"How Incredibly Awesome!"- Click Here to Read More

Hyung-il Ahn
MIT Media Lab
20 Ames Street, Cambridge, MA 02139
hiahn@media.mit.edu

Werner Geyer, Casey Dugan, David R. Millen
IBM T.J. Watson Research
One Rogers Street, Cambridge, MA 02142
{werner.geyer, cadugan, david_r_millen}@us.ibm.com

Abstract

We investigate the impact of a discussion snippet's overall sentiment on a user's willingness to read more of a discussion. Using sentiment analysis, we constructed positive, neutral, and negative discussion snippets using the discussion topic and a sample comment from discussions taking place around content on an enterprise social networking site. We computed personalized snippet recommendations for a subset of users and conducted a survey to test how these recommendations were perceived. Our experimental results show that snippets with high sentiments are better discussion "teasers."

Introduction

Popular social media sites attract millions of users who contribute enormous amounts of content, such as photos, videos, bookmarks, etc. Helping users identify content they would likely find interesting is a challenging problem. Typically sites surface content to users in a number of ways, such as presenting lists of recently added content for serendipitous discovery, lists of matching content after a search, and personalized content recommendations made to a given user. In all of these cases, screen space is limited and the system can only present a "*snippet*" or "teaser" view of the content. This snippet must include the information most likely to help the user make an informed decision as to whether or not to click and view the content in its entirety. One approach is to display a snippet that contains the most relevant keywords for a user, as Google does in presenting its search results.

In our research we are not only interested in the content itself but also the discussions that evolve around the content, represented as comments from different users. For a discussion, one approach is to show the topic and most recent comments, which is often done in forums. But for discussions taking place around the content shared on social media sites, it is unclear which of any of the hundreds of possible comments left on a given item should be included in the snippet.

In this paper, we are interested in the impact of sentiment on a user's click-through behavior and in particular how snippets designed by taking sentiment into account are received by users. We studied the effect of sentiment in the context of a discussion recommendation system that leverages the sentiment information of discussion topics and comments when presenting snippets to users. Sentiment analysis has become an important tool in mining the web for personal feelings and opinions, and shedding light on the collective consciousness of users (Wright 2009) but, to our knowledge, has not yet been applied to personalized discussion recommendations. We studied whether affective snippets are more persuasive than neutral snippets in helping users decide whether to read discussions.

This paper presents the findings from an online survey of 1,200 users of an enterprise social networking site, SocialBlue, formerly known as Beehive (DiMicco et al. 2008). We asked participants to rate the recommendations we computed for them. Our goal was to address the following question: Does the snippet sentiment (i.e., the sentiment of the topic and sample comment of the discussion) influence how likely the user is to go to the discussion page and read the entirety of the discussion (i.e., the *willingness to read or willingness to click through*)?

We hypothesized that high-sentiment snippets (those containing positive or negative sentiment) would increase the user's willingness to read them, compared to snippets with low-sentiment (neutral). Our main hypotheses are:

H1) For recommendations of discussions that match a user's interests/keywords (i.e., when the user tends to be interested in the discussion topic), presenting snippets containing high sentiment (positive, negative) will increase the user's likelihood of reading the entirety of the discussions, compared to the low-sentiment snippets (neutral).

H2) Even for the recommendations of non-topic-matched discussions (i.e., when the user tends to be uninterested in the discussion topic), high-sentiment snippets (positive, negative) will still increase the user's likelihood of reading the entirety of the discussions, compared to low-sentiment snippets (neutral).

Design and Algorithm Evaluation

We designed and implemented a discussion recommender system for SocialBlue, using sentiment analysis and interest-matching algorithms. SocialBlue, formerly known as Beehive, is an IBM internal social networking site which was launched in September 2007 (DiMicco et al. 2008). The site had more than 62,000 users in August 2009. SocialBlue has three different content types: photos, hive fives (top 5 lists) (Geyer et al. 2008), and events, with nearly 26,000 discussions taking place across those. Note that each piece of content has a title (which we used as the discussion topic), a content description, tags and one or more comments following the description.

Sentiment-based Snippet Design

For each discussion recommendation, our system creates a simple snippet that is composed of the topic of the discussion (i.e. the social media object's title) and one sample comment from the discussion. This simplifies both the interface as well as the analysis of the overall sentiment of the snippet (versus showing multiple comments). Note that even for the same topic the overall sentiment of the snippet may differ (and be adjusted) according to which sample comment is used in the snippet. For instance, for a topic with no sentiment (or neutral sentiment), when a snippet has a negative sample comment, the overall sentiment of the snippet (the topic plus a sample comment) becomes negative (Figure 1).

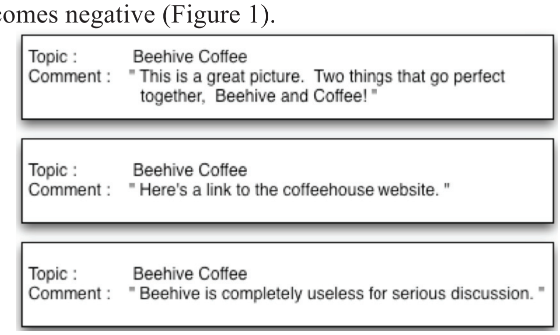

Figure 1: Different snippet examples recommending the same topic: positive sentiment (top), neutral sentiment (middle), and negative sentiment (bottom).

Personalized Topic-Matching

The core discussion recommender made use of a personalized topic-matching algorithm to compute a set of potentially interesting recommendations for each user. Our algorithm uses TF-IDF (term-frequency inverse-document-frequency) based word vectors and cosine similarity as a similarity measure. For each user, matched discussions (i.e., the similarity score with the user's profile is greater than a threshold value) and non-matched discussions (i.e., the similarity score is below a threshold value) were then classified according to the overall sentiment of each potential snippet (discussion topic and comment pair) using our sentiment analysis algorithm and presented to users as discussion recommendations in our experiment.

Sentiment Analysis Algorithm Evaluation

For each discussion topic and comment pair on SocialBlue, we classified the overall sentiment into a positive, neutral or negative category. Prior to our experiment, we evaluated different sentiment analysis algorithms: an IBM eClassifier-based algorithm (Cai et al. 2008), a SentiWordNet-based algorithm (Denecke 2008, Esuli and Sebastiani 2006), a LIWC-based algorithm (Pennebaker et al. 2001), various machine-learning algorithms using the WEKA package (Witten and Frank 2005), and the LingPipe package (Carpenter and Baldwin 2008) using n-gram character-based language models.

In order to be able to train a classifier and also to validate the accuracy of each algorithm, we prepared a human-labeled training set of 600 randomly selected sample snippets, 200 snippets from each content type (photos, hive fives, events). For each snippet in this training set, two human coders individually classified it into a positive, neutral or negative sentiment class. When the two coders disagreed on the sentiment class of a snippet, they talked with each other and reached an agreement. Note that before the coders started the individual coding, they did a level-setting session in which they coded and discussed a set of example snippets. Among the 600 random snippets in the training set, 62.7% were coded with positive sentiment, 20.7% neutral, and 16.6% negative.

	Precision			Recall			Total Accuracy
	+	o	-	+	o	-	
LIWC	72%	53%	54%	91%	42%	13%	68%
SentiWNet	68%	50%	39%	93%	28%	5%	65%
eClassifier	68%	41%	41%	91%	26%	7%	64%
LingPipe	68%	44%	36%	92%	18%	12%	64%
WEKA	73%	56%	40%	88%	44%	21%	67%

Table 1. The classification performance (from 10 fold cross-validations)

The LIWC-based sentiment analysis algorithm using the logistic regression classifier based on positive and negative emotion scores obtained the best classification performance on average (see Table 1). We obtained a relatively higher precision/recall rate for positive sentiments, compared to neutral and negative sentiments. This is mainly due to the fact that our underlying domain (and thus, our human-labeled training set constructed by random selection) is highly unbalanced, i.e. far more positive topic-comment pairs than neutral and negative ones. When we applied the LIWC-based algorithm to all possible snippets on SocialBlue (i.e., all possible topic-comment pairs), we found that among a total of 80,647 possible snippets, there were 78.4% with positive, 17.2% with neutral, and 4.4% with negative sentiment.

Method

We conducted a personalized survey on SocialBlue in order to understand the impact of the snippet sentiment when recommending a content discussion, i.e. how willing is the user to read the discussion and how much is the user generally interested in the discussion.

We designed a mixed-condition experiment for 1,200 active SocialBlue users (see Figure 2). There are two experimental variables that influence each discussion snippet: topic (match/no-match) and sentiment (positive/neutral/negative). The topic condition was tested between-subjects, and the snippet-sentiment condition was tested within-subjects.

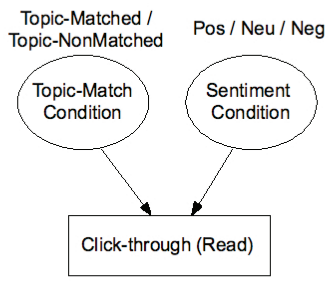

Figure 2. The mixed design for our survey experiment: topic match condition (between subjects), sentiment condition (within subjects)

Participants were randomly selected from all users who had logged into SocialBlue at least once during the two weeks preceding the start of the survey. In one group of 600 active users, each user was recommended 12 matched discussions (topic-matched group); in the other group, each was recommended 12 non-matched discussions (non-topic-matched group). Also, for both matched and non-matched conditions, there were three within-subject conditions related to the sentiment of the snippet (positive, neutral, negative sentiment), so every user was exposed to 12 snippets in total, four snippets for each sentiment category.

Given that each snippet can be described as a (T, C) pair where T is the topic and C is the selected comment from the discussion, our strategy for selecting four snippets for a target sentiment category (positive / neutral / negative) for each user was as follows:

1) Find all possible (T, C) pairs for the topic-matched (or non-topic-matched) discussion topics.

2) Reorder the list of (T, C) pairs by the predicted probability of the target sentiment category.

3) Select four (T, C) pairs randomly from the pool of (T, C) pairs whose predicted probability of the target sentiment category is within the top 25%.

Applying the top 25% criterion to select four snippets for each sentiment category increased our total accuracy of finding the snippets with the target sentiment by about 5%.

The survey was composed of two pages. On the first page we asked the following question (Q1) for each of the 12 recommendations. The goal was to measure the willingness of a participant to read the discussion.

(Q1) *Considering the topic of the discussion and the sample comment shown, how likely is it you would go to the discussion page and read the discussion?*

Responses included: very unlikely, somewhat unlikely, somewhat likely, very likely.

On the second page of the survey we asked the question (Q2) for each of those recommendations the participant responded to on the first page. The question measures the user's perceived interest in the recommended discussion.

(Q2) *In general, how interesting is the recommended topic and sample comment to you?*

Responses included: very uninteresting, somewhat uninteresting, somewhat interesting, very interesting.

Results

Of the total 1,200 users, 754 logged in and 448 submitted their survey. Because of missing responses in the survey, the actual sample size of users was 222 for the topic-matched and 169 for the non-topic-matched condition.

Willingness to Read Discussions

A general linear model (GLM) repeated measures ANOVA analysis to assess participants' willingness-to-read acceptance response rate data showed a significant main effect of the topic-match condition ($F[1, 389] = 40.7$, $p < .001$), a significant main effect of the sentiment condition ($F[2, 778] = 11.6$, $p < .001$) and no significant interaction effect between match and sentiment conditions ($F[2, 778] = 0.2$, $p = .82$).

Post-hoc comparisons (LSD) showed (see Table 2) that the matched group of users (Mean=53.6%) have a significantly higher willingness to read discussions than the non-matched group of users (Mean=37.8%) (p<.001). Snippets with positive sentiment (Mean=48.6%) and snippets with negative sentiment (Mean=47.3%) drove a significantly higher willingness to read discussions than snippets with neutral sentiment (Mean=41.2%) across both the topic-matched and non-topic-matched groups (p<.001). However, there was no significant difference in willingness to read between positive and negative sentiment snippets.

	Pos	Neu	Neg	Average
Overall (N=391)	48.6%	41.2%	47.3%	45.7%
Matched (N=222)	57.3%	48.2%	55.2%	53.6%
Non matched (N=169)	40.2%	33.8%	39.3%	37.8%

Table 2. The average acceptance rates for the willingness to read responses

For the topic-matched group, there was a significant main effect of sentiment ($F[2, 442] = 7.6$, $p < .01$). Post-hoc comparisons (LSD) showed that snippets with positive sentiment (Mean=57.3%) and snippets with negative sentiment (Mean=55.2%) drove a significantly higher willingness to read the discussions than snippets with neutral sentiment (Mean=48.2%) ($p < .001$). Figure 3 shows the willingness-to-read response distribution over all the recommendations in each sentiment category for this group.

The high-sentiment (positive, negative) conditions had a higher percentage of "somewhat likely" and "very likely" responses and a lower percentage of "somewhat unlikely" and "very unlikely" responses, compared to the low-sentiment (neutral) condition. Moreover, for the non-topic-matched group, we still found a significant main effect of sentiment ($F[2, 336] = 4.59$, $p < .05$). The above results from Q1 verify our main hypothesis *H1* and *H2*.

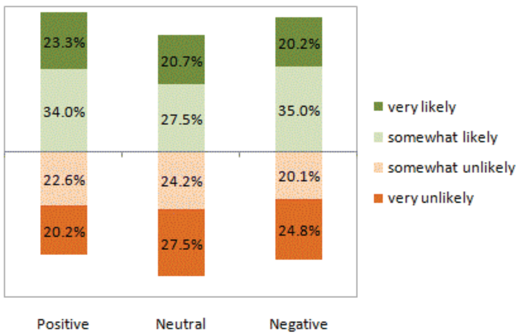

Figure 3. The willingness to read response distribution in each sentiment category for the topic matched group

Perceived Interest

Analyzing Q2 answers, post-hoc comparisons (LSD) showed that the topic-matched group of users (Mean=55.1%) have a significantly higher interest in discussions than the non-topic-matched group of users (Mean=42.7%) ($p < .001$), where "interest" was defined as a user choosing either "somewhat interesting" or "very interesting". This confirms that our personalized topic-matching algorithm worked well enough to significantly determine the general discussion interest of participants.

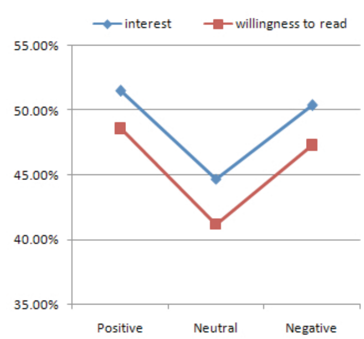

Figure 4. The average acceptance rates for the willingness to read and the discussion interest responses across all groups

Also, similar to the findings on the willingness to read discussions, snippets with positive (Mean=51.5%) or negative sentiment (Mean=50.4%) had a significantly higher perceived interest than snippets with neutral sentiment (Mean=44.7%) across all groups ($p < .001$).

As expected the self-reported interest in discussion topics is significantly greater than the willingness to read discussions for all sentiment categories as shown in Figure 4 (paired-sample t-tests: $T(390) = 2.4$, $p < .05$ for positive sentiment, $T(390) = 3.0$, $p < .01$ for neutral sentiment, $T(390) = 2.46$, $p < .05$ for negative sentiment).

Conclusion

The overall sentiment of discussion snippets influenced users' willingness to read discussions independent of a user's interest in the discussion. In both conditions (topic-matched, non-topic-matched), high-sentiment (positive, negative) snippets drove a 5~9% increase in the average acceptance response rate of the willingness to read discussions, compared to low-sentiment (neutral) snippets.

Given that there was about 16% difference in the average acceptance response rate of the willingness to read discussions between topic-matched and non-topic-matched conditions, the sentiment factor can be considered more than half as important as the topic-match factor in terms of the influence on the willingness to read. This suggests that sentiment could play a very essential role in the design of snippets, i.e. presenting matched content in the best possible way to users when making recommendations.

Future work focuses on improving the sentiment classification performance. Affective profiles, for example, could be used to model individual preferences. If we are able to understand and represent whether a person prefers negative, positive, or neutral sentiment, or potentially other facets of emotions, the snippets could be created in a more targeted fashion.

References

Cai, K., Spangler, S., Chen, Y., Zhang, L. (2008). Leveraging Sentiment Analysis for Topic Detection. In Proc Web Intelligence and Intelligent Agent Technology.

Carpenter, B. and Baldwin, B. (2008). LingPipe, http://alias-i.com/lingpipe.

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentiment Analysis. In Proc ICDEW 2008.

DiMicco, J., Millen, D., Geyer, W., Dugan, C., Brownholtz, B., Muller, M. (2008). Motivations for Social Networking at Work. In Proc. ACM CSCW '08.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proc LREC-06, Genova, IT, 2006.

Geyer, W., Dugan, C., DiMicco, J., Millen, D.R., Brownholtz, B., Muller, M. (2008). Use and reuse of shared lists as a social content type. In Proc ACM CHI '08.

Pennebaker, J. W., Francis, M. E., Booth, R.J. (2001). Linguistic inquiry and word count: LIWC. Mahwah: Lawrence Erlbaum Associates.

Witten, I. H. and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques, San Fransisco, Morgan Kaufman Publishers.

Wright, A. (2009). "Mining the Web for Feelings, Not Facts ." http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html.