Soc. Net. Anal. Min. manuscript No.
(will be inserted by the editor)

An Exploration of Submissions and Discussions in Social News

Mining Collective Intelligence of Reddit

Tim Weninger

Received: date / Accepted: date

Abstract Social news and content aggregation Web sites have become massive repositories of valuable knowledge on a diverse range of topics. Millions of Web-users are able to leverage these platforms to submit, view and discuss nearly anything. The users themselves exclusively curate the content with an intricate system of submissions, voting and discussion. Furthermore, the data on social news Web sites is extremely well organized by the user-base, which, like in Wikipedia, opens the door for opportunities to leverage this data for other purposes. In this paper we study a popular social news Web site called Reddit. Our investigation looks at the dynamics of hierarchical discussion threads, and we ask three questions: (1) to what extend do discussion threads resemble a topical hierarchy? (2) Can discussion threads be used to enhance Web search? and (3) what features are the best predictors for high scoring comments? We show interesting results for these questions on a very large snapshot several sub-communities of the Reddit Web site. Finally, we discuss the implications of these results and suggest ways by which social news Web sites can be used to perform other tasks.

Keywords social news · reddit · online discourse · comment threads · popularity prediction

1 Introduction

Social news Web sites are platforms in which (1) users generate or submit links to content, (2) submissions are voted on and ranked according to their vote totals, (3) users comment on the submitted content, and (4) comments are voted on and ranked according to their vote totals. These platforms provide a type of *Web-democracy* that is open to all comers. Social news Web sites, including Digg, Reddit, Slashdot, HackerNews, etc., have become exponentially more popular during the past few years.

Social media frameworks represent a stark departure from traditional media platforms in which a news organization, *i.e.*, a handful of television, radio or newspaper producers, sets the topics and directs the narrative. Social news sites increasingly set the news agenda, cultural trends, and popular narrative of the day. Leskovec {em et al. demonstrated with the MemeTracker project that Web logs drive the media narrative [Leskovec et al. (2009)]. This trend shows no signs of waning. Furthermore, the number of blogs, news outlets, and other sources of user generated content has outpaced the rate at which Web users can consume information. Social news sites and their many subtopic pages collectively curate, rank and provide commentary on the top content of the day by harnessing the power of the masses.

One of the most interesting and important features of social news sites is the ability for users to comment on a submission. These comment threads provide a user-generated and user-curated commentary on the topic at hand. Unlike message boards or Facebook-style comments that list comments in a mostly-flat, chronological order, or Twitter discussions that are person-to-person and oftentimes difficult to discern, comment threads in the social news paradigm are public, permanent (although editable), well-formed and hierarchical. The hierarchical nature of comment threads, where the discussion structure resembles a tree, is especially important because this allows divergent sub-topics resulting in a more robust overall discussion.

The result of a robust discussion may yield more information about the topic than the actual linked-content.

This work is funded by the National Science Foundation Graduate Research Fellowship.

Tim Weninger
353 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46656
E-mail: tweninger@nd.edu

For example, a submitted link that points to a New York Times article about George Zimmerman may contain comments about gun control, self-defense laws, attorneys, jury trials, and so on. These comments create a hierarchically self-organized context that can act as a supplement to the content of the news article. The vote-scores of each comment are also helpful in gauging the community's opinion on the topic, sub-topic, etc. This is not unlike social hierarchies [Gilbert et al. (2011)] wherein certain influential actors move the dynamics of a social network. Except, in this domain, the influence of a post or comment spreads via comment and page views, which is determined by the post or comment's relative ranking, which largely based on vote-scores.

Ranking systems vary widely across social media aggregation sites [Bross et al. (2012)], but generally the vote totals for a particular post indicate the community' opinion on the general topic. Popular opinions, responses and rebuttals are likely to be voted to the top, while unpopular opinions are unlikely to be highly scored. Just as the vote total indicates the community's opinion of a post's topic or assertion, the vote totals on comments are indicative of opinion in a more fine-grained manner. For example, using the George Zimmerman example article from above, the article itself is highly voted because it is a topic of wide interest; but in the comment section, a comment branch (*i.e.*, a sub-thread) about racism in America may be highly scored, and a comment branch about increasing gun control laws may be poorly scored. The case in this example would indicate that the community favors the discussion about racism, while disagreeing with the need for more gun control laws.

In this paper we explore the social news site Reddit in order to gain a deeper understanding on the social, temporal, and topical methods that allow these types of user-powered Web sites to operate. This paper presents first-of-a-kind, large-scale study of posts and comments on the social news site. The specific research questions we address are:

- User sentiment is often complex and multi-faceted. On social news sites, user sentiment is codified into comments which are organically organized into hierarchies. The first part of this paper investigate the extent to which comment hierarchies threads represent a topical hierarchy. A positive correlation would validate hierarchical topic modeling as well as provide a numerical confirmation to the hypothesis that comment hierarchies are sub-divided topically.
- A topically diverse comment section is likely to contain information and opinions that supplement the content of the linked-article. The second part of the paper investigates the amount of supplemental information comment threads add to the posted articles content. If comment sections are found to contain a great deal of supplemental information, then they may also contain information useful to enhance Web search and retrieval.
- Comments and submissions are displayed by rank according to their vote totals. Thus, as users vote on their favorite comments, the comment section is constantly being re-adjusted to accommodate rising and falling comments indicating the community's general sentiment. The third part of this paper investigates if popular opinion is the sole driving force behind a comment's vote total. If not, this section seeks to identify variables which are indicative of the future score of a given comment.

The remainder of this paper is organized as follows. The next section briefly surveys related works. We follow with a description of the Reddit dataset and the method by which it was obtained. In Section 5 we investigate the structure and topical hierarchies of comment threads. In Section 6 we study to extent to which comment threads provide supplemental content that can be used to improve Web search. In Section 7 we look at comment and content popularity and develop a model to predict the future score of a comment. We conclude by discussing various insights that we gained during this study and suggest topics for further research.

2 Related Work

Despite the booming popularity user-generated content aggregation Web sites like Reddit, which is listed as the 33rd most popular Web site in the United States and the 99th most popular Web site globally (and climbing)[1], this paper is the among the first to explore its dynamics. Previous studies have looked at similar, yet smaller Web sites and forums like Slashdot [Gómez et al. (2008), Lampe and Resnick (2004)], Usenet [Fisher et al. (2006)], Digg [Lerman and Galstyan (2008), Zhu (2010)], 4chan [Bernstein et al. (2011)], etc. However, the previous works focus mainly on the friendship dynamics of the Web site. The social network of a Web site play an important role in promiting content. Lerman [Lerman (2007)] found that users with larger social networks are more likely to have their posts highly scored on Digg. Lerman [Lerman (2007b)] also found that certain social recommendation systems, like Digg, that allow or encourage on social networks to form can lead to a small number of well-connected users to dominate the site. One of the problems that developed in the Digg platform is that "voting rings" began to form; as a result, in order for a users' post to have any chance at success required a large number of friends to vote on a submission. Although this has not been studied conclusively, our nom-scientific opinion is that the so called "tyranny of the minority," arguably, is among the main reasons why Digg eventually failed.

[1] According to Alexa.com, accessed Sept 27, 2013

Reddit, on the other hand, does not annotate friendship, and a brief investigation into comment reply relationships did not indicate a noticeable number of hidden friendships; furthermore, revealing a user's real identity is strictly and emphatically forbidden by both the terms of service and the user-base. Furthermore, forming voting rings is also emphatically forbidden by the site's terms of service.

This line of work also has similarities in recent work that mines knowledge from question answering sites like Yahoo Answers [Adamic et al. (2008)], Stack Overflow [Anderson et al. (2012)], Quora [Paul, et!al. (2012)], etc. In particular, Anderson *et al.*'s study developed a model to predict the future score of an answer. They found that the best answers were typically given by those who have answered other questions well. Questions Answering (QA) sites are similar to Reddit because they are made entirely of user-generated content and because of the voting system that QA sites employ. In fact, Reddit has organically evolved a question answering component, which could be studied independently, but the general composition of Reddit is much broader than question answering.

A study by Muchnik *et al*. [Muchnik, et al. (2013)] found that random votes on a social media platform resulted in wide swings in the final score of a random post. Although this study was not performed on Reddit, it raises questions of the susceptibility of social news sites to outside, or non-organic influence.

One of the most fascinating properties of user-curated social news Web sites is their ability to perform organic crowdsourcing. These Web sites, Reddit especially, are largely immune to spam and marketing campaigns because non-relevant, or uninteresting submissions are quickly identified by the users. Social news sites are a type of implicit crowdsourcing network [Doan et al. (2011)] because they ask the crowd to indirectly solve a problem: to rank content submissions and comments. This is interesting because, although users are never asked nor are required to explicitly rank submitted content, the crowd is able to organically generate sets of topical, relevant, non-redundant, high-quality content.

Several recent studies have indicated that the news agenda is increasingly dominated by blogging services and other types of "citizen journalism" rather than by professional media organizations. The Memetracker project, for example, found that several popular phrases found on mainstream or cable news channels first appeared online [Leskovec et al. (2009)]. Aside from the tracking of topics and memes, there has been work on news content in particular. The standard line of research in algorithmic curation and filtering of news is featured in automatic news aggregation Web sites like Google News or services like Twitter's Trends. It is widely believed that social media sentiment can be used to forecast public opinion [Mukherjee and Liu(2012)]. However, a recent study found no correlation between Twitter sentiment and the results of the 2012 US GOP primary [Mejova et al. (2013)]; yet a similar study found that Twitter sentiment was able to predict box-office revenues for movies [Asur and Huberman (2010)]. Some of the research topics involved here include identifying temporal topics [Hong et al. (2011), Kawamae and Higashinaka (2010)], and cascades of news and information [Leskovec et al. (2007)], among many others.

Research on Web log comments and discussion threads includes: mining hierarchies from linear discussions [Wang et al. (2011), Cong et al. (2008)], exploring hierarchies in online discussions [Laniado et al. (2011)], and popularity prediction [Tsagkias et al. (2009)]. While there is utility in these research efforts for linear (Facebook-style) discussion threads, many new comment systems, including Reddit and the recently popular Disqus system, are explicitly hierarchical.

Comment threads have also been useful in enhancing information retrieval models. In these retrieval models text in comment threads are added to the background of the overall language model. Researchers have found that the adaption of user comments can substantially increase retrieval performance [Duan and Zhai (2011), Seo et al. (2009)].

Predicting the future popularity of a post or comment is also an area of growing interest because users typically wish for their submission to be scored highly so that their opinion or insight might be viewed by more users. This topic has been approached in many different ways. One prediction mechanism measures a post's immediate popularity, such as page views on YouTube and Digg, to predict future popularity. The researchers find that early patterns of access can indicate the long term popularity of content [Szabo and Huberman (2010)].

It is also possible to predict the popularity of a submission before it is submitted by looking at features engineered from the post's content, such as the subjectivity of the content, the source of the article, the number 'tweets' which mention the named entities in the article, and the amount of 'tweets' that mention the article in question [Bandari et al. (2012)]. Similarly, textual and semantic features engineered from a submitted article can be used to predict the number of comments a post will have, thereby indirectly predicting popularity of a post [Tsagkias et al. (2009)].

Using Reddit image-posts specifically, Lakkaraju *et al*. found that the words used in the titles of posts are very indicative of its ultimate score [Lakkaraju et al. (2013)]. That is, even though the same image may be posted to Reddit dozens of times, usually only one of the image-posts will become popular. Lakkaraju *et al* find that posts with original titles that are specific to the target community are more likely to be popular. Only a few other studies use Reddit as a source of data. Among these is a study on the 'under-

provision' of Reddit, which notes that many popular posts where unsuccessfully submitted many times prior (probably by different users) before eventually becoming popular [Gilbert (2013)]. The researchers argue that this is because only a small number of users actually vote on a post or comment. Instead, most users rely on *everyone else* to rank the information on the site, thereby allowing relatively few people to control the information viewed by the millions of daily visitors. Another study finds similarities among the comment sections on Reddit, Digg and Epinions by analyzing the growth of conversations in discussion threads [Wang et al. (2012)]; this work is similar to ours in that it investigates discussion threads, but the work done by Wang *et al* focuses on the temporal and structural dynamics of when and how users make comments rather than the topicality of user comments as studied in this paper.

3 Dataset Description

User-powered social news sites such as Reddit, Slashdot and others have similar setups and user interaction schemes. Web users may access these sites anonymously (without an account) in read-only mode where they can browse postings and comments, but not contribute, vote or comment. Account creation typically only requires a username, password, and the passage of a challenge-response test (*e.g.*, Captcha-test); thus users typically remain anonymous. Registered users may contribute posts, comment and vote.

We chose to study Reddit in particular because (1) the user-community is very active, (2) the Web site has a soaring popularity, and (3) *all* posting, comment and aggregate user data is publicly accessible.

Reddit, in particular, is beginning to influence the world in ways that both the mainstream media and research community do not yet fully understand. The Reddit community is able to bring a higher order of organization to online content, and is changing the methods of discourse online. Recent posts by presidents, including Barack Obama, Nobel laureates, A-list actors, singers, astronauts, scientists, CEOs, and so on, *c.f.* `http://www.reddit.com/r/iama/top/`, reinforce this trend.

Before we introduce the experimental dataset, we describe the basic framework for the Reddit system:

Subreddits. Reddit is comprised of thousands of user-created and user-moderated *subreddits*, which are topical forums for content. For example, there is a general POLITICS subreddit as well as CONSERVATIVE, LIBERAL, PROGRESSIVE, etc., subreddits. Any user can create and moderate a subreddit at any time, and Reddit administrators rarely interfere with or censor subreddits. New users are auto-subscribed to a handful of popular subreddits, and other subreddits can be subscribed to according to the user's interests.

Table 1 Statistics of the Reddit dataset

Capture Dates	7/25/2012 – 11/19/2012
Users	1,154,184
Posts	369,833 (across 25 subreddits)
Post Votes	488,555,185 (58% Upvotes)
Comments	16,540,321
Comment Votes	371,439,104 (79% Upvotes)

Certain subreddits have specific rules that determine what can and can not be posted, for example, PICS requires posts to be only pictures. It is unclear, and outside the scope of this paper, if these rules play any part in this study's results. There used to be a general subreddit called REDDIT.COM, but it was removed to encourage topical discussion.

Posts. Regardless of subreddit subscription status, any registered user can contribute to any subreddit by submitting a link to external content or by creating a self-post. Self-posts are Wiki-style text with a generous 10,000 character limit.

Comments. Registered users can also comment on posts. The comment pages of Reddit are hierarchically threaded, *i.e.*, a comment can be in response to the post in general (a root comment), or in reply to another comment. This creates a discussion hierarchy and facilitates discussion subtopics.

Voting. Registered users are able to *upvote* or *downvote* posts and comments; one vote per post/comment per user, +1 point per upvote, -1 point per downvote. Posts and comments are displayed on the site in sorted order according to a time and vote total ranking function. Popular posts may trigger "vote fuzzing", which is an anti-spam mechanism and the only closed-source part of Reddit. According to the Reddit FAQ[2] the vote fuzzing mechanism changes the number of up and down votes; the vote score *i.e.*, upvotes - downvotes, is not changed.

Karma. When a post or comment receives votes, the user who contributed the post or comment receives *karma*. For example, if a user submits a link to an article that receives a total of 10 upvotes and 2 downvotes, then that user will receive 8 karma points. Post-karma and comment-karma are counted separately. Self-posts do not receive karma points. Users with a large amount of karma are allowed to contribute more frequently. This rewards users who contribute high quality content and make insightful, amusing or otherwise interesting comments.

To gather a dataset sufficient for a large-scale exploration, we crawled the Reddit API four times daily: at 0:00, 6:00, 12:00 and 18:00 CST. During each crawl we retrieved the 100 top-scoring posts from the 25 most popular subreddits[3], as well as the 100 top-scoring posts of the day from across all subreddits. From each post we retrieve the 500

[2] `http://www.reddit.com/wiki/faq`

[3] `http://www.reddit.com/reddits/`, accessed on 7/24/2012

top-scoring comments, with a depth limit of 10. Each post and comment has submission time, text, username, and vote totals. To ensure we gather complete voting results, comments, and full set of edits, we initially only *note* the top posts and comments; we actually *collect* the complete text, votes, etc. after 48 hours has elapsed. Results presented later in this paper demonstrate that 48 hours is a sufficient waiting period; in fact, we find that the vast majority of activity occurs within the first 4 hours of a post's life-cycle. We also collect the registration date and aggregate karma scores for each user we encounter. Of course, we would like to collect the full set of data, but Reddit asks that crawlers limit the number of API requests to one per second making this full dataset impossible to collect without violating the terms of service. Table 1 has statistics of the collected data.

Unfortunately, this method of data gathering introduces a severe bias into the data and thus may skew the results. The introduction of bias comes from the fact that the system only captures the *top* 25 subreddits. The top subreddits are far more active than most subreddits; this bias will likely result in an inflated number of comments and votes simply because more users are likely to see posts and comments from top subreddits on the frontpage. Another source of bias due to the Reddit API's 500 comment maximum. It is possible to download the complete set of comments for comment-threads containing more than 500 top comments, but this process involves multiple (*i.e.*, hundreds or thousands) API-calls. Because Reddit asks users to limit requests to 1 every 2 seconds a choice had to be made to either a) get the whole comment thread or b) get lots of different comment sections. For the purposes of this study we opted for variety instead of completeness. As a result of capturing the top 500 comments, many low-scoring comments are not considered in the following experiments. Careful consideration was given to these biases, and any conclusions are formed with these biases in mind.

Posts and comments are frequently deleted. However, our data capture system does not make any effort to delete a comment or post from the captured dataset if it has been deleted on Reddit post hoc. Obviously, if a post is deleted before the crawl, then it cannot be captured. However, if a comment received replies before it was deleted prior to the crawl, then the Reddit API will return `[deleted]` as the author and text. Deleted comments are ignored in all evaluations, but children of deleted comments are not ignored.

We mentioned earlier that Reddit has experienced remarkable growth in the past several years. In August 2013 Reddit reported 4.8 billion page views over 73 million unique visitors. This data is up from a reported 3.4 billion page views over 42.9 million unique visitors the prior year, August of 2012, according to the reddit blog[4].

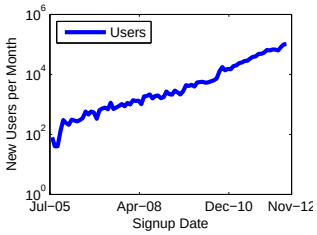

Fig. 1 Signup dates among currently active users in month-sized buckets (log scale).

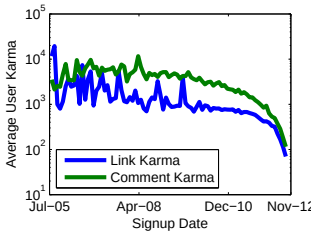

Fig. 2 Signup dates by karma rates for currently active users. (log scale)

Our dataset, however, only captures data a subset of registered users that contribute at least one post or comment in a top 25 subreddit during our crawl period. The crawling system captures the state of the each author/user at the time of the post or comments retrieval. This results in many users being recorded multiple times. Post and comment history, as well as the karma scores and other meta-data are associated with each registered user, and is frequently updated (*e.g.*, karma scores change with every vote), while certain meta-data, such as username and registration date, remain constant. Figure 1 shows the registration date of the users captured in our crawl; this figure demonstrates that either a) recently registered users were more active during the crawl period or b) the number of users is increasing dramatically or both. Note that we did not retrieve information from every user; instead, we only retrieved information from those users which were captured in the subset of popular posts and comments during in the crawl period. As a indication of the completeness of the user-data, we retieved data from more than 1.1 million unique users in total while Reddit reported that 1.6 million registered users logged in on the last day of the crawl period [5].

Among the users retrieved, it is reasonable to expect that users with earlier registration dates ought to have higher karma scores than newer users simply because they've had more time to accumulate karma and because karma is never spent. Figure 2 show that this is in fact the case because the karma rates for new users are lower than the karma rates for older users (from among all users captured).

4 Description of the Tasks

Here we describe the tasks that motivate our analysis. The first task is to analyze the topical structure and evolution of a comment thread; the second task looks to use comment threads as supplemental information to enhance Web search; and the third task attempts to distill features from the Reddit dataset in order to predict the final vote score of a comment.

[4] `http://blog.reddit.com/`

[5] `http://reddit.com/about` accessed 11/19/2013

Each task can stand alone, but viewed in aggregate the results may be able to illustrate the nature of Reddits discussion threads. In the final section of this paper, we describe how this analysis can be used to enhance future studies and systems.

4.1 Topical hierarchies and the evolution of a comment thread

Topical clustering algorithms, such as LDA [Blei et al. (2003)] and its hierarchical cousin hLDA [Blei et al. (2010)], have received a lot of recent attention both in research literature and in commercial system development. Hierarchical LDA, in particular, clusters words into hierarchical topics such that general words appear towards the top of the hierarchy, and specific words appear at the leaves of the hierarchy. Comment threads on Reddit are hierarchical, that is, a comment can be a reply to the post (a root comment) or a comment can be in reply to another comment. In this section, we investigate the extent to which topical hierarchies exist within comment threads. If we find that comment threads are topically hierarchical as we expect, then perhaps comment threads could be used to enhance future developments in topic models. On the other hand, if we find little or negative correlation between topic and discussion hierarchies, then we would need to rethink our assumptions about hierarchical topic models, discussion threads or both. We are also interested in how discussion topics evolve temporally and structurally. In temporal terms, we ask the question: does the discussion diversify as time passes? or does the discussion diversify immediately and then stay topically disjoint? In structural terms, we investigate the effect that a comments thread depth has on its topical granularity and its ultimate vote score.

4.2 Comment threads as supplemental information

The text of a comment thread is almost always relevant to the posted article or content. For example, if a user posts an article about the Obama versus Romney presidential debate, then its comments will most likely be about the presidential debate, the candidates positions, user opinion, etc. Under most circumstances the set of terms in the comment thread is much larger and generally more robust than the set of terms in the posted article or content especially when the posted content is a tweet or an image.

In this second task we ask two questions: (1) how much extra information do comment threads provide to the posted article or content, and (2) how does the comment thread effect Web search on the Reddit dataset. To answer the first question we create three term-document indexes: (1) content only, (2) comment only, and (3) a combined index made up of the first two indexes. We evaluate the degree to which comments supplement the content by measuring the number of results returned by various queries. To answer the second question we perform a user study to determine the average relevance, measured by normalized discounted cumulative gain (nDCG) and mean average precision (MAP), of the retrieved documents to a query set.

4.3 Predicting comment scores

Using the analysis from the first two tasks, Section 7 distills several features from post data, user information, and comment threads in order to develop a model capable of predicting the final vote score of a given comment. This section emphasizes feature development over predictive performance because we are most interested in performing a statistical analysis of Reddit, rather than building a robust prediction system.

5 Topical hierarchies and the evolution of a comment thread

This first subsection investigates the extent to which comment hierarchies exhibit a topical hierarchy.

This task is clearly important to the social media community, but it is also important to the topic modeling community because, to date, there is very little real-world data to collaborate the claims made by the topic modeling community, especially with respect to hierarchical topic models [Chang et al. (2009)]. If we find that comment threads are topically hierarchical as we expect, then perhaps comment threads could be used to enhance future developments in topic models. On the other hand, if we find little or negative correlation between topic and discussion hierarchies, then we would need to rethink our assumptions about hierarchical topic models.

We are also interested in how discussions topics evolve temporally and structurally. In temporal terms, we ask the question: does the discussion diversify as time passes? or does the discussion diversify immediately and then stay topically disjoint? In structural terms, we investigate the effect that a comment's thread depth has on its topical granularity and its ultimate vote score.

Previous studies have examined the structure of comment threads by analyzing the radial tree representation of thread hierarchies [Gómez et al. (2008)], via a text classification problem [Mishne and Glance (2006)], and by examining discussion *chains* [Laniado et al. (2011)]. A relevant study by Kaltenbrunner *et al.* on the hierarchical comments of Slashdot found that the volume of comments over time represented a lognormal distribution [Kaltenbrunner et al. (2008)].

Very little is known about the topical distribution of comment hierarchies. We hypothesize that comment threads are topically similar to the contributed content, and that subtopics emerge as discussion progresses and the thread hierarchies deepens.

5.1 Comment Threads over Time

Recall that our dataset contains the top-scoring posts from the most popular subreddits; thus the values in this section are likely to be inflated in comparison to less popular subreddits. In our dataset, posts received an average of 53 comments, and half of all posts receive 10 comments or fewer. A small number of highly discussed posts, however, can receive tens-of-thousands of comments, although in these cases we only collect the 500 highest scoring comments.

Figure 4 shows the number of distinct users and comments per posting. This figure shows a heavily tailed distribution similar to the findings of Laniado *et. al* on Wikipedia's discussion dataset [Laniado et al. (2011)].

However, a major difference is found in the tail of the distribution: there is a drastic uptick in the number of articles having between 475 and 500 comments (blue points). This is an artifact of how Reddit handles large numbers of comments and our data collection method. As a comment section grows and receives more votes the Reddit comment system hides many comments with low or negative scores from view. Furthermore, the maximum number of comments the Reddit API allows to be downloaded per comment section (without issuing prohibitively-many extra API calls) is 500. Thus, as the number of comments approaches 500 there is a higher likelihood of some comments being hidden because of poor vote totals until only the top/best-scoring 500 comments are shown.

The number of users per discussion (green points) exhibits a moderate deviation in the tail, that is, there are more discussions with 400 distinct users than with 350 distinct users. This is also a result of the data collection method. Top-scoring posts are ranked higher in the listing order; posts towards the top of the listing order are seen by more people; the more a post is seen, the more likely someone will read and comment on the post; thus, highly scored posts receive more comments than poorly scored posts. These reasoning is empirically observed in Figure 3, which clearly demonstrates that high scoring posts have, on average, a higher number of comments. Because the data collection step looks at the top-scoring posts every 6 hours, we are more likely to collect data from top-scoring posts (posts on the right side of Figure 3), which are more likely to have many comments. The steep decline in postings with between 480-500 distinct users solely is an artifact of the 500 comment collection limit. Simply put: it is rare to find a post with 500 comments from 500 distinct individuals.

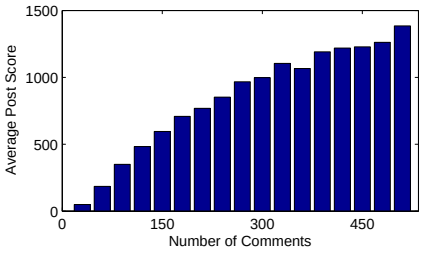

Fig. 3 Average number of comments as a function of the average post score (ups-downs). Higher scoring posts generally have more comments.

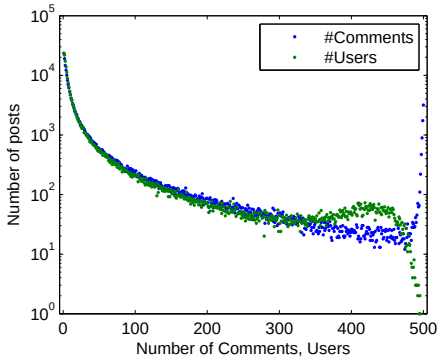

Fig. 4 Distribution of the number of comments and users per discussion thread

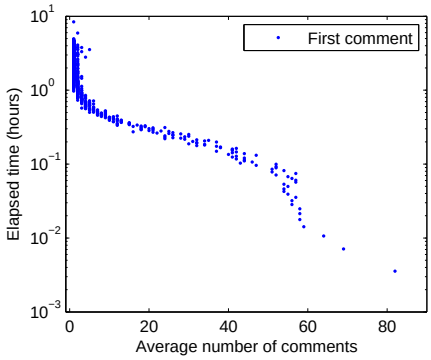

Fig. 5 Average number of total comments as a function of the elapsed time to the first comment

Timeliness matters. Figure 5 shows the average number of comments as a function of the elapsed time to the first comment. We find that when the first comment is submitted early-on in the post's life-cycle, then the post is likely to receive a large amount of comments. Conversely, when the first comment is submitted later in the post's life-cycle, then the post is not likely to have a large number of comments. This echos the results demonstrated by Szabo and Huberman on Youtube and Digg datasets [Szabo and Huberman (2010)]. This effect is causal because posts having a large (or small) number of comments must start with the first comment.

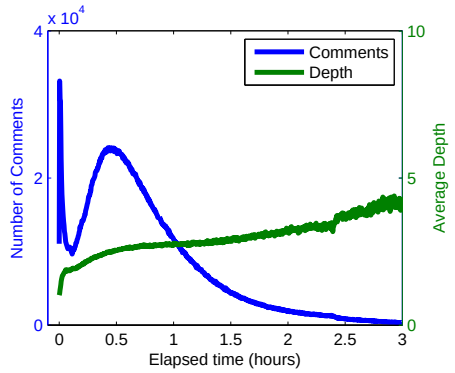

Fig. 6 Number of comments and average comment depth as a function of time

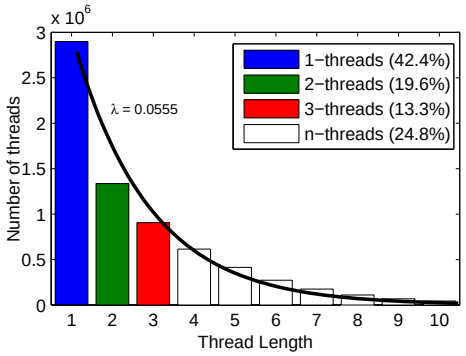

Fig. 7 Number of discussions at different depths

As time passes the number of comments ought to increase. Figure 6 shows the rate of commenting as a function of the elapsed time in hours (blue). We see that, in aggregate, there is a spike in extremely early commenting; these early comments come as soon as 1 to 5 seconds after the posting. After the initial surge the comment rate gradually rises and falls over the aggregate lifetimes of all posts. Except for the initial spike, our result are represent a lognormal distribution (with μ=4.618, σ=.2494) which are consistent with the results reported by Kaltenbrunner *et al* [Kaltenbrunner et al. (2008)].

The depth of a comment in the discussion hierarchy refers to the number of ancestors the comment has. Also in figure 6 we find that the average depth (green) steadily increases as the discussion progresses. The next subsection discusses the topicality of comments given their time and depth.

The density of discussions at progressive depths is illustrated in figure 7. Clearly, most comments are situated at the top level (depth of 1), and the number of comments at each successive depth trails off exponentially (λ=0.0555).

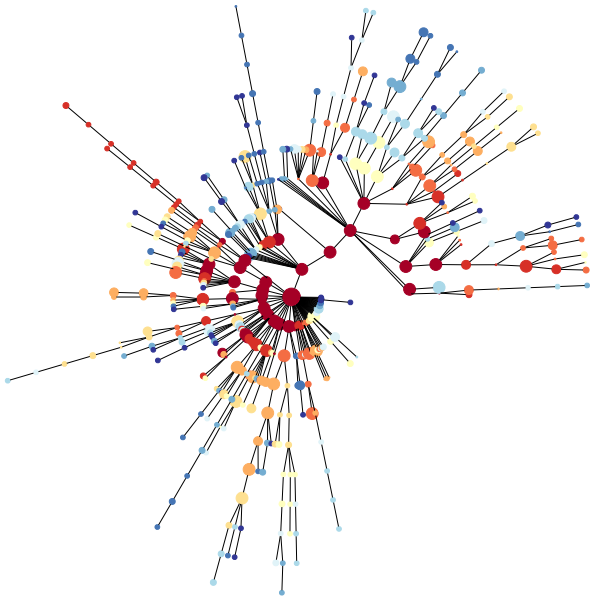

Fig. 8 Structure of a randomly selected comment thread[7]. Early comments are in bright colors, later comments are in dark colors. Node sizes indicate each comment's final vote score.

5.2 Structure of Comment Threads

As a comment thread evolves new comments are added in response to parent-comments, and users vote on older comments. The previous subsection showed aggregate statistics for thread depth and timeliness. Figure 8 illustrates a discussion thread for a randomly chosen post. In this illustration bright/red colors indicate early comments while dark/blue colors indicate later comments, and large circles indicate higher vote scores, while smaller circles indicate low (and sometimes negative) vote scores.

We see that many of the first-level comments are early comments, and the comments tend to become darker as their depth increases. Likewise, first-level comments are typically high scoring, and the comments tend have lower vote scores as their depth increases. Figure 8 also hints and an answer to one of our original questions: does the discussion diversify as time passes, or does the discussion diversify immediately? Observations from the radial comment thread illustration and Figure 6 show that subthreads (and presumably their subtopics) are started early in a post's life-cycle and *also* diversify further, creating sub-subthreads, later in the post's life cycle.

One particular sub-discussion on the right-hand side of the radial comment thread illustration in Figure 8 developed quickly, and has a comparatively broad fanout along with relatively high scores. In general, we find that Reddit discussions typically have one or two sub-threads that receive the most attention, by way of comments and votes,

Table 2 Truncated discussion thread showing topically narrow thread (top) and topically diverse thread hierarchy (bottom).

12 hottest years on record have come in the last 15 years
This is the best site to discredit climate deniers...
The reason people are skeptical is because they should be...
There is not one item in this response that even makes a serious attempt at making an argument...
The problem with skeptics of all kinds is that their approach is...
The [problem] in that argument is that facts show...
Too bad his “solution” is fracking and “clean” coal.
Clean coal lol
And a vast expansion in solar and wind energy over the past several years...
Wind and solar energy are inefficient, nuclear energy is where it is at.
I think people underestimate the influence of big oil over governments.
People also underestimate the influence of big oil over their own lives.

and these high-attention sub-threads usually develop relatively quickly.

5.3 Topical hierarchies

Previous figures show that as time progresses the average comment depth increases. We believe that this is, in part, an artifact of the nature of online discourse. More concretely, the results from the previous subsection suggests that when an online discussion first begins users contribute top-level comments that often initiate various threads of discourse. Based on these observations we ask: do hierarchical threads, like those on Reddit a) demonstrate a hierarchy of topics; or b) do hierarchical threads present a flat or narrowing topical representation.

For example, an illustration of the two types of threads is found in Table 2. The first discussion is a debate between and about climate change skeptics - a relatively narrow topic with back-and-forth rebuttals, etc. The second discussion is more topically diverse, and the topics continue to diversify into subtopics as the comment hierarchy deepens. Specifically, the root comment talks about the article’s proposed solution, this topic is then subsumed by discussion on wind and solar energy in one subthread and oil in another subthread, which is further diversified into nuclear alternatives instead of solar/wind, etc.

Unlike this small, truncated example, actual comment threads can contain thousands of comments and deep and broad thread-trees. In this subsection we investigate the extent to which threads trees are topically hierarchical. Fortunately, recent advances in hierarchical topic models allow for a systematic, quantitative evaluation of the topical distributions in text hierarchies.

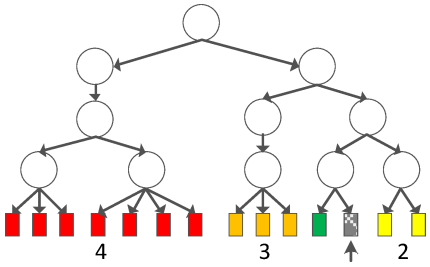

Fig. 9 Illustration of 4 level hLDA output. Green, yellow, orange, red indicate most topically similar to least topically similar.

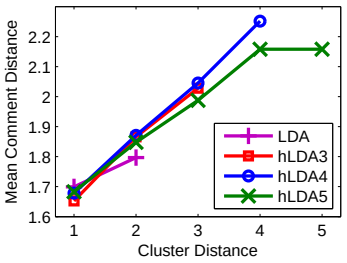

Fig. 10 Average distance between comments as a function of cluster distance

Latent Dirichlet Allocation (LDA) [Blei et al. (2003)] and its nonparametric/hi-erarchical extension (hLDA) [Blei et al. (2010)] are two commonly used probabilistic topic models. Given a set of documents hLDA hierarchically clusters comments/documents so that topically similar documents share the same topic-parent, less-similar comments share topic-grandparents, etc. In essence, the topical distance between two comments can be measured by the tree-distance in the hLDA output; sibling-comments have more in common than cousins, who have more in common than second-cousins, etc.

Figure 9 shows an example output of the hLDA algorithm. This figure illustrates, with respect to a given document/comment (indicated by the arrow at center-right), that comments that are most topically similar are siblings, colored in green. The next most similar set of documents are first-cousins, colored in yellow. Followed by less-similar second-cousins in orange, and most dissimilar third-cousins in red. In general, comments that are topically dissimilar share distant ancestors in the hLDA output tree.

The goal, therefore, is to measure if and how topics diverge as discussion threads deepen. This measurement is accomplished by a straightforward methodology. First, we randomly sample 10,000 postings resulting in 429,041 comments. For each post we extracted all of the comments (up to the limit of 500 if necessary). hLDA was run on each set of comments for 5,000 Gibbs iterations and the hLDA output tree with the highest log likelihood was captured as the

[7] `http://redd.it/100icq` – “Former National Security Agency official Bill Binney says US is illegally collecting huge amounts of data on his fellow citizens — The Guardian”

[6] Full discussion available at `http://redd.it/18l9je/`

output model. This was done with hLDA at varying heights of 2, 3, 4 and 5.

At this point, for each height, we have 10,000 different hLDA output trees dist_c with an average of 42 comments at each trees' leaves. For each output tree, we measured the distance between each pair of comments in the hLDA output tree, where a sibling (green) has a distance of 1, a cousin (yellow) has a distance of 2, and so on. This resulted in n^2 distance measurements for each comment thread (*i.e.*, on average 42^2 measurements).

Recall that each comment originally had a place in the discussion thread, which is also a tree structure. Unlike the hLDA output tree, in the a comment thread comments can live at inner-nodes as well as leaf nodes. Thus pairwise-distance is calculated by the distance to the least common ancestor $dist_s$ – a similar, yet slightly different measure than sibling, cousin, etc. Therefore, each comment thread also has n^2 distance measurements.

Each pair of comments now has a hLDA-based cluster/topical distance $dist_c$ of 2, 3, 4 or 5 (where the maximum possible distance depends on the manually defined depth of the hLDA output tree, that is, the maximum distance is a tree of depth x is x) and a structural-based thread distance $dist_s$. For each manually-defined hLDA depth we average all $dist_s$ for each dist_c and plot the results. For example, we average all of the $dist_s$ where $dist_c = 1$, and then all of the $dist_s$ where $dist_c = 2$ and so on.

If discussion threads exhibit a topic hierarchy, then topically similar comments should appear in the same or similar hLDA clusters. If comments threads do not exhibit a topical hierarchy, then we expect to find a low correlation between the comment thread distance and the topical cluster distance, and vice verse.

Figure 10 shows the results aggregated from all 10,000 posts of these measurements. Recall that hLDA with a depth of 3 can only show results for cluster distances of 1, 2 and 3 because the maximum cluster distance is 3; in general hLDA with a manually defined depths of x can have a $dist_c$ of at most x. Comments that are siblings (green) thus having a low $dist_c$ in the hLDA output trees have, on average, a small $dist_s$ in the structured discussion threads. This shows that, in the aggregate, comments in a discussion thread that are structurally near each other are also topically similar. These results seem to show that thread structures correlate to thread topicality. In other words, thread hierarchies tend to exhibit a topical hierarchy in the general case. We stress that these measurements are for the general case; there are certainly cases in which the opposite is true.

6 Comments as Supplementary Information

The previous section provides some insight into the nature of user comments in a social news site. Next, we focus on evaluating what effect, if any, comments have on search quality. For this search evaluation task we collected a set of 88 queries from the New York Times Web site's most frequent queries list during a 6 day period from Oct. 11 through Oct. 17, 2012, a temporal subset of the entire Reddit-crawl. Example queries from this set include "lance armstrong", "health care", "felix baumgartner", "nobel prize", and "obama romney debate".

An initial analysis of Web log comments by Mishne and Glance [Mishne and Glance (2006)] found that search recall can be improved by indexing user comments as well as the blog or post text. This is not a surprising result because any amount of extra text would almost certainly result in more results. They further argue that recall is more important than precision in the context of Web log retrieval because search results are typically sorted by most recent, rather than by relevance. Unlike previous Web log studies, the Reddit dataset mostly contains posted external content rather than self-authored blogs.

We adopted the recall evaluation from Mishne and Glance by crawling and indexing the external content and creating three different indexes 1) a content-only index, 2) a comment-only index, and 3) a combined index of both content and comment data. For each query, we compared the list of results from each of the three indexes. For example, the query "health care" retrieved 63,871 total results from the combined index. Of these, 33,366 (55.8%) were retrieved from the content index, 40,175 (62.9%) were retrieved from the comment index. Among these, 11,670 (18.7%) results were retrieved from both indexes.

Table 3 Example of Recall Contribution Setup. Letters A–F indicate a total of 6 results from the combined index; 5 of which are from the content index, 3 of which are from the comment index, and 2 were from both indices.

Content	Comments	Overlap
A	D	D
B	E	E
C	F	
D		
E		
5 (83.3%)	3 (50.0%)	2 (33.3%)

In this way, we are able to determine the amount of supplemental information that exists about a post in its comment thread. A high overlap would indicate that the content and comments are very similar, while a low overlap would indicate that the content and comments contribute different sets of information (via the terms/words that are used) to the user.

Table 4 shows the aggregate results over all 88 queries. We find that the comments make a large contribution to the raw the number of search results. We show a 36% average

Table 4 Recall Contribution of content and comments

	Content	Comments	Overlap
Mean	73.42%	35.87%	9.29%
StdDev	14.87%	18.78%	5.10%
Median	74.03%	37.54%	9.19%
Minimum	35.04%	0.36%	0%
Maximum	99.64%	75.50%	23.12%

comment contribution while previous results on a Web log corpus from the Mishne and Glance study showed only a 6.4% average comment contribution [Mishne and Glance (2006)]. These major differences in results are either due to differences between Web log and social news Web sites and/or because of an increased rate in user-engagement in recent years.

There is a stark difference in the minimum and maximum contributions too. The query "Ebay" resulted in the maximum share of comment contribution, and thus the lowest content contribution, and the query "Rosneft" resulted in the the minimum share of comment contribution, and thus the highest content contribution. These min/max results can be attributed to the general popularity of Ebay, as well as the relative obscurity of Rosneft, a Russia-based oil company, especially among Reddit's young, tech-savvy demographic.

We also note that the high standard deviation of the comment contribution indicates that comment content is especially important for some queries to achieve complete search results.

6.1 Comments to improve general search

The previous section shows that the inclusion of comments significantly boosts search recall. Previous experience suggests that as recall increases the precision ought to decrease. In general, this is because larger result sets provides a greater opportunity include spurious entries, thereby decreasing precision. In this subsection we evaluate the effect comment threads have on search results using standard nDCG and MAP metrics.

For the evaluation we use the same set of 88 queries as before, and employ the BM25F ranking function, which is a straightforward modification of the original BM25 [Robertson and Walker (1994)] ranking function that weights two or more fields with different degrees of importance. For our purposes we consider a "document" to consist of a content-field and a comment-field; both fields use the bag-of-words model. As a result we have the following ranking function known as BM25F [Zaragoza et al. (2004)]:

$$\text{BM25F}_{\text{mix}} = \lambda \text{BM25}_{\text{content}} + (1-\lambda)\text{BM25}_{\text{comment}} \quad (1)$$

where BM25F is essentially a weighted combination of *fields* within Robertson and Walker's original BM25 heuristic. By changing the λ value we can evaluate the contributions each field makes towards the search results. BM25-specific parameters were manually set to $k = 1.2$, $b = 0.75$ and were not empirically tuned.

To measure query performance, we obtained the top 100 results for each query using $\lambda = 0.0$; this weighting effectively ignored the comment field and used only information from the post's content. Mechanical turk was used to generate relevance scores. We use the same experimental setup used in other, similar studies. Each query result was judged by 5 separate turkers on a scale of 1 to 4, with 1 being not relevant at all and 4 being very relevant. In order to receive quality judgments we manually judged 75 easy results (gold results); if a turker did not judge 90% of the gold results correctly, then all of his judgments are thrown out and he is not paid. We obtained the median score from the 5 judges for each result. This resulted in 8,800 median judgments for $\lambda = 0.0$.

Next, top 10 results for $\lambda = 0.05, 0.10, \ldots, .95$ were generated, and relevance judgments from the 8,800 original judgments were applied when possible. We found 443 new results that were not judged in the original mechanical turk evaluation. Therefore a second mechanical turk evaluation was conducted to generate relevance scores for the 443 additional results using the same methodology as the first mechanical turk evaluation. The result of this setup is a set of high quality relevance judgments for the top 10 results of the BM25F ranking function for 20 λ values. Results for $\lambda = 1.0$ were not evaluated because there was very little overlap between the results from $\lambda = 1.0$ and the $\lambda = 0.0$ results that were manually evaluated by turkers; a proper evaluation of the $\lambda = 1.0$ results would require an extra set of mechanical turk evaluations thereby doubling the expense of the overall experiment.

We measure the performance of each query using mean average precision (MAP) at k and normalized discounted cumulative gain (nDCG) at k [Jarvelin and Kekalainen (2002)].

The mean average precision at k is the ratio of the number of relevant documents found in the top k results to the total number of relevant documents or k, whichever is smaller, averaged over all queries:

$$\text{MAP}_k = \frac{\sum_{q=1}^{Q} \left(\frac{1}{k}\sum_{i=1}^{k} i/r_i\right)}{Q}, \quad (2)$$

where r_i is the rank of the ith relevant document in the result list, and Q is the set of queries. One disadvantage of MAP is that it cannot measure differences in relevance, so we assume that judgments of 3 or 4 are relevant, and judgments of 1 or 2 are not relevant. The nDCG measure generalizes the MAP-score to account for the 1 to 4 relevant scores used by the turkers. The general form of nDCG is:

$$\text{nDCG}_k = \frac{\sum_{q=1}^{Q} \frac{r_1 + \sum_{i=2}^{k} \frac{r_i}{\log_2 i}}{IDCG}}{Q}, \quad (3)$$

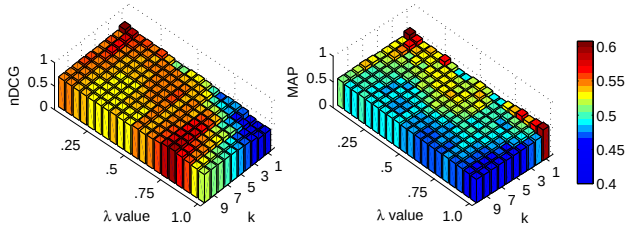

Fig. 11 NDCG scores (left) and MAP scores (right) per variations of λ and k. $\lambda = 0$ means content-only, $\lambda = .95$ means comment-only. Higher is better.

where IDCG is the *ideal* discounted cumulative gain, which assumes a perfect ordering of the top k best results.

Figure 11 (at left) shows the nDCG_k scores for each value of k as λ alternates between 0.0 and 0.95. Figure 11 (at right) shows the MAP_k scores for each value of k as λ alternates between 0.0 and 0.95.

For nDCG results, we find that although content-heavy weights ($\lambda \approx 0.0$) results in the best scores, λ weights near 0.8 also perform very well at higher k values. Results from MAP metric are less encouraging except that the Prec@1 scores at $\lambda = .95$ are almost as high as the content-only Prec@1 score.

These results demonstrate that the inclusion of comments are indeed detrimental to the precision of search results. However, the nDCG and MAP metrics are unable to communicate some interesting properties of search results from comment-heavy weightings. For example, we find that top results in comment-heavy search rankings ($\lambda \approx .95$) have (1) a greater likelihood of being images and (2) are more likely to be from non-mainstream media sources.

For practical purposes the inclusion of comments in a search index can be helpful when ranking is based on timeliness or in other instances when recall is most important. When ranking based on general query relevance a very low, yet non-zero, comment weight would dramatically increase recall without hurting precision too much.

7 Predicting Comment Value

The previous sections presented several statistical observations that may be able to aid in the development of a model that predicts a comment's value. In this section we extract and explore several pertinent features that are correlated to the final score of a given comment, where a comment's *final score* is the number of upvotes minus the number of downvotes received after 48 hours. Higher comment scores are generally viewed as having a higher value to the Reddit community than low-scoring comments; as such, Reddit, by default, lists comments on its Web site ordered by the score.

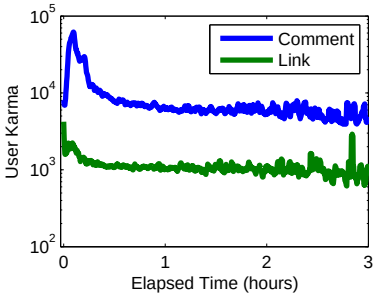

Fig. 12 User karma as a function of elapsed time to comment.

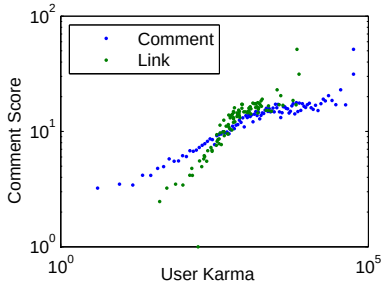

Fig. 13 Comment score as a function of user post and comment karma

7.1 Features used for learning

We explore four different sets of features (18 features in all) that describe various facets of a given comment. Recall that our goal here is not necessarily to develop a robust prediction system, but rather to explore the space of features and their relative predictability.

The first set of features we consider are **Commenter Features** (S_A), 8 features total: number of days user has been registered, link karma, comment karma, total number of comments, total number of upvotes and downvotes, average upvotes and downvotes. Commenter features encapsulate information about the specific user who is submitting the comment. The intuition behind this set of features is that highly reputable commenters are likely to contribute high quality, and therefore high scoring, comments, while unknown or poorly reputed commenters will contribute average or poor quality comments.

Recall that a user's karma is the summation of the user's previous scores. Specifically, comment karma is the total score of the user's comments, and link karma is the total score of the user's posted links (self-posts, *i.e.*, user generated posts without an external link, do not count towards post-karma). Figure 13 (at right) shows comment scores as a function of user link and comment karma. They are both clearly correlated in log-space: users' comment karma has a tighter correlation with comment score ($\rho = 0.957$) than post karma ($\rho = 0.923$). A modest correlation also exists in linear space where comment karma is correlated with comment score at $\rho = 0.775$, and post karma is correlated with

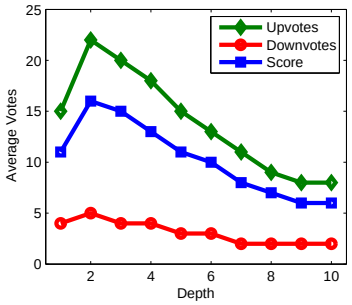

Fig. 17 Average votes as a function of thread depth

comment score at $\rho = 0.765$. In general, this means that comments are better than posts at indicating future comment scores.

The second set of features are called **Post Features** (S_B), 2 features total: total upvotes, total downvotes. Post features encapsulate the vote totals of the post at the time that the comment is submitted. The intuition behind post features is that popular posts attract more comments and more votes because popular posts are ranked higher resulting in a larger readership.

The third set of features are called **Comment Structure Features** (S_C), 6 features total: number of seconds after parent comment, number of seconds after posting, depth of the comment, parent's upvotes, parent's downvotes, parent's score (upvotes-downvotes). This set of features contains information regarding the proposed comment's place within the entire thread. The context of a comment is likely to be an important indicator of its final score. Intuitively, if a comment is surrounded by high quality comments, then it may "ride the coattails" of its ancestor and/or neighbor comments and receive many upvotes.

Figure 14 illustrates a comment's final upvotes, downvotes and score as a function of it's parent's score at the time of submission. This illustration shows a clear correlation: comments with a large number of downvotes have parents with a large number of downvotes, while comments with a large number of upvotes have parents with a large number of upvotes. Because we only looked at the parent's score at the time of comment submission instead of the parent's *final* comment score, we can also deduce a causal property from this graph: high parental scores cause high comment scores. There may also be a mutual causal effect, but it cannot be determined from Figure 14 alone.

Figure 17 shows the final number of upvotes, downvotes and score of the average comment as a function of its depth in the comment thread. Interestingly, the first-level comments received a lower score than the second-level comments, but after the second level the scores diminish as the comment depth grows deeper. Also notice that the number of downvotes decreases as the depth increases as well. From this graph we deduce that deeper comments are not necessarily of lower-quality, instead they simply receive fewer votes presumably because readers don't read an entire discussion thread and/or users read a discussion one time before all the comments have been posted as indicated by Figure 16.

Aside from depth, the timeliness of a comment is essential to its ultimate score. Figure 15 shows two distinct trends over the same time. The y-axis at left indicates the comment volume; we find that comments frequently occur very early in the life-cycle of a post, slow down for the next 15 minutes, and then increase again. This second "bump" in the comment volume can be attributed to a post making the "front page" of the subreddit. This is akin to virility on the Web wherein more users are likely to view and comment on a post once it reaches a certain critical mass; within the confines of Reddit, a post reaches its critical mass when it is listed on the front page. Recall that our dataset contains the top 100 posts for a given 6 hours time period. Therefore, many of the collected posts will exhibit this type of comment distribution.

Interestingly, comments which receive the highest score are most frequently submitted during the 15 minute low-point in the comment volume. In other words, Figure 15 shows that the best comments are submitted at the time of fewest submissions. The graph also shows, counter-intuitively, that the first comment(s) are not always the highest rated. We have several possible, yet unstudied, explanations for these observations.

One plausible explanation is that Reddit contains a small set of *power-users* who frequently check the queue for new content. When an interesting new post arrives the power-users are among the first (but perhaps not the actual first) to upvote and comment what will eventually be a popular, front-page post. Upon further investigation, we find that the Reddit community affectionately titles these power-users the "knights of new" because they are assumed to be the ones who sift through the vast numbers of low-quality posts and collectively upvote worthy posts. Of course, once a post receives enough votes to be listed on the front-page, then the broader user community will vote on the post's ultimate fate.

The fourth set of features are called **Comment Syntax Features** (S_D), 2 features total: number of characters, and number of words. In a given comment the number of words and characters might also have some predictability. Perhaps pithy comments receive high scores, or perhaps lengthy, detailed comments receive high scores.

7.2 Results

We use the four sets of features to induce a linear regression model that predicts the final score of a given comment.

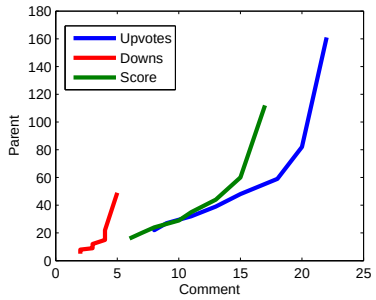

Fig. 14 Comment votes as a function of parent votes

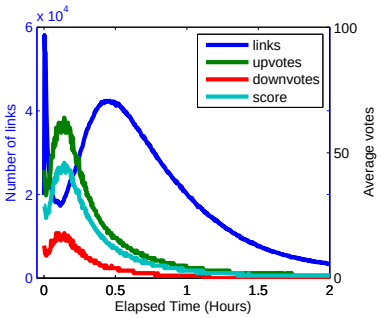

Fig. 15 Volume of comments and average votes as a function of time.

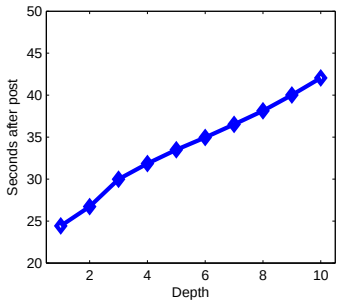

Fig. 16 Average elapsed time of comment per depth.

Table 5 Relative importance of features for predicting comment score.

Feature	Coefficient
Mean num. upvotes of author's comments	+0.5974
Mean num. downvotes of author's comments	-1.1495
Depth of current comment	+0.7296
Number of upvotes of parent	+0.1706
Length of comment (# characters)	+0.0269
Length of comment (# words)	-0.1422

In order to avoid biases while learning the model, we made a special effort to separate the feature set from the learnable score (class variable). For example, we made sure to use data from the commenter, post, structure, and syntax as it appeared at the time the comment was being submitted. Concretely, if a comment c for post X was submitted at time t, then the feature set of c is created from X's data at time $t-1$. This separation simulates a real world prediction system and keeps the experiment setup realistic.

The training set is comprised of 5000 randomly selected comments created on or before August 23, 2012, the test set contains 5000 similarly selected comments created after August 23, 2012.

Due to the quick lifespan of a post and its comment thread, we consider post complete after 48 hours have passed. This is a fair assumption because the Reddit system removes posts from the front-page after only 24 hours, and Figure 15 shows that most comments are made within the first four hours.

First we formulate the task of predicting the comment score as a linear regression task, and report the results using mean squared error (MSE). A linear regression classifier was learned using the full feature set ($S_A \cup S_B \cup S_C \cup S_D$) and found a core set of 6 features that are statistically correlated with the comment's final score. Table 5 shows these 6 features and their coefficients. Of these features, the mean number of commenter's previous comment upvotes, depth of the current comment, comment parent's upvotes, and the comment character length are positively correlated; the mean number of commenter's previous comment downvotes, and the comment word length are negatively correlated.

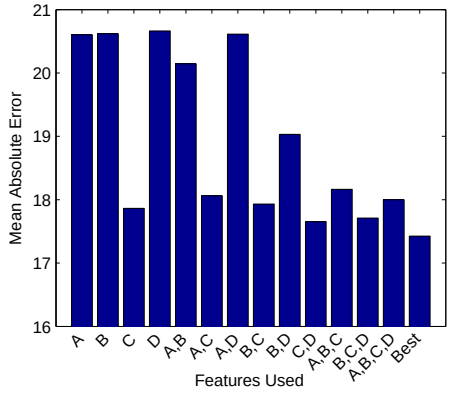

Fig. 18 Error results of comment score prediction. Lower is better

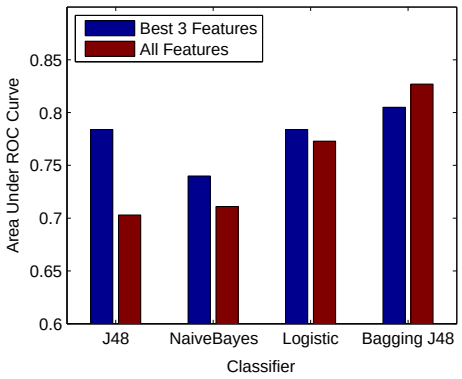

Fig. 19 Area Under ROC Curve for different classifiers with all features and best 3 features. Higher is better.

One interesting result is that the length of a comment in total number of characters is positively correlated, but the length measured in number of words is negatively correlated. We deduce from these statistics that comments that contain big words are more likely to have higher final score than comments that contain smaller words.

Figure 18 shows the mean absolute error of a linear regression model trained with different combinations of features. The mean absolute error (MAE), in general, measures

how close the predictions are to the acutal outcomes. MAE takes the following form:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|, \quad (4)$$

where f_i is the forecasted value and y_i is the actual outcome. The difference between forecast and actual can also be called the error $e_i = f_i - y_i$. Thus, the MAE is the mean average of the absolute value of the error.

Clearly the set S_C, containing thread structure features, contains the most predictive features. An exhaustive search of all possible feature combinations found that the combination of individual features receiving the lowest mean square error contains: (1) the mean upvotes of the authors previous comments, (2) the elapsed time since the post's submission, and (3) the number of upvotes of the submitted comments' parent. This is indicated by "best" in Figure 18.

Next we discretize the class variable, final comment score, by labeling the class variable 'low' for comment scores less than 5, 'medium' for scores less than 10, and 'high' for all other scores. Nominal class variables allow a larger set of classifiers to be used, as well as measurements using precision and recall metrics.

Figure 19 shows the area under the ROC curve scores for J48, Naive-Bayes, Logistic and Bagged J48 classifiers each with the full feature set and the best 3 features determined earlier. We find that the best three features outperformed the full set in all cases except for the bagged J48 classifier. We were not surprised to find that bagging significantly improves the classifier trained on all features because bagged decision trees with several features generally show significant improvement from the non-bagged classifier.

Results show that the structure features of a comment's thread are good indicators for future value. Furthermore, the commenter's past comment scores are also good indicators for future value, a result shared by a study of question-answering sites [Anderson et al. (2012)].

Recall that the data set used in these experiments are baised towards successful posts and comments. Thus, the conclusions drawn from these results must be made with the biases in mind. In the case of learning a regression model or decision tree, the specific values for each feature are not shown in Table 5 because they are sure to be biased by the data set. Instead, we show only the coefficients to give demonstrate the relative correlation of the most correlated features (both positive and negative correlation).

8 Conclusions

We conclude by revisiting the original questions raised at the beginning of this work.

Regarding the structure and evolution of a comment thread, we observe that, in general, hierarchical comment threads consist of top level comments that start a subtopic. We also observe that these top level comments, especially those which receive a large number of replies, are usually created during the early stages of the post's life cycle. From among the early, top-level comments/subtopics further sub-subtopics are created as a natural part of online discourse. In plain terms, we present strong evidence that hierarchical comment threads on Reddit represent a topical hierarchy. An anecdote to topic divergence is the rise of the Internet-slang, *thread hijacking*, in which a group of users deviate so far off topic as to warrant the creation of an entirely new post.

We also demonstrate that comments can be used to substantially enhance the recall of Web search without severely degrading the precision. Interestingly, we found that the degree to which comments increase the recall is substantially greater than those reported in previous work [Kaltenbrunner et al. (2008)]. These results demonstrate that comment threads do contain a large amount of supplemental information.

We show that certain features are excellent predictors of a comments eventual vote score. The context and timing of a submitted comment are found to be the most indicative of its final score. Our experience with Reddits comment threads indicate that this is no secret: astute Reddit users are sometimes known to comment on the highest scoring subthread instead of the most topical. This practice increases the comments visibility because comments are listed by the order of their scores, thereby rendering the comment more likely to receive votes.

Finally, we encourage readers to use the information presented in this paper to inform their future works. For example, the discussion threads and edit history of Wikipedia have been used in role-finding [Welser et al. (2011)], quality assessment [Kittur and Kraut (2008)], content enhancement [Schneider et al. (2011)], and for dozens of other purposes. We believe that the comment threads from Reddit can serve a similar role by annotating its linked-content. One important aspect of the Reddit site that we did not address in this paper is the topical differences among different subreddits. We believe that different subreddits can serve to inform separate language and network models for further community detection, document labeling, and so on.

A recent decision by Popular Science to turn off its comment section

The data and source code used in these experiments is available from the author's Web page.

Acknowledgements We thank Reddit for allowing us to crawl and curate their user data. The author is not affiliated with Reddit in any way.

References

[Adamic et al. (2008)] Adamic LA, Zhang J, Bakshy E, Ackerman MS (2008) Knowledge sharing and yahoo answers. In WWW, ACM Press, p 665

[Anderson et al. (2012)] Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Discovering value from community activity on focused question answering sites. In: SIGKDD, ACM Press, p 850

[Asur and Huberman (2010)] Asur S, Huberman BA (2010) Predicting the Future with Social Media. In: WI-IAT

[Bandari et al. (2012)] Bandari R, Asur S, Huberman BA (2012) The Pulse of News in Social Media: Forecasting Popularity. In: ICWSM

[Bernstein et al. (2011)] Bernstein MS, Monroy-Hernández A, Harry D, André P, Panovich K, Vargas G (2011) 4chan and /b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In: ICWSM, pp 50–57

[Blei et al. (2003)] Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. The Journal of Machine Learning Research 3:993–1022

[Blei et al. (2010)] Blei DM, Griffiths TL, Jordan MI (2010) The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM 57(2):1–30

[Bross et al. (2012)] Bross J, Richly K, Kohnen M, Meinel C (2012) Identifying the top-dogs of the blogosphere. Social Netw. Analys. Mining 2(1):53–67

[Chang et al. (2009)] Chang J, Gerrish S, Wang C, Blei DM (2009) Reading Tea Leaves : How Humans Interpret Topic Models. In: NIPS 31:1–9

[Cong et al. (2008)] Cong G, Wang L, Lin CY, Song YI, Sun Y (2008) Finding question-answer pairs from online forums. In: SIGIR, ACM Press, p 467

[Doan et al. (2011)] Doan A, Ramakrishnan R, Halevy AY (2011) Crowdsourcing systems on the World-Wide Web. Communications of the ACM 54(4):86

[Duan and Zhai (2011)] Duan H, Zhai C (2011) Exploiting thread structures to improve smoothing of language models for forum post retrieval. In: ECIR, pp 350–361

[Fisher et al. (2006)] Fisher D, Smith M, Welser H (2006) You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In: HICSS, IEEE, pp 59b–59b

[Gilbert (2013)] Gilbert E (2013) Widespread underprovision on Reddit. In: CSCW, ACM Press, p 803–808

[Gilbert et al. (2011)] Gilbert F, Simonetto P, Zaidi F, Jourdan F, Bourqui R (2011) Communities and hierarchical structures in dynamic social networks: analysis and visualization. In: Social Netw. Analys. Mining 1(2):83–95

[Gómez et al. (2008)] Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in slashdot. In: WWW, ACM Press, p 645

[Hong et al. (2011)] Hong L, Yin D, Guo J, Davison BD (2011) Tracking trends. In: SIGKDD, ACM Press, p 484

[Jarvelin and Kekalainen (2002)] Jarvelin K, Kekalainen J (2002) Cumulated gain-based evaluation of IR techniques. In: ACM Transactions on Information Systems 20(4):422–446, ACM Press

[Kaltenbrunner et al. (2008)] Kaltenbrunner A, Gómez V, Moghnieh A, Meza R, Blat J, López V (2008) Homogeneous temporal activity patterns in a large online communication space. International Journal on WWW/INTERNET 6(1)

[Kawamae and Higashinaka (2010)] Kawamae N, Higashinaka R (2010) Trend detection model. In: WWW, ACM Press, p 1129

[Kittur and Kraut (2008)] Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in wikipedia. In: CSCW, ACM Press, p 37

[Lakkaraju et al. (2013)] Lakkaraju H, McAuley J, Leskovec J (2013) What's in a name? Understanding the Interplay between Titles, Content and Communities in Social Media. In: ICWSM, p 311–320

[Lampe and Resnick (2004)] Lampe C, Resnick P (2004) Slash(dot) and burn. In: SIGCHI, ACM Press, pp 543–550

[Laniado et al. (2011)] Laniado D, Tasso R, Volkovich Y, Kaltenbrunner A (2011) When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In: ICWSM, pp 177–184

[Lerman and Galstyan (2008)] Lerman K, Galstyan A (2008) Analysis of social voting patterns on digg. In: WOSP, ACM Press, p 7

[Lerman (2007)] Lerman K (2007) Social information processing in social news aggregation. In: IEEE Internet Computing: special issue on Social Search, 11(6):1628

[Lerman (2007b)] Lerman K (2007b) User participation in social media: Digg study. In: WI-IAT Workshop on Social Media Analysis

[Leskovec et al. (2007)] Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007) Cascading Behavior in Large Blog Graphs. In: SDM

[Leskovec et al. (2009)] Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: SIGKDD, ACM Press, p 497

[Mejova et al. (2013)] Mejova Y, Srinivasan P, Boynton B (2013) GOP primary season on twitter. In: WSDM, ACM Press, p 517

[Mishne and Glance (2006)] Mishne G, Glance N (2006) Leave a Reply: An Analysis of Weblog Comments. In: WWE

[Muchnik, et al. (2013)] Muchnik L, Aral S, Taylor S (2013) Social Influence Bias: A Randomized Experiment. In: Science, 341(6146):647-651

[Mukherjee and Liu(2012)] Mukherjee A, Liu B (2012) Mining contentions from discussions and debates. In: SIGKDD, ACM Press, p 841

[Paul, et!al. (2012)] Paul SA, Hong L, Chi, EH (2012) Who is Authoritative? Understanding Reputation Mechanisms in Quora. In: Collective Intelligence

[Robertson and Walker (1994)] Mukherjee A, Liu B (2012) Some simple effective approxmiations to the 2-Poisson model for probabilistic weighted retrieval. In: SIGIR, ACM Press, p 232–241

[Schneider et al. (2011)] Schneider J, Passant A, Breslin JG (2011) Understanding and improving Wikipedia article discussion spaces. In: SAC, ACM Press, p 808

[Seo et al. (2009)] Seo J, Croft WB, Smith DA (2009) Online community search using thread structure. In: CIKM, ACM Press, p 1907

[Szabo and Huberman (2010)] Szabo G, Huberman BA (2010) Predicting the Popularity of online content. In: Comm. of the ACM 53(8):80–88, ACM Press

[Tsagkias et al. (2009)] Tsagkias M, Weerkamp W, de Rijke M (2009) Predicting the volume of comments on online news stories. In: CIKM, ACM Press, p 1765

[Wang et al. (2011)] Wang H, Wang C, Zhai C, Han J (2011) Learning online discussion structures by conditional random fields. In: SIGIR, ACM Press, p 435

[Wang et al. (2012)] Wang C, Ye M, Huberman BA (2012) From user comments to On-line Conversations. In: SIGKDD, ACM Press, p 244–252

[Welser et al. (2011)] Welser HT, Cosley D, Kossinets G, Lin A, Dokshin F, Gay G, Smith M (2011) Finding social roles in Wikipedia. In: iConference, ACM Press, p 122–129

[Zaragoza et al. (2004)] Zaragoza H, Craswell N, Taylor M, Saria S, Robertson S (2004) Microsoft Cambridge at TREC-13: Web and HARD tracks. In: TREC, ACM Press

[Zhu (2010)] Zhu Y (2010) Measurement and analysis of an online content voting network. In: WWW, ACM Press, p 1039