA [%]College Park, MD, USA ^Athens, OH, USA
*vdb5@cornell.edu **marc.smith@telligent.com %getoor@cs.umd.edu ^welser@ohio.edu

Abstract

Social media communities (e.g. Wikipedia, Flickr, Live Q&A) give rise to distinct types of content, foremost among which are *relational content* (discussion, chat) and *factual content* (answering questions, problem-solving). Both users and researchers are increasingly interested in developing strategies that can rapidly distinguish these types of content. While many text-based and structural strategies are possible, we extend two bodies of research that show how social context, and the social roles of answerers can predict content type. We test our framework on a dataset of manually labeled contributions to Microsoft's Live Q&A and find that it reliably extracts factual and relational messages from the data.

1. Introduction

The deluge of data generated in social media communities (Wikipedia, Flickr, Yahoo!Answers) underscores the need for users to rapidly identify the information that is valuable to them, and for developers to provide tools for automating the evaluation of content. There are many possible strategies for automating content identification, but only a handful that leverage the inherently *social* nature of the data creation.

Several recent studies have shown that particular types of content, types of contributors, and social settings can be identified by distinctive signatures in the network structure of interaction associated with the creation of that content (Fisher et al 2006; Welser et al 2007; Adamic et al 2008). This study extends that work by focusing on how the context of where contributions are made and the types of contributors involved can be used to automatically distinguish factual from relational content. In the context of Q&A (Question and Answer) services, factual answers address questions like “how do I fix this specific technical problem:…” and generate knowledge capital in the form of factual information. Relational exchanges, which often begin with non-factual questions like "are atheists good people?" or even "how are you?", generate social capital by fostering relationships between users. Although other content types are generated in Q&A services, we will focus on relational and factual content exclusively in this work.

We construct a ranking framework that predicts the likelihood of a particular contribution to Live Q&A being factual or relational based on structural indicators of social roles and the structural features of tagged online contexts. These two types of features successfully extract both relational and factual content from a repository of Q&A messages.

We evaluate our framework on a dataset of user contributions to Microsoft Live Q&A system, where each contribution was manually labeled as relational or factual. The relational ranking function performs very well on this dataset, achieving 88% precision (vs. 80% baseline) at 50% recall. The factual ranking function performs worse, but still significantly exceeds the baseline at up to 25% recall. These results are especially significant given the complexity of the problem and the significant skew (80% relational posts) of the labeled data. We note the simplicity of our ranking function which makes use of just two variables.

2. Background

Recent studies suggest that network signatures can be used to identify distinctive social roles in the production of online content. The work of (Fisher et al. 2006) and (Welser et al. 2007) on roles in Usenet provides the basis for our operational definitions of “answer people” and “discussion people” in Q&A services, and for the link between the role a user plays and the type of content he or she produces. Researchers, (Adamic et al. 2008, and Mendes et al. 2008) examine tagging systems in Yahoo!

Answers and Live Q&A and find that different tags are associated with different levels of social interaction between users. Their findings motivate our use of roles and context to predict content type.

In addition to structural approaches, several recent papers in information retrieval have used textual attributes to automatically distinguish between facts and opinions, for a review (Wiebe et al. 2004). Harper, Moy and Konstan (2009) studied samples from three Q&A services and implemented a variety of content and structural indicators to distinguish between informational and conversational questions. Their work provides an important complement to our current study, and focuses on question content exclusively. We seek to extend the earlier research on the structural attributes of those who provide both questions and answers (Welser et al. 2007).

3. Roles and Context

3.1 Roles

Previous investigations of social media (Fisher et al. 2006; Welser et al. 2007) have discovered that users often follow very distinctive patterns of activity, playing roles in their online community. The identification and cataloguing of these patterns in spaces like Usenet has extracted two roles relevant to this work: the role of the answer person, who provides factual and technical contributions and the role of the discussion person, who provides opinions and chats (relational contributions) with other users. Answer people engage in many threads, but usually post one reply per thread, and their egocentric social networks are very sparse, resembling stars. Discussion people engage in fewer threads than answer people, post many messages per thread, and have very dense egocentric social networks resembling cliques.

In order to perform automatic identification of the roles users play in Q&A services, we extend previous work to define a ranking that indicates how closely a given user matches the activity pattern of a particular role. We make our ranking approach precise by defining a user's "answer person score" *Ans(u)* and "discussion person score" *Dis(u)*. First, we define three quantities for each user: *NT(u)*, the number of threads posted to (regardless of whether the post is a question, answer, best answer or comment), *MPT(u)*, the average number of messages posted per thread, and *LND(u)*, the user's local network density. In order to calculate local network density, we construct the social network graph *N=(V,E)* as follows: first, we bin data about all contributions by week. Then, for every week *t*, we construct a directed edge from *u* to *v* if *u* has answered *v*'s question or commented on *v*'s answer, provided both *v's* question/answer and *u's* answer/comment happened in *t.* We then calculate local network density for a user *u* as follows:

$$LND(u) = \frac{|(a,b): a \in Nbr(u) \wedge b \in Nbr(u)|}{k(u) * (k(u) - 1)} \quad \text{(Eq.1)}$$

where *Nbr(u)* is the set of *u*'s alters, the union of *u*'s in- and out-neighbors, and *k(u)*= $|Nbr(u)|$. This quantity is a tie ratio similar to the Watts and Strogatz(1998) clustering coefficient. Finally, we average *LND* over all *t* for the same user, ignoring time periods when the user had no alters.

Given these three quantities, we can define the answer and discussion scores of user *u*, *Ans(u)* and *Dis(u),* to match the descriptions of these roles' activity patterns given above.

$$Ans(u) = \alpha_{ANS} f_1\big(NT(u)\big) - \beta_{ANS} f_2\big(MPT(u)\big) - \gamma_{ANS}\, f_3\,(LND(u)\,) \quad \text{(Eq.2)}$$

Here $\alpha_{ANS}, \beta_{ANS}, \gamma_{ANS} \alpha_{DIS}, \beta_{DIS}, \gamma_{DIS}$ are weighting constants between 0 and 1, and f_1, f_2, f_3 are non-decreasing functions.

$$Dis(u) = -\alpha_{DIS} f_1\,\big(NT(u)\big) + \beta_{DIS}\, f_2\,(MPT(u)) + \gamma_{DIS} f_3\big(LND(u)\big) \quad \text{(Eq.3)}$$

These two scores impose two rankings on all users. The first ranking corresponds to the likelihood of user *u* contributing factual content and the second ranking corresponds to the likelihood of user *u* contributing relational content.

3.2 Context

From the perspective of roles, every user (or at least the prototypical users) in a Q&A service behaves in a consistent manner, producing content of a particular type regardless of the circumstances. It is, however, more likely that users behave differently in different contexts within a Q&A service. The tagging systems of Q&A services provide just such a context. These systems allow users to assign tags to questions and apply the assigned tags to all answers and comments on their parent question. Previous work on tags in Q&A spaces (Adamic et al. 2008) suggests that different tags induce different relationship patterns, for instance the "Marriage" tag (which has a lot of relational content) induces a dense relationship network, while the "Programming" tag (which has a lot of factual content) induces a sparse relationship network.

This observation leads to a hypothesis about tags and content type. We propose that tags that are characterized by very sparse relationship patterns are much more likely to feature factual content, while tags that are characterized by very dense relationship patterns are much more likely to feature relational content. Our operational definition of the relationship network density of some tag *g* is expressed as a tag density score *TG(g)*, defined as follows:

$$TG(g) = \frac{\sum_{u \in V(N(g))} LND_{N(g)}(u)}{|V(N(g))|} \quad \text{(Eq. 4)}$$

Here *N(g)* is a tag network for *g*, generated in the same way as *N,* but restricted only to messages tagged with *g*. Accordingly, *V(N(g))* is the set of vertices of *N(g)* and $LND_{N(g)}(u)$ is the local network density of a vertex *u* in *N(g)*. As with *LND(u),* we average over all *t,* ignoring periods when the user had no alters.

We can make use of *g* to create contextualized versions of answer and discussion person scores. To do so, we calculate $NT_g(u)$ and $MPT_g(u)$ for a user *u* considering only posts (questions, answers, comments) made by *u* that are tagged with *g*. Then, we calculate $LND_{N(g)}(u)$ using *N(g)*. Finally, we substitute these three quantities in place of *NT*, etc. into equations 2 and 3 above to calculate $Ans_g(u)$ and $Dis_g(u)$, the tag-specific answer and discussion person scores of user *u*:

$$Ans_g(u) = \alpha_{ANS_g} f_1\left(NT_g(u)\right) - \beta_{ANS_g} f_2\left(MPT_g(u)\right) - \gamma_{ANS_g} f_3\left(LND_{N(g)}(u)\right) \quad \text{(Eq. 5)}$$

$$Dis_g(u) = -\alpha_{DIS_g} f_1\left(NT_g(u)\right) + \beta_{DIS_g} f_2\left(MPT_g(u)\right) + \gamma_{DIS_g} f_3\left(LND_{N(g)}(u)\right) \quad \text{(Eq. 6)}$$

Here $\alpha_{ANS_g}, \beta_{ANS_g}, \gamma_{ANS_g} \alpha_{DIS_g}, \beta_{DIS_g}, \gamma_{DIS_g}$ are weighting constants between 0 and 1, and f_1, f_2, f_3 are nondecreasing functions.

4. Ranking Framework

Given the role and context scores calculated above, we translate them into a unified framework for predicting the content type of a particular contribution, as follows:

1. For any contribution *c,* determine the user *u* who made the contribution and the set of tags *SG* assigned to the contribution.
2a. Calculate *Ans(u)*, *Dis(u)*, *TG(g)*, $Ans_g(u)$, $Dis_g(u)$ for user *u* and all tags $g \in SG$.
2b. Set: $TG(SG) = Agg_{g \in SG}(TG(g))$, $Ans_{SG}(u) = Agg_{g \in SG}\left(Ans_g(u)\right)$, $Dis_{SG}(u) = Agg_{g \in SG}\left(Dis_g(u)\right)$ where *Agg(x)* is a function from $\mathbb{R}^{|SG|}$ to $\mathbb{R}$. Examples of *Agg(x)* are the average, maximum, and minimum functions.
3. Interpolate subsets of these quantities into two ranking scores for *c*, $R_{Fact}(c)$ and $R_{Rel}(c)$ as follows:

$$R_{Fact}(c) = \delta_{FACT} Ans(u) - \epsilon_{FACT} TG(SG) + \eta_{FACT} Ans_{SG}(u) \quad \text{(Eq. 7)}$$

$$R_{Rel}(c) = \delta_{REL} Dis(u) + \epsilon_{REL} TG(SG) + \eta_{REL} Dis_{SG}(u) \quad \text{(Eq. 8)}$$

We can apply these scores to empirical data and present top lists of contributions of each content type.

5. Evaluation

5.1 Data

The Microsoft Live Q&A system (qna.live.com) is a public question answering community. We studied the database of all messages posted on Live Q&A during a five-month period (September 2007 through January 2008). , The studied sample had roughly 950,000 messages posted by roughly 30,000 users. The total tag count in the sample was around 65,000.

5.2 Results

In order to evaluate the performance of our ranking functions, we applied them to a labeled dataset of Live Q&A messages. This dataset, reported on by (Welser et al. in progress) consists of almost 6000 messages in Live Q&A hand-labeled by human judges. Each message was labeled as belonging to zero or more of the following categories: Factual, Technical, Advice, Opinion, Support, Joke, Flame. We separated these categories into a factual set, consisting of Factual, Technical and Advice, and a relational set, consisting of Opinion, Support, Joke and Flame. We placed Advice in the factual set, because the latter label indicates the presence of specific instructions for solving a problem, as opposed to general supportive statements (which are labeled Support). We assigned each message in the dataset the label "Factual" if it belonged to at least one category in the factual set and no categories in the relational set, and "Relational" if it belonged to at least one category in the relational set and no categories in the factual set, removing all messages that belong to both or neither category. The resulting dataset showed a heavy class skew (around 80% are labeled "Relational").

For each message *c*, we calculated $R_{Fact}(c)$and $R_{Rel}(c)$ as described above. This process involves setting forms for the functions *Agg(x)* and $f_1, \ldots f_3$ and values for the 18 constants $\alpha_{ANS\backslash DIS} \cdots \gamma_{ANS\backslash DIS}$, $\alpha_{ANS_g\backslash DIS_g} \cdots \gamma_{ANS_g\backslash DIS_g}$, $\delta_{FACT\backslash REL} \cdots \eta_{FACT\backslash REL}$. We found that a restricted set of these parameters led to drastic improvements in performance over more complex settings. Specifically, we set *Agg(x)* to be the "relational-skew" function which selects the maximum *TG(g)*, $Dis_g(u)$ and the minimum $Ans_g(u)$ for all $g \in SG$, which returns the values of all three quantities that most strongly indicate that *c* is a relational contribution. We also imposed the following constraints:

	Relational Content										Factual Content									
	Precision at					Precision / Recall						Precision at					Precision / Recall				
Function	10	50	100	250	500	.05	.1	.25	.5	.8	Function	10	50	100	250	500	.05	.1	.25	.5	.8
LND,TG(g)	**1.0**	.94	**.94**	**.91**	**.90**	**.89**	**.91**	**.90**	**.88**	**.86**	*LND,TG(g)*	**.9**	.68	**.58**	.44	**.37**	**.74**	**.58**	**.40**	**.31**	**.23**
MPT,LND,TG(g),MPT$_g$, LND$_g$	**1.0**	**.98**	.90	.90	.85	.88	.89	.85	.84	.84	*LND, TG(g), LND$_g$*	**.9**	.68	**.58**	.45	**.36**	**.74**	**.58**	.38	.29	**.23**
MPT,LND, MPT$_g$, LND$_g$	**1.0**	**.98**	.90	.90	.85	.88	.90	.85	.84	.83	*LND, TG(g), MPT$_g$, LND$_g$*	**.9**	**.69**	**.58**	.4	.32	.70	**.58**	.35	.27	.20

Table 1. Precision at *k* and Precision / Recall for Ranking functions of Relational and Factual Content

set $f_1, \ldots f_3 = I$, the identity function; impose constant symmetry, i.e. $\alpha_{ANS} = \alpha_{DIS} \ldots \gamma_{ANS} = \gamma_{DIS}$ and $\alpha_{ANS_g} = \alpha_{DIS_g} \ldots \gamma_{ANS_g} = \gamma_{DIS_g}$,$\delta_{FACT} = \delta_{REL} \ldots \eta_{FACT} = \eta_{REL}$; and restrict the values of all constants are either 0 or 1.

These constraints make the ranking functions symmetrical, so $R_{Fact}(c) = -R_{REL}(c)$, and restrict the parameter space to $2^7 = 128$ possible combinations. Each combination of parameters corresponds to a ranking function, and for each such function we calculated precision-at-*k* and precision for various values of recall, for each label separately.

We found that certain simple combinations of parameter values significantly boosted values of precision at *k* and precision / recall for both labels. We report these two measures for the three functions that performed best on ranking the relational and factual content (Table 1 above). The baseline performance in all cases is 80% precision for the relational content and 20% for the factual content. The function column lists the quantities that were non-zero for that particular combination of parameter values.

The very simple function pair: $R_{Fact}(c) = -LND(u) - TG(SG)$, $R_{Rel}(c) = -R_{Fact}(c)$ achieves maximum precision among all 127 functions examined, for many values of *k* and recall. Averaging precision at *k* over all values of *k*, and precision over all values of recall, confirms that this simple function pair achieves the best performance in this parameter space. Further, we see the features *LND(u)* and *TG(g)* in most of the top ranking functions for both labels. This suggests that egocentric network density is a strong predictor of content type, at least at extreme values. Note that both a role element (*LND(u)*) and a context element (*TG(g)*) are present in the best-performing function pair, indicating that both role and context are crucial for predicting content type.

6. Conclusion

Our work proposes a new approach to ranking content in Q&A and related social media services. Our approach is reasonably effective (at least at low values of *k* and recall), suggesting that it would be appropriate for navigation and search tasks, where a user may be interested in finding some, but not all, messages of a particular type. Our approach is simple – the best-performing ranking function uses just two variables. At the same time, our approach relies on both role- and context-related features to rank content. This suggests that the sociological concepts of role and context capture interesting patterns of user behavior, which may be effectively used to predict the content a user is likely to contribute to a Q&A service. Finally, our approach is flexible: we can apply the ranking framework to any system with posts, users, reply relationships and message tags. We hope to extend this work to other domains and further explore how different patterns of user interaction affect the content created by these users in the rapidly growing space of social media.

Acknowledgements

We would like to thank Microsoft Research, the Live Q&A Product team and users, LLNL and NSF #0746930, which helped support this work. We would like to thank Stephen Purpura for helpful feedback.

References

1. Adamic, L.A.; Zhang, J; Bakshy, E.; and Ackerman, M.S. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. *WWW2008*.
2. Fisher, D.; Smith, M; and Welser, H.T. 2006. You Are Who You Talk to: Detecting Roles in Usenet Groups. *In Proceedings of the 39th Hawaii International Conference on Systems Sciences (HICSS)*.
3. Harper, F.M.; Moy, D.; and Konstan, J.A. 2009. Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites. Forthcoming.
4. Mendes, E.; Milic-Frayling, N.; and Fortuna, B. 2008. Social Tagging Behaviour in Community-driven Question Answering. *In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence.*
5. Watts , D.J, and Strogatz, S.H. 1998. Collective Dynamics of 'Small-world' Networks. *Nature* 393(668): 409-10.
6. Wellman, B. 2001. Computer Networks as Social Networks. *Science* 293(14): 2031-2034.
7. Welser, H. T.; Gleave, E.; Fisher, D.; and Smith, M. 2007. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure* 8(2).
8. Welser, H. T.; Smith, M., Gleave, E.; Barash, V.; and Meckes, J. 2009. Unanswered Questions: Social Affordances and the Cultivation of Experts in Q&A Systems. Forthcoming.
9. Whittaker, S.; Terveen , L.; Hill, W.; and L. Cherny, L. 1998. The Dynamics of Mass Interaction." *In Proceedings of the 1998 ACM conference on Computer supported cooperative work,* p.257-264, 11/14-18.
10. Wiebe, J.M.; Wilson, T.; Bruce R.; Bell, M.; and Martin, M. 2004. Learning Subjective Language. *Computational Linguistics* 30:277-308.