PREDICTING COMMUNITY PREFERENCE OF COMMENTS ON THE SOCIAL WEB

A Thesis

by

CHIAO-FANG HSU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2009

Major Subject: Computer Science

PREDICTING COMMUNITY PREFERENCE OF COMMENTS ON THE SOCIAL WEB

A Thesis

by

CHIAO-FANG HSU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	James Caverlee
Committee Members,	Yoonsuck Choe
	Patrick Burkart
Head of Department,	Valerie E. Taylor

December 2009

Major Subject: Computer Science

ABSTRACT

Predicting Community Preference of Comments on the Social Web. (December 2009)

Chiao-Fang Hsu, B.S., National Tsing Hua University

Chair of Advisory Committee: Dr. James Caverlee

Large-scale socially-generated metadata is one of the key features driving the growth and success of the emerging Social Web. Recently there have been many research efforts to study the quality of this metadata – like user-contributed tags, comments, and ratings – and its potential impact on new opportunities for intelligent information access. However, much existing research relies on quality assessments made by human experts external to a Social Web community. In the present study, we are interested in understanding how an online community itself perceives the relative quality of its own user-contributed content, which has important implications for the successful self-regulation and growth of the Social Web in the presence of increasing spam and a flood of Social Web metadata.

We propose and evaluate a machine learning-based approach for ranking comments on the Social Web based on the community's expressed preferences, which can be used to promote high-quality comments and filter out low-quality comments. We study several factors impacting community preference, including the contributor's reputation and community activity level, as well as the complexity and richness of the comment.

Through experiments, we find that the proposed approach results in significant improvement in ranking quality versus alternative approaches.

DEDICATION

To my beloved Qing and parents for their encouragement, support and love.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Caverlee, and my committee members, Dr. Choe and Dr. Burkart, for their guidance and support throughout the course of this research. I would also like to thank my senior student, Elham Khabiri and other labmates for their constructive suggestions and discussions in the completion of my research.

I also want to extend my gratitude to the 2009 IEEE International Conference on Social Computing reviewers as well as the 2009 International AAAI Conference on Weblogs and Social Media reviewers for their valuable feedback and suggestions.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

NOMENCLATURE

SVR	support vector regression
TFIDF	term frequency – inverted document frequency
MI	mutual information
Digg	a social news aggregator website. Our case study target.
digg	to vote for an item posted to Digg
SWCP	social web comment prediction
NDCG	normalized discounted cumulative gain

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

1. INTRODUCTION

1. 1 Opportunity

The Social Web is one of the early successes in the emerging social computing paradigm. Prominent Social Web examples include large-scale information sharing communities (e.g., Wikipedia), social media sites (e.g., YouTube), and web-based social networks (e.g., Facebook). Beyond these popular successes, the emergence of Web-based systems incorporating social computing features is promising to fundamentally transform what information we encounter and digest, how businesses market and engage with their customers, how universities educate and train a new generation of researchers, how healthcare and medical advances are managed and disseminated, how the government investigates terror networks, and even how political regimes interact with their citizenry (e.g., the use of Twitter and Facebook in the recent Iranian election controversy). One of the key features driving the growth and success of the Social Web is large-scale user participation in content annotation via user-contributed tags, comments, and ratings. Tagging, rating, and commenting functionalities have enabled people around the world to interact rapidly via mass collaboration in the exploration and discovery of community-based information. Unlike traditional metadata annotation by a handful of domain experts, this **socially-generated metadata** builds on the crowd intelligence of the Social Web,

---

This thesis follows the style of IEEE Transactions on Services Computing.

enabling new community-based information access, organization, and retrieval. Indeed, several recent research efforts have begun steps in this direction (e.g., [1], [2]) to leverage the mass amount of socially-generated metadata.

1. 2 Challenge

While tags and ratings provide succinct metadata about Social Web content (e.g., a tag is often a single keyword), **user-contributed comments** offer the promise of a rich source of contextual information about Social Web content but in a potentially "messier" form, considering the wide variability in quality, style, and substance of comments generated by a legion of Social Web participants. To illustrate, Figure 1 and Figure 2 display typical Social Web content (in this case, a video and a photo); the comments associated with these objects are displayed in Figure 3 and Figure 4. While the comments themselves can help other users obtain more information about the Social Web object or the community's view of the object, the un-restricted free-style writing results in a wide variety of comments.

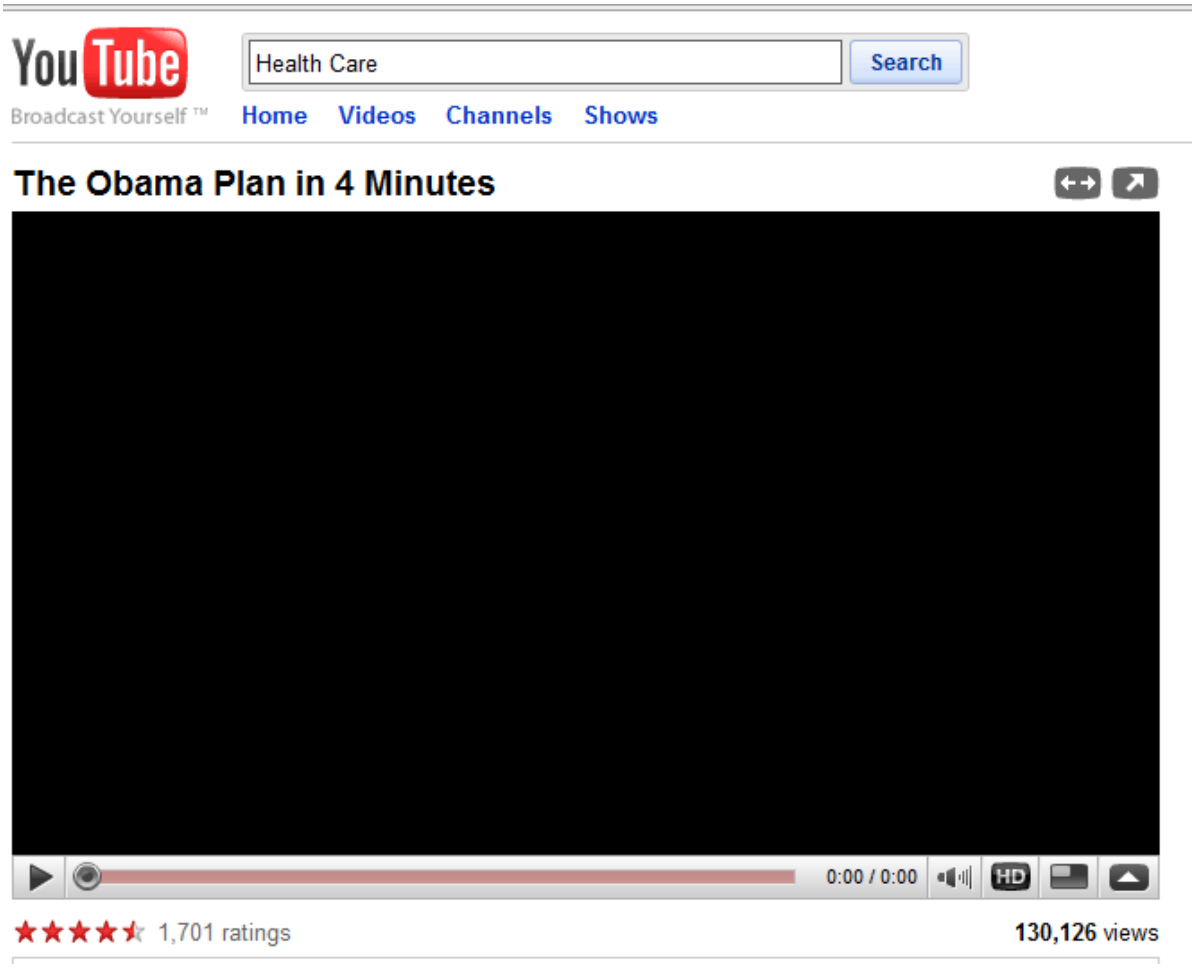

Figure 1 A Youtube video page titled “The Obama Plan in 4 Minutes”

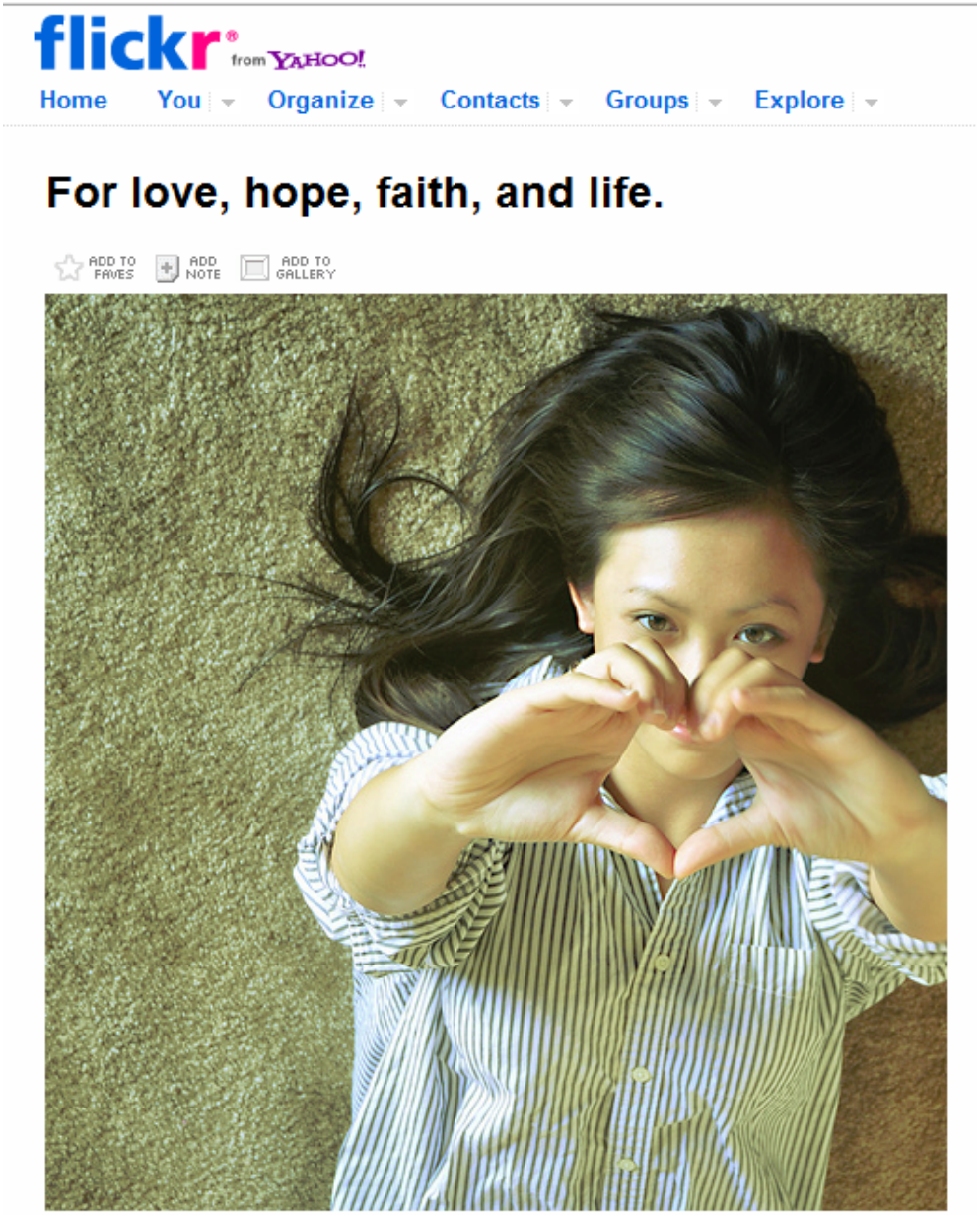

Figure 2 A Flickr photo page titled “For love, hope, faith, and life”

Figure 3 Part of the comments for the Youtube video “The Obama Plan in 4 Minutes”

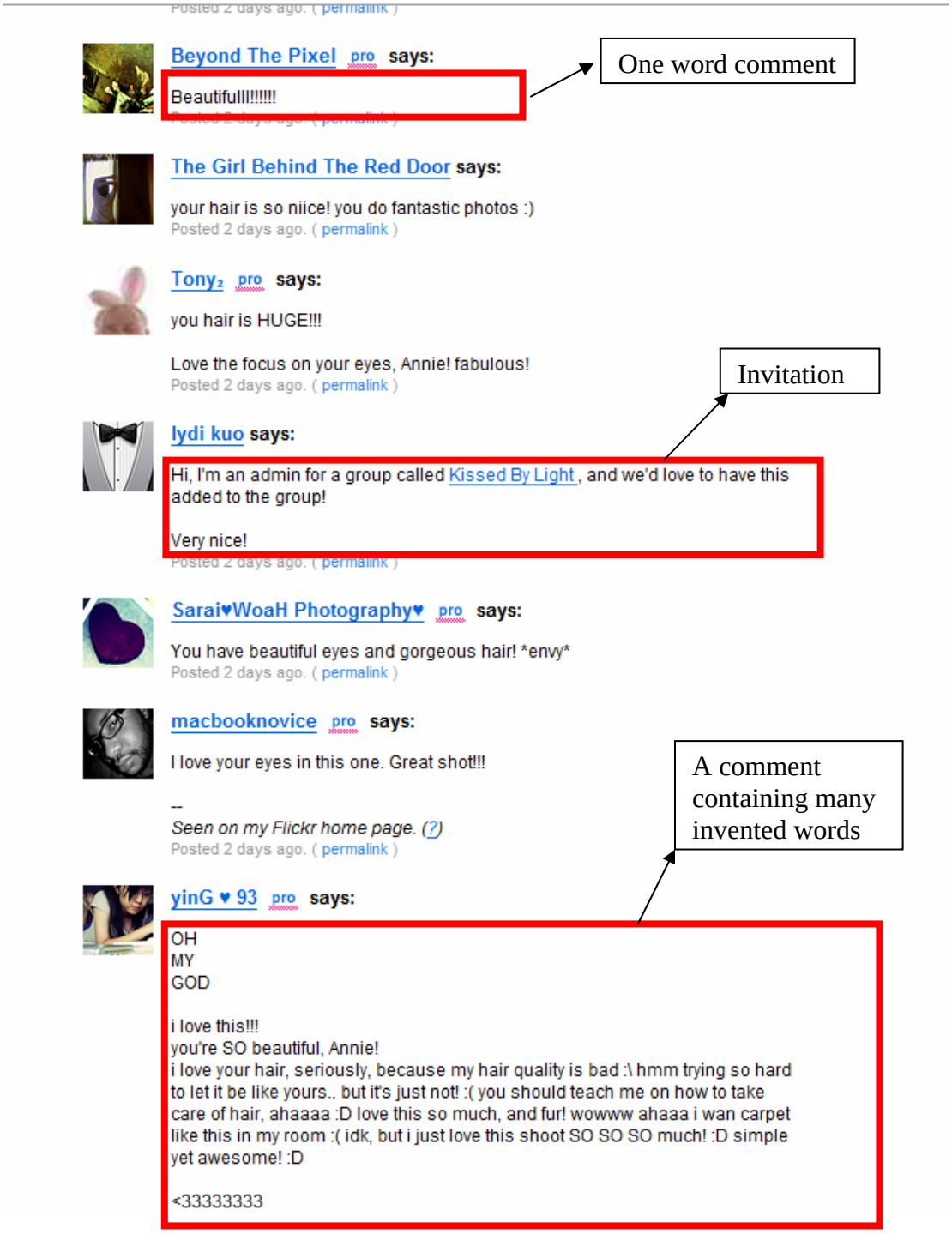

Figure 4 Part of the comments for Flickr photo “For love, hope, faith, and life”

Leveraging the social collective intelligence embedded in these user-contributed comments requires first a study of the factors impacting the quality of these comments, especially in the presence of untrusted users, spammers, and other disruptions to the quality of user-contributed content. *Studying the quality of user-contributed comments* can help ensure the continued growth of the Social Web and mitigate the potentially negative impact of spam and low-quality comments on the sustainability of the Social Web. Compounding the comment quality assessment is the inherently subjective and variable nature of quality. That is, the perceived quality of a user-contributed comment may vary from user to user and from community to community. Dealing with this variation in perceived quality is a difficult and important challenge.

1. 3 Contribution

This thesis research studies how a Social Web community itself perceives the quality of user-contributed comments within the community, so that the community is the final arbiter of quality. By studying how a community can self-regulate, we may gain insights into what a community values and how to sustain the positive growth of the community.

In particular, this thesis research makes two contributions:

- The first contribution is a framework for modeling and measuring comment quality on the Social Web. One of the salient features of the framework is a regression-based learning approach for automatically ranking comments based

on the expressed preferences of the community. By learning ranking functions for user-contributed comments, we may (i) automatically score new comments as they arise in the community; (ii) promote high-quality comments; (iii) filter out low-quality comments, so that user attention is not wasted; (iv) provide a sound basis for enhanced comment-based Social Web applications like summarization, content retrieval, visualization, and so on.

- The second contribution is a large-scale experimental study of comment quality on the popular Digg social news aggregator. We study several factors impacting the community's preference for user-contributed comments on Digg, including the contributor's reputation and community activity level, as well as the complexity and richness of the comment. Through experiments, we find that the proposed approach results in significant improvement in ranking quality versus alternative approaches. Additionally, we study an extension to the model for balancing the visibility of a comment with its intrinsic quality.

The rest of this thesis is organized as follows: In Section 2 we discuss related work. Section 3 describes the general Comment Preference Framework that can be applied to any social website with commenting functionality. In Section 4, a large scale experimental study on Digg is presented. Section 5 shows the empirical results of our evaluation. Finally, we conclude our work and discuss some possible extensions.

2. RELATED WORK

2. 1 Evaluating Quality of User-Generated Metadata

Recently there have been several research efforts to study the quality of socially-generated metadata, including the quality of user-contributed tags [3], blog comments [4], user-contributed answers on Question-Answering forums [5], product reviews on Amazon [6], and so forth. In many cases, these quality assessments rely on experts external to the Social Web community (e.g., a panel of human experts declares that a blog comment is “spam” or “not-spam”). This thesis approaches the problem of quality from a different angle, by considering the community’s preference as the baseline quality indicator. That is, any user is eligible to express his/her opinion on whether a particular comment is good or bad. The aggregated rating is the indication of the quality of the comment perceived by the community (i.e., the group of web users who care to read and rate the comment). We believe that this approach will result in a more democratic environment that people with different taste are able to enjoy their preferred style of comments.

2. 2 Studies of News Aggregators

This thesis research is inspired by some previous studies of comments in message forums and newsgroups, including [7] and [8]. In particular, the Slashdot community – one of the acknowledged forebears of Digg and related social news aggregators – has attracted much attention. Several researchers have examined Slashdot’s moderation policy for rating and filtering user-contributed comments, including [9] and [10]. Gomez et al. has studied the statistical properties of Slashdot discussion threads to identify

degrees of controversial topics [11]. Other researchers have developed some machine learning approaches for semi-automating or fully automating the moderation of comments on Slashdot, including [12] and [13]. In a separate direction, Lerman has studied Digg and its article rating system in some detail, e.g., [14], [15], [16].

It is important to note that Digg and most new Social Web commenting systems differ from Slashdot in a number of critical dimensions. First, Slashdot offers a restricted form of comment rating (moderation) in which only a fraction of all users are selected to moderate a given comment. This restriction is in direct opposition to the Social Web philosophy, in which all users are eligible to rate a comment. Second, Slashdot's comment rating policy restricts the ratings of a comment from -1 to 5, unlike Digg's comment rating system, which is (potentially) unbounded, allowing for a wide variety of scores to be applied to comments. This purely democratic system could be potentially more problematic for sustaining the growth and quality of the community comment rating system, hence motivating the need for this work.

2. 3 Learning to Rank

The comment preference model developed in this thesis draws on related approaches in the Web community, where "learning-to-rank" approaches for automatically ranking Web search results, e.g., [17], [18], have recently shown great promise. Researchers have typically relied on one of two types of ranking approaches:– point-wise and pair-wise.

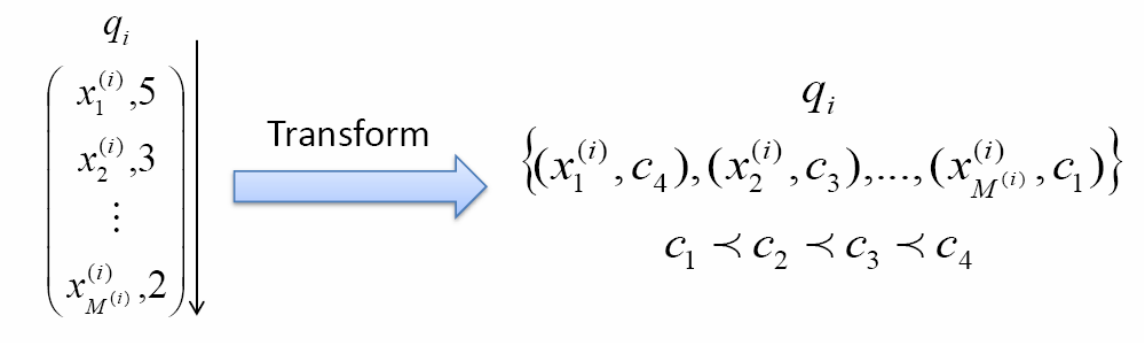

Figure 5 Point-wise approach for preference learning

2. 3. 1 Point-Wise Ranking

The point-wise approach uses ordinal regression to learn the ranking order (say, of a set of Web documents returned in response to a particular query). Figure 5 shows an example of transforming the query result document set $\{x_1, x_2, \ldots x_m\}$ and the corresponding relevance scores into four categories$\{c_1, c_2, c_3, c_4\}$ representing the relative level. This ordinal regression approach can be regarded as smoothing between regression and classification. Recall that classification problems usually assign each object a class label from a set of non-ordered categories; regression problems usually produce continuous values using some function. Ordinal regression is a balance between these approaches for datasets that have not only discrete values but also ordered categories. Several point-wise ranking algorithms (ordinal regression) such as Pranking [19], OAP-BMP [20], Ranking with Large Margin Principle [21], Constraint Ordinal Regression [22] have been proposed. We applied the point-wise approach in our work to rank comments based on community preference.

2. 3. 2 Pair-Wise Ranking

While point-wise ranking has shown promise, there has been some recent work in the Web domain for adopting partial order (pair-wise) preferences into a learning model. Pair-wise ranking considers the pair-wise preferences over documents (e.g., document 1 is preferred to document 2).

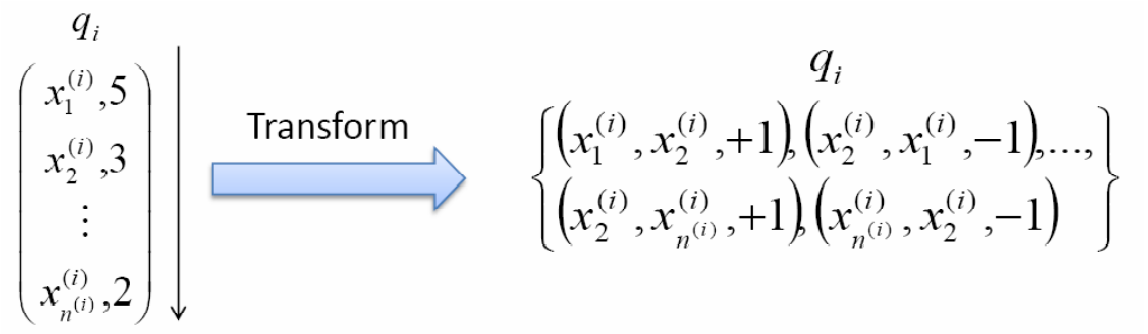

Figure 6 Pair-wise approach for preference learning

Figure 6 shows the transformation from system generated ranking list to document pair preference representation where +1 indicates x_i is preferred over x_j and -1 indicates x_j is preferred over x_i. Several efforts have studied techniques for automatically learning this preference, including Learning to Retrieve Information [23], Learning to Order Things [24], RankNet [25], Frank [26], RankBoost [27] and RankSVM [28][29]. Other algorithms in this category include the linear discriminant model for information retrieval [30], preference learning with gaussian process [31], a regression framework for learning ranking functions using relative relevance judgments [32], a general boosting method and its application to learning ranking functions for web search [33]

and magnitude-preserving ranking algorithms [34]. Advances in the “learning-to-rank” domain could inform future progress on learning to rank comments on the Social Web.

3. COMMENT PREFERENCE FRAMEWORK

In this section, we present the framework for ranking comments on the Social Web by community preference. A preliminary architecture and its problems will be briefly discussed first. We will then show the detailed architecture of our comment preference framework. We present the design of the feature extraction component as well as the regression based machine learning approach for community preference learning.

3. 1 Architecture

The first contribution of this thesis is a framework for assessing comment preference on the Social Web. In particular, we have developed two complementary systems: (i) The initial system uses a comment quality classification approach; (ii) Based on our observations of the strengths and weaknesses of this initial design, we developed and evaluated a more robust comment community preference ranking approach. In the rest of this section, we highlight the initial design before focusing our attention in the rest of the thesis on the full comment community preference ranking framework and experimental evaluation.

3. 1. 1 Initial Framework -- Comment Quality Classification

In the first trial [35], the basis architecture with feature selection component, ground truth target define, and a learning model was proposed. In our first trial, the feature set is composed of 13 features only. We did not do term based feature extraction until our second trial. The prediction framework relies on a classification approach for building a predictive model as shown in

Figure **7**. The two classifiers used in the first trial are Linear regression and Quadratic classifier. The goal is to predict for an unseen comment one of four different labels: Excellent, Good, Fair, and Bad.

Algorithm 1 Process of building a model for comment rate prediction
1: Get Features 2: Preprocess Data 3: Train the classifiers using train data 4: Give labels to test data 5: Evaluate results by classification rate 6: Apply Feature selection

Figure 7 Process of building a model for comment rate prediction

Based on evaluation of this initial comment classification approach, we found that Digg users prefer short, simple, and readable comments, and that so-called power users in the community do not, in fact, wield considerable influence over the scores of comments in the community. However, the Comment Quality Classification framework raised some significant issues with respect to the choice of the quality boundary selection and community diversity concern. An example shown in Figure 8 explains about the Boundary Selection issue. If we have a comment with score 11, why it is true that we say it is a good comment but not a fair one? No matter how we select the boundary to define the quality bucket, it’s never convincible.

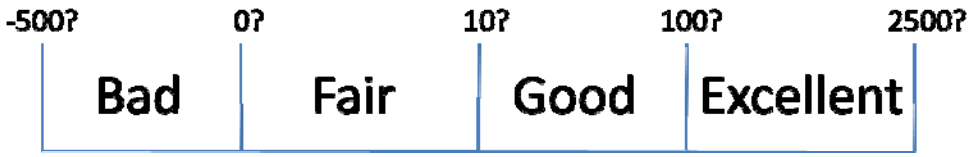

Figure 8 Example for boundary selection issue

Another example shown in Figure 9 is about community diversity issue. Comments in different category usually have different range of score. For example, a 25 rating comments in the category of science is said to be excellent whereas if a comment with the same score resides in the offbeat category would be just “Good”. So, considering these problems, I switch gears to Comment Community Preference Ranking framework.

3. 1. 2 Comment Community Preference Ranking

We thus developed a new advanced approach to cover the disadvantages[36]. As a result the final system can automatically rank the comments associated with a news article on the popular social news aggregator Digg, and potentially many other similar social websites with aggregated comment rating functionality, based on the expressed preferences of the community itself.

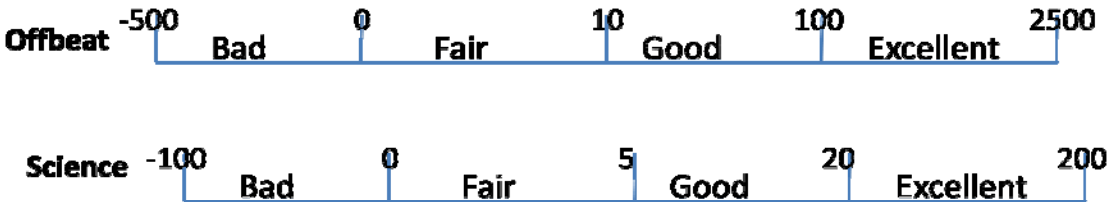

Figure 9 Example for community diversity issue

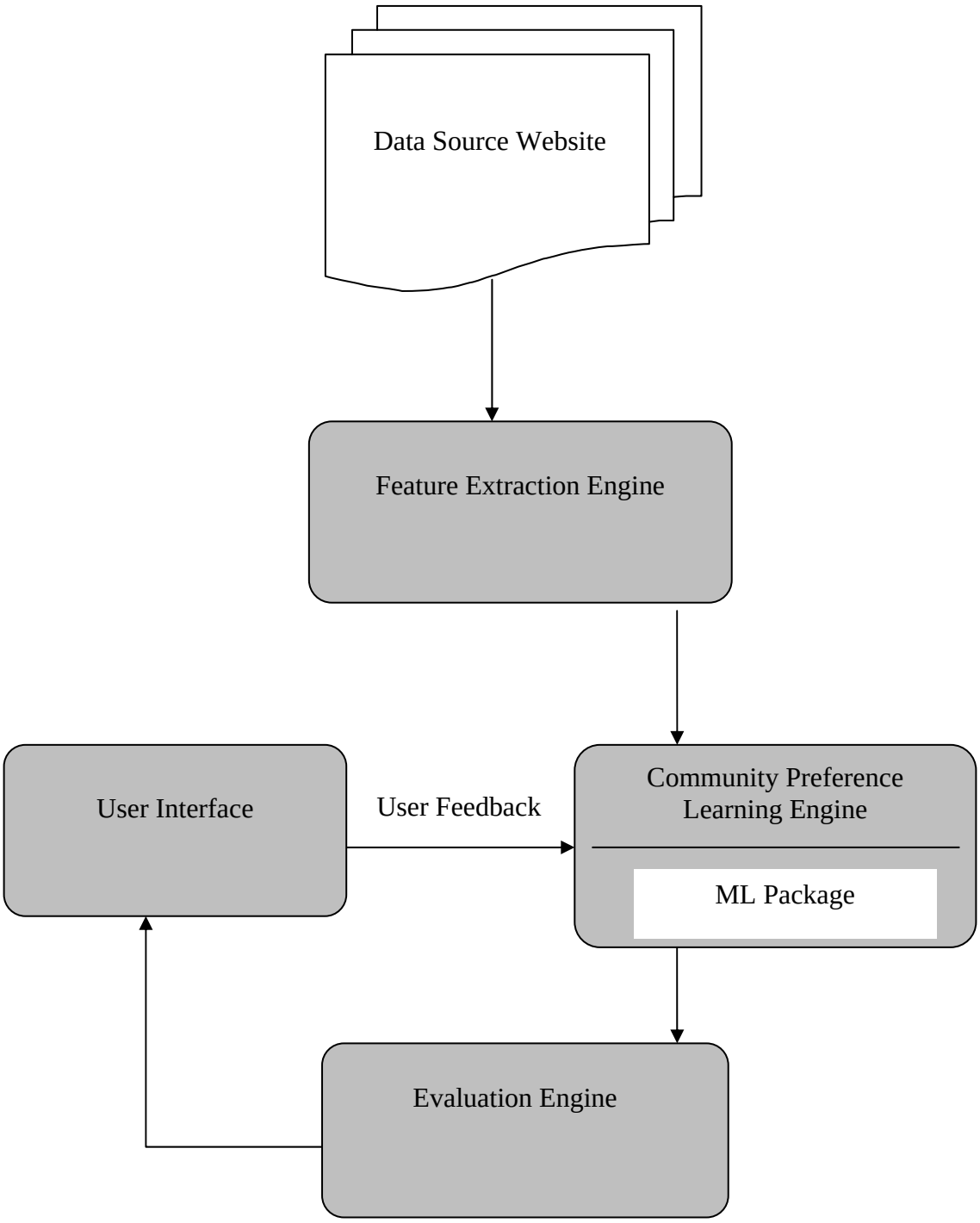

Figure 10 Comment preference framework

The system as shown in Figure 10 includes a feature extracting component, a community preference learning model and a performance evaluation interface. The gray components are of those we developed. We incorporated some existing resources such as the machine learning package as shown in the white area. This graph illustrates a general framework of the Social Web Metadata Community Preference Predictor. First of all, the original dataset is collected from the source website. Then, the feature extraction engine will go and parse any useful information that can be used to describe our target metadata. This is a domain dependant component that needs to be customized when we switch our target domain. For this thesis research, we have developed a feature extraction engine specifically for Digg comments. A diagram shown in Figure 11 illustrates the workflow in the Feature Extraction Engine. Generally, the Feature Extraction Engine is composed with two elements. Besides those properties that can be parsed from the metadata and objects that the metadata attached to, a dictionary generation engine is constructed for term related property extraction.

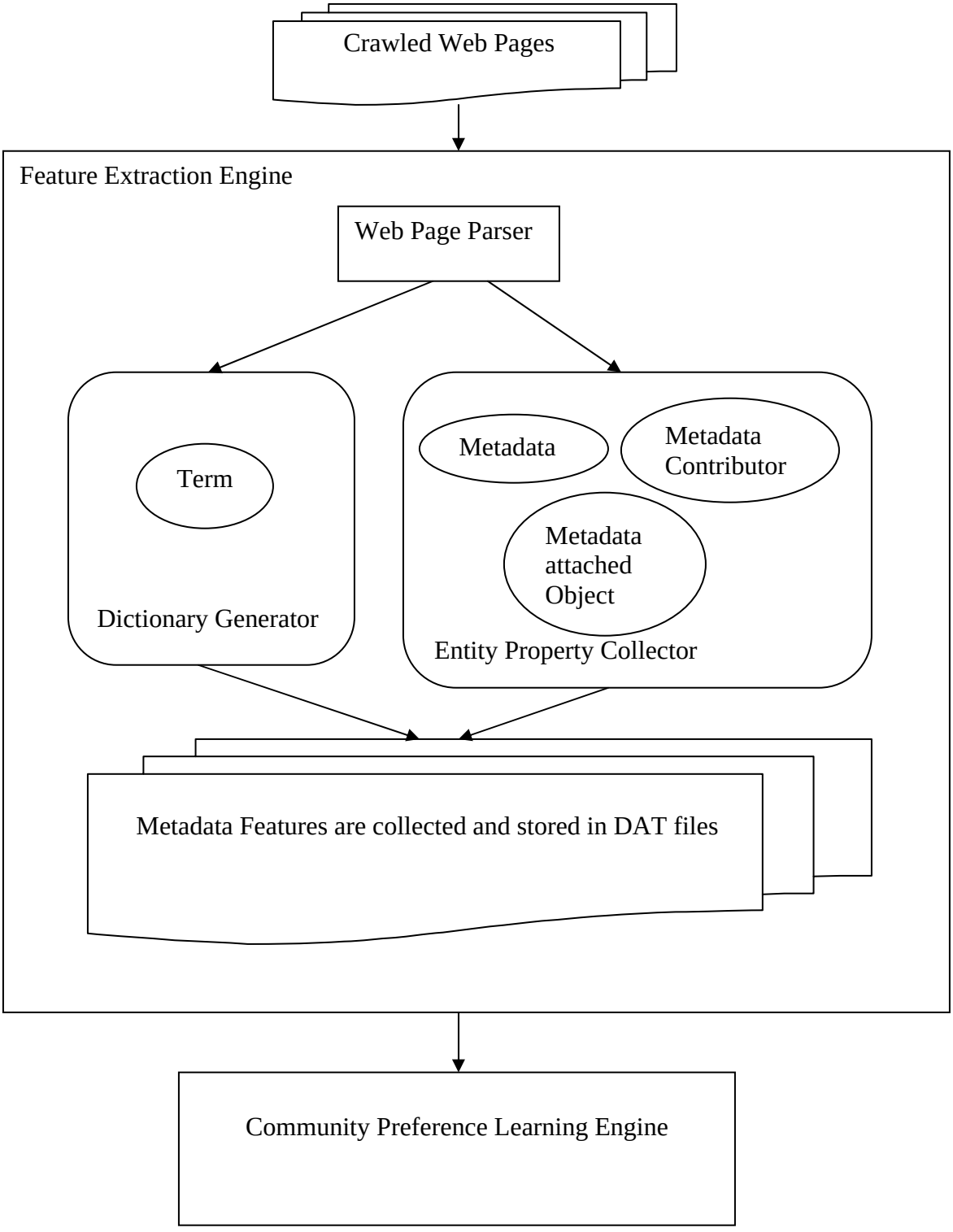

Figure 11 Feature extraction workflow

Later on, the feature set will serve as the input to the Learning Engine as shown in Figure 12. In the Community Preference Learning Engine, we first select the features to use and the target community preference that we want to learn. We then apply some effective machine learning package to learn the model and apply the learned model on the test dataset. The score obtained from the Virtual Score Prediction component is used to find the community preference order of a group of metadata that is within the object that they attached to. That is to say, the objective of the learning system is to find the ranking model that can best describe the community preference view of the metadata within each object.

3. 2 Metadata Attribute Extraction

Note that our study subject, comment, is itself a metadata to some other object such as an article, a video clip or a photo on social web. Based on this fact, we can find many relationships between the metadata and the environment it attaches to or resides in.

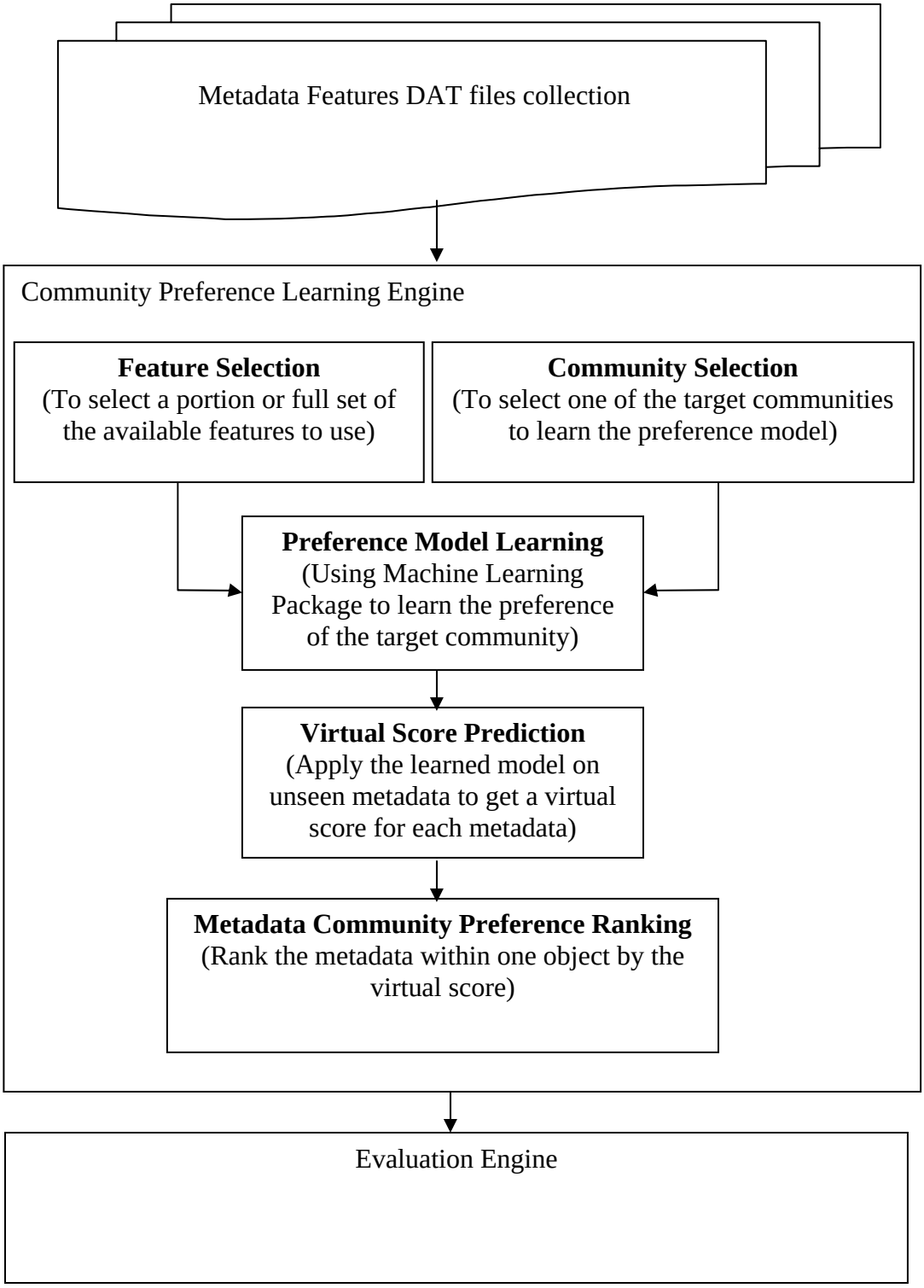

Figure 12 Community preference learning workflow

Although deciding on what are the attributes we can extract to describe a comment is domain dependent, we can generally provide three categories of attribute sources:

1. the metadata content, in our case, the comment itself.

2. The metadata creator, in our case, the commenter.

3. Attributes related to the object that metadata attaches to, in our case, the story and the story category.

3. 3 Learning Community Preference

This section presents the formal model for ranking comments on the Social Web by community preference. We approach the problem of ranking comments as a regression problem.

Consider a set of k Social Web objects (e.g., Web documents, images, videos) $O = \{o_1, o_2, \dots, o_k\}$. Each object o_i has a set of up to n comments associated with it $C_i = \{c_{i1}, c_{i2}, \dots, c_{in}\}$. Each comment c_{ij} has a set of m features $F_{c_{ij}} = \{f_1, f_2, \dots, f_m\}$. Each feature refers to some quality measure with respect to the comment. We assume there exists some training data that have the form:

$$\{(F_{c1,1}, r_{c1,1}) \dots (F_{c1,n}, r_{c1,n}), (F_{c2,1}, r_{c2,1}) \dots (F_{c2,n}, r_{c2,n}), \dots,$$

$$(F_{ck,1}, r_{ck,1}) \dots (F_{ck,n}, r_{ck,n})\} \subset F \times R$$

where the pair $(F_{ci,j}, r_{ci,j})$ corresponds to the feature set for comment c_{ij} and the comment community rating $r_{c_{ij}}$ for comment c_{ij}.

To tackle the community preference-based ranking problem, we can train a regression model over this training data. Concretely, we build the model through

(i) a selection of features, as we will discuss in the following section

(ii) the application of Support Vector Regression [37], a state-of- the-art regression model similar-in-spirit to the popular Support Vector Machine classifier that has proven successful across many domains, e.g., [38]. We have applied a free available online package called LibSVM [39] in our research. The theoretical explanation of SVR is as below.

Support Vector Regression is a state-of- the-art regression model similar-in-spirit to the popular Support Vector Machine classifier. Support Vector Regression uses an ε-insensitive loss function that defines a tube with radius ε around the hypothetical regression function. If the data is placed within this tube, the loss function can be regarded as 0. By introducing the positive slack variables ξ_i and ξ_i^*, the SVR regression can be formulated as the constrained optimization problem:

$$Minimize\ \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i + \xi_i^*$$

$$Subject\ to\ \begin{cases} r_i - w^T\phi\left(F_{c_{ij}}\right) - b \le \varepsilon + \xi_i \\ W^T\phi\left(F_{c_{ij}}\right) + b - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, l, \varepsilon > 0 \end{cases}$$

where $\phi\left(F_{c_{ij}}\right)$ is the feature mapping for each comment in the high dimensional feature space, *w* and *b* are the slope and offset of the regression line, and C > 0, called the

regularization parameter, is a positive constant. The positive slack variables ξ_i and ξ_i^* are to measure the deviation of training samples outside the tube ε zone. The constrained optimization problem given by the equation can be reformulated into a dual problem formalism by introducing Lagrange multipliers. Based on the Karush-Kuhn-Tucker conditions, the function is given by:

$$f(F_c) = \sum_{g=1}^{k*n} \sum_{h=1}^{k*n} (\alpha_g - \alpha_g^*) K\left(F_{c_g}, F_{c_h}\right) + b$$

where α_g, α_g^* are the Lagrange multipliers corresponding to the training data. Note that for those comments that serve as support vectors, the $\alpha_g > 0$ and $\alpha_g^* > 0$ whereas all the other comments must have $\alpha_g = 0$, $\alpha_g^* = 0$. $K\left(F_{c_g}, F_{c_h}\right) = \phi\left(F_{c_g}\right)\phi\left(F_{c_h}\right)$ denotes the kernel function, which satisfies the Mercers conditions. The kernel function we used in this work is the radial basis function:exp ($\gamma * \left|F_{c_g} - F_{c_h}\right|^2$).

In the testing phase we use this model to predict a rating for the unseen comments associated with an object $S = \{s_1, s_2, \ldots, s_n\}$ (e.g., *S = {30,100,40}*). Based on these ratings we can determine the relative rank order of the unseen comments: $R = \{r_1, r_2, \ldots, r_n\}$ (e.g., *R = {3, 1, 2}*). Note that our goal here is not to precisely estimate the actual comment community rating for a comment. Since comments may be continually rated, a predicted rating may quickly become stale. Instead, our goal is to predict the relative order of comments, so that even as new ratings are made on the comments, the model will be able to capture the relative quality.

4. EXPERIMENTAL STUDY: DIGG SOCIAL NEWS AGGREGATOR

The second contribution of this thesis is a large-scale experimental study over the popular social news aggregator Digg and the socially-generated comments that Digg users can annotate news articles with. We study several factors impacting the community's preference for user contributed comments, including the contributor's reputation and community activity level, as well as the complexity and richness of the comment. Through experiments, we find that the proposed approach results in significant improvement in ranking quality versus alternative approaches. Additionally, we study an extension to the model for balancing the visibility of a comment with its intrinsic quality.

4. 1 Background on Digg

We begin with an introduction to Digg. Digg is a prominent Web 2.0 news aggregation service in which users can submit stories to the community, rate stories that have been submitted by others (to "digg" a story is to cast a positive vote for it) and comment on stories.

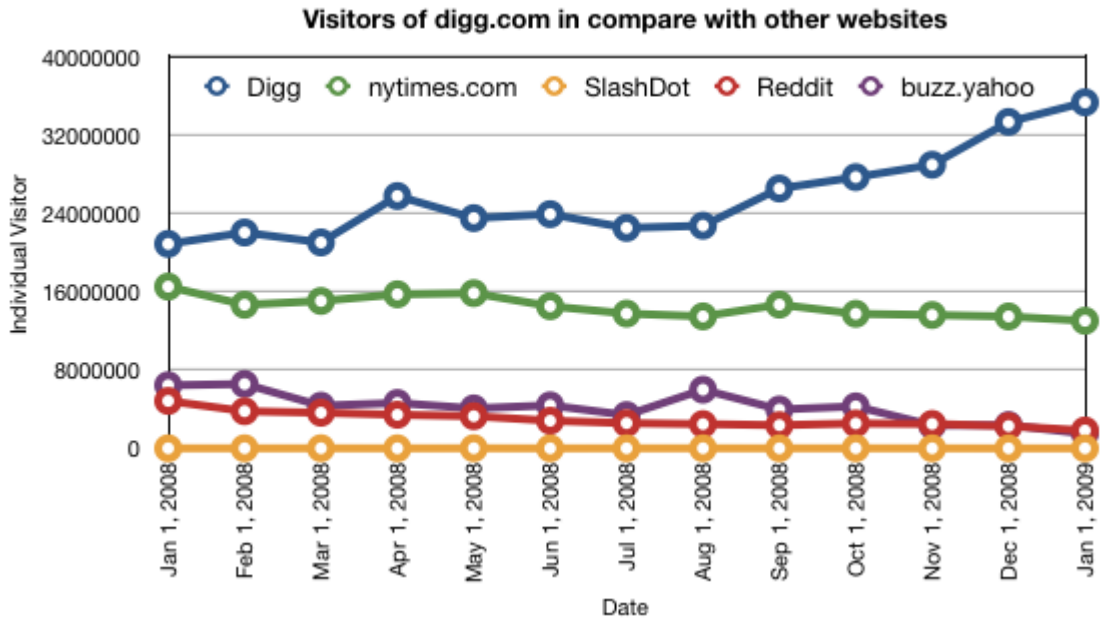

Figure 13 Comparison between Digg and other social news aggregator websites

As shown in Figure 13 with more than 27 million visitors in the past year (according to statistics from Compete.com), Digg is one of the most successful social news aggregators among its rivals such as Reddit, NYtimes, mixx5, Buzz!Yahoo, and Slashdot. Figure 14 illustrates an example submission to the Digg community. Socially generated metadata (i.e., comments), which are included in every news articles submitted and posted on Digg, play a significant part of the Digg community. Our interest in this thesis is to study this metadata in particular. Each comment may be rated by members of the community using a simple thumbs-up or thumbs-down rating system. The system aggregates all ratings applied to a comment so that users can filter comments by rating. Comments on Digg range in style and perceived quality within the community; some examples include the publicly interesting and highly-rated, to the poorly received.

Figure 15 shows the first 4 comments submitted for the story “Google’s April Fools Joke for 2009”. The first comment had the chance to show itself to many users and hence received a large number of 338 diggs in this case. However the second one was not liked by the Digg community and has received a total score of -14. Comparison between the comments that are sorted by time (Figure 15 and **Figure 16**) and the comments sorted by Digg score (**Figure 17**) reveals that earlier comments have a greater chance to be dugg by the community and there is a strong overlap between position and the number of received thumbs up. Late arriving comments typically receive fewer votes than earlier arriving comments. As a result, some possibly high-quality comments with valuable content may be overlooked by the community if they arrive relatively late (and hence, attract smaller social attention). Later we will propose an adjustment that lightens the effect of position on the digg score to overcome such a strong bias due to comment arrival time.

Figure 14 Example story on Digg

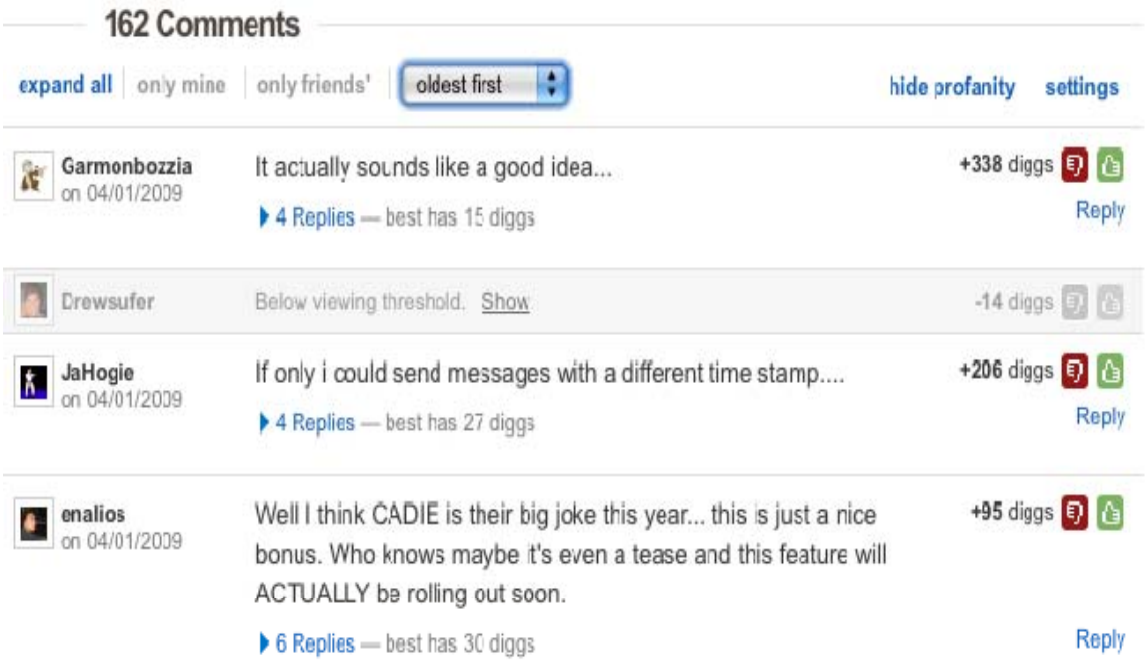

Figure 15 Comments sorted by time (oldest first)

Figure 16 Comments sorted by time (newest first)

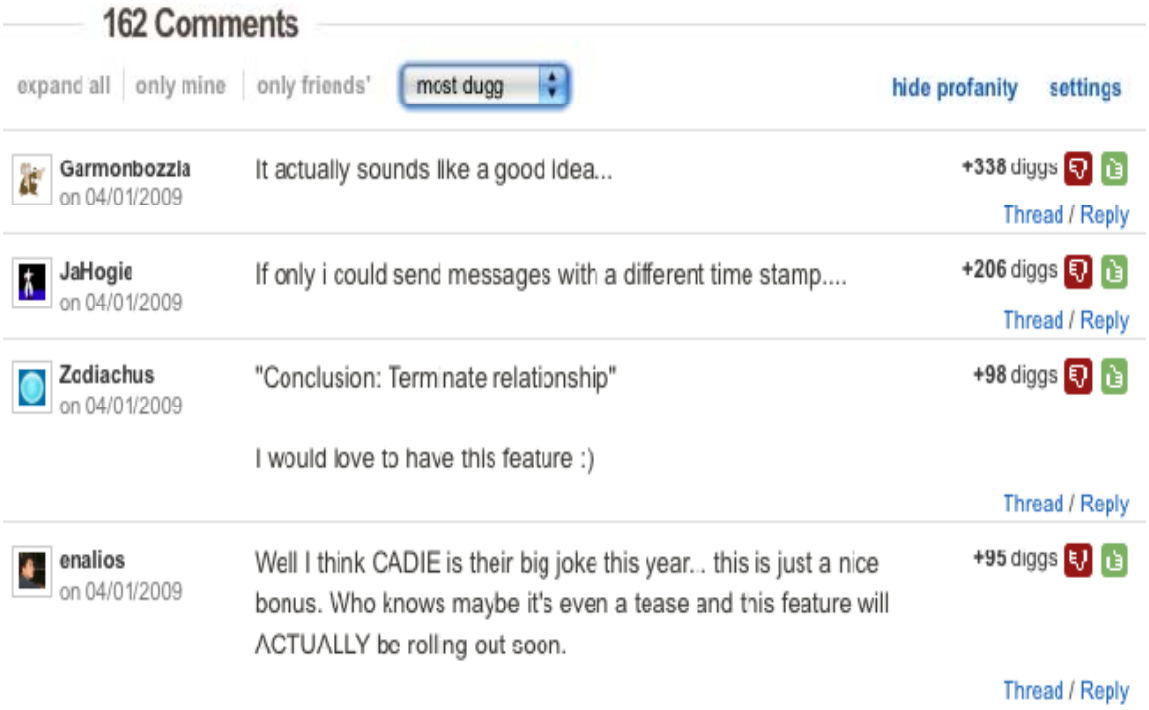

Figure 17 Comments sorted by Digg score

Users in Digg can vote (thumb-up/down) on the comments they read only once while staying anonymous. They can also comment in each story. There are eight top-level categories in Digg – **Technology**, **World & Business**, **Science**, **Gaming**, **Lifestyle**, **Entertainment**, **Sports** and **Offbeat**. Each top-level category contains several sub-categories called topics. In this thesis research, the top-level categories are used to define eight social communities. There are some additional Digg user activity studies shown in Appendix C.

4. 2 Data Source

For our dataset, we crawled the most-Dugg stories of the past 365 days in November 2008, resulting in a corpus of 9,000 Digg stories containing 247,004 comments

submitted by 47,084 unique contributors. We focused our collection on these older pages since the commenting and rating activity has most likely stabilized for these stories, leading to a more reliable analysis of the comments.

4. 3 Distribution of Digg Score

Figure 18 shows the distribution of aggregated digg score for all of the comments in our dataset. Note that the majority of comments receive an aggregate positive score, though with some outliers at both the extreme negative and positive ends. The maximum comment score is 2357, the minimum is -861, and the mean of comment score is 2.

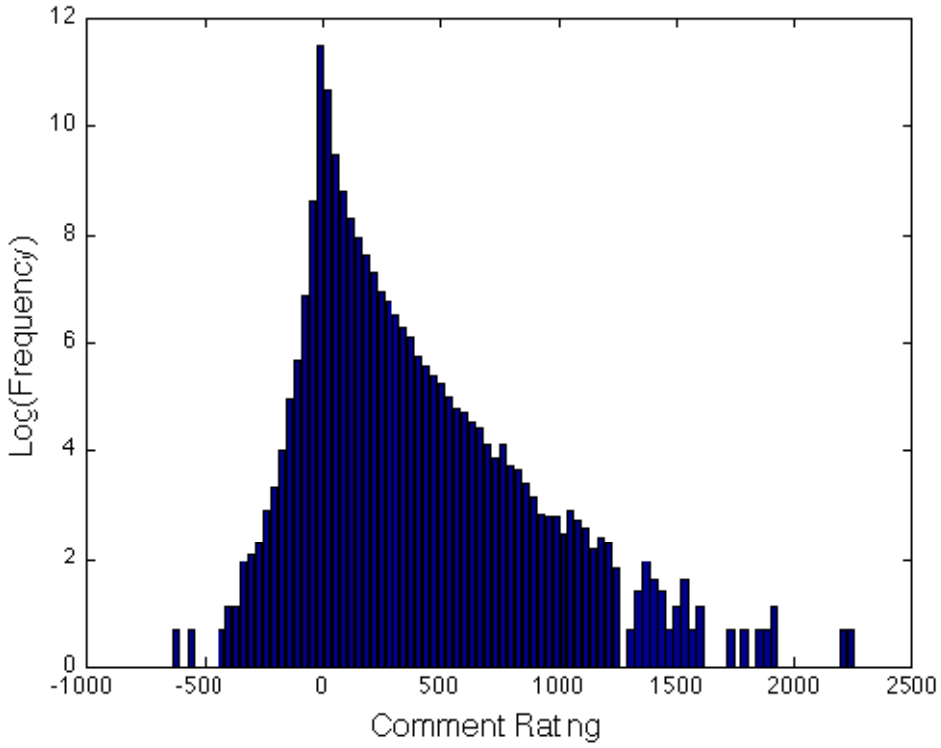

Figure 18 Distribution of Digg comment score (log)

4. 4 Comment Representation

In this section, we discuss the choice of features to represent the comments. The quality of a ranking model is strongly influenced by the quality of the features used to model the domain. In this case, we study several factors that we hypothesize may influence comment community ratings – the visibility of the comment, the influence and reputation of the user contributing the comment, and the content of the comment itself. They involve different kinds of techniques such as the natural language processing [12], parsing, existing metric (such as SMOG [41]) applying and statistical analysis.

4. 4. 1 Comment Visibility

The first factor we consider is comment visibility within the community. Intuitively, if more users in the community view a comment, it is more likely to attract a larger community rating. On the other hand, a comment that is either related to a story that is of little interest to the community or it is placed at the very late position will have less capacity to attract a large community rating.

We measure the visibility of a comment through two factors:

(i) the article community rating of the article that the comment is attached to

(ii) the comment posting time, since earlier comments may have the capacity to be viewed by more community members than later arriving comments.

Figure 19 shows that the number of diggs for each article is highly correlated with the comment digg (The correlation coefficient is 0.93). It shows that comments made to a popular article under a popular category potentially retain higher digg scores. Figure 20 shows that the mean score of comments that arrive earlier is greater than the mean score of comments arriving later, though with greater variability for early comments. In the figure, comments are arranged in order of their posting time (e.g, 1st, 2nd, ...). An early comment has greater visibility, and hence, greater capacity for a high community rating.

Recall that our overall goal is to automatically find the relative rankings of the comments associated with an article, even in cases when the community has not yet made its aggregate community preferences known. Hence, the first visibility feature (article community rating) will not necessarily be available for our prediction framework. As a result, we train the regression models with the article community rating feature to control for the article visibility bias across articles.

For the testing phase we ignore the article community rating since it may not be known in practice and since all comments for an article would share the same feature value. The second visibility feature – comment posting time – is known in the testing phase, and so we can use it as a prediction feature. Of course, it may be reasonable to try to control for comment posting time in much the same way we have controlled for the overall article visibility – so that potentially high-quality comments that happen to arrive late (and hence, may receive a low score due to low visibility within the community) are boosted to a higher position.

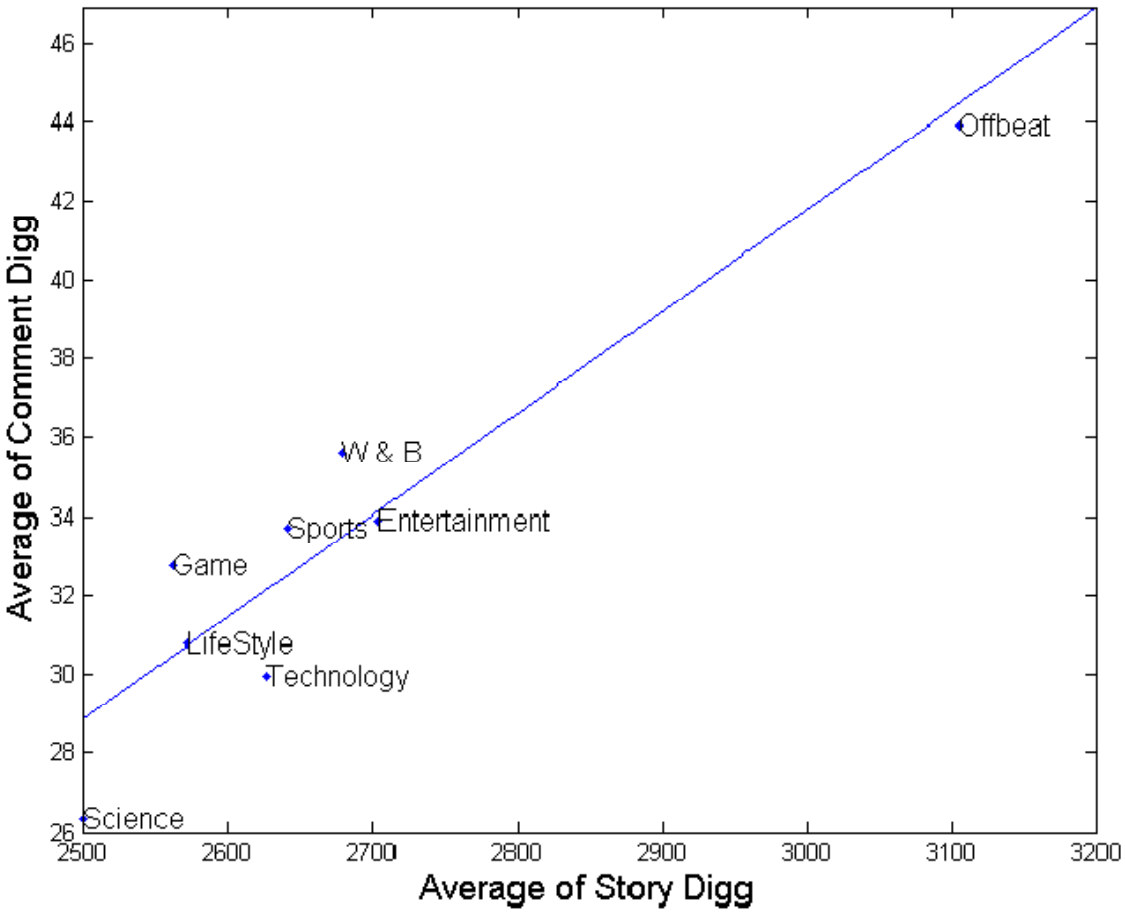

Figure 19 Story Digg vs. comment Digg

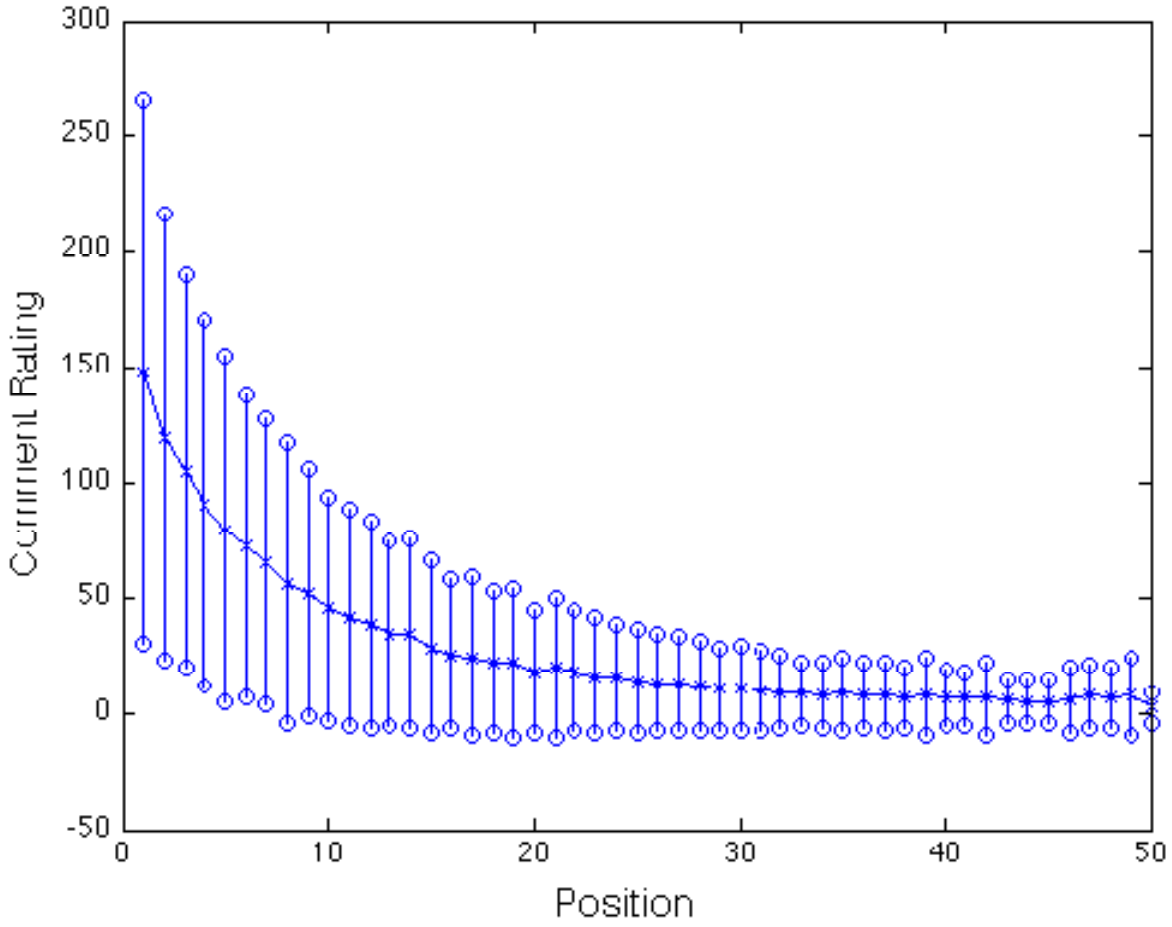

Figure 20 Comment posting time (by position) versus comment community rating. We report the mean comment rating +/- one standard deviation

4. 4. 2 User Reputation and Influence

The second factor we study is the reputation and influence of the user contributing the comment. We want to know if a power user's comments will be more interesting and valuable to the Digg community. Here are some per-user features.

The first set of user-based features gives insight into each user's activity and interest level within the community:

- Number of articles submitted: This measures a user's activity in the community by the number of articles the user has submitted to the Digg community.
- Community membership date: This feature indicates how long the user has belonged to the community and the digg world experience of the commenter. For smoothing purposes, the account starting date (yyyymmdd) of each user is normalized into the range of 0 to 1, with higher values indicating newer members.
- Category activity: We calculate the percent of that user's article ratings to articles from each of the eight top-level Digg categories (e.g., Sports, Technology). For a comment from this user on a particular article, we take the user's category activity percentage for the article's category. The assumption is that for users who comment in an area of their expertise, their comments may have a higher likelihood of being appreciated by the community.

The second set of user-based features measures user popularity in the community:

- Number of articles appearing on the Digg front page: Digg uses a proprietary promotion algorithm to determine which stories submitted by its users reach the front page of Digg (and hence, reach the largest audience). A user who has had success submitting stories that reach the front page is an influential member.
- Number of profile views: How many times has the commenter's Digg profile been viewed? Is this a popular person on Digg?

- Number of friends: The number of friends of the commenter is recorded. Users with many friends may be more appreciated as commenters. This feature tells us how important the social impact is on the comment preference expression.

The final set of user-based features considers how well each user has participated in commenting in the past:

- History of received comment ratings: This feature measures the aggregate (sum) rating of a user's past comments. Does this user tend to make highly-rated comments? Or low-rated comments?
- History of received comment replies: This feature measures the number of replies that the commenter has received to past comments and can be viewed as a reflection of the interestingness of his comments.

In Figure 21, and Figure 22, we report the relationship between three of these user-based features and the comment score. Note that when the number of submitted posts and front page posts by a commenter increases, no increase can be observed for the Digg score for the comment. Similar relationships hold for the other user-based features. Based on these observations in our Digg dataset, it would seem that being an active and influential member of the Digg community is not a good predictor of comment score.

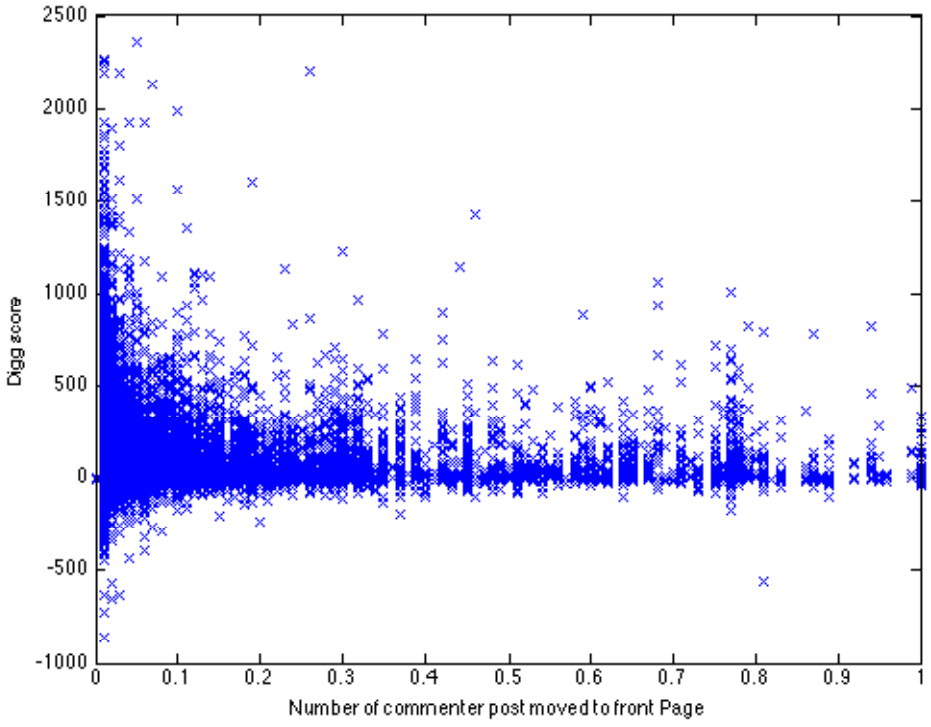

Figure 21 Number of articles appearing on the Digg front page versus comment score

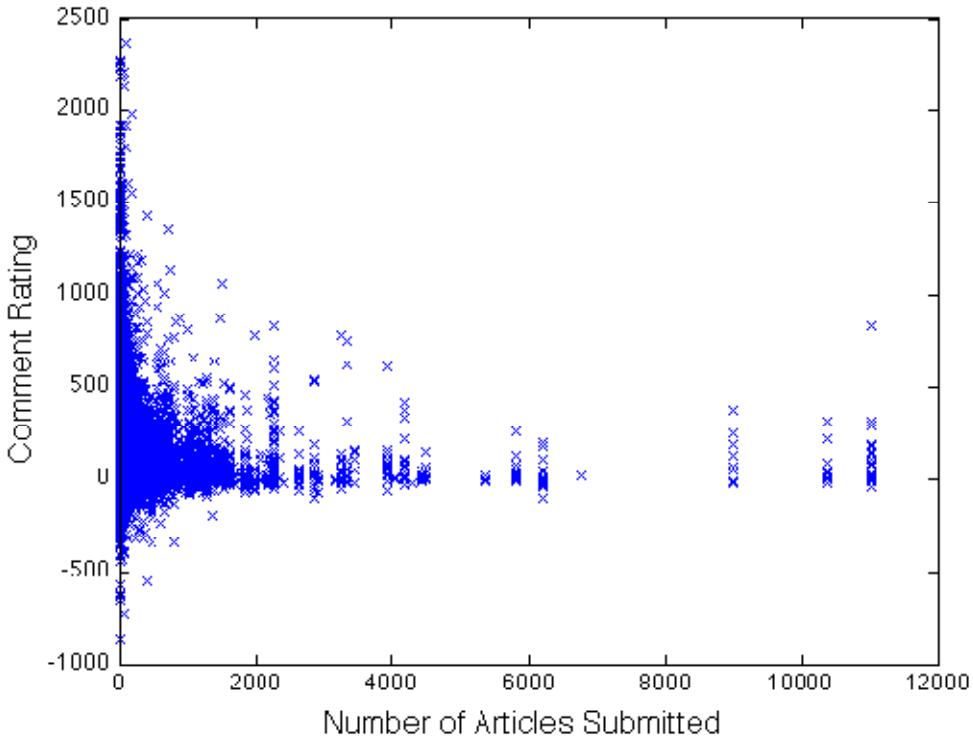

Figure 22 Number of articles submitted versus comment score

4. 4. 3 Content Based Features

The third factors we study are features related to the content of the comment itself. Since Digg and other Social Web sites attract comments from users with a wide-range of educational backgrounds, ages, and interests, the comments these users contribute may vary greatly in word choice, grammar, use of novel phrases, and so on. To capture the impact of these content-based attributes, we consider several semantic and statistical features of the comment text.

The first set of content-based features reflect some statistical properties of the text:

- Comment length: The first feature measures the number of words in the comment text. There may be a tradeoff between longer comments compared with the community's time and effort spent to appreciate the comment. We hypothesize that the Digg community values average-length comments rather than extremely short or extremely long comments. Although a long comment may be more informative, the community may not appreciate the effort to read and understand it.
- Comment complexity: We measure the complexity of a comment by the entropy of the words in the comment. The entropy of a comment reflects the richness of the comment by measuring the variety of words in the text. In our experiments we found that comments with less complexity get higher Digg scores. The equation below shows that for a comment c_j with number of words what is the entropy of c_j when each word has frequency p_i.

$$entropy(cj) = \frac{1}{\lambda}\sum_{i=1}^{n} p_i[log_{10}(\lambda) - log_{10}(p_i)]$$

- Number of upper case words: This is a simple count of upper case words.
- Comment informativeness: Informativeness is meant to capture the uniqueness of the content in a comment relative to other comments attached to the same Social Web object. We measure the informativeness of comment *cj* using a variation of the standard TFIDF approach from information retrieval, where we sum over the TFIDF values for all terms in a single comment:

$$inform(c_j) = \sum_{t_i \in c_j} tf_{i,j} \times idf_i$$

  The *tf* component values terms that occur frequently within a comment: $tf_{i,j} = n_{i,j}/(\sum_k n_{k,j})$ where $n_{i,j}$ is the number of occurrences of the considered term in comment c_j, and the denominator is the sum of number of occurrences of all terms in comment c_j. The *idf* component values terms that occur infrequently across comments $idf_i = \log(|C|/(|\{c: t_i \in c\}| + 1))$ where $|C|$ is the number of comments and $|\{c: t_i \in c\}|$ is the number of comments in which t_i appears.
- Category cohesion: This feature measures the commenter's word choice with respect to the other comments within a particular category. The hypothesis is that each category has its own sub-community that uses particular jargon. Commenter who use those terms more frequently indicates the relatively long term involvement or better familiarity with the particular community. Hence, comments that have high cohesion with the rest of the category are more likely to

receive high ratings. We measure category cohesion using the sum of the Mutual Information (MI) between all terms in the comment and the category (cat) of the article:

$$cohesion(c_j; cat) = \sum_{t \in c_j} MI(t, cat)$$

MI measures the amount of information each term t tells us about category cat: $MI(t, cat) = p'(t|cat)p(cat)\log(p(t|cat)/p(t))$. $p(t|cat)$ is the probability that term *t* appears in comments in cat. $p'(t|cat)$ is a correction to $p(t|cat)$ that gives every term a non-zero probability of occurrence across all categories. Therefore we have $p'(t|cat) = \alpha p(t|cat) + (1-\alpha)p(t)$ as a smoothed probability that a comment contains term t given that it belongs to category cat is between 0 and 1. In practice we select a smoothing factor of $\alpha = 0.9$. p(t) is the fraction of all comments containing t; and p(cat) is the fraction of comments belonging to category cat. To prevent comments with more terms from receiving higher cohesion values, we also considered a version that divides cohesion by the number of terms in c_j. Experimentally, we find that this normalized version yields qualitatively similar results.

The Comment Informative and Category Cohesion features were made possible only with the availability of the Digg Comment Dictionary. In order to gain more understanding of a comment, we decided to build a dictionary for Digg comments in our dataset. Due to the free style writing format from different types of users, words used in

the comments can be very creative. That is, people invent words, phrases, grammar and expressions to make their post more attractive to other users within the community. Technically, we went through all comments and record the terms occur in each of them. An inverted dictionary from terms to comments that contains the term is also recorded separately. In the inverted dictionary, we also count the collection frequency (*cf*) and document frequency (*df*) of each term. Using the dictionary, we can extract a new feature set specifically for describing terms in each comment.

The next set of content-based features rely on NLP-style analysis of the comments:

- Readability: We measure the readability of a comment by its SMOG score [11], which estimates the years of education needed to understand a piece of writing. SMOG considers the number of poly Syllables and the number of sentences in a text. Based on what we observed, comments with higher readability SMOG scores receive higher ratings.

$$SMOG = \sqrt{polySyllables * 30.0/sentences}$$

- Subjectivity vs. objectivity: Subjective comments refer to unjustified personal opinions, in contrast to knowledge and justified belief. We measure the subjectivity/objectivity of each comment using the open source NLP tool LingPipe [40].
- Verb+Noun count: A simple count of verbs and nouns.

The last set of content-based features compare the comment text to the article the comment is attached to:

- Comment-article overlap: This feature measures the overlap between terms in the article abstract and the comment.
- Comment-article polarity: Finally, we measure if the polarity of each comment (positive or negative) matches the polarity of the article (using LingPipe [12]): 1 for agreement; 0 for disagreement. Our hypothesis is that the community will tend to favor those comments where their polarities match the polarity of the story.

In Figure 23, Figure 24, and Figure 25, we report the relationship between three of these content-based features and the comment score. Note that the comment score is maximum for short comments. In Figure 24 we see that comments with higher readability SMOG scores receive higher ratings and Figure 25 shows that comments with less complexity get higher Digg scores. Of course, the variance in comment scores is much greater for shorter, simple, and more readable comments, so we will need to revisit these features when we construct our comment score predictor in the following section.

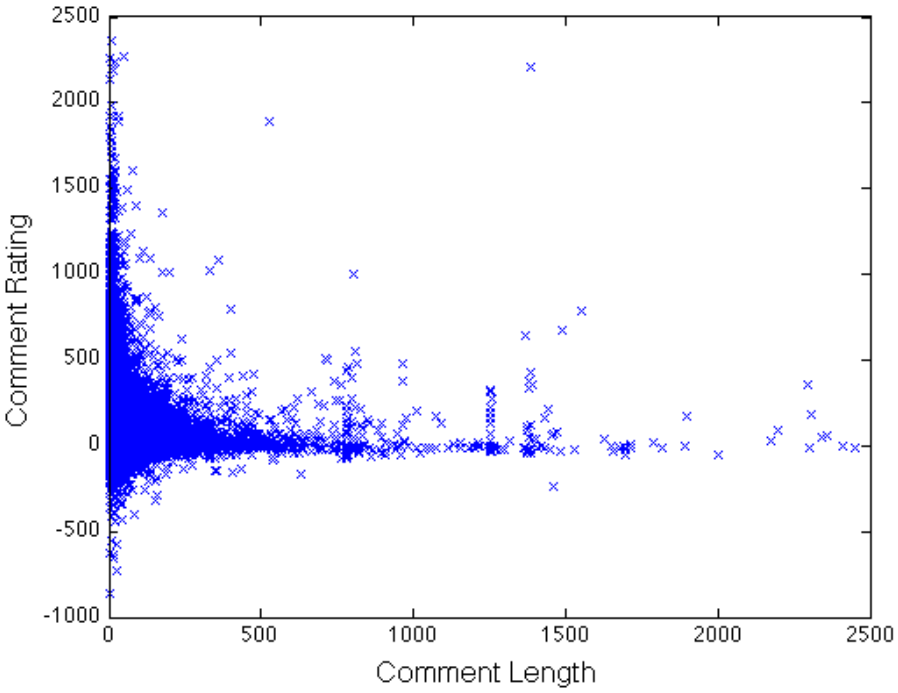

Figure 23 Length of comment versus comment score

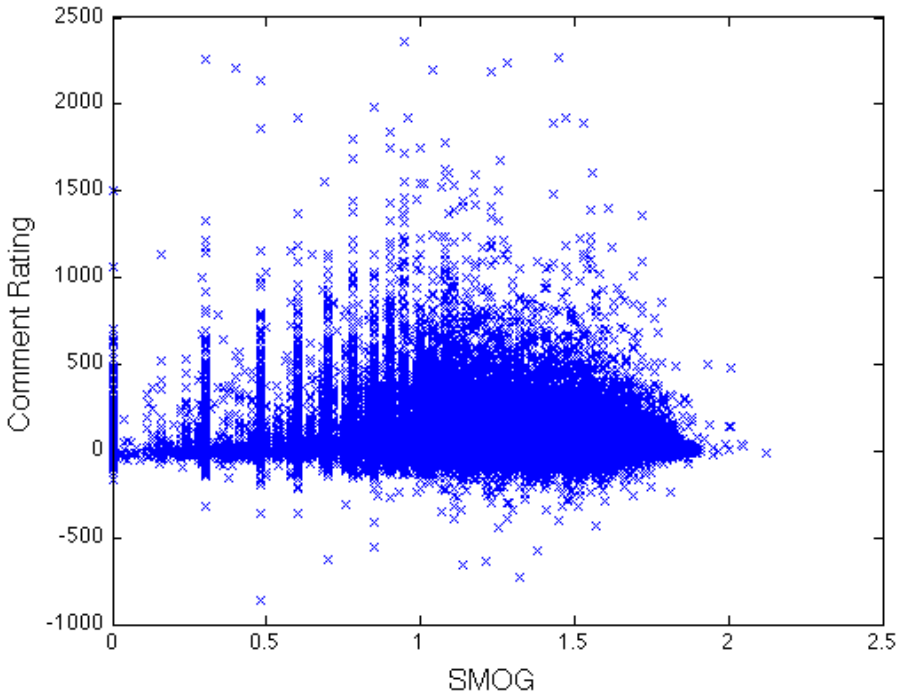

Figure 24 Readability (SMOG) versus comment score

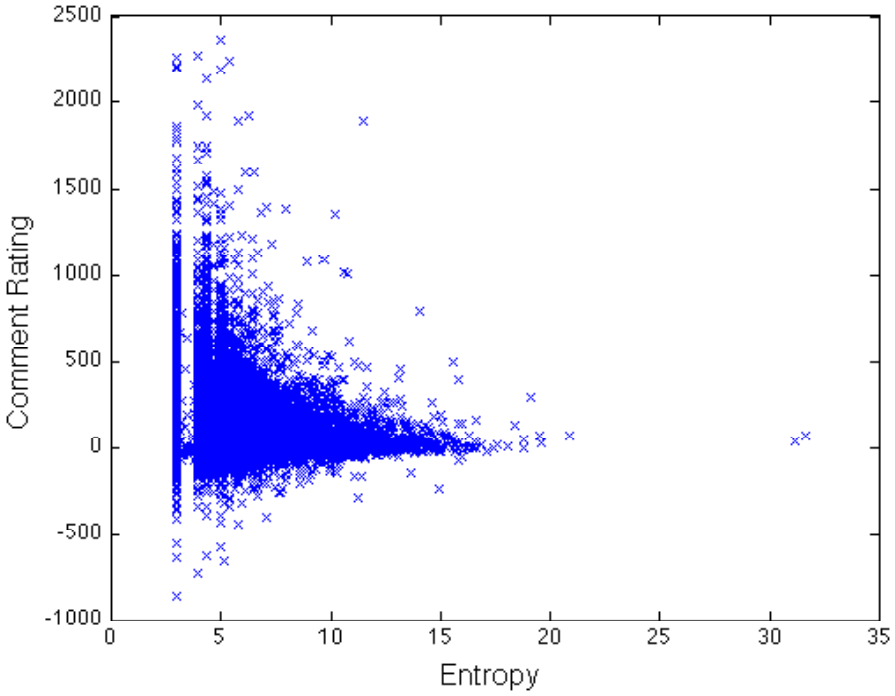

Figure 25 Comment entropy versus comment score

4. 5 Evaluation Metric

As a baseline, we can measure the effectiveness of the learned model by comparing the predicted rank order of the comments to the ground truth rank order, as determined by the ground truth comment community ratings. Since users and applications are typically most interested in these high-quality comments, it is important that the predicted comment rankings be of especially high-quality for the top-k comments for small k. For this reason, we evaluate the quality of the predictions using the well-known Normalized Discounted Cumulative Gain (NDCG) measure for evaluating the quality of top-k lists [42]. The origin and introduction of NDCG can be found in Appendix C. In this work, we had modified the NDCG a little bit to suit our need.

Formally, the discounted cumulative gain (DCG) is found for a top-k list as:

$$DCG_k = \sum_{i=1}^{k} fav_i / \log_2(1+i)$$

where fav_i is a favorability score for the comment at position i. We define the favorability score as its rank complement: $fav_i = N - Rank_i + 1$. For comparison across top-k lists for different articles, DCG is normalized by the ideal discounted cumulative gain at *k*. The ideal DCG (iDCGk) is found by sorting the comments in order of their comment community rating and calculating DCG as above, resulting in $NDCG_k$:

$$NDCG_k = DCG_k / iDCG_k$$

NDCG ranges from 0 to 1, with higher-scores indicating greater agreement between the predicted rank order and the ideal rank order (based on the comment community ratings).

5. EVALUATION ANALYSIS

Our evaluation is designed with three goals in mind. First, we aim to compare the learning-based ranking approach versus alternative approaches, to understand if the model does indeed capture salient features for predicting community preference. Second, we isolate the features used by the model to gain a better understanding of which comment features are good predictors of community preference. Finally, we explore an extension to the model for identifying and promoting high quality comments that may have been overlooked. As a baseline, we can measure the effectiveness of the learned model by comparing the predicted rank order of the comments to the ground truth rank order, as determined by the ground truth comment community ratings. Recall that it is important that the predicted comment rankings be of especially high-quality for the top-k comments for small *k*, since users and applications are typically most interested in these high-quality comments. Errors in ranking prediction at lower ranks are of less importance (e.g., swapping the 200^{th} and the 201^{st} comment).

In all of the experiments reported here, we train and test the model using 10-fold cross validation and a 20-80 train-test split. After randomly sampling 24,000 comments from the dataset, the data is randomly split into 10 parts. We train the model over two of the parts (including the ground truth comment community rating) and then test the model over the remaining eight parts (for which the model has no access to the ground trust comment community rating). This procedure is repeated 10 times; the results are averaged over the 10-folds.

5. 1 Model Comparison

First, we compare the proposed model – denoted here as the Social Web Comment Prediction SWCP model – against two alternatives: a random ranking model and a time-of-posting based ranking model. In the random ranking model, comment order is purely random. This simplistic model provides us with a baseline against which to compare the developed models. The second model is a time-of-posting ranking model. Recall that in Figure 20, we saw how comment posting time has a strong impact on its community rating, since earlier comments have greater visibility in the community. It might be reasonable to conjecture that posting time is all that matters. Concretely, this model assigns rank order to comments based solely on time-of-posting, i.e., comments arriving in the order $\{c_1, c_2, \ldots, c_n\}$ are ranked $\{1, 2, \ldots, n\}$.

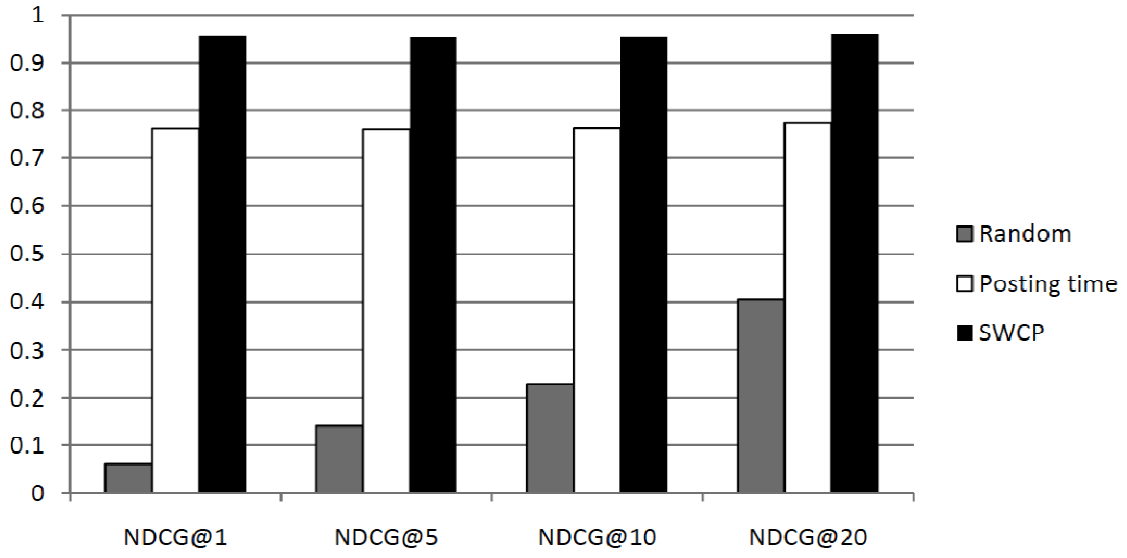

Figure 26 Comparing the SWCP model versus alternatives

Figure 26 shows the performance of the three models across four different NDCG k-values: NDCG@1, NDCG@5, NDCG@10, and NDCG@20. First note that both the comment-posting time model and the SWCP model outperform the random model for all NDCG metrics. Second, although the comment-posting time performs reasonably well, it alone is an insufficient determiner of comment community preference. We see that the inclusion of the user-based and comment-based features results in around a 25% improvement across all NDCG metrics. What is especially encouraging is that the model performs extremely well for the top-1 comment, meaning the model almost always correctly identifies the top- 1 comment regardless of its posting time. The similarly good results for 5, 10, and 20 are also encouraging, validating the premise that comments, although a “messier” form of user based annotation (compared to tags and ratings), do contain implicit quality signals that can be mined and used for automatic comment extraction by community preference. This has strong positive implications for the success of new comment based applications (e.g., enhanced information organization, summarization, content retrieval, and visualization), as well as the continued success of the Social Web in the presence of growing spam and low-quality comments.

5. 2 Feature Study

Given the performance of the Social Web Comment Prediction model, what impact do the user-based and content-based features have on the prediction quality? Since evaluation of all possible feature combinations would be computationally expensive, we isolate the features in groups to better understand which features are good predictors.

First, we train two models – one using only user-based features and one using only content-based features. Figure 27 shows the performance of the user-based model, the content-based model, and the full feature model for NDCG@20. We find qualitatively similar results for other values of NDCG@k (k=1, 5, 10). The user based features alone do a better job than content-based feature alone, however, both approaches perform significantly less well than the full combination of features. We view the user based features as a "prior" on the preference of the community for the user's comments. Only in combination with the actual comment text can we predict the community preference with good success. This negates the hypothesis that power users wield excessive control over comments (unlike the article promotion feature of Digg, which many presume is heavily influenced by power users).

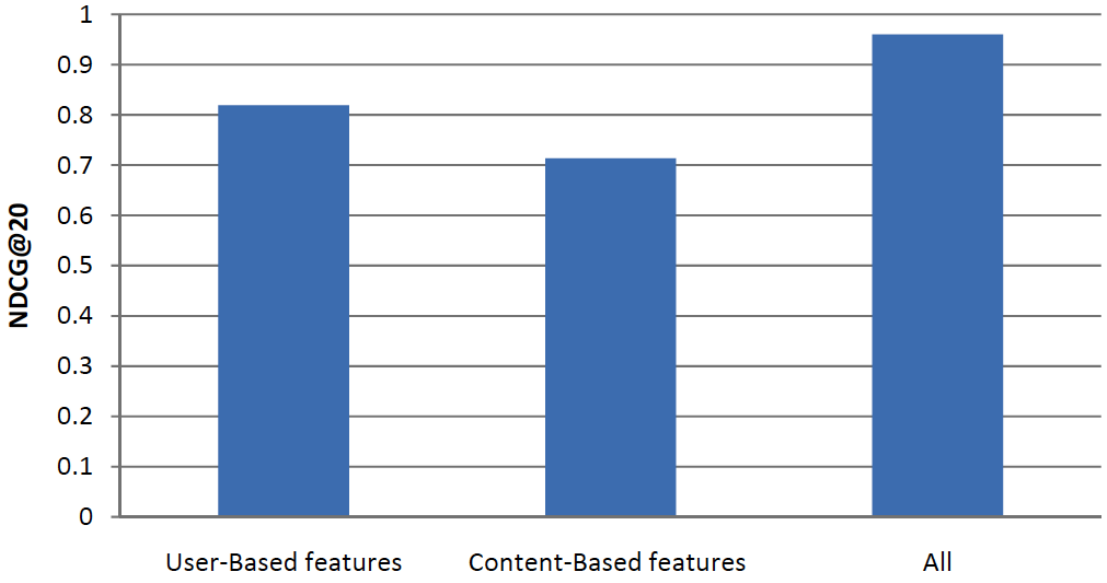

Figure 27 Comparing feature sets

To better understand the relative impact of particular user based and content based features, we next train and evaluate six models – one for each of the three user-based feature groups, and one for each of the three content-based feature groups. Table 1 reports the NDCG@k for k=1, 5, 10, and 20 for each of these six feature groupings.

Table 1 NDCG for six feature groupings

Feature group	NDCG@1	@5	@10	@20
User activity and interest	0.61	0.62	0.65	0.70
User popularity	0.64	0.65	0.67	0.72
User comment history	0.66	0.69	0.71	0.73
Content statistics	0.62	0.65	0.67	0.71
Content NLP features	0.64	0.67	0.68	0.72
Comment-article	0.66	0.68	0.70	0.73

For the user-based feature, the user comment history feature group (recall that this includes the history of a user's previous comment ratings and the number of replies those comments have received) shows the strongest impact. This indicates that some users have a specialty for writing comments that are appreciated by the community; again, we can interpret this feature as a "prior" on a given comment's quality. The hypothesis is that commenter history feature can partially compensate for the lack of comment style capture ability of our content features. We believed that there are some certain types of writing style or background knowledge of those commenters that can usually grab public attention. However, our content features are not good enough to identify the popular writing style yet. On the other hand, we understand it is not necessary true that users digg on comments after they know who is the commenter. The argument is: considering the phenomenon that people do not change their writing style easily, if a commenter made good comments and obtained high digg score before, they are more likely to make comments that will be appreciated by public.

Also note that content-based features are important; two of the top-three feature groups are content-based. We find it interesting that user activity and interest level – based on articles submitted, length of community membership, and category activity – is the single weakest performing feature group. Authoring comments that are perceived as high-quality by the community is largely independent of the user's activity level. Our hypothesis is that there are fundamentally different user types within a Social Web community: article submitters, article raters, commenters, etc. Exploring these different user types and their inter-relationship is an area deserving of further study.

In the final feature study, we explore the importance of content-based features for appropriately modeling the domain. We begin by assuming that our model has access to all user based features. Could it be that comments are not really "messy"? And would it be true that by adding a single content-based feature we can equal the performance of the full feature model? Intuitively, this would mean that the comments contain some clear quality indicators once we factor in the "prior" for the user contributing the comment.

Table 2 NDCG for user-based feature as baseline combined with single content-based features

	NDCG@1	@5	@10	@20
All user-based features (A)	0.74	0.74	0.75	0.81
A + Text length	0.76	0.76	0.77	0.83
A + Upper case	0.74	0.74	0.75	0.81
A + Entropy	0.73	0.74	0.75	0.81
A + Informativeness	0.73	0.74	0.75	0.82

Table 2 reports the NDCG values for the baseline model considering only user-based features, plus four models that consider the baseline plus a single content-based feature only (text length, upper case, entropy, informativeness). In all, however, the content-based features are quite valuable. This indicates that comment content is complex, and that the community's preference for a comment is not driven by a simple feature. Instead, we see the need for full content analysis to capture this complexity.

5. 3 Rank Boosting

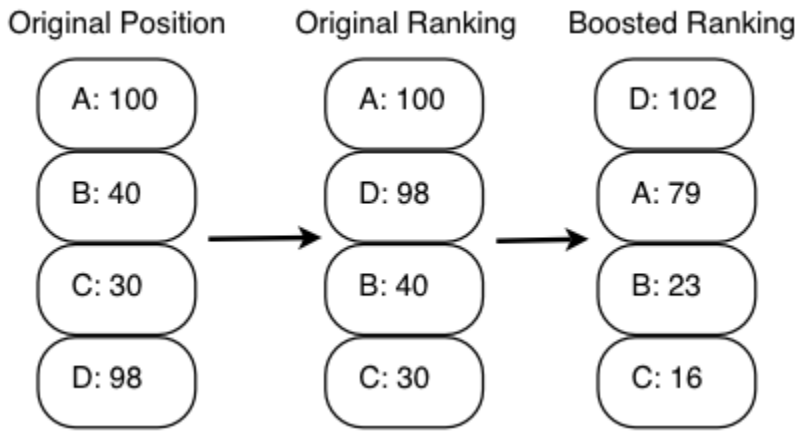

Figure 28 Example illustrating the original time-of-posting position for each comment, the predicted ranking according to the SWCP model, and the boosted ranking using the positional boost modification

As we have seen in Figure 20, the comment posting time has a strong influence on the visibility of a comment and the resulting comment community rating. In this experiment, we are interested in further exploring this phenomenon, as a first step toward breaking the rich-get-richer visibility cycle. As an example, consider the four comments A, B, C, and D and their actual comment community ratings as illustrated in Figure 28. Applying a simple comment posting time ranking to these comments results in the rank order *{A, B, C, D}*. After applying the Social Web Comment Prediction model, we would ideally find the rank order *{A, D, B, C}*. This rank order is in strict order of the community ratings. Indeed, we have seen how the proposed model in this paper performs well on this problem.

It might be reasonable, however, to claim that comment D is the most preferred comment. Based on its late arrival time, but high community rating, we could assert that comment D has been most appreciated by the community relative to its smaller community visibility. This intuition motivates this last exploratory experiment. Referring back to Figure 20, we propose to re-scale the comment community ratings for each training instance with respect to the average community rating for other comments posted in the same order position. In this way, we can evaluate a post arriving 4th (as in the example with comment D) against all other comments in our training data arriving 4th. The intuition is the further a comment's rating is from the average relative to other comments in the same position, then the more the comment's rating should be rewarded or punished. Concretely, for a comment in the jth position attached to a Social Web object i, we can define the boosted comment community rating $\hat{r}_{c_{ij}}$ with respect to all k comments at this same position as:

$$\hat{r}_{c_{ij}} = r_{c_{ij}+}r_{c_{ij}} \times (r_{c_{ij}} - \bar{r}_{c_{ij}}) / \sqrt{\frac{1}{k}\sum_{i=1}^{k}(r_{c_{ij}} - \bar{r}_{c_{ij}})^2}$$

Where $\bar{r}_{c_{ij}}$ is the mean comment score at position j ($\bar{r}_{c_{ij}} = (1/k)\sum_{i=1}^{k} r_{c_{ij}}$) and the denominator is the variance of these comment scores. So a comment with a large rating in a position with a small average rating and small variance would be promoted to a new boosted rating. Returning to our example, suppose the (average, variance) pairs of all comments at positions 1 to 4 are: (148, 235), (119, 193), (105, 169), and (91, 158).

Applying the boosting formula results in the rank order *{D, A, B, C}*. Since comment D's original rating is much higher than the average rating for other comments at the same position, it is boosted from a score of 8 to 102. More importantly, comment A underperforms for its position and is penalized from 100 to 79. In Figure 29 we compare this "boosted" version against the alternative random and time-of-comment ranking models. As in our original model, we see significant enhancement.

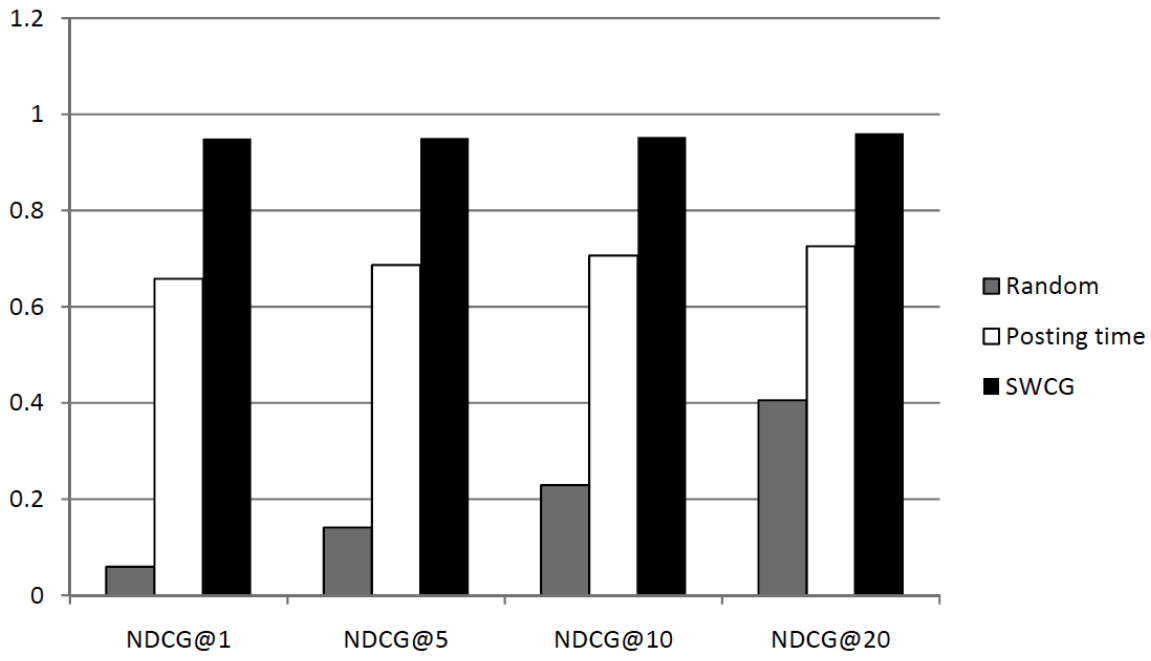

Figure 29 Comparing the rank boosted SWCP model versus alternatives

6. CONCLUSIONS AND DISCUSSION

6. 1 Conclusion

In this thesis, we have proposed and evaluated a regression-based learning model for automatically identifying comment quality within a Social Web community based on the community's preferences. In particular, we have examined the impact of different comment features like visibility, reputation of the comment's author, and the content of the comment itself to understand the influence of these features on the overall community's preference for comments. Aside from the general framework, we have also conducted a broad investigation of the Digg community and its community preference for user-contributed comments to understand some of its general social behavior and the differences across different type of communities. For example, we have seen that Digg users prefer short, simple, and readable comments, and that so-called power users in the community do not, in fact, wield considerable influence over the scores of comments in the community.

6. 2 Potential Extension Work

As part of our future work, we are interested to integrate these results as part of our broader research effort to build enhanced Social Web information management applications that leverage this social collective intelligence.

6. 2. 1 Comment Preference Personalization

We are interested in using this system for personalization purposes. The idea is to train the model to identify the pattern of comments that only the user and/or his friends are

interested in. In addition, we may also train different orientated models for people to choose from such as {informative comments, funny comments, controversial comments} or {comments with joy, comments with anger, comments with encouragements}. As people have varied taste on different characteristics of the comments, this personalized model will provide users with a pleasant comment viewing space.

6. 2. 2 Cross-Community Comment Personalization

While this study has focused on the Digg community, the proposed framework is flexibly designed so that the comment prediction model can be easily applied to other web-based systems incorporating social comment rating features. As mentioned before, the only component that needs customization is the feature extraction component. In addition to targeting other communities, it may be interesting to see how well it may work if we apply a model trained in one community (the source of personalization) and apply the same model to a different community (the target). For example, we may have a group of friends in Digg that like Japanese drama related, funny comments a lot. If we acquire this particular model (the source of personalization) and apply it to YouTube (the target), will the system extract comments with the same taste for me (i.e., Japanese drama related, funny comments)?

6. 2. 3 Quality-Driven Comment Cloud

Finally, we are interested in exploring novel social information exploration and discovery frameworks that leverage the rich socially-generated metadata embedded in comments. As one step in this direction, we are interested in comment clouds, inspired

by the popular word cloud concept. A word cloud is generally a visual depiction of a bag of words. In many applications, the importance of a term is shown with font size or salient color. Word clouds are good for visualization and navigation. In our case, we want to generate comment-based clouds that are based on high quality comments extracted by the SWCP system. Concretely, we can use the learned model to extract valuable keywords from the high quality comments to form a comment cloud that enhances the visualization and topic navigation in the Social Web. Figure 30 and Figure 31 show the differences between the Comment Cloud generation results using a full set of comments attached to the original story and using only those comments that have high community preference rate. In this example, the story title is “Kids Who Don’t Play Video Games are at Risk.” The abstract of the story is “Lawrence Kutner and Cheryl Olson, researchers at Harvard University and authors of 'Grand Theft Childhood' discuss some of their findings.” We can see that the comment cloud with community preferred comments includes constructive terms (e.g., research, evidence, mom) without the meaningless words such as “br” you can easily find in the full set cloud. As part of our ongoing investigation, we are exploring more fuller this comment-cloud navigation framework.

Figure 30 Comment cloud generated from all comments of the story “Kids Who Don’t Play Video Games Are at Risk”

Figure 31 Comment cloud generated from comments with high community preference rate of the story “Kids Who Don’t Play Video Games Are at Risk”

REFERENCES

[1] M. Hu, A. Sun, and E.-P. Lim, "Comments-oriented Document Summarization: Understanding Documents with Readers' Feedback," *Proceedings of the 31st Annual International ACM Special Interest Group on Information Retrieval Conference (SIGIR),* pp. 291-298, 2008.

[2] G. Mishne, "Using Blog Properties to Improve Retrieval," *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2007. Available online: http://www.icwsm.org/papers/3--Mishne.pdf

[3] S. Sen, M. F. Harper, A. Lapitz, and J. Riedl, "The Quest for Quality Tags," *Proceedings of the 2007 International ACM Conference on Supporting Group Work(GROUP)*, pp. 361-370, 2007.

[4] G. Mishne and D. Carmel, "Blocking Blog Spam with Language Model Disagreement," *Proceedings of the Workshop on Adversarial Information Retrieval on the Web(AIRweb)*, 2005. Available at CiteSeer^{X}: http://airweb.cse.lehigh.edu/2005/mishne-.pdf

[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding High-quality Content in Social Media," *ACM International Conference on Web Search and Data Mining(WSDM),* pp. 183-194, 2008.

[6] N. Jindal and B. Liu, "Opinion Spam and Analysis," *ACM International Conference on Web Search and Data Mining(WSDM),* pp. 219-230, 2008.

[7] D. Goldberg, D. Nichols, B. M. Oki , D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM,* vol. 35, no. 12, pp. 61-70*,* 1992.

[8] G. Mishne and N. Glance, "Leave a Reply: An Analysis of Weblog Comments," *Workshop on the Weblogging Ecosystem*, 2006. Available on CiteSeer^{X} : http://www.blogpulse.com/www2006-workshop/./papers

[9] C. Lampe and P. Resnick, "Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space," *Proceedings of the 2004 Conference on Human Factors in Computing System,* ACM Press, pp. 543-550, Vienna, Austria, 2004.

[10] C. A. C. Lampe, E. Johnston, and P. Resnick, "Follow the Reader: Filtering Comments on Slashdot," *ACM Conference on Human Factors in Computing Systems*, pp. 1253 - 1262, 2007.

[11] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical Analysis of the Social Network and Discussion Threads in Slashdot," *International World Wide Web Conference (WWW),* pp. 645-654, 2008.

[12] A. Arnt and S. Zilberstein, "Learning to Perform Moderation in Online Forums," Proceedings of the 2003 IEEE/WIC International Conference on *Web Intelligence*, pp. 637-641*,* 2003.

[13] A. Veloso, W. Meira, T. Macambira, D. Guedes, and H. Almeida, "Automatic Moderation of Comments in a Large On-line Journalistic Environment," *Proceedings of International Conference on Weblogs and Social Media*, 2007. Available online : http://www.icwsm.org/papers/4--Veloso-Meira-Macambira-Guedes-Almeida.pdf

[14] K. Lerman, "Social Networks and Social Information Filtering on Digg," *Proceedings of International Conference on Weblogs and Social Media*, 2007. Available online : http://www.icwsm.org/papers/4-Lerman.pdf

[15] L. Krista, "Social Information Processing in News Aggregation," *IEEE Internet Computing: Special Issue on Social Search*, vol. 11, no. 6, pp 16-28*,* November 2007.

[16] K. Lerman and A. Galstyan, "Analysis of Social Voting Patterns on Digg," *Proceedings of the ACM SIGCOMM workshop on Online Social Networks*, pp. 7-12, Jun 2008.

[17] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature Selection for Ranking," *Special Interest Group on Information Retrieval (SIGIR), ACM Press*, pp.407-414, 2007.

[18] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A Regression Framework for Learning Ranking Functions Using Relative Relevance Judgments," *Special Interest Group on Information Retrieval (SIGIR),* pp. 287-294, *ACM*, 2007

[19] K. Crammer, Y. Singer, "PRanking with Ranking," *Neural Information Processing Systems (NIPS),* pp.641-647, 2002.

[20] F. Harrington. "Online Ranking/Collaborative Filtering Using The Perceptron Algorithm," *International Conference On Machine Learning* (*ICML)*, 2003. Available online: http://www.hpl.hp.com/conferences/icml2003/papers/93.pdf

[21] A. Shashua, A. Levin, "Ranking with Large Margin Principle: Two Approaches," *Neural Information Processing Systems (NIPS)*, 2002. Available online : http://books.nips.cc/papers/files/nips15/AA58.pdf

[22] W. Chu, and Z. Ghahramani. “Preference Learning with Gaussian Processes,” *International Conference On Machine Learning* (*ICML)*, pp.137 -144, 2005.

[23] T. Bartell, G. W. Cottrell, “Learning to Retrieve Information,” *International Conference on Services Computing(SCC)*, 1995. Available online : http://www-cse.ucsd.edu/users/gary/pubs/sncc95.ps

[24] W. W. Cohen, R. E. Shapire, Y. Singer, “Learning to Order Things,” *Neural Information Processing Systems (NIPS),*pp. 451-457, 1998.

[25] C.J.C. Burges, T. Shaked, E. Renshaw, M. Deeds, N. Hamilton, G. Hullender, “Learning to Rank using Gradient Descent,” *International Conference On Machine Learning* (*ICML),* pp.89-96, 2005.

[26] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, W.-Y. Ma, “FRank: A Ranking Method with Fidelity Loss,” *Special Interest Group on Information Retrieval (SIGIR), ACM,* pp. 383-390, 2007.

[27] Y. Freund, R. Iyer, R. Dharmarajan, “An Efficient Boosting Algorithm for Combining Preferences,” *Journal of Machine Learning Research (JMLR),* pp. 170-178, 1998.

[28] R. Herbrich, T. Graepel, K. Obermayer, “Large Margin Rank Boundaries for Ordinal Regression,” *Advances in Large Margin Classifiers*, pp. 115-132, 2000.

[29] T. Joachims, “Optimizing Search Engines Using Click through Data,” *Knowledge Discovery and Data Mining (KDD),* pp.133-142, 2002.

[30] J. Gao, H. Qi, X. Xia, J. Y. Nie, “Linear Discriminant Model for Information Retrieval,” *Special Interest Group on Information Retrieval (SIGIR),* pp. 290-297, *ACM*, 2005.

[31] W. Chu, and Z. Ghahramani. “Preference Learning with Gaussian Processes,” *International Conference On Machine Learning* (*ICML),* pp. 137-144, 2005.

[32] Z. Zheng, K. Chen, G. Sun, H. Zha, “A Regression Framework for Learning Ranking Functions using Relative Relevance Judgment,” *Special Interest Group on Information Retrieval (SIGIR), ACM*, pp. 287-294, 2007.

[33] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, G. Sun, “A General Boosting Method and its Application to Learning Ranking Functions for Web Search,” *Neural Information Processing Systems (NIPS)*, 2007. Available online : http://www.stat.rutgers.edu/~tzhang/papers/nips07-ranking.pdf

[34] C. Cortes, M. Mohri, A. Rastogi, "Magnitude-preserving Ranking Algorithms," *International Conference On Machine Learning* (*ICML),* pp. 169-176, 2007.

[35] E. Khabiri, C.-F. Hsu, and J. Caverlee. "Analyzing and Predicting Community Preference of Socially Generated Metadata: A Case Study on Comments in the Digg Community". Poster presented at *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009. Available online : http://infolab.tamu.edu/static/papers/digg-paper.pdf

[36] C.-F. Hsu, E. Khabiri, and J. Caverlee. "Ranking Comments on the Social Web," *IEEE International Conference on Social Computing (SocialCom)*, May 2009. Available online : http://faculty.cs.tamu.edu/caverlee/pubs/hsu09socialcom.pdf

[37] H. Drucker, Chris, B. L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," *Advances in Neural Information Processing Systems* 9, vol. 9, 1997. Available at CiteSeer^{X} : http://mlg.anu.edu.au/~smola/./papers/DruBurKauSmo

[38] D. Sculley and G. M. Wachman, "Relaxed Online SVMs for Spam Filtering," *Proceeding of the International ACM of Special Interest Group on Information Retrieval (SIGIR) Conference,* pp. 415-422, 2007.

[39] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” 2001. Available at CiteSeer^{X} : http://cramer.cs.nmt.edu/~kdd/libsvm.pdf

[40] B. Carpenter, “Phrasal Queries with LingPipe and Lucene,” *Proceedings of the 13th Meeting of the Text Retrieval Conference (TREC)*, 2004. Available online : http://www.colloquial.com/carp/Publications/TREC2004.pdf

[41] G. H. McLaughlin, “SMOG Grading: A New Readability Formula,” *Journal of Reading,* vol. 12 no. 8 pp. 639-646, 1969

[42] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, 1st ed. Addison Wesley, February 2009.

APPENDIX A

MEASURING COMMENT CONTROVERSY

As part of our experimental study of Digg, we observed that community users may be interested in top-rated comments as well as in comments for which there is no clear consensus. These controversial comments often receive many up-votes and many down-votes, indicating a high-level of community interest even though the aggregate vote may be close to 0. For example, an attention-lacking comment with no votes is very different from a controversial comment with 50 positive votes and 50 negative votes that cancel each other out. We thus want to examine the prevalence of controversial comments in different communities and explore how these controversial comments may impact comment quality prediction.

A. 1 Comment Controversy in Digg

We say *controversy* is a state of public debate concerning a matter of opinion. Due to the digging functionality in Digg, the controversy of a comment or a story or even the category can be revealed by the aggregated thumbs up and thumbs down score of the comments. We say that a comment is highly controversial if there are nearly equal amount of thumbs up and thumbs down applied to it. For a story, it is controversial when the summation of thumbs ups and the summation of thumbs down for all comments attached to the story are nearly equal. Similarly, we can define the controversy of a category by aggregating all of the thumbs up and thumbs down for comments in all stories within a single category

A. 2 Using Entropy for Controversy Measurement

Entropy is used to define the controversial level of one comment. A formula for calculating the entropy is derived from the Bayesian expected entropy [3]. It represents the idea that the entropy measures the uncertainty of random variable. In our case the entropy of one comment X is the summation of both thumbs up and thumbs down:

$$H(X) = -(p(x_p)\log_2 p(x_p) + p(x_n)\log_2 p(x_n))$$

However, the amount of disagreement should be different for different amount of positive and negative score even they have the same ratio. The Bayesian expected entropy treats the thumbs up and thumbs down ratio itself a random variable. Thus, given u positive digg and d negative digg for each comment, the probability of a particular ratio q being f is:

$$\int_{f=0}^{1} p(q = f|u,d)(-f \cdot \log(f) - (1.0 - f) \cdot \log(f)\, df$$

Where $p(q = f|u,d)$ is the probability of binomial distribution defined as

$$\frac{(u+d)!}{u! \cdot d!} \cdot f^u \cdot (1-f)^d$$

As mentioned in section 4.2, there are eight top-level categories in Digg: Technology, World and Business, Science, Lifestyle, Entertainment, Sports, Gaming, and Offbeat. The controversy of each category can be calculated by the percentage of comments in that category that is over a threshold of entropy.

Table 3 shows the percentage of controversial comments after we applied different thresholds. The relative value is similar across different threshold. The table shows that comments in Science category has high possibility of becoming a controversial comment,

where World and Business is relatively abnormal to have both side of people debating on one comment.

Table 3 Percentage of controversial comments after applying different thresholds

threshold	Tech.	W & B	Science	Gaming	LifeStyle	Ent.	Sports	Offbeat
0.9	0.17	0.07	0.19	0.17	0.19	0.17	0.13	0.14
0.5	0.17	0.07	0.19	0.17	0.19	0.17	0.13	0.14
0.3	0.42	0.21	0.45	0.41	0.42	0.41	0.33	0.37
0.1	0.96	0.91	0.97	0.97	0.96	0.97	0.95	0.96

APPENDIX B

NORMALIZED DISCOUNTED CUMULATIVE GAIN

Discounted Cumulative Gain is an effectiveness measurement for many information retrieval related applications. It emphasizes the correctness of earlier retrieved documents of a result set from the system we want to evaluate. On the other hand, we want to penalize the system performance when there are highly relevant documents appearing lower in a result list. One of the ways to do so is to reduce the relevance value logarithmically proportional to the position of the document in the result list. The DCG accumulated at a particular rank position p is defined as:

$$DCG_p = rel_1 + \sum_{i=1}^{p} \frac{rel_i}{\log_2 i}$$

We want to normalize the *DCG* since each calculated DCG is for a particular result list generated from a particular system. In order to make cross-system and cross-result-set comparison, the DCG should be normalized. This is done by producing an ideal *DCG.* The ideal DCG is obtained by sorting documents of a result list by relevance. With that, the normalized discounted cumulative gain, or *nDCG*, is computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

The *nDCG* values for all result sets can be averaged to obtain a measure of the average performance of an information retrieval system's ranking algorithm. We can see that *nDCG* is a relative value in the interval 0.0 to 1.0.

APPENDIX C

ACTIVITY OF THE DIGG USERS

We define the aggregate thumbs up for each comment as the comment popularity, which reveals how much a comment was seen and liked by Digg community. This leads to the definition of community popularity as the summation of comment popularity of all the comments in one category (like technology) is called community popularity. Secondly, the number of comments for each category shows how much the users are eager to communicate their ideas with others in that category. This feature is called community communication. Lastly, the combination of popularity and communication degree gives us a basic idea of the involvement level in different categories. After summing up the popularity and community communication of each category we normalize the result with the formula : $(x - \min(X))/(\max(X) - \min(X))$. We can see from the Table 4 that World & Business has the highest user involvement, following with technology. Sports, on the other hand, is the least involved community among all.

Table 4 Community popularity, community communication and involvement for different categories in Digg

	Technology	W&B	Science	Game	Lifestyle	Ent.	Sports	Offbeat
Popularity	0.55	0.91	0.10	0.20	0.16	0.27	0.00	1.00
Communication	0.00	1.00	0.26	0.02	0.25	0.09	0.11	0.20
Involvement	0.24	1.00	0.14	0.06	0.17	0.14	0.00	0.60

Figure 32, Figure 33 and Figure 34 reveal a similar pattern across eight categories. Note that the amount of comments is highly correlated with the number of stories under each

category. Similarly, the total digg count is highly correlated with the number of comments.

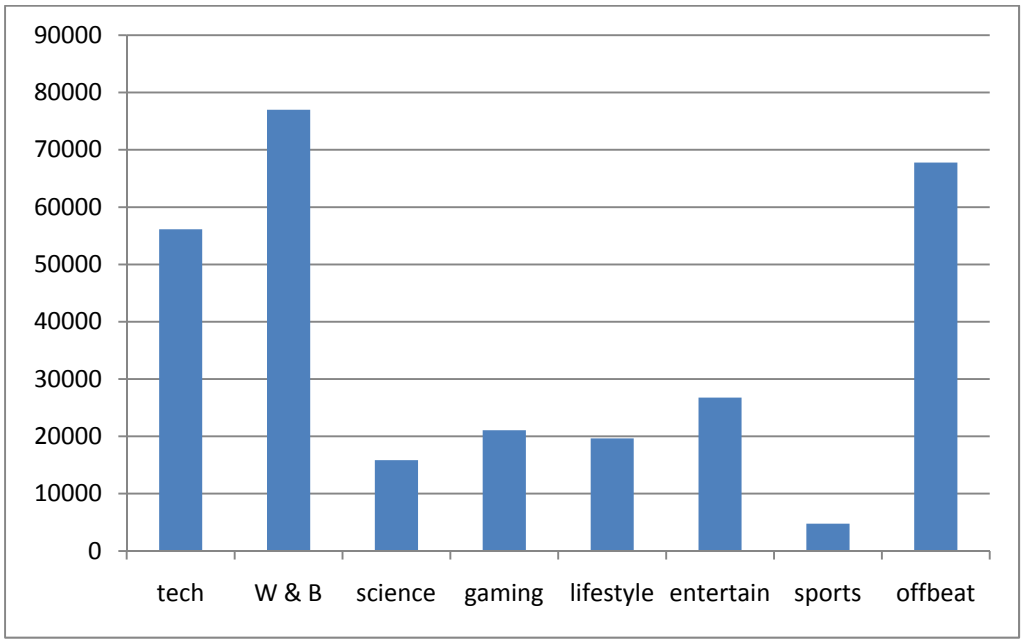

Figure 32 Total comment count in each category

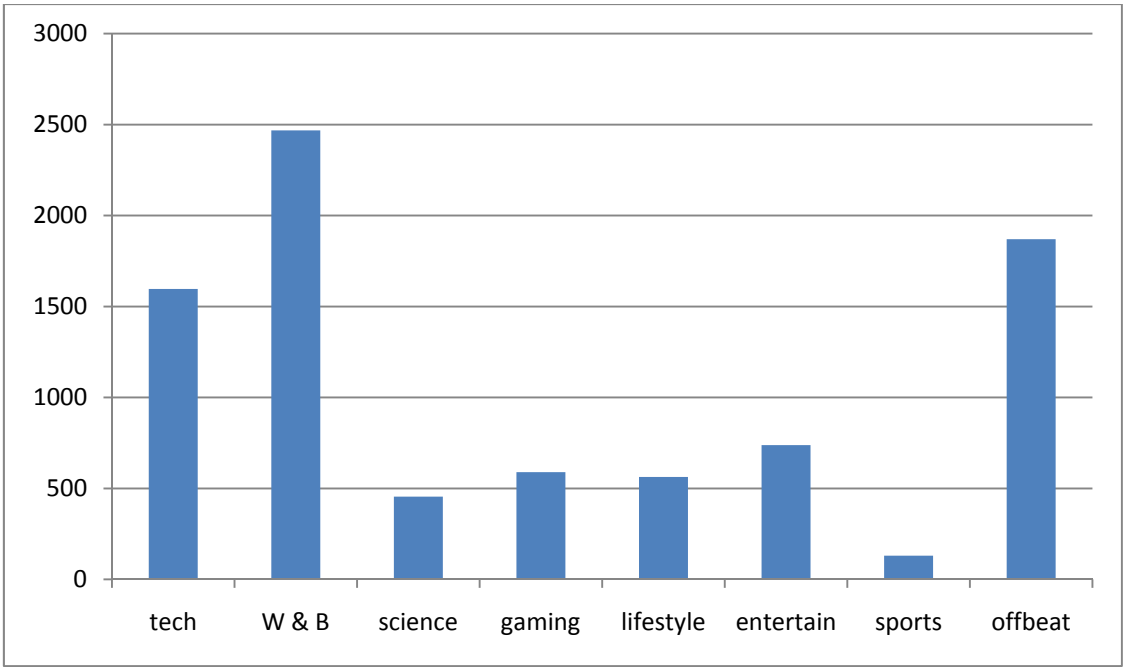

Figure 33 Story count in each category

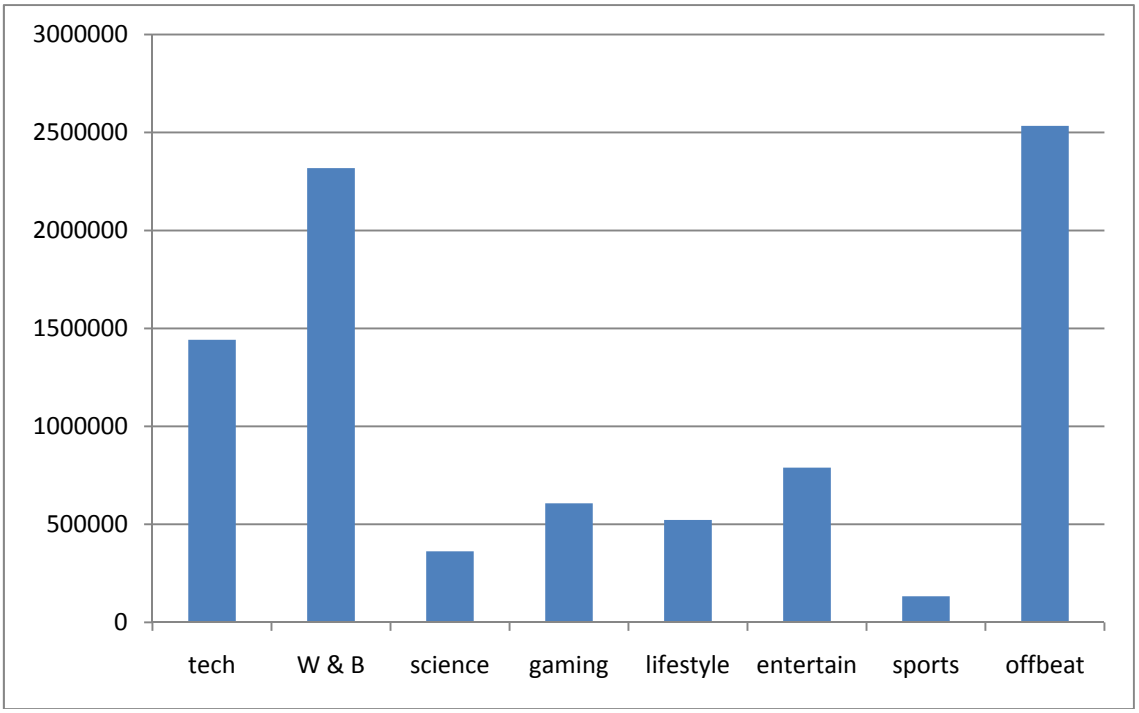

Figure 34 Total digg count in each category

VITA

Chiao-Fang Hsu received her Bachelor of Science degree in computer science from National Tsing Hua University in Taiwan in 2007. She entered the Computer Science and Engineering program at Texas A&M University in September 2007 and is a member of the TAMU infolab focused on Web and Distributed Information management. She received her Master of Science degree in computer science from Texas A&M University in December 2009. Her research interests include information retrieval, text mining and machine learning.

Ms. Hsu may be reached at Department of Computer Science and Engineering, Texas A&M University, TAMU 3112, College Station, TX 77843-3112.