# A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web

ELAHEH MOMENI, Faculty of Computer Science, University of Vienna
CLAIRE CARDIE, Departments of Computer Science and Information Science, Cornell University
NICHOLAS DIAKOPOULOS, College of Journalism, University of Maryland

User-generated content (UGC) on the Web, especially on social media platforms, facilitates the association of additional information with digital resources; thus, it can provide valuable supplementary content. However, UGC varies in quality and, consequently, raises the challenge of how to maximize its utility for a variety of end-users. This study aims to provide researchers and Web data curators with comprehensive answers to the following questions: What are the existing approaches and methods for assessing and ranking UGC? What features and metrics have been used successfully to assess and predict UGC value across a range of application domains? What methods can be effectively employed to maximize that value? This survey is composed of a systematic review of approaches for assessing and ranking UGC: results are obtained by identifying and comparing methodologies within the context of short text-based UGC on the Web. Existing assessment and ranking approaches adopt one of four framework types: the community-based framework takes into consideration the value assigned to content by a crowd of humans, the end-user–based framework adapts and personalizes the assessment and ranking process with respect to a single end-user, the designer-based framework encodes the software designer's values in the assessment and ranking method, and the hybrid framework employs methods from more than one of these types. This survey suggests a need for further experimentation and encourages the development of new approaches for the assessment and ranking of UGC.

Categories and Subject Descriptors: H.3.m [**Information Storage and Retrieval**]: Miscellaneous; H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.2.m [**Database Management**]: Miscellaneous

General Terms: Algorithms, Design, Documentation, Experimentation, Reliability

Additional Key Words and Phrases: User-generated content, ranking, assessment, social media, adaptive, interactive, machine centered, human centered

**ACM Reference Format:**
Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A survey on assessment and ranking methodologies for user-generated content on the Web. ACM Comput. Surv. 48, 3, Article 41 (December 2015), 49 pages.
DOI: http://dx.doi.org/10.1145/2811282

## 1. INTRODUCTION

User-generated content (UGC) on the Web, and on social media platforms in particular, is a vital part of the online ecosystem [Asselin et al. 2011; Rotman et al. 2009; Rangwala and Jamali 2010]. UGC is the foremost mechanism for participants to comment on, enhance, and augment social media objects ranging from YouTube videos, Flickr

Authors' addresses: E. Momeni, University of Vienna, Waehringer Strasse 29/5, A-1090 Vienna, Austria; email: elaheh.momeni.roochi@univie.ac.at; C. Cardie, Cornell University, 417 Gates Hall, Ithaca, USA; email: cardie@cs.cornell.edu; N. Diakopoulos, University of Maryland, 1100 Knight Hall, College Park, USA; email: nad@umd.edu.

DOI: http://dx.doi.org/10.1145/2811282

images, and SoundCloud audio fragments to more classic news articles. Perhaps not surprisingly, the growing popularity and availability of UGC on the Web has generated exciting new opportunities for actively using information technologies to understand the opinions of others as well as to benefit from the diversity of their knowledge. UGC can moreover be employed to aid and improve machine-based processes such as recommendation, retrieval, and search systems. However, managing and hosting this content can be costly and time consuming. As a result, the owners of platforms that host UGC wish to sort and filter contributions according to their *value*—their credibility, helpfulness, diversity, and so forth—so as to create the best experience possible for viewers. And as the volume of UGC increases, the ability to perform this assessment and ranking *automatically* becomes increasingly important.

The task of assessing and ranking UGC is a relatively complex one. This is because (1) UGC, a relatively broad term, can encompass different *application domains* (e.g., tags, product reviews, postings in the questions and answers (Q&A) platforms, and comments on digital resources), and each type of UGC has different characteristics; (2) the definition of *value* varies with regard to different characteristics of application domains and specific tasks in hand (e.g., extracting relevant posts—such as tweets—related to a specific news topic is an important value in microblogging platforms, whereas extracting truthful product reviews is a value in product reviews); and (3) a particular value can be assessed and maximized in different ways due to the different characteristics of UGC. For example, assessing the credibility of product reviews requires different features and methods compared to extracting credible postings in microblogging platforms. Product reviews can be long, and authors can write false reviews on purpose to deceive the reader. Therefore, the features related to the text of a review are important features to assess the credibility of a review [Ott et al. 2012]. Instead, postings in microblogging platforms are sometimes short, and features related to texts alone cannot help to assess the credibility of postings. Hence, features need to be included that relate to the activities and backgrounds of authors for a more accurate assessment [Castillo et al. 2011].

This article aims to explore and shed light on the methods and frameworks for assessment and ranking of different types of UGC by presenting a unifying scheme that includes the commonly used definitions of values and methods for maximizing the value in existing research. This is achieved by answering the following general research questions: What are the existing approaches and methods for assessing and ranking UGC? What are effective features and metrics used to assess and predict UGC value across a range of application domains? What methods can be effectively employed to maximize that value? The findings of a systematic review of existing approaches and methodologies for assessing and ranking UGC are put forward to answer these questions. The focus is, in particular, on the short, text-based UGC typically found on the Web.

What counts as a value can be defined and assessed by (1) *a crowd of humans*, such as giving each end-user the possibility to assess the quality and vote on the content provided by other end-users, rank content with regard to the accumulated value of all judgments and votes, or use computational methods to train a ranking function that learns from a crowd of humans (which can be a community of end-users or an external crowd) to assess and rank content; (2) *a single end-user*, such as a system enabling each end-user to interact with the platform and rank content with regard to the value in the mind of the user and task at hand or using computational methods to train a ranking function with regard to a particular end-user's preferences, background, or online social interactions; and (3) *a platform designer*, such as a design that provides balanced views of UGC around an issue (e.g., a review site that explicitly samples from the diverse positive and negative reviews), a design that maximizes diversity among the displayed UGC items so that certain elements are not redundant, or a design that ranks relevant content to a particular topic.

Nevertheless, it is important to note that a decision of the platform's designer partially influences the definition of the value for any type of assessment and ranking system. This is because the designer decides which type of human-centered (crowd of humans or a single end-user) processes should be taken into consideration for an assessment process, how to design the system to leverage interactions of end-users, how to involve the end-users to define the value of content, or how to develop a machine-centered function to assess content. For example, in many product review platforms (e.g., the Amazon platform), the designer of the platform enables the involved community of end-users to assess helpfulness of reviews by providing a helpfulness voting button beside each review. This means that the considered value is helpfulness that is assessed and evaluated by a crowd of end-users. Another example is the exploration of the design space through iterative prototyping of various rating interfaces, which can be utilized by the human-centered method for developing the advanced community-based framework [Nobarany et al. 2012].

Furthermore, the main methods utilized for assessment and ranking decisions can be categorized in two groups:

(1) *Human centered:* This method enables (1) a crowd of end-users to interact with the system and assess and rank the content with regard to a particular definition of a value (e.g., assessment of quality of content by providing a vote button beside each element of content), (2) an end-user to interact with the system to specify his or her own notion of value and to adapt the ranking of content with regard to his or her preferences and specific task at hand, which is also called *interactive* and *adaptive*; and (3) a platform designer to specify default rankings or settings. In fact, these three correspond to the three entities that were mentioned earlier for defining values.

(2) *Machine centered:* This method utilizes computational methods, in particular machine-learning methods (supervised, semisupervised, or unsupervised), to (1) develop a ranking and assessment function that learns from the assessment and ranking behavior of a crowd of humans (external or end-users), such as training a classification function, which learns from helpfulness votes of the involved community of end-users; (2) develop a ranking and assessment function, which learns from a particular end-user's preferences, background, or online social interactions, which is also called *personalization*; and (3) develop a ranking and assessment function with regard to the designer's definition of value, such as providing balanced views of UGC around a political issue on online news platforms.

With regard to these high-level observations, in this study we categorize available frameworks related to assessment and ranking of UGC into the following groups:

(1) *Community-based framework*: Approaches that fall under this group use the human-centered or machine-centered methods to classify, cluster, and rank UGC based on the majority preferences (or an appropriate metric of preference) of the crowd of humans mainly with regard to a particular definition of value and for a particular domain of an application. Examples include distinguishing helpful versus nonhelpful product reviews, classifying useful and nonuseful comments on social media objects (e.g., YouTube videos, News articles), or identifying credible postings in online forums.

(2) *End-user–based framework*: Approaches that use this framework aim to accommodate individual differences in the assessment and ranking of UGC through human-centered or machine-centered methods, thus offering an individual user the opportunity to explore content, specify his or her own notion of value, or interact with the system to modify the display of rankings and assessments in accordance

with preferences expressed, behaviors exhibited implicitly, and details explicitly indicated by individual users. Examples include generating a set of content rankers by clustering subcommunities of the user's contact (based on the common content produced) to help users find content more relevant to their interest on their feeds without using explicit user input [Burgess et al. 2013].

(3) *Designer-based framework*: Approaches that fall under this group do not utilize the community's assessment, and they are not intended to be personalized or adaptive for an end-user. Instead, they mainly use machine-centered methods to encode the software designer's values in the ranking scheme. Examples include an approach that provides balanced political views around an issue [Munson et al. 2013].

(4) *Hybrid framework*: The three previous groups of approaches are not necessarily exclusive and often overlap each other. Therefore, there are bodies of assessment and ranking approaches that do not fall explicitly under any of the previous groups. Nevertheless, they take advantage of different categories and are hybrid approaches. Examples include an approach that learns from community behaviors to develop higher-level computational information cues that are useful and effective for adaptive and interactive systems for a single end-user [Diakopoulos et al. 2012]—a combination of community-based and end-user–based approaches.

Furthermore, it is observed that approaches employing machine-centered methods for different application domains use similar sets of content and context features related to three different entities of social media platforms. These three entities are *Author*, *User-Generated Content*, and *Resource* (the media object or topic on which authors generate content). Relationships exist between these entities. Thus, for different application domains, many approaches, particularly those that employ machine-centered methods, utilize similar sets of features related to these entities to assess UGC. However, the influence of the features changes with regard to the application domain and definition of the value to be maximized. Figure 1 shows a taxonomy of influential features that were referenced and demonstrated by available approaches that are influential for training the assessment and ranking functions for various application domains and values. We group influential features into nine different groups:

—*Text-based features (related to the Content entity)*: They include characteristics founded upon aggregate statistics derived from the text of a posting, such as the readability, informativeness, average sentence length, number of punctuation marks, number of different links, and part-of-speech (POS) tagging of the words in the text.

—*Semantic features (related to the Content entity)*: They include features related to meaning and semantics of the text of a posting, such as number of name entities, number of different types of name entities, subjectivity tone, sentiment polarity, and psychological characteristics of the content of postings.

—*Time-related features (related to Content and Recourse entities)*: These features are related to time, such as the time period associated with the object or topic under discussion or the time a posting was posted. For example, earlier postings may attract more attention by community members than later postings [Szabo and Huberman 2010].

—*Topic-based features (related to Content, Recourse, and Author entities)*: They include standard topic modeling features that measure the topical concentration of the author of posts, topical distance of a post compared to other postings on an object, or topical distance of a post compared to other postings on a particular topic.

—*Community-based features (related to Content, Recourse, and Author entities)*: These include features related to the relationship between content (or author) and the community with which the content is shared. For example, a user might be more likely to pay attention and reply to a post that is posted by a member of a community

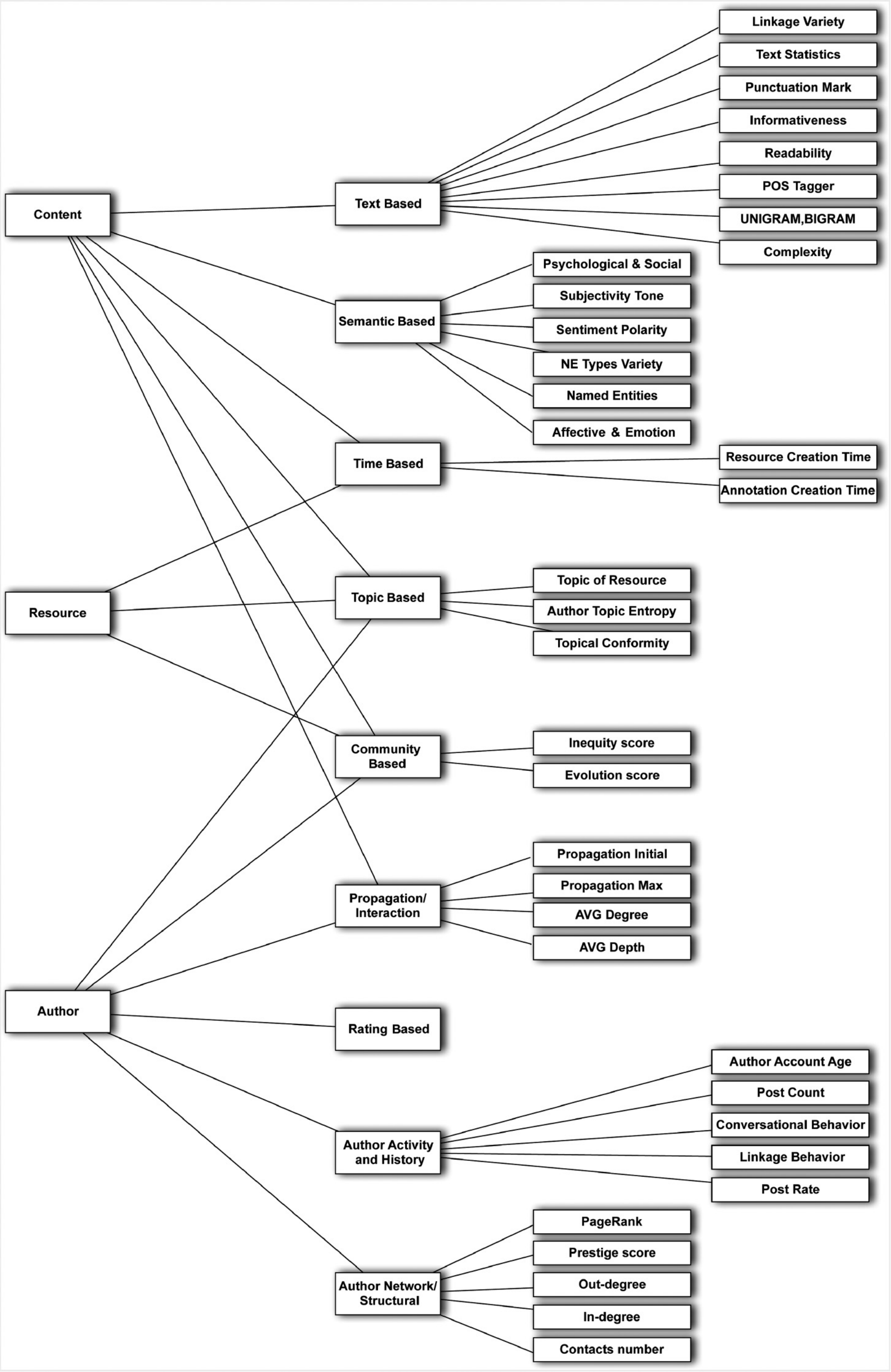

Fig. 1. Taxonomy of influential features, which are examined and demonstrated by available approaches.

in which the user has membership, and it therefore matches topics in which the user is interested.

—*Propagation/Interaction features (related to Content and Author entities)*: These include features related to the depth of the sharing tree and propagation tree of a posting (e.g., retweets).

—*Rating-based features (related to Content and Author entities)*: These features are related to the rating a post is given by a community, such as average number of thumbs-up/thumbs-down or number of helpfulness votes on a posting.

—*Author activity and background features (related to Author entity)*: These features describe the author's previous activities, behavior, and characteristics, such as registration age, number of contacts (e.g., number of followers), the number of postings the author has posted in the past, and the reputation of the author (average rating the author received from the community).

—*Author's network/structural features (related to the Author entity)*: These features capture the author's engagement and status in the social network (e.g., in-/out-degree, PageRank degree).

*Survey scope and methodology.* A survey is performed to realize this study. So, given the fact that the topic of assessing and ranking UGC is a vast area and "ranking" is used in some form in almost every system that uses UGC, an exhaustive survey would be overly ambitious. Therefore, it becomes very important to clarify what the survey is really intended to cover. The main focus of this survey is the assessment and ranking of short free textual UGC on the Web, and the scope of this survey encompasses comparing and analyzing UGC assessment approaches related to the following four main application domains: (1) product reviews, (2) questions and answers in Q&A platforms, (3) postings and discussions in microblogs and forums (e.g., Twitter, Slashdot), and (4) comments on social media objects (e.g., photos in Flickr or YouTube videos).

It is worth noting that by short free textual UGC, we mean all free textual content that was provided on a topic or a social media object at a specific point in time by a single user. Therefore, all research related to collaborative content (an article in a wiki platform, etc.) that also has temporal evolution and more than one author is excluded from the review process. Furthermore, considering the scope of the article that particularly focuses on free textual content, user-generated tags on social media objects are also one type of free textual content. However, as they contain mainly one to two keywords, they exhibit other characteristics compared to other types of free textual content. Therefore, this article gives less consideration to this type of content. However, in some sections, we give a short overview of highlighted available approaches related to this type of content. Finally, all research on the assessment of users who provide content, roles in online communities, and nontextual UGC (photos, video, etc.) are also excluded from the review process.

This survey was first conducted by using popular digital library search services (ACM Digital Library[1] or IEEE Xplore Digital Library[2]).We searched articles related to assessment and ranking methods of UGC based on their titles and main keywords. The main search keywords include "user-generated content," "quality assessment," "social media," "online community," "question answering," "ranking," "user interactions," among others. Collected articles were published in the most influential and pioneer proceedings and journals (e.g., Proceedings of the International Conference on World Wide Web, Proceedings of the Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings of the SIGCHI Conference on Human Factors

[1] http://dl.acm.org.

[2] http://ieeexplore.ieee.org/.

in Computing Systems, Proceedings of the International AAAI Conference on Weblog and Social Media, and Proceedings of the International Conference on Web Search and Data Mining). Second, for each relevant articles retrieved, all relevant articles that had been cited therein were collected. Third, the most relevant articles are filtered by reviewing their abstracts and discussion sections, resulting in the retrieval of a corpus of 90 relevant articles published between 2003 and 2015. The approaches proposed by these articles are compared in detail and sorted with respect to commonly utilized methods. Fourth, the systematic review procedures described by Kitchenham [2004] are adhered to in conducting the survey. However, the advantages and disadvantages of these various approaches are not compared.

*Survey structure and flow.* The survey is structured in five sections. Section 2 through 5 respectively overview and discuss available approaches and methods related to the four mentioned frameworks: community-based, end-user–based, designer-based, and hybrid frameworks. As the designer-based framework mainly used machine-based methods, only the first two sections related to community-based and end-user–based frameworks were structured into two subsections. These subsections are related to machine-centered and human-centered methods.

More precisely, Section 2 gives an overview of available approaches related to community-based assessment and ranking of UGC. This section includes two subsections related to machine-centered and human-centered methods and mainly focuses on values related to quality or a dimension of quality. The machine-centered subsection itself is structured into three subsections. The first subsection is Machine-Centered Approaches for Assessing Various Dimensions of Quality. As many proposed approaches that utilize machine-centered methods primarily focus on a highly general definition of value to extract high-quality UGC from different platforms or particular dimensions of quality (credibility, usefulness, etc.), this subsection briefly overviews important factors for assessing high-quality content concerning various application domains. The second subsection is Machine-Centered Approaches for Assessing a Particular Value of Interest. For some domains, especially in Q&A platforms, there are values that are not examined in the majority of assessment approaches but are beneficial to platform owners and facilitate development of other machine-based approaches, such as search or recommendation processes. Such values include distinguishing between posts such as editorials from news stories, subjective from objective posts, or conversational from informational posts. The third subsection is Machine-Centered Approaches for Assessing High-Quality Tags, in which user-generated free textual content has different characteristics compared to user-generated tags. User-generated free text is longer and has an informal structure so that users can converse and express their subjective opinions and emotions, and describe informative useful information about a media resource. However, tags are short, and, as a result it is more challenging to assess and rank their quality. Therefore, there is a range of available approaches only related to this type of content. Although a detailed discussion of such available approaches is beyond the scope of this work, this subsections gives a short overview of available approaches related to assessing high-quality tags.

Similarly, Section 3 gives an overview of available approaches related to the end-user–based framework. Two subsections related to machine-centered methods (or personalized approaches) and human-centered methods (or interactive and adaptive approaches) are also included. Section 4 gives a short overview of available approaches related to the designer-based framework. Approaches in this category mainly utilize machine-centered methods to rank content that mainly focuses on approaches for providing balanced or diverse views of UGCs around an issue. Section 5 outlines approaches related to the hybrid framework, such as approaches that leverage a community-based framework for developing an advanced end-user framework. Finally, Section 6 summarizes our observations, open challenges, and future opportunities with regard to our survey study. This section includes two subsections: first, Observations and Important Factors to Consider, which summarizes important factors observed by reviewing each framework, and second, Challenges and Opportunities for Future Work, which are based on the aforementioned observations and analyses of results; additionally, several open issues and limitations of the available approaches are listed. Addressing these issues and limitations provides natural avenues and opportunities for future work. Appendix gives a short overview of main contributions, evaluation methods, or experimental datasets of each discussed approach and study.

## 2. COMMUNITY-BASED FRAMEWORK

Approaches related to community-based assessment and ranking of UGC use different methods to classify, cluster, and rank UGC in accordance with the particular definition of the value expected to be maximized relying on majority-agreement sources of ground truth received from the crowd of humans. More precisely, they assess and rank content with regard to judgments on content received from the community of end-users (human centered, e.g., giving each user possibilities to judge the quality of content provided by other users) or with regard to a trained assessment and ranking function (machine centered), which learns from the assessment and ranking behavior of a community of end-users or external crowd (independent human raters). For example, some available approaches utilize machine-learning methods to train a function for learning from helpfulness votes on product reviews and develop a classifier to predict helpfulness of product reviews.

Therefore, the main methods proposed by the available approaches can be grouped in two categories: human-centered and machine-centered approaches. In recent years, the number of community-based approaches that utilize the machine-centered method has been increasing. Nevertheless, the prevalent default method utilized by many platforms is the human-centered approach. An overview is found later, which outlines available approaches related to different human-centered and machine-centered methods for different application domains and values expected to be maximized.

Figure 2 provides an overview of available community-based assessment and ranking approaches of UGC. The majority of approaches that use the machine-centered method for assessing and ranking UGC focus on a value related to quality (or a dimension of quality e.g., credibility or helpfulness). These approaches address the principal question of which content and context features can help predict accurately quality (or a dimension of quality) of content, where the gold standard for quality is based either on external crowd ratings (ratings by independent human raters) or ratings of a community of end-users (e.g., thumbs-up and thumbs-down votes on the platforms). Furthermore, for different application domains and values, various machine-learning methods (supervised, semisupervised, or unsupervised) are appropriate (see Figure 2). Accordingly, several features are found to be predictors of quality (or an appropriate dimension of quality) in various application domains, such as textual content, review length, star rating, and product category for product reviews; sentiment of comments on online sharing platforms; topic of discussion; and amount of payment on online Q&A platforms.

In contrast to these approaches, less literature focuses on the behavior of the community of end-users. These approaches provide better strategies and perspectives for more sophisticated development of the human-centered method, such as investigating how social factors influence users' ratings of content, how closely an opinion expressed by an element of content agrees with other opinions on the same issue (i.e., the product being reviewed), or how a user's consciousness of previous judgments on an element of content impacts the user's own judgment. Finally, in line with these

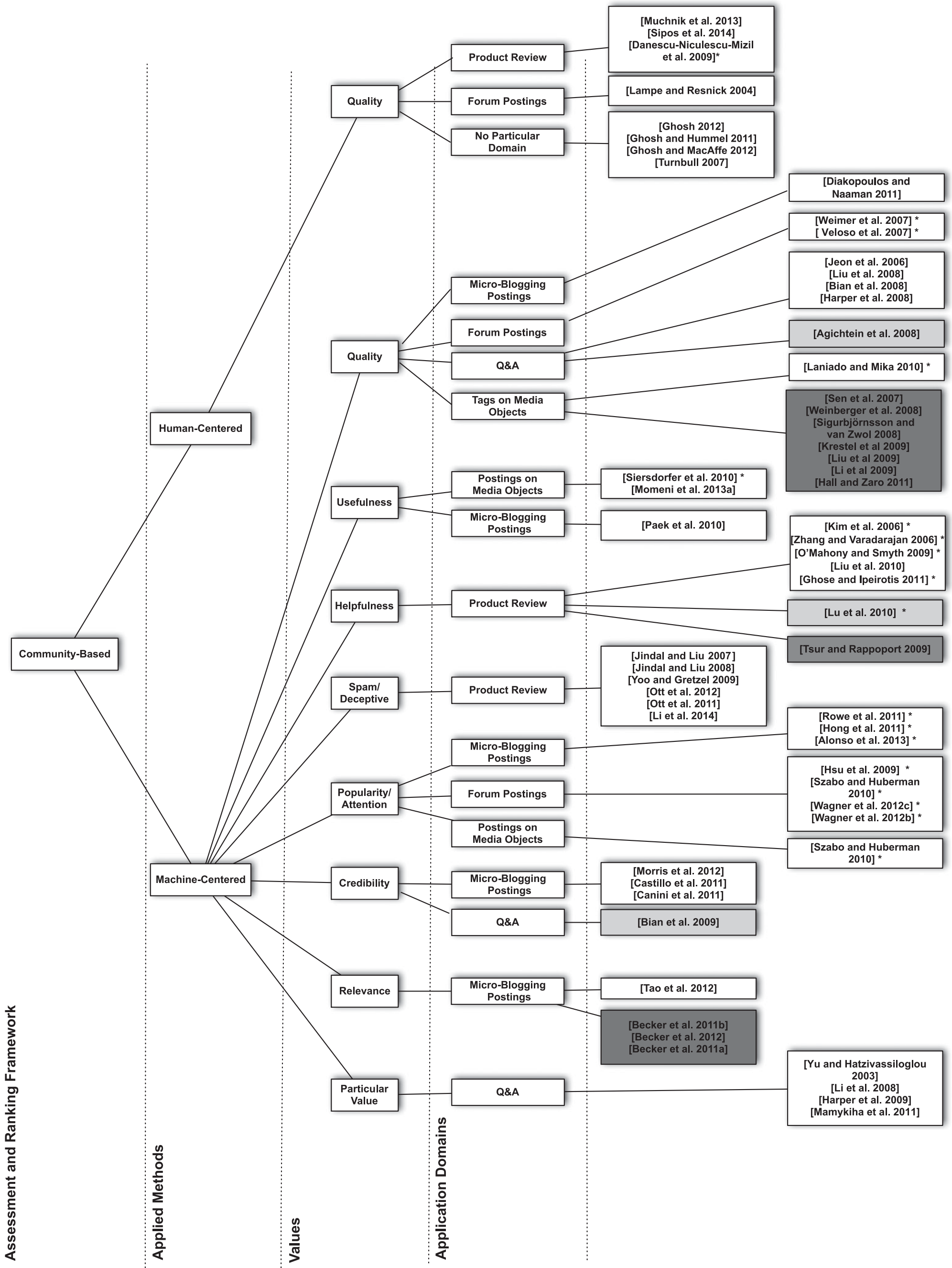

Fig. 2. Overview of community-based assessment and ranking of UGC approaches. At the lowest level, related to citations, dark grey boxes show approaches that utilize unsupervised learning, light grey boxes show approaches that utilize semisupervised learning, and white boxes show approaches that utilize supervised learning. An asterisk "*" beside the citation indicates that the approach utilizes judgments of end-users for creating the ground truth.

works, there are some approaches available that intend to find a mechanism that incentivizes high-quality contributions and maintains a high level of participation for the human-centered method.

In the following section, we give an overview of available approaches and works, first related to the human-centered method and second related to the machine-centered method.

### 2.1. Human-Centered Method

The prevalent default ranking method of many platforms is a human-centered method that attempts to classify UGC by allowing all users to vote on the contributions by others. This *wisdom-of-the-crowd* approach simply allows all users to vote on (thumbs-up or thumbs-down, stars, etc.) or rate UGC. This method, which is also called *distributed moderation* or *crowd based*, attempts to rank content according to the value estimates provided by the viewers' votes, such as the thumbs-up/thumbs-down style. Accordingly, the platforms display contributions that have attracted more votes by placing them near the top of the page and pushing those that have attracted fewer votes to the bottom of the page. Nevertheless, the crowd-based mechanism elicits higher quality when a system achieves high participation [Ghosh and Hummel 2011]. Moreover, the lowest quality that can arise in any mixed strategy equilibrium of the crowd-based mechanism becomes optimal as the amount of available attention diverges [Ghosh and Hummel 2011].

Popular examples of the distributed moderation and usage of the crowd-based method are used by Yelp, Slashdot, YouTube, Reddit, Facebook, and Digg. The Yelp platform permits all viewers to judge if a review written on an item is "Useful," "Funny," or "Cool." The Slashdot platform is another example that filters out abusive comments by using a crowd-based moderation system. First, every comment is awarded a score of −1 to +2. Registered users receive a default score of +1, anonymous users (Anonymous Coward) receive 0, users with high "karma" receive +2, and users with low "karma" receive −1. While reading comments on articles, moderators click to moderate the comment. In addition, adding a particular descriptor to the comments, such as "normal," "off-topic," "troll," "redundant," "interesting," "informative," "funny," "flamebait," and so forth, with each corresponding to a −1 or +1 rating, is an option for moderators. This means that a comment may have a rating of "+1 insightful" or "−1 troll." A user's karma increases with moderation points, and a user must have a high karma to become a moderator. Being a regular user does not mean that one becomes a moderator, but instead the system gives five moderation points at a time to users based on the number of comments they have posted. To moderate the moderators and help reduce the number of abuses in the moderation system, the *meta-moderation system* is implemented. The meta-moderator examines the original comment and the arguments given by the moderator (e.g., troll, funny) for each moderation and can judge moderations based on the context of comments. The YouTube, Digg, and Reddit platforms give viewers the opportunity to judge thumbs-up/thumbs-down of comments or textual postings written on a video or article. The vote is used for ordering the post and discovering its place in the front-end representation. For product reviews, Amazon.com gives users possibilities to vote on the helpfulness of product reviews. More highly voted reviews are displayed more prominently by placing them near the top of the page.

Lampe and Resnick [2004] indicate in a summary statistic the extent to which users contribute to the crowd-based method (especially on Slashdot.com). The distribution of scores for comments shows that the dispersal of scores for comments is reasonable and agreement on the part of the community exists on the fairness of moderations. Analyzing Slashdot.org from a statistical perspective confirms the validity of the concept that underlies distributed moderation. However, a closer analysis reveals that identifying comments may require considerable time, especially for valuable comments. In addition, comments that have been incorrectly moderated are often not reversed, and comments that have low starting scores are often not treated by moderators in the same manner as other comments are. Thus, it is important to take into consideration how timely the moderation is, how accurate or inaccurate the moderation is, how influential individual moderators are, and how the input on the part of individual moderators can be reduced.

It is important to consider that context or a user's awareness of previous votes on a review impacts her own voting decision [Muchnik et al. 2013; Sipos et al. 2014; Danescu-Niculescu-Mizil et al. 2009]. Danescu-Niculescu-Mizel et al. [2009] assert that helpfulness votes on product reviews are influenced by social factors, such as how closely an opinion of a review agrees with other opinions on the same product. In addition, they show that the perceived helpfulness ratings correlate with other evaluations of the same product of a review and not necessarily with the content of reviews. Furthermore, Sipos et al. [2014] observe the relationship between voting behavior and context and assert that voting behavior cannot be captured by a principal voting model, where users make absolute and independent judgments on a review without the context in which it is presented at the time of voting. Therefore, it is proposed that the voting system should incorporate context in addition to inherent quality of a review.

Finally, another key factor to be considered is that participation and contribution in the human-centered method is voluntary—contributors may decide to take part or not [Ghosh 2012]. It should also be noted that many crowd-based approaches fail, either immediately or eventually, because of very sparse contributions. Moreover, having decided to participate does not necessarily mean that contributors will put effort into their contributions [Ghosh 2012]. Therefore, methods to incentivize contributors need to be developed so as to allocate rewards such as monetary and nonmonetary (attention, reputation [Beenen et al. 2004; Huberman et al. 2009], and virtual points), which appear to motivate contributors contrary to what may be expected [Nam et al. 2009; Yang et al. 2011]. Despite there being such a need for these kinds of methods to incentivize high-quality UGC, there are few approaches that focus on the development of these approaches.

Ghosh and McAfee [2011] propose a game-theoretic model in the context of diverging attention rewards with high viewership. Strategic contributors are the focus of the model that is motivated primarily by exposure or viewer attention. The model allows the endogenous determination of both the quality and the number of contributions in a free-entry Nash equilibrium. The importance of making choices to contribute endogenously is underlined because the production of content, and not only incentivizing high quality, is necessary in UGC. Additionally, Ghosh and McAfee [2012] explore the design of incentives in environments with endogenous entry for finite rewards. In the context of limited attention rewards in Q&A platforms such as Quora[3] or StackOverflow,[4] the choice of which answers to display for each question, the choice whether to display all answers to a particular question, or the choice whether to display only the best ones and suppress some of the weaker contributions remains with the mechanism designer or platform owner [Ghosh and McAfee 2012].

OpenChoice [Turnbull 2007] is another example for incentivizing end-users. It encourages everyone to take an active role in crafting OpenChoice's configuration. Users of the portal will see a ranked list of resources most in need of human review and vote

[3] Quora.com is a question-and-answer Web site where questions are created, answered, edited, and organized by its community of users.

[4] StackOverflow.com is a Web site, the flagship site of the Stack Exchange Network.

on as many of these URLs as the user desires. Once the votes on a particular URL reach a critical mass of consensus, that URL is added to the canonical OpenChoice blacklist. This system offers two distinct advantages: first, using the collective efforts of a user community to improve the community's control over information, and second, incentivizing members of the community to participate in the system's improvement by allocating social capital to those community members who participate meaningfully.

Finally, for Q&A as an application domain, Jain et al. [2009] propose a game-theoretic model of sequential information aggregation. When an asker posts a question, each user decides whether to aggregate a unique piece of information with existing information or not. When a certain threshold has been exceeded with regard to quality, the asker closes the question and allocates points to users. Taking into consideration the effect of different rules for allocating points on the equilibrium, it is found that a best-answer rule provides a unique, efficient equilibrium in which all users respond in the first round. On the other hand, the best-answer rule isolates the least efficient equilibrium for complements valuations.

### 2.2. Machine-Centered Method

Many approaches related to the community-based framework that use machine-centered methods, mainly employ a machine-learning method (classification, clustering, etc.) by precisely defining what is considered as valuable UGC for the application domain of interest. Examining these approaches more closely shows that many available machine-centered assessment approaches use and include judgments of a community of end-users to create a ground truth. On the other hand, others due to various biases arising from the community of end-users completely exclude these judgments and employ external crowd judgments instead. For example, many assessment approaches for classification of product reviews with regard to helpfulness as the value have used crowd votes—helpfulness votes—to create the helpfulness ground truth, whereas approaches related to deception as the value exclude crowd votes and employ independent coders for creating the ground truth.

Next, we provide an overview of approaches that use machine-centered methods for assessment and ranking of UGC for different dimensions of quality as values. Finally, we give a short outline of approaches for assessing the quality of a particular type of UGC—user-generated tags on multimedia objects.

*2.2.1. Machine-Centered Approaches for Assessing Various Dimensions of Quality.* Many proposed approaches that utilize machine-centered methods primarily focus on a highly general definition of value, extracting *high-quality* UGC from different platforms. In the following, a short overview of important factors for assessing high-quality content concerning various application domains is provided.

For finding high-quality questions and answers in Q&A platforms, a combination of different types of features is likely to increase the assessment's accuracy, and adding knowledge about the author is important when assessing the quality of questions or answers [Jeon et al. 2006; Agichtein et al. 2008]. Bian et al. [2008] show that textual, community, and user feedback (while they are noisy) features are important to improve the training of the ranking functions. Nevertheless, the reputation of the authors submitting the answers is not as important as many other features. This suggests that the authority, expertise, and history of the author are only important for some, not all, of the predictions [Liu et al. 2008].

Furthermore, for postings in the Q&A domain, Harper et al. [2008] explore influential factors on answer quality by conducting a comparative, controlled field study of answers posted across different types of Q&A platforms: digital reference services, ask an expert services, and Q&A sites. "Digital reference" services enable users to access library

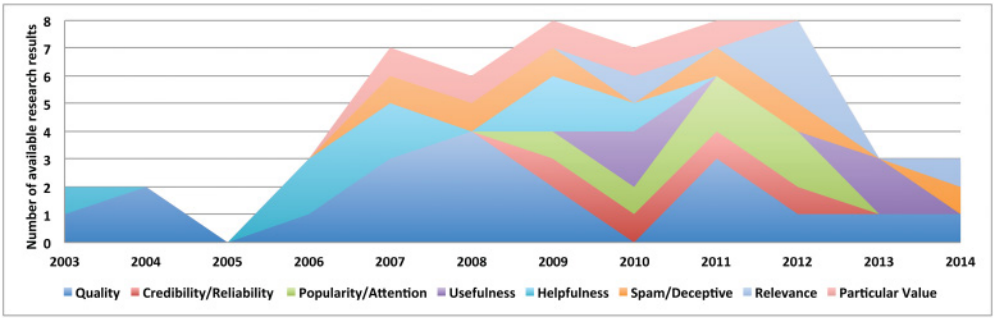

Fig. 3. Evolution of approaches for assessing various dimensions of quality. Over time, more dimensions of quality have been taken into consideration.

reference services. "Ask an expert services" is manned by "experts" in different topic areas, such as science (e.g., at "MadSci Network"[5]) or oceanography (e.g., at "Ask Jake, the SeaDog"[6]). First, they show "you get what you pay for" [Harper et al. 2008]. For example, answer quality is better in Google Answers than on the free platforms, and paying more money for an answer has a positive impact on the likelihood of receiving high-quality answers. Second, Q&A platforms with different types of users are more successful. For example, Yahoo! Answers, which is open to the public for answering questions, outperforms platforms that depend on specific users to answer questions.

For posting on microblogging platforms such as an application domain, Diakopoulos and Naaman [2011] examine the correlation between comment quality and consumption and production of news information. They also describe and explore what motivates readers and writers of news comments. Their results have shown (1) how much low-quality comments influence users and journalists; (2) how perceptions of quality can be influenced by various reading motivations of the individual; and (3) how flagging, moderation, and engagement can be used as policies for enhancing quality. Furthermore, they show that aspects peculiar to many online communities include unpredictable participation patterns (e.g., interaction between regular users and other actors in different situations).

Finally, for posting in forum platforms as an application domain, Weimer et al. [2007] and Veloso et al. [2007] present supervised approaches to assess the quality of forum posts in online forums that learn from human ratings. Weimer et al. use the Nabble[7] platform as a data source, whereas Veloso et al. use a collection of comments posted to the Slashdot[8] forum.

Over time, the value expected to be maximized has been defined more particularly and more sophisticatedly with more application domains being taken into consideration (Figure 3). Initially, quality was considered an important value. However, quality is a very general term, and it has a vague definition in the context of many application domains. Therefore, the requirements to assess UGC have evolved, and more dimensions of quality have become important, such as credibility and usefulness.

It is important to note that these values are not necessarily exclusive and often overlap with each other, because different dimensions of quality are ambiguous and blend into one another. For example, the "usefulness" is ambiguous and has relationships

[5] http://www.madsci.org.

[6] http://www.whaletimes.org.

[7] Nabble.com provides an embeddable forum, embeddable photo gallery, embeddable news, embeddable blog, embeddable mailing list, and archive.

[8] Slashdot.org is a news forum.

with other values, such as “helpfulness.” Therefore, it is challenging to categorize these values.

*Approaches for assessing usefulness.* Usefulness is generally defined as “the quality or fact of being able to be used for a practical purpose or in several ways.”[10]For posting on multimedia objects (e.g., comments on YouTube videos), Siersdorfer et al. [2010] define usefulness as “community acceptance of new comments (community feedback for comments).” On the other hand, for an explicit definition of usefulness, Momeni et al. [2013a] define usefulness as “a comment is useful if it provides descriptive information about the object beyond the usually very short title accompanying it.” Furthermore, Liu et al. [2007] define an answer as useful in Q&A platforms “when the asker personally has closed the question, selected the best answer, and provided a rating of at least 3 stars for the best answer quality.” In the context of the microblogging platforms, Becker et al. [2011b, 2012] define usefulness as “the potential value of a Twitter message for someone who is interested in learning details about an event. Useful messages should provide some insight into the event, beyond simply stating that the event occurred.”

Many approaches are related to usefulness. Momeni et al. [2013a] and Siersdorfer et al. [2010] use a supervised learning method to classify useful from nonuseful content on social media objects. These approaches show that what counts as useful content can depend on several factors, including the practical purpose at hand, the media type of the resource (if the object is a document, video, art object, photo etc.), topic type of the resource (if the video that is commented on is associated with a person, place, event, etc.), the time period associated with the resource (it is about the 20th century or the 1960s, etc.), or even the degree of opinion polarity around the resource.

For comments on social media resources (YouTube videos, Flickr photos, etc.) as an application domain, semantic and topic-based features play an important role in the accurate classification of usefulness comments, and especially important are those features that capture subjective tone, sentiment polarity, and the existence of named entities [Momeni et al. 2013a]. In particular, comments that mention named entities are more likely to be considered useful, whereas those that express the emotional and affective processes of the author are more likely considered to be nonuseful. Similarly, terms indicating “insight” (think, know, consider, etc.) are associated with usefulness, whereas those indicating “certainty” (always, never, etc.) are associated with nonuseful comments. With regard to different topics of media objects—people, places, and events—the classifier more easily recognizes useful comments for people and events regardless of the social media platform [Momeni et al. 2013a]. In addition, negatively rated comments by a crowd that are considered as nonuseful content [Siersdorfer et al. 2010] contain a significantly larger number of negative sentiment terms. Similarly, positively rated comments that are considered as useful content contain a significantly larger number of positive sentiment terms [Siersdorfer et al. 2010]. Therefore, all of these results suggest that training “topic-type–specific” usefulness classifiers generally allows improved performance over the “type-neutral” classifiers [Momeni et al. 2013a]. For prediction of useful comments, text-based and semantic-based features play important role [Paek et al. 2010; Momeni et al. 2013a].

Usefulness is very closely related to helpfulness.

*Approaches for assessing helpfulness.* Helpfulness is generally defined as “giving or being ready to give help.”[10] Helpfulness is mainly defined in the product review domain and is primarily explained as the number of helpfulness votes a review received on platforms (e.g., Amazon.com) [Kim et al. 2006; Ghose and Ipeirotis 2007; Lu et al. 2010].

Helpfulness is largely prevalent in the product reviews domain. This is because many online shopping and online booking platforms explicitly ask their users to vote on the helpfulness of product reviews. Accordingly, many machine-centered approaches utilize and learn from these votes to train and develop an assessment model.

Many approaches demonstrate that a few relatively straightforward features can be used to predict with high accuracy whether a review will be deemed helpful or not. These features include length of the review [Kim et al. 2006], mixture of subjective and objective information, readability such as checking the number of spelling errors, conformity (the helpfulness of a review is greater when the star rating it has received is more similar to the aggregate star rating of the product) [Danescu-Niculescu-Mizil et al. 2009; Kim et al. 2006], and author reputation and social context features [O'Mahony and Smyth 2009; Lu et al. 2010]. However, the effectiveness of features related to social context depends on there being sufficient training data to train these extra features [Lu et al. 2010], and features related to social context are less successful in comparison to author reputation features [O'Mahony and Smyth 2009]. Furthermore, it can be asserted that helpfulness of a product review is based on properties actually found in the review itself and is not necessarily consistent with its similarity to the corresponding product description [Zhang and Varadarajan 2006]. In addition, it is shown that the helpfulness of a product reviews has a slight correlation with the subjectivity or sentiment polarity of a reviewed text [Zhang and Varadarajan 2006].

The majority of the available approaches use supervised learning methods based on user votes as the ground truth [Kim et al. 2006; O'Mahony and Smyth 2009; Zhang and Varadarajan 2006; Ghose and Ipeirotis 2007, 2011]. However, there are few studies based on the semisupervised [Lu et al. 2010] and unsupervised learning [Tsur and Rappoport 2009]. With regard to semisupervised learning methods, Lu et al. [2010] exploit information gleaned from social networks and propose a semisupervised approach by adding regularization constraints to the linear text-based predictor. Four constraints are defined: (1) Author Consistency, (2) Trust Consistency, (3) Co-Citation Consistency, and (4) Link Consistency [Lu et al. 2010]. With regard to unsupervised learning methods, Tsur and Rappoport [2009] propose supervised learning approaches, such as the REVRANK algorithm. The REVRANK algorithm first created a virtual optimal review by identifying a core of dominant words found in reviews, achieved in two stages. First, dominant words are identified by how often they are used, then, words that are used less often but provide pertinent information on the specific product are identified. Second, by using these words, a definition of the "feature vector representation" of the most desired review is created. Finally, reviews are rearranged to this representation and ordered with regard to their similarity with the "virtual core" review vector.

So far, many proposed approaches have utilized a crowd of end-users (user ratings) for developing and training a prediction model. However, Liu et al. [2007] show that users ratings at Amazon have three kinds of biases: (1) imbalance vote bias, (2) winner circle bias, and (3) early bird bias [Liu et al. 2007]. Therefore, they propose a specification—a guideline for what a good review consists of to measure the quality of product reviews—and a classification-based approach developed from manually annotated product reviews that complies with the proposed specification.

Juxtaposed to helpfulness in product review application domains, there are two values, namely Spam and Deceptive. These are expected to be minimized.

*Approaches for assessing spam and deceptive content.* Spam and deceptive content are generally defined as "giving an appearance or impression different to the true one."[10] They can also be irrelevant or inappropriate messages sent on the Internet to a large number of recipients. Yoo and Gretzel [2009] define a deceptive product review as "a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver," and following this definition, Ott et al. [2011, 2012] and Li et al. [2014] define deceptive product reviews as "fictitious reviews that have been deliberately written to sound true, to deceive the reader." Jindal and Liu [2008] consign reviews to the category of spam when they are based upon dubious opinions and are, as a result, very damaging.

Similar to helpfulness, assessing spam and deceptive content is mainly discussed in the product review domain. Approaches in these areas can be basically categorized into two groups: (1) approaches for assessing spam product reviews [Jindal and Liu 2008] (product reviews on brands, duplicates, and nonreviews such as advertisements, other irrelevant reviews) and (2) approaches for assessing deceptive product reviews [Yoo and Gretzel 2009; Ott et al. 2011, 2012; Li et al. 2014].

Approaches related to both groups apply supervised learning methods and mainly use text- and content-related features. For assessing spam product reviews, three types of features are used [Jindal and Liu 2008]: (1) review-centric features, which include rating- and text-based features; (2) reviewer-centric features, which include author-based features; and (3) product-centric features. The highest accuracy is achieved by using all features. However, it performs as efficiently without using rating-based features. Rating-based features are not effective factors for distinguishing spam and nonspam because ratings (feedback) can also be spammed [Jindal and Liu 2008].

With regard to deceptive product reviews, deceptive and truthful reviews vary concerning the complexity of vocabulary, personal and impersonal use of language, trademarks, and personal feelings. Nevertheless, linguistic features of a text are simply not enough to distinguish between false and truthful reviews [Yoo and Gretzel 2009]. N-gram–related features have the highest impact, but an approach that combines psycholinguistically related features and n-gram features can achieve slightly improved results. Moreover, there is a reasonable correlation between deceptive opinion and imaginative writing based on similarities of distributions of POS tags [Ott et al. 2011].

*Approaches for assessing popularity and attention.* Popularity and attention is "the state or condition of being liked, admired, or supported by many people."[10] For postings in forums, Wagner et al. [2012a, 2012b] define attention as "the number of replies that a given post on a community message board yields as a measure of its attention," whereas Szabo and Huberman [2010] define it as "the number of votes (diggs) a story collected on Digg.com[9] and the number of views a video received on YouTube.com." For posting on microblogging platforms, Hong et al. [2011] measure popularity as the number of retweets.

Many approaches related to popularity and attention use a supervised learning method to classify content into popular (or seed) and nonpopular categories [Hong et al. 2011; Rowe et al. 2011; Wagner et al. 2012a, 2012b; Hsu et al. 2009]. The temporal and author-related features are shown as important features for assessment and ranking of popular content.

Unlike popular posts that receive lots of attention (such as retweets, reshares), normal posts only attract a small audience and users lose interest in them quickly [Hong et al. 2011]. Therefore, temporal features have a stronger effect on posts with a low and medium volume of attention compared to highly popular messages. Furthermore, the social network provided by the service does not influence users to look at the content once the content has become visible to a huge number of viewers [Szabo and Huberman 2010], although during situations with a low number of views, they are still important. Furthermore, based on experiments on two well-known social media platforms, Digg and YouTube, Szabo and Huberman [2010] show that in Digg, assessment of access to given stories during the first 2 hours after posting enables us to estimate their popularity within the next 30 days with a relative error margin of 10%, whereas predicting the polarity of YouTube videos (with regard to the download rate of YouTube videos) needs

[9] Digg.com is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Web audience, such as science and trending political issues.

to be followed for 10 days to achieve the same relative errors. The influence of time on predictions is due to differences in how content is consumed on the two platforms. Posts on Digg become outdated very fast, whereas posts on YouTube videos become outdated much later. Therefore, predictions are more accurate for content with a short life cycle, whereas for predictions for content with a longer life cycle, greater statistical errors are more likely.

It is also empirically demonstrated that the use of author features for identifying seed or popular posts has more effect [Rowe et al. 2011; Hsu et al. 2009] than the use of text-based features.

Finally, the manner in which attention is created varies relating to different community forums. How particular features are associated positively with the start of discussions in one community may differ in another community [Wagner et al. 2012b]. The influential factors for predicting whether a discussion begins around a post may vary depending on the factors that impact how long the discussion lasts [Wagner et al. 2012a, 2012b]. Therefore, in forums, Wagner et al. [2012a] argue that the unawareness of a user is not advantageous, since understanding the behavioral patterns peculiar to individual communities is influenced by posts that attract a community and stimulate long dialogues in a forum.

*Approaches for assessing credibility or reliability.* Credibility is generally defined as the "quality of being convincing or believable."[10] For postings on microblogging platforms, Castillo et al. [2011] define credibility as "credibility in the sense of believability: offering reasonable grounds for being believed." For postings on discussion forums, Canini et al. [2011] define credibility as being "associated with people who not only frequently publish topically relevant content but also are trusted by their peers."

Examining approaches for assessing credibility or reliability more closely indicates that most of the available approaches use supervised learning and are mainly based on external sources of ground truth [Castillo et al. 2011; Canini et al. 2011]—features such as author activities and history (e.g., a bio of an author), author network and structure, propagation (e.g., a resharing tree of a post and who shares), and topical-based affect source credibility [Castillo et al. 2011; Morris et al. 2012]. Castillo et al. [2011] and Morris et al. [2012] show that text- and content-based features are themselves not enough for this task. In addition, Castillo et al. [2011] indicate that authors' features are by themselves inadequate. Moreover, conducting a study on explicit and implicit credibility judgments, Canini et al. [2011] find that the expertise factor has a strong impact on judging credibility, whereas social status has less impact. Based on these findings, it is suggested that to better convey credibility, improving the way in which social search results are displayed is required [Canini et al. 2011]. Morris et al. [2012] also suggest that information regarding credentials related to the author should be readily accessible ("accessible at a glance") due to the fact that it is time consuming for a user to search for them. Such information includes factors related to consistency (e.g., the number of posts on a topic), ratings by other users (or resharing or number of mentions), and information related to an author's personal characteristics (bio, location, number of connections).

For questions and answers as an application domain, Bian et al. [2009] propose a semisupervised approach for assessing content credibility and author reliability based on a coupled mutual reinforcement framework that requires only a very small number of trained samples. The proposed framework elaborates on the mutual reinforcement between the connected entities (beginning with a set of known labels for two entities, authors, or answers) in each bipartite graph to assess the credibility and reputation. Bian et al. [2009] state the mutual reinforcement principle as follows:

[10] *New Oxford American Dictionary*, 2011.

> An answer is likely to be of high quality if the content is responsive and well-formed, the question has high quality, and the answerer is of high answer-reputation. At the same time, an author will have high answer-reputation if the user posts high-quality answers, and obtain a high question-reputation if the user tends to post high-quality questions. Finally, a question is likely to be of high quality if it is well stated, is posted by an author with a high question reputation, and attracts high-quality answers.

*Approaches for assessing relevant content around an issue.* For many application domains, ranking relevant content to particular issues (e.g., an event or a topic) is an important value, which is defined by the designer of the platform. Relevance is generally defined as "closely connected or appropriate to the matter in hand."[10] Usually it is driven by an explicit/implicit information need or query, and therefore the generic framing is almost impossible. For postings on microblogging platforms, Becker et al. [2011b, 2012] define relevance as "relevant social media documents for a specific event." Instead, for postings on micro-blogging platforms, Tao et al. [2012] define relevance as "interesting and relevant micro posts for a given topic."

For postings on microblogging platforms as an application domain, Becker et al. 2011b, 2012] explore approaches for finding representative posts among a set of Twitter messages that are relevant to the same event, with their aim being to identify high-quality, relevant posts that provide useful information about an event. The problem is approached in two concrete steps: first by identifying each event and its associated tweets using a clustering technique that clusters together topically similar posts, and second, for each cluster of event, posts are selected that best represent the event. Centrality-based techniques are used to identify relevant posts with high textual quality and are useful for people looking for information about the event. Quality refers to the textual quality of the messages—how well the text can be understood by any person. From three centrality-based approaches (Centroid, LexRank [Radev 2004], and Degree), Centroid is found to be the preferred way to select tweets given a cluster of messages related to an event [Becker et al. 2012]. Furthermore, Becker et al. [2011a] investigate approaches for analyzing the stream of tweets to distinguish between relevant posts about real-world events and nonevent messages. First, they identify each event and its related tweets by using a clustering technique that clusters together topically similar tweets. Then, they compute a set of features for each cluster to help determine which clusters correspond to events and use these features to train a classifier to recognizing between event and nonevent clusters.

With regard to relevancy for a topic, Tao et al. [2012] explore if additional micropost characteristics exist that are more predictive of the relevance of a post rather than its keyword-based similarity when querying in microblogging platforms such as Twitter. Based on an investigation of 16 features along two dimensions—topic-dependent and topic-independent features—they showed the higher influence of topic-dependent features rather than topic-independent features for this task. Furthermore, Chen et al. [2010] demonstrated that both topic relevance and the social voting process are helpful in providing URL recommendation on Twitter as a means to better direct user attention in information streams.

Concerning simple measures of relevance—"article" or "conversational"—from the perspective of editors for comments on news articles, Diakopoulos [2015] investigated if simple measures of relevance correlate to editors' selections of comments on news articles and demonstrated that editors' selections of comments are correlated with the relevancy of comments to the related article as well as to the other comments on the article.

*2.2.2. Machine-Centered Approaches for Assessing a Particular Value of Interest.* For some domains, especially in Q&A platforms, there are values that are not examined in the majority of assessment approaches but are beneficial for platform owners and facilitate development of other machine-based approaches, such as search or recommendation processes. Among these works in the Q&A domain, there are approaches for distinguishing between posts, such as editorials from news stories, subjective from objective posts, or the conversational from informational posts. Many of these approaches also employ machine-centered methods for classifying content concerning a particular value.

For distinguishing between question and answer postings with a very large number of opinions written about current events, Yu and Hatzivassiloglou [2003] present a classifier. They show that at document level, a Bayesian classifier can differentiate between "factual" and "opinion" posts by using lexical information. However, the task is significantly more difficult at sentence level. Furthermore, features such as words, bigrams, trigrams, polarity, and POS play an important role for this task [Yu and Hatzivassiloglou 2003].

For predicting a question's subjectivity or objectivity in a Q&A site, Li et al. [2008] present the CoCQA model, which is based on the concept of co-training [Blum and Mitchell 1998] (semisupervised learning approach). It is expected that objective questions are answered with well-founded information. Instead, subjective questions result in answers disproving personal, emotional states. For creating an experimental dataset, they download questions from every top-level category of Yahoo! Answers and randomly choose a set of questions from each category to be labeled by coders from Amazon's Mechanical Turk Service. With regard to the feature set, they compute question and answer content and three term weighting schemes separately (e.g., Binary, TF, and TF-IDF[11]). By applying CoCQA to this task, they show that they can significantly decrease the amount of the required training data.

For distinguishing between "conversational" questions and "informational" questions in Q&A platforms, Harper et al. [2009] propose a classifier, defining conversational questions and informational questions as follows:

> Informational questions are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. Conversational questions are asked with the intent of stimulating discussion. They may be aimed at getting opinions, or they may be acts of self-expression.

They develop an online coding tool and use data from three well-known Q&A sites (Yahoo Answers, Answerbag, and Ask Metafilter) for human coding. Based on their human coding evaluation, they show that people are able to reliably differentiate between questions that are part of a conversation and questions that ask for information and demonstrate that the archival value of the former is lower than that of the latter. For training a classifier, they evaluate several structural properties and features related to the social network model. They show that features related to structure of the text are important to distinguish conversational from informational questions. With regard to the social network features, they show that none of these features improves performance, despite there being potentially more indicators to be extracted from the text [Harper et al. 2009]. Furthermore, they show that taking into consideration only questions is not simply enough for classifying a Q&A thread.

For the success factors of some Q&A platforms (e.g., greater than 92% of StackOverflow questions about expert-related topics are answered in a median time of 11 minutes)

[11] Term frequency–inverse document frequency.

based on a qualitative study, Mamykina et al. [2011] argue that daily involvement and high visibility of the design team within the community is more important than just a superior technical design and argue the benefit of using gamification mechanisms (e.g., leaderboards and pointsystem).

For postings on online forums, Burke et al. [2007], by using posts from Usenet,[12] conduct a series of studies related to the impact of two rhetorical strategies on community responsiveness: introductions (e.g., "I've been lurking for a few months") and requests (show a request of the author). They show that requests attract more community responses and which community responses have a higher correlation with detection of requests compared to other contextual and text-based features, such as length of post and number of posts and contributions in a group.

For product reviews, Gilbert and Karahalios [2010] investigate why some reviews resemble earlier reviews and discover roughly 10% to 15% of reviews considerably resemble previous ones. They argue that motivations for reviewing, and reactions to seeing "deja reviews" (who reflect what others said), varied remarkably between these two groups of reviewers: amateurs and pros. Amateurs reviewed only occasionally by reviewing fewer than 30 products and hardly received helpful votes, whereas pros reviewed many hundreds of products. Where amateurs do not mind being part of community, pros write to advance a personal opinion and want to stand out.

Finally, for comments on social news, Hullman et al. [2015] present the results of a qualitative study of commenting around visualizations published on a mainstream news outlet: the EconomistÕs Graphic Detail blog.[13] Their results show that 42% of the comments discuss the visualization and/or article content; greater than 60% of comments discuss matters of context, including how the issue is framed; and greater than 33% of total comments provide direct critical feedback on the content of presented visualizations.

*2.2.3. Machine-Centered Approaches for Assessing High-Quality and Relevant Tags.* User-generated free textual content has different characteristics from user-generated tags. User-generated free text is longer and has an informal structure, so users can converse and express their subjective opinions and emotions, and describe informative useful information about a media resource. Tags are short, and therefore it is more challenging to assess and rank their quality. In this section, we give a short overview of available approaches related to assessing high-quality and relevant tags. However, reviewing this type of content is not the main focus of this survey. A further extended review of these works can be found in Wang et al. [2012] and Gupta et al. [2010].

For assessing high-quality tags on media resources (e.g., online photos), Weinberger et al. [2008] propose a method that assesses the ambiguity level of a tag set, and to supplement this method they propose two additional tags to resolve the ambiguity. Weinberger et al. [2008] define a tag as ambiguous as such: "A tag set is ambiguous if it can appear in at least two different tag contexts." The tag contexts are defined as "the distribution over all tag co-occurrences."[14] They use 50 different tags (the ambiguity evaluated by users) for evaluating and examining parameters of the algorithm. They show that the majority of the ambiguous tags are found within one of three dimensions: temporal, geographic, or semantic. Sen et al. [2007] explore implicit (behavioral) and

[12] Usenet is a worldwide distributed Internet discussion system.

[13] http://www.economist.com/.

[14] A prime example is "Cambridge," a city found both in Massachusetts and England. A tag such as "university" makes sense if it is used in both contexts, but the ambiguity remains unresolved. Thus, in the case of the tag "Cambridge," the method notes that this tag contains ambiguity and recommends "MA" or "UK" [Weinberger et al. 2008].

explicit (rating) feedback to analyze and devise methods for identifying high-quality tags.

They investigate different lightweight interfaces used for collecting feedback from members about tags to identify which interfaces result in the richest metadata for determining the quality of individual tags. Implicit system usage data and explicit feedback by members are then employed to devise a method for predicting tag quality. As a result, Sen et al. [2007] propose guidelines for designers of tagging systems:

(1) Use systems that both support positive and negative ratings,
(2) Use tag selection methods that normalize each user's influence,
(3) Incorporate both behavioral and rating-based systems, and
(4) Assume that a user's rating for a particular tag application extends to other applications of the tag.

For the same domain, Sigurbjörnsson and van Zwol [2008] present a characterization of tag behavior in Flickr that might be useful for the tag recommendation system and evaluation. They take a random set of Flickr photos to analyze how users tag their uploaded media objects (e.g., photos) and what types of tags are created. Their results show that the tag frequency distribution is associated with a perfect power law and indicate that the middle part of this distribution contains the most interesting tags, which can be used for tag recommendation systems. Furthermore, they find that the generality of the photos are included with only a few tags. On the same platform and to deal with the same problem, Krestel et al. [2009] propose an approach based on latent Dirichlet allocation (LDA) for recommending tags of resources to improve search.

For the interpretation of the relevance of a user-generated tag with respect to the visual content the tag describes, many available approaches are based on intuition that if different people label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. For example, Li et al. [2009] propose a neighbor voting algorithm that accurately and efficiently learns tag relevance by accumulating votes from visual neighbors. Liu et al. [2009] estimate initial relevance scores for the tags based on probability density estimation and then perform a random walk over a tag similarity graph to refine the tags' relevance scores.

In line with these works, Hall and Zarro [2011] compare the metadata created by two different communities: the ipl2 digital library[15] and the social tagging system Delicious.com.[16] Their results show that user-contributed tags from Delicious that have the potential to be used as additional access points for ipl2 digital potentially benefit from user-library resources. The intersection area between the tags applied to ipl2 resources and indexing indicates that the two groups are similar enough to be helpful but are nevertheless dissimilar enough for new access points and description. Furthermore, Nov et al. [2008] present a quantitative study and examine what motivations are associated with tagging levels. Conducting a study of tagging on Flickr, they discover that two of the three motivation categories (Self, Family & Friends, and Public) impact users' tagging levels. They find that the levels of Self and Public motivations, the social presence indicators, and the number of photos have positive impact on tagging level, whereas the Family & Friends motivation is found not to be significantly correlated with the tagging level [Nov et al. 2008].

Finally, an alternative work related to assessment of quality of UGC is proposed by Laniado and Mika [2010]. They analyze the extent to which a hashtag can act as

[15] ipl2 was born as the Internet Public Library in 1995 in a library and information science class taught by Joe Janes at the University of Michigan, with the central motivating question of "what does librarianship have to say to the networked environment and vice-versa?"

[16] Delicious (formerly del.icio.us) is a social tagging Web service for sharing and exploring Web tags.

an identifier for the Semantic Web. By using the vector space model (VSM), Laniado and Mika [2010] propose four metrics to measure this: (1) Frequency refers to a hashtag being used reasonably often by a community of users; (2) Specificity refers to how the usage of a word may differ, depending on whether a hashtag is used or not; (3) Consistency refers to the meaning that may be attributed to a word as a result of the consistent usage of a hashtag by different users in various contexts; and (4) Stability over Time refers to meaning acquired by a hashtag as a result of it being used repeatedly and relentlessly over time.

## 3. END-USER–BASED FRAMEWORK

End-user–based framework approaches use different methods to allow for the differences between individual end-users for adaptive, interactive, or personalized assessment and ranking of UGC. They utilize computational methods to personalize the ranking and assessment process or give an individual end-user the opportunity to interact with the system, explore content, personally define the expected value, and rank content in accordance with individual user requirements. These approaches can also be categorized in two main groups: human centered approaches, also referred to as interactive and adaptive approaches, and machine-centered approaches, also referred to as personalized approaches. The main difference between interactive and adaptive systems compared to personalized systems is that they do not explicitly or implicitly use users' previous common actions and activities to assess and rank the content. However, they give users opportunities to interact with the system and explore the content space to find content suited to their requirements. Figure 4 provides an overview of end-user assessment and ranking of UGC approaches. In the following, an overview of these approaches is provided.

### 3.1. Human-Centered Method

The human-centered method enables an end-user to interact with the system to specify user's own notion of value and to adapt ranking of content according to preferences and the specific task at hand. The term *adaptation* refers to a process in which an interactive system (adaptive system) adapts its behavior to individual users based on information acquired about its end-users and their environment.

Interactive and adaptive approaches create interfaces that enable users to more efficiently browse their feed by providing a browsable access to all content in a user's feed and allowing users to more easily find content relevant to their interests. For example, OpinionSpace [Faridani et al. 2010], based on the commenters' responses to a short value-based questionnaire, visualizes the individual comments in a Web forum on a two-dimensional map. By exploring this space, readers are able to access a range of comments and find information, thus engaging with the opinion and view of someone with different values.

In particular, some adaptive ranking solutions have focused on topic-based browsing, which groups the comments into coherent topics and creates interfaces that allow users to browse their feeds more efficiently. Abel et al. [2011] propose strategies for inferring topic facets and facet values on Twitter by enriching the semantics of individual Twitter messages. Topic modeling–based methods (both on users and content) feature prominently in this space [Ramage et al. 2010; Chen et al. 2010; Sriram et al. 2010]. Bernstein et al. [2010] propose a more focused approach for ordering a user's feed into consistent clusters of topics. This means that the proposed framework clusters tweets in a user's feed into topics that have been discussed explicitly or implicitly, enabling users to browse for subjects that appeal to them. For clustering comments into coherent topics, an algorithm has been created for recognizing topics in short status

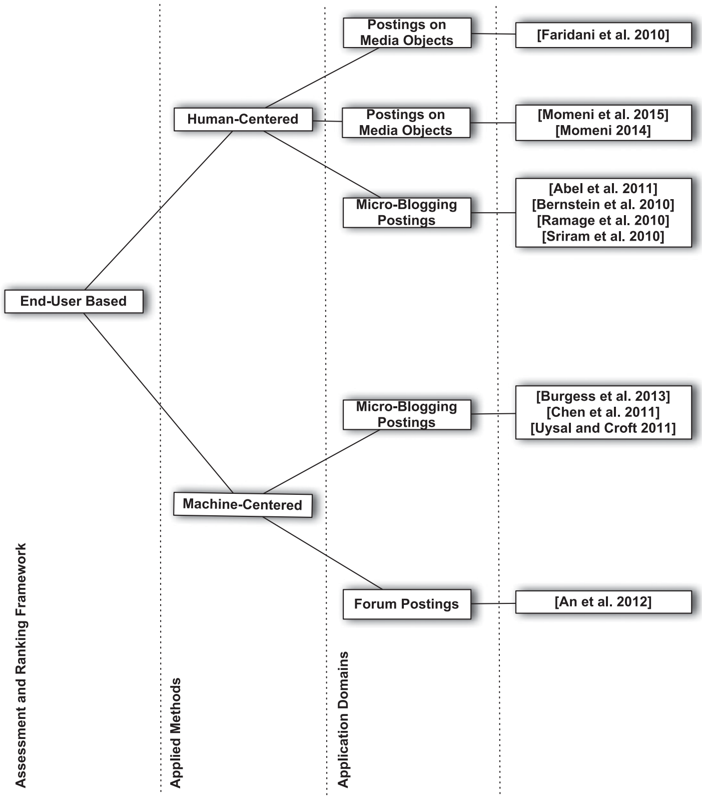

Fig. 4. Overview of approaches related to an end-user-based framework.

updates. Evaluating the algorithm reveals that enrichment of text (by calling out to search engines) outperforms other approaches by using simple syntactic conversion.

Finally, most of these works propose topic-based browsing for microblogging platforms such as Twitter, which group the microblogging postings into coherent topics and create interfaces that enable users to browse their feeds more efficiently. However, there are still difficulties: user-generated comments are longer and have an informal structure. Users can converse, express their subjective opinions and emotions, and describe informative useful information about a media resource. Thus, the topics discussed alongside comments can be unorganized/noisy. Furthermore, as comments have multiple explicit dimensions (language tone, physiological aspects, etc.), grouping them exclusively based on topic results in a single imperfect faceted ranking that does not allow users to rank comments with regard to other potentially useful facets. Therefore, by extracting different types of semantic facets from comments, the solution proposed by Momeni et al. enables the ranking of comments with regard to different dimensions of comments and not only with regard to topics [Momeni 2014; Momeni et al. 2015].

### 3.2. Machine-Centered Method

The machine-centered method utilizes computational methods, particularly machine-learning methods, to develop a ranking and assessment function that learns from a particular end-user's preferences, background, or online social interactions and connections to personalize the ranking and assessment process.

Personalization approaches assess and rank UGC relevant to the individual user, taking into account how the user acted previously, in what activities the user participated, what implicit behavior and preferences can be observed, and what details were explicitly provided. Accordingly, Chen et al. [2011] demonstrated the diversity in usage purpose and preference among end-users—some subjects use microblogging platforms for social purposes, whereas others report high informational usage. Furthermore, they found that the performance of the same assessment and ranking methods can be significantly different for various users whose usage purposes and preferences are different. Methods utilizing tie strength performed significantly better for subjects with high social purpose than for subjects with high informational usage. Examples of personalized approaches for different application domains are listed next.

With regard to postings on microblogging platforms, Burgess et al. [2013] propose BUTTERWORTH, which is a service that helps users find content more relevant to their interest on their feeds without using explicit user input. BUTTERWORTH automatically generates a set of rankers by clustering subcommunities of users' contact based on the common content they produce. The proposed service is composed of three main components. First, the "list generator" groups friends into lists by examining their social contact. Second, the "list labeler" generates a human-readable label representing the topic of the list. Third, the "topic ranker" trains ranking models for core topics. The models can then be utilized to order the user's feed by the selected topic [Burgess et al. 2013]. For the same application domain, Uysal and Croft [2011] propose a personalized ranking of tweets by exploiting users' retweeting patterns and conduct a pilot user study to explore the correlation between retweeting and the interestingness of the tweets for an individual user.

For postings on online media objects, An et al. [2012] propose a model that uses the co-subscriptions relationships inferred by Twitter links and maps the news media sources along a dimensional dichotomous political spectrum. Their result reveals extreme polarization among media sources, which indicates that the political dichotomy naturally appears on Twitter in the media subscription patterns of users.

## 4. DESIGNER-BASED FRAMEWORK

Approaches that fall under the designer-based framework encode the software designer's values in the ranking method, such as a design that maximizes diversity among the displayed UGC items so that certain elements are not redundant, a design that provides "balanced" views of UGC around an issue (e.g., a review site that explicitly samples from the diverse positive and negative reviews), or a design that ranks relevant content to a particular topic or event.

Approaches in this category mainly utilize machine-centered methods to rank content. An overview of these approaches is found next.

*Approach for providing balanced or diverse views of UGCs around an issue.* When applying an information filtering system (e.g., recommender systems, aggregators, search engines, and feed ranking algorithms), users sometimes explicitly choose information filters that isolate themselves in information bubbles. This only partly happens because of their own volition—some of them hardly even notice. Therefore, their views on a particular issue will be influenced and, even worse, will be more blurred by these filters, which may be difficult to correct.

Thus, developing systems and algorithms that encourage users toward more diverse exposure or developing diversity-aware aggregators have increasingly attracted attention in recent years. It is shown by Freelon et al. [2012] that users take significant advantage of three key opportunities to engage with diversity regarding political views: accessing, considering, and producing arguments on both sides of various policy proposals.

In line with this category of work, Park et al. [2009] show that the presentation of multiple articles about the same news event, which emphasize different aspects of the event, enables users to read more diverse news stories. Furthermore, they present the benefit of showing agreeable items on the front page with challenging items on the same topic linked to the agreeable item page.

Munson et al. [2013] have developed a browser extension. This extension adds feedback on the left-right balanced views of news articles. This feedback shows the norm of balanced exposure and furthermore creates accountability. Their experimental results suggest a small but obvious change in reading behavior among users by seeing the feedback. Furthermore, Munson and Resnick [2010] investigate the relationship between readers' satisfaction and the number of supporting and challenging items in a collection of political opinion items. More precisely, they evaluate whether highlighting agreeable items or showing them first can increase satisfaction when fewer agreeable items are present. Their results show that some users are "diversity seeking," whereas others are "challenge averse." For challenge-averse readers, highlighting does not increase overall satisfaction, although it appears to give satisfaction with sets of mostly agreeable items. However, generally ordering agreeable content first appears to decrease satisfaction rather than increase it.

Finally, it is shown by Liao and Fu [2013] that even when diverse views are presented side by side, information selection leads to more noticeable selective exposure to their current viewpoints. Furthermore, users selection of information is influenced by various factors, such as perceived threat and topic involvement. A perceived threat induces selective exposure to viewpoints consistent with information on topics in which participants had low involvement.

## 5. HYBRID ASSESSMENT AND RANKING APPROACHES

Recently, there have been bodies of assessment and ranking approaches that do not fall explicitly under any of the introduced categories. Nevertheless, they take advantage of different categories and are combined approaches.

We believe that combined and hybrid approaches have lots of potential for further development, as they can benefit from the strengths of various previously discussed frameworks to develop more sophisticated and useful techniques for assessment and ranking of UGC. For example, the development of systems that use crowd behaviors (by utilizing the human-centered method) for individual behaviors (learn from crowd behaviors for personalized assessment of content) or learn personalized models for a smaller group (e.g., geographical or other demographic-driven factors that provide individualized content for a group). Nevertheless, there is still less consideration of the combined and hybrid approaches. Different examples of available hybrid approaches can be found next.

***Leveraging a community-based framework for an end-user framework.*** Hu et al. [2013] propose Whoo.ly, a Web service that provides "neighborhood-specific" information based on Twitter posts for an individual end-user (a personalized approach). By utilizing activities of a crowd of end-users, the service provides four types of hyperlocal content: (1) active events (current events in the locality by using a statistical event detector that identifies and groups popular features in tweets), (2) top topics (most-used terms and phrases from recent tweets using a simple topic modeling method), (3) popular places (most popular checked-in/mentioned places using both template-based and learning-based information extractors), and (4) active people (Twitter users mentioned the most, using a ranking scheme on the social graph of users) [Hu et al. 2013].

Another example in this group is a platform proposed by Diakopoulos et al. [2012], where they examine how journalists filter and assess the variety of trustworthy tweets found through Twitter. The proposed platform gives an individual end-user (a journalist) the chance to interact with the system and explore several computational information cues, which were trained using a crowd of humans. They have introduced three types of cues: (1) two classifiers, in which the first classifier classifies users into three types—organizations, journalists, or ordinary people [Choudhury et al. 2012]—and the second classifier identifies users who might be eyewitnesses to the event; (2) characteristics of the content that are shared by the sources; and (3) characteristics that refer to the event location. With regard to the second classifier, detecting the presence of eyewitnesses is achieved by using supervised learning with manually labeled training examples that include text features.

In particular, some of these approaches leverage the patterns of assessment and ranking settings by end-users to minimize the cost of changing settings for an end-user, generally leveraging ideas from collaborative filtering and recommender systems [Lampe et al. 2007; Hong et al. 2012; Uysal and Croft 2011]. For postings on online forums, it is recommended by Lampe et al. [2007] that for ranking comments, patterns recognized by setting filters of users can be used to minimize the cost of settings for other users. One suggested strategy is creating static schema that take into consideration the filtering patterns of different groups of viewers. Another strategy is the setting of filtering thresholds for each conversational tree dynamically, based on the selections of previous viewers. This shows that selections previously made by readers are much more helpful than content of postings for this task (e.g., the ratings of those comments). Moreover, it is discovered that users can be grouped in three categories: "those who never change the default comment display," "those who use ratings to modify the display," and "those who change the comment display to suppress ratings" [Lampe et al. 2007]. In addition, a large number of users do not change from system set default setting. For the same application domain, Hong et al. [2012] explore the creation of ranking systems by proposing a probabilistic latent factor model for social feeds from the perspective of LinkedIn.[17] Principally, they convey this task as an intersection of "learning to rank," "collaborative filtering," and "clickthrough modeling."

***Leveraging and end-user framework for a designer-based framework.*** ConsiderIt [Kriplean et al. 2012a] encodes designer value by leveraging personal interest of an individual end-user (by enabling the end-user to interact with the system). It enables end-users to create personal pro/con lists for an issue and to browse other people's pro/con lists for identifying items they might want to include in their own lists. End-users can see ranked lists of items that were popular on pro or con lists of both "supporters" and "opponents" of a proposition. The system encourages people to listen to others. Reflect [Kriplean et al. 2012b] is another example of such a system where a listening box is added next to each comment, enabling users to briefly and clearly express the points that the commenter makes and modify the comment sections of Web pages at the same time. This is a motivation to listen to other users. Other users can then read the original comment and the listeners' explanations of what has been said. This supports a wider understanding of the discussion.

Giannopoulos et al. [2012] investigate an approach that encodes designer value by leveraging personal interest of an individual end-user (by leveraging previous activities of the end-user). The proposed approach diversifies user comments on news articles by extracting the respective diversification dimensions in the form of feature vectors. These involve "commenting behavior of the respective users," "content similarity," "sentiment expressed within comments," and "article's named entities also found within comments." A preliminary qualitative analysis demonstrates that the diversity criteria result in distinctively diverse subsets of comments compared to a baseline of diverse comments only with regard to their content (textual similarity).

[17] LinkedIn.com is a social networking Web site for people in professional occupations.

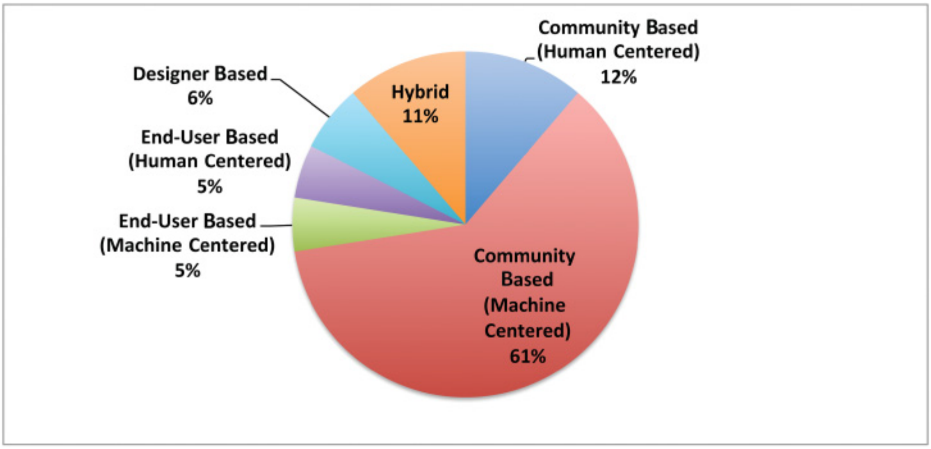

Fig. 5. Percentages of different frameworks related to available approaches for ranking and assessment approaches of UGC.

Finally, Memeorandum,[18] the political news aggregator, may be also mentioned in this group, as it groups items by topics. The front page includes abstracts for top items with links to other items on the same topic below the abstract. To appeal to individual end-users' diversity-seeking behavior, the display can be adapted to user preferences with more challenging items appearing in the links to show a top-level item and the abstract for any topic from a supportive source.

## 6. FINDINGS: OPPORTUNITIES AND CHALLENGES FOR ONGOING RESEARCH

The results of a systematic review of approaches for assessing and ranking UGC are now presented. In this section, we will discuss the main observations, factors to consider, challenges, and opportunities for future research that have transpired from this study.

### 6.1. Observations and Important Factors to Consider

The existing approaches generally adopt one of four frameworks: the community-based framework, which employs human-centered or machine-centered methods; the end-user–based framework, which also employs human-centered or machine-centered methods; the designer-based framework, which mainly employs machine-centered methods; and the hybrid framework, which leverages and combines the advantages of different frameworks and methods to enable more advanced assessment and ranking of UGC. Next, we discuss the main observations with a focus on these four frameworks concerning three aspects: values, applied methods, and application domains.

Figure 5 shows percentages of different frameworks related to available approaches for ranking and assessment approaches of UGC. In recent years, the number of approaches related to end-user-based and designer-based frameworks has increased. Despite this increase, end-user and designer-driven frameworks have received little consideration compared to community-based frameworks, whereas hybrid frameworks have received more attention recently.

*6.1.1. Observations for Community-Based Assessment and Ranking Framework.* We have observed thatmost of the available research approaches related to community-based

[18] http://www.memeorandum.com.

ranking and assessing of UGC utilize machine-centered methods. Nevertheless, default methods utilized by many platforms are human centered.

*Important factors for human-centered methods.* It is important to consider that when human-centered methods (distributed moderation or crowd-based methods) are utilized for ranking and assessment systems, participation and contribution in the human-centered methods are basically voluntary, and accordingly, methods to incentivize contributors need to be developed to allocate rewards [Ghosh 2012]. In addition, the context or users' awareness of previous votes by a crowd of end-users on particular elements of content (e.g., a product review) needs to taken into consideration in that it affects the quality of the new vote [Muchnik et al. 2013; Sipos et al. 2014; Danescu-Niculescu-Mizil et al. 2009]. These issues of social engineering, such as motivation and social feedback loops, can make designing effective human-centered methods challenging in practice.

*Bias of judgments of a crowd of end-users.* Examining machine-centered methods more closely reveals that some machine-centered assessment approaches use judgments of a crowd of end-users on the content to create a ground truth, whereas other machine-centered assessment approaches completely exclude such end-user ratings. Understanding the nature of biases in such human-produced training data is essential for characterizing how that bias might be propagated in a machine-centered assessment approach. Three reasons that have been articulated in the literature for excluding such human ratings include the following:

(1) Different biases of crowd-based approaches, such as "imbalance voting," "winner circle," and "early bird," [Liu et al. 2007; Muchnik et al. 2013].
(2) A lack of an explicit definition of value that may be requested by the crowd to assess some application domains. For example, many assessment approaches for classification of product reviews with regard to helpfulness as the value have used either a human-centered or a combination of human- and machine-centered approaches. This is because many product review platforms have explicitly defined and asked a crowd of end-users to assess the helpfulness of product reviews. However, most approaches related to assessment of credibility exclude judgments of a crowd of end-users because no platforms have asked them for credibility judgments.
(3) Human judgments cannot be as precise as machine-centered judgments in the case of some application domains and values, such as in rating the truthfulness of a product review [Ott et al. 2012].

*Different methods for creating ground truths for machine-centered methods.* Approaches that exclude judgments of crowds of end-users mainly utilize two methods to create a ground truth or training set: (1) using an external crowd (e.g., using crowdsourcing platforms) that independently judges content with regard to a particular value and (2) developing their own coding system for collecting independent judgments from a closed set of users. Both of these methods may of course introduce their own sets of biases into the training data.

*Different machine-centered methods are appropriate for different values and application domains.* With regard to the application domain, a more detailed examination leads us to discover that many proposed machine-centered assessment approaches utilize supervised methods. However, when interconnectedness and interdependency between sets of entities in an application domain (e.g., interdependency between Questions, Answers, and Users in a Q&A domain) occur, assessment and ranking approaches mainly utilize semisupervised learning methods such as co-training or mutually reinforcing approaches [Li et al. 2008]. As supervised and semisupervised methods require adequate amounts of labeled data for an accurate training, the development of adaptive machine-centered methods that can be utilized in different application domains

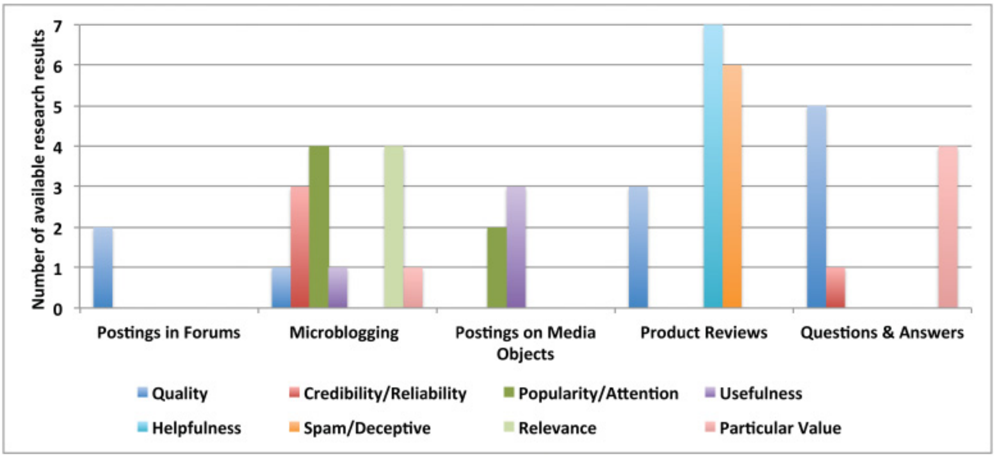

Fig. 6. Values that are important and assessed by different application domains.

is challenging in practice. Therefore, finding a way to optimize the process of labeling and improve the accuracy of hard machine-centered judgments is essential.

*Approaches for providing relevant content around an issue mainly employ unsupervised learning approaches.* A set of approaches related to community-based frameworks that provide relevant content mainly utilize unsupervised machine learning methods. This is because relevancy is influenced by textual features. Therefore, applying unsupervised text clustering methods is effective to minimize the effort of labeling a huge amount of unlabeled content and for maximizing this value.

*The importance of different dimensions of quality (as a value) varies according to the application domain.* With regard to different values that are expected to be maximized, many approaches appear to maximize quality in general, applying a human-centered method as a default ranking method. Nevertheless, with quality being a very general value, some approaches focus on more sophisticated definitions of value and take into consideration different dimensions of quality. In addition, what is defined as value varies with regard to different application domains and specific tasks at hand because different application domains of UGC have different characteristics. Figure 6 shows which values are important and assessed for which application domains.

Many approaches that aim to maximize helpfulness are mainly discussed in the domain of the product review, where judgments of a crowd of end-users are predominantly used as the ground truth to build the prediction model of these approaches. Similar to helpfulness, spam and deception are mainly discussed in the domain of the product review, but they differ from helpfulness in that they mainly exclude judgments of a crowd of end-users. Additionally, approaches related to the assessment of popularity mainly develop their identification and prediction models based on votes by end-users and ratings by crowds (in the case of Twitter, retweets).

Finally, it is observed that most of the available approaches focus on maximizing different dimensions of quality for microblogging platforms, particularly for Twitter, perhaps due to the very simple and structured characteristics of these platforms.

These observations provide opportunities and openings for future work. For instance, in a domain such as Q&A, values related to quality and credibility are clearly useful and have been explored in research, but other values such as helpfulness or usefulness could also be of value, or for a domain such as posting on a media object (e.g., comments on news article), credibility could also be an important value.

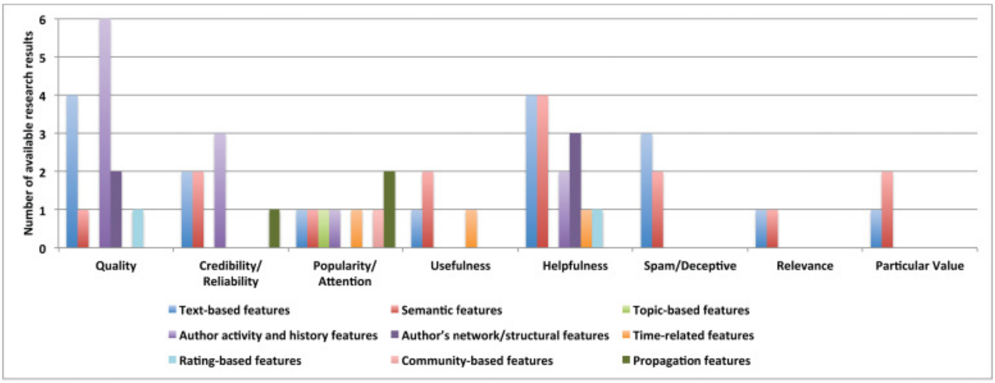

Fig. 7. Influential features sets for assessment and ranking of different values related to various dimensions of quality.

*Influential features vary for different values and application domains.* Some features have high impact when assessing a particular value based on our feature analysis. Figure 7 provides a short overview of the examination of influential features for various values, which are demonstrated by available research results.

A more detailed examination of features leads us to discover that many text- and semantic-based features are important for classifying and clustering UGC in all application domains. It should be noted that the features to be used depend on the notion of the value that is expected to be maximized. For quality, almost all features are helpful to achieve higher assessment accuracy, because quality is in itself a very general notion. Similar to the assessment of quality, popularity requires many features to be used in its assessment, and some of the more important ones include authors' activities, background, networks and structures, and propagation and interactive features. These features related to authors' activities and networks also play an important role when assessing credibility, because features simply related to texts cannot help to assess the credibility of postings. Thus, we require more contextual features to be included. However, in the case of assessing spam and deceptive content, authors can write fake reviews that have been written to appear true and deceive the reader. Accordingly, the features related to the text and semantics of a review are important features to assess spam and deceptive content. Similar to the assessment of spam and deceptive content, text- and semantic-based features are very often influential when assessing relevancy.

In platforms where a particular value is explicitly asked for from the crowd of end-users, rating-based features naturally play important roles for assessment of the value. An example is helpfulness of product reviews in many platforms when judgments on the helpfulness are requested by the crowd of end-users. It is worth noting that time-based features play an important role for assessing helpfulness, usefulness, and popularity. Finally, community-based features are mainly taken into consideration for assessment of postings in forums that include different communities [Wagner et al. 2012a, 2012b].

Based on these observations, our recommendations for the user interface and system designers are as follows:

—Some features have high impact for assessment of a particular value based on our feature analysis. Therefore, for maximizing some values, systems should take into consideration an easier way to build influential features at the design phase. For example, when maximizing value related to usefulness for comments on online media objects (e.g., YouTube videos), the system should encourage users and provide them with the opportunity to define references for enriching the content semantically

[Momeni et al. 2013a]. In addition, value related to credibility should take authors' profile pages into consideration [Morris et al. 2012].

—Systems should provide an explicit definition of values that are expected to be precisely judged and assessed by a crowd of end-users. For example, when maximizing value related to usefulness, the system should explicitly define for end-users what is considered to be useful content. By explicitly articulating value definitions, this should improve the consistency and reliability of such crowd ratings.

*6.1.2. Observations for End-User Assessment and Ranking Framework.* A detailed exploration of available approaches for end-user assessment and ranking reveals that most of the available end-user assessment and ranking approaches focus on maximizing different values mainly for two application domains: postings in microblogging platforms and forums. These approaches can be divided into two groups: interactive and adaptive (human-centered) approaches and personalized (machine-centered) approaches. The main difference between these two categories is that interactive and adaptive approaches in contrast to personalized approaches, which utilize machine-centered methods, do not explicitly or implicitly use a user's previous common actions and content to assess and rank the content. However, they use human-centered methods to rank content and provide individual end-users with opportunities to interact with the system and explore the ranked content to find content to match their requirements.

*Interfaces for enabling users to more efficiently browse their feeds*. Interfaces created by interactive and adaptive approaches permit users to browse their feeds more efficiently by providing ready access to all content in a user's feed and also enabling users to find content related to their own interests [Faridani et al. 2010].

At the backend of these interfaces, there is an algorithm that extracts a set of computational information cues with regard to the context and social feed of a user. In particular, some interactive and adaptive ranking solutions have focused on topic-based [Bernstein et al. 2010; Abel et al. 2011] browsing, which groups the comments into coherent topics and creates interfaces that allow users to browse their feeds more efficiently by exploring clusters of topics. However, computation of these approaches is costly and noisy, and requires too much adjustment to work effectively across a large number of users because users prefer to remove superfluous words from a short posting (e.g., tweets) to save space. Furthermore, as UGCs have multiple explicit dimensions (language tone, physiological aspects, etc.), grouping them exclusively based on topic results in a single imperfect single-dimension ranking that does not allow users to rank content with regard to other potentially useful dimensions. Therefore, the exploration of different types of semantics from content to enable the ranking of content concerning different dimensions provides opportunities for future work.

*Algorithms for learning from an end-user's preferences, background, and connections.* Personalized approaches are based on an algorithm that learns from a particular end-user's preferences, background, or online social interactions and connections to personalize the ranking and assessment process [Burgess et al. 2013; An et al. 2012; Uysal and Croft 2011]. Nevertheless, the number of research results for development of advanced personalized approaches is very low. Furthermore, available approaches are mainly based on the concept of *collaborative filtering*, which isolate users in information bubbles. Therefore, their views on a particular issue will be influenced and, even worse, more distorted by these filters, which may be difficult to correct. Thus, developing systems and algorithms that provide users more diverse and balanced information may be an interesting challenge for future work.

*6.1.3. Observations for Designer-Based Assessment and Ranking of UGC.* As described previously, the decision of the platform's designer partially influences the *definition* of the value for every type of assessment and ranking framework (community based or end-user based). Designers choose definitions for values either because they can be understood and rated by a community or because they can be operationalized for machine-centered methods. Designers can also introduce other objectives into rankings, including desires to try to optimize more than one value simultaneously. For instance, approaches identified under the designer-based assessment and ranking framework mainly focus on the development of systems that encourage users toward more diverse exposure to content, essentially diversity-aware rankings. Although rankings seek to optimize a set of objects along a dimension of interest, diversity instead seeks to introduce another dimension or dimensions into that ranking so that a *set* of content achieves balance across the values of interest.

Approaches in this category mainly utilize machine-centered methods and primarily focus on (1) development of alternative measures of diversity for UGC sets, (2) development of algorithms and systems for selecting content sets that jointly optimize diversity of presentation, or (3) development of alternative selection and presentation methods on users' desire to apply their exposure or an aggregation service to diverse opinions.

*6.1.4. Observations for Hybrid Assessment and Ranking of UGC.* The advantage of the hybrid framework is that it has different categories and combined approaches. Accordingly, it has high potential for developing more sophisticated and useful techniques. Nevertheless, it has received inadequate attention. Current approaches mainly focus on the following: (1) leveraging a community-based framework for an end-user framework, such as (a) a Web service that provides "neighborhood-specific" information based on Twitter posts for an individual end-user by utilizin activities of a crowd of end-users, or (b) a platform that examines how journalists filter and assess the variety of trustworthy tweets found on Twitter and gives an individual end-user (a journalist) the chance to interact with the system and explore a number of computational information cues, trained using a crowd of humans; (2) leveraging the end-user framework for the designer-based framework, such as (a) approaches that leverage the patterns of assessment and rank settings by end-users to minimize the cost of changing settings for another end-user, (b) a framework that encodes designer value by leveraging personal interest of an individual end-user, thus enabling end-users to create personal pro/con lists for an issue and browse other people's pro/con lists to identify items they might want to include in their own lists, (c) a framework that codes designer value by leveraging previous activities of the end-user and diversifies user comments on news articles, or (d) a system that adds a listening box next to each comment, enabling users to briefly and clearly express the points that the commenter makes and modify the comment sections of Web pages at the same time. This is a motivation to listen to other users.

### 6.2. Challenges and Opportunities for Future Work

Based on the aforementioned observations and analyses of results, next we list several open issues and limitations of the available approaches. Addressing these issues and limitations creates natural avenues and opportunities for future work:

—Bridging the conceptual gap between human-centered and machine-centered approaches receives little attention, triggering many technical challenges. These include how to develop algorithms and methods for mitigating biases of the crowd, how to take advantage of semisupervised learning such as active learning for efficient integration of the crowd into machine-centered approaches, or how to utilize a crowd to optimize the process of labeling large amounts of unlabeled UGC and improve the accuracy of hard machine-centered judgments.

—Maximizing some values related to various dimensions of quality for some application domains receives less consideration. In other words, some dimensions of quality are analyzed only for one application domain. For example, credibility is mainly discussed and analyzed in the domain of microblogging platforms. This may be due to the very simple and structured characteristics of these platforms. Nevertheless, credibility, for example, may be an important value for other application domains, such as commenting systems for news articles or answers in Q&A platforms. Therefore, it is important to find out to what extent the impact of the influential features for different dimensions of quality vary with regard to various application domains.

—The development of methods to incentivize high-quality UGC has not been completed, thus triggering challenges such as how advancement of computational methods (e.g., game-theoretic foundations) can help incentivize high-quality UGC and advanced development of assessment and ranking approaches [Turnbull 2007]. In line with this challenge, several avenues for future development of the game-theoretic approaches can be found [Ghosh 2012; Jain et al. 2009]. For example, a multidimensional model of quality is a more realistic representation of the value of a single contribution. Users' decisions, at various times after monitoring the existing set of contributions from other users, are then influenced by the content that they have contributed. Therefore, for a more accurate incentivizing model, the temporal aspect of UGC may be taken into consideration (a sequential model may be better suited to many UGC environments) [Ghosh 2012]. Recently, some approaches have become available [Anderson et al. 2013; Cavallo and Jain 2012] for incentivizing users on the Web or crowd-sourcing platforms. For short free textual content, similar cases can also be implemented.

—Approaches aiming to accommodate individual differences in the assessment and ranking of UGC, and in general end-user-based frameworks, have received inadequate attention. In other words, how can we help people make personal assessments of a particular value rather than rely on particular sources as authorities for ground truth? Most of the available approaches rely on particular sources of ground truth and do not enable users to make personal assessments of a particular value. For example, most of the work on identification of helpfulness of product reviews creates and develops prediction models based on a set of majority-agreement labeled reviews. However, helpfulness is a subjective concept that can vary for different individual users, and therefore it is important that systems help individuals make personal assessments of a particular value.

—The additional presentation techniques receive scant attention. These include more sophisticated displays of challenging and diverse content for supporting balanced or diverse views around an issue. In line with this challenge, Munson and Resnick [2010] suggest that rather than trying to increase the percentage of challenging information in the collections shown to challenge-averse readers, it may be more effective when answering the needs of those who seek diversity to provide them with the means to spread insights they have gained from challenging content to the people who avoid such exposure in their everyday news reading. Furthermore, there are bodies of works in the context of information retrieval that maximize diversity among the displayed items so that certain elements are not redundant. However, there is lack of attention given to such work, particularly for UGC.

—Approaches that focus on particular values and take into consideration requirements of platform owners have received insufficient consideration. A few approaches are related to particular values, such as distinguishing subjective from objective [Li et al. 2008] or conversational from informational content [Harper et al. 2009]. These approaches will enable further development of advanced machine-centered approaches (e.g., advanced recommendation or retrieval systems), thus helping end-users to access more appropriate content.

—Approaches focusing on system objectives (e.g., increasing the throughput, retrievability, and navigability), or the so-called system-centered framework, have received inadequate attention. Nevertheless, they are still heavily influenced by the designers. Recently, some approaches have become available for other types of content. Adding another textual link to an article could have an effect on the quantifiable navigability of the article graph [West et al. 2015] or assessing the effect of adding more correct but generic tags on retrievability and useful images [Robertson et al. 2009] are examples of such approaches. Considering these examples, for short free textual content, similar cases can be implemented, such as encouraging users and providing them with the opportunity to define references for enriching the content semantically. These may have an effect on the quantifiable retrievability of content.

—Finally, combined and hybrid approaches deserve more attention, as we believe that combined and hybrid approaches have significant potential for further development because they can benefit from the advantages of various frameworks discussed in this article to develop more sophisticated and advanced techniques for assessment and ranking of UGC, such as the development of systems that learn from crowd behaviors to personalize assessment and ranking of content or the development of personalized models for a smaller crowd (geographically or other demographically driven measures that produce individualized/adapted content for a crowd).

## APPENDIX

Table I provides a short overview of main contributions, evaluation methods, or experimental datasets of each discussed approach and study. In the table, “C” indicates community-based framework, “E” indicates end-user–based framework, “D”indicates designer-based framework, “H” indicates hybrid framework, and “CS” indicates a case study. For approaches related to the end-user–based framework, the third column of Table I, instead of the value, shows the related proposed method.

Table I. Overview of Main Contributions and Experimental Datasets

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| | | Postings in Microblogging | |
| Castillo et al. [2011] | C | Credibility | **Contribution:** Identifying credible information on Twitter.<br>**Dataset:** Collected a set of 10,000 tweets related to events and used the Mechanical Turk coders for labeling credibility of tweets. |
| Canini et al. [2011] | C | Credibility | **Contribution:** Finding credible information sources in social media.<br>**Dataset:** Selected 5 domains of expertise and then manually selected 10 users from Twitter (using WeFollow service) with high relevancy and expertise for those domains. |
| Morris et al. [2012] | C | Credibility | **Contribution:** Understanding microblog credibility perceptions.<br>**Dataset:** Conducted a survey with selected participants. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| Becker et al. [2011b, 2012] | C | Relevance | **Contribution:** Identifying quality content for planned events across social media sites.<br>**Dataset:** Compiled a dataset of events from 4 different platforms—Last.fm events, EventBrite, LinkedIn events, and Facebook events—and also gathered social media posts for the events from 3 social media platforms—Twitter, YouTube, and Flickr—between May 13, 2011, and June 11, 2011. |
| Hong et al. [2011] | C | Popularity | **Contribution:** Predicting popularity of tweets.<br>**Dataset:** Collected 10,612,601 tweets and social contexts of 2,541,178 users who were active in November and December 2009. Popularity is calculated by the number of retweets. |
| Laniado and Mika [2010] | C | Quality | **Contribution:** Identifying identical and representative tweets.<br>**Dataset:** Collected a dataset of 539,432,680 tweets during November 2009. |
| Tao et al. [2012] | C | Relevance | **Contribution:** Predicting relevancy of tweets.<br>**Dataset:** Used the Twitter corpus that had been used in the microblog track of TREC 2011. |
| Becker et al. [2011a] | C | Relevance | **Contribution:** Identifying real-world events from Twitter.<br>**Dataset:** Used a dataset of 2,600,000 tweets in February 2010 and used human coders to label clusters for both the training and testing phases of the experiments. |
| Chen et al. [2010] | C | Relevance | **Contribution:** Experiments on recommending content of social feeds.<br>**Dataset:** Conducted a pilot interview to obtain qualitative feedback and then conducted a controlled field study to gather quantitative results (conducted a field experiment on the publicly available news recommender Web site based on Twitter zerozero88.com). |
| Rowe et al. [2011] | C | Attention | **Contribution:** Predicting discussions, which attract high attention.<br>**Dataset:** Used two online datasets of tweets (available at http://infochimps.com/). |
| Alonso et al. [2013] | C | Interestingness | **Contribution:** Predicting interestingness of tweets.<br>**Dataset:** Collected 9,990 tweets, sampled at random from Twitter firehose between August 7, 2011, and October 1, 2011. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| Bernstein et al. [2010] | E | Interactive & Adaptive | **Contribution:** Providing interactive topic-based browsing of social status streams.<br>**Dataset:** Conducted a laboratory study for evaluating to what extent the proposed framework (Eddi) performs better for browsing a personal feed than the standard reverse chronological ranking strategy. |
| Abel et al. [2011] | E | Interactive & Adaptive | **Contribution:** Providing adaptive faceted search for Twitter.<br>**Dataset:**Collected more than 30 million tweets by monitoring the Twitter activities of more than 20,000 Twitter users for more than 4 months. |
| Sriram et al. [2010] | E | Personalized | **Contribution:** Classification of tweets to improve information filtering.<br>**Dataset:** Composed a collection of 5,407 tweets from 684 authors and then manually labeled them. |
| Burgess et al. [2013] | E | Personalized | **Contribution:** Leveraging noisy lists for ranking of social feeds.<br>**Dataset:** Collected a set of 10 lists (with different topics) from Muckrack.com (each list includes up to 500 users), found the creator of each list, provided a set of nearly 400,000 users, and randomly sampled 100 users from the follower set. |
| An et al. [2012] | E | Personalized | **Contribution:** Visualizing media bias through Twitter.<br>**Dataset:** Collected profiles of 54 million users, 1.9 billion directed follow links among these users, and all 1.7 billion public tweets that were ever posted by the collected users. |
| Uysal and Croft [2011] | E | Personalized | **Contribution:** Ranking user-oriented tweets.<br>**Dataset:** Crawled 24,200 tweets; for each seed user, randomly selected 100 tweets that would appear on the user's Twitter feed. |
| Hu et al. [2013] | H | Community Based | **Contribution:** Providing personalized information seeking for hyperlocal communities using social media.<br>**Dataset:** Used a within-subjects comparison of Whoo.ly and Twitter where users completed tasks to search for information on each platform and then provided feedback. |
| Diakopoulos et al. [2012]; Choudhury et al. [2012], | H | Community Based and Interactive & Adaptive | **Contribution:** Assessing social media information sources in the context of journalism.<br>**Dataset:** Collected 3 sets of 13,423 tweets related to events in 2011. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| Chen et al. [2011] | CS | Personalized | **Contribution:** Recommending conversations in social streams.<br>**Dataset:** Conducted a user study using zerozero88.com. |
| | | Product Review | |
| Jindal and Liu [2007, 2008] | C | Spam | **Contribution:** Analyzing and detecting review spam.<br>**Dataset:** Crawled 5.8 million reviews written on 6.7 million products by 2.14 reviewers from Amazon.com. |
| Ott et al. [2011, 2012] | C | Deceptive | **Contribution:** Finding deceptive opinion spam.<br>**Dataset:** Created a balanced set of 800 training reviews. The gold-standard deceptive reviews were collected using Amazon Mechanical Turk coders. |
| Li et al. [2014] | C | Deceptive | **Contribution:** Finding deceptive opinion spam.<br>**Dataset:** Collected a dataset of deceptive opinions from different domains using Amazon Mechanical Turk coders and also asking domains' experts. |
| Yoo and Gretzel [2009] | C | Deceptive | **Contribution:** Finding deceptive opinion spam.<br>**Dataset:**Crawled 40 deceptive hotel reviews from students who studied tourism marketing and extracted truthful reviews from TripAdvisor.com. |
| Ghose and Ipeirotis [2007, 2011] | C | Helpfulness | **Contribution:** Estimating the helpfulness and economic impact of product reviews.<br>**Dataset:** Compiled a dataset of product reviews and related information about product prices and sales rankings from Amazon.com. |
| Kim et al. [2006] | C | Helpfulness | **Contribution:** Assessing product review helpfulness.<br>**Dataset:** Collected product reviews related to two product categories: "MP3 Players" and "Digital Cameras" from Amazon.com. |
| Liu et al. [2007] | C | Helpfulness | **Contribution:** Detection of low-quality product reviews.<br>**Dataset:** Collected 4,909 reviews from Amazon.com and then hired two human coders to label the reviews. |
| Lu et al. [2010] | C | Helpfulness | **Contribution:** Exploiting social context for predicting quality of product reviews.<br>**Dataset:** Collected reviews, reviewers, and ratings until May 2009 for all products in three groups. For measuring a value of review quality, average rating of the reviews was used. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| O'Mahony and Smyth [2009] | C | Helpfulness | **Contribution:** Assessing product review helpfulness.<br>**Dataset:** Compiled 2 datasets by crawling all reviews before April 2009 from TripAdvisor.com. Reviews were selected from users who had reviewed at least 1 hotel in "Chicago" or "Las Vegas" and had received a minimum of 5 (either positive or negative) opinion votes. |
| Tsur and Rappoport [2009] | C | Helpfulness | **Contribution:** Unsupervised algorithm for selecting the most helpful book reviews.<br>**Dataset:** Tested their system on reviews written for 5 books with 5 different genres from Amazon.com. Labeled each review by three different human coders. |
| Zhang and Varadarajan [2006] | C | Helpfulness | **Contribution:** Scoring utility of product reviews.<br>**Dataset:** Used Amazon.com to obtain a set of reviews. |
| Sipos et al. [2014] | C | Quality | **Contribution:** Studying helpfulness of product reviews.<br>**Dataset:** Selected a set of 595 products from Amazon.com and tracked their reviews daily for a period of 5 months. |
| Danescu-Niculescu-Mizel et al. [2009] | CS | Helpfulness | **Contribution:** Studying helpfulness of product reviews.<br>**Dataset:** Compiled a dataset that contained 4 million reviews (which received at least 10 helpfulness votes) on 675,000 books from Amazon.com. |
| Gilbert and Karahalios [2010] | CS | Deja Reviews | **Contribution:** Understanding deja reviewers.<br>**Dataset:** Downloaded all reviews (in total, 98,191 product reviews) from the 200 best-selling products (which attract relatively too large/small numbers of reviews) for 15 of Amazon's product categories. |
| Comments on Media Objects and Online Forums | | | |
| Siersdorfer et al. [2010] | C | Usefulness | **Contribution:** Identification of usefulness of comments.<br>**Dataset:** Created a test collection by obtaining 756 keywords, searched for "related videos," and gathered the first 500 comments for the video, along with their authors, timestamps, and comment ratings for each video. |
| Momeni et al. [2013a, 2013b] | C | Usefulness | **Contribution:** Prediction of usefulness of comments. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| | | | **Dataset:** Collected 91,778 comments from YouTube videos and 33,273 comments from Flickr photos related to 3 types of topics (topics were extracted from the history timeline of the 20th century provided by About.com) and used CrowdFlower coders for labeling useful comments. |
| Hsu et al. [2009] | C | Quality | **Contribution:** Ranking comments on the social media.<br>**Dataset:** Compiled a corpus from Digg that contained 9,000 Digg stories and 247,004 comments posted by 47,084 individual users in November 2008. |
| Wagner et al. [2012a, 2012b] | C | Attention | **Contribution:** Conducting an empirical analysis of attention patterns in online communities.<br>**Dataset:** Used all data published in 2006 from boards that contained 10 datasets from 10 different community forums. |
| Weimer et al. [2007] | C | Quality | **Contribution:** Assessing the quality of a post in online discussions.<br>**Dataset:** Collected 1,968 rated posts in 1,788 threads from 497 forums on the “Software” category of Nabble.com. |
| Lampe and Resnick [2004] | C | Quality | **Contribution:** Distributed moderation in a large online conversation space.<br>**Dataset:** Created a dataset from usage logs of Slashdot.org between May 31, 2003, and July 30, 2003. The dataset contained 489,948 comments, 293,608 moderations, and 1,576,937 metamoderations. |
| Szabo and Huberman [2010] | C | Popularity | **Contribution:** Predicting the popularity of online content.<br>**Dataset:** Assembled a dataset that contained 29 million Digg stories written by 560,000 users on 2.7 million posts. Also gathered “view-count time series” on 7,146 selected YouTube videos. |
| Veloso et al. [2007] | C | Quality | **Contribution:** Moderation of comments in a large online journalistic environment.<br>**Dataset:** Collected 301,278 comments on 472 stories that were published in Slashdot. |
| Muchnik et al. [2013] | C | Quality | **Contribution:** Analyzing social influence bias.<br>**Dataset:** For manipulating the votes, 101,281 comments on news articles were randomly assigned to 1 of 3 treatment groups: up-treated, down-treated, or control. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| Faridani et al. [2010] | E | Interactive & Adaptive | **Contribution:** Providing a tool for browsing online comments.<br>**Dataset:** Created three interfaces—"List," "Grid," and "Space"—and presented each of the interfaces in random order to 12 study participants in a within-subject study. |
| Momeni et al. [2014, 2015] | E | Interactive & Adaptive | **Contribution:** Proposing a framework for adaptive faceted ranking of social media comments.<br>**Dataset:** Collected 91,778 comments from YouTube videos related to 3 types of topics (topics were extracted from the history timeline of the 20th century provided by About.com) |
| Munson and Resnick [2010]; Munson et al. [2013] | D | Diversity | **Contribution:** Study on diverse political opinions.<br>**Dataset:** Created a user-installable extension to the Firefox Web browser that augmented Digg and Reddit and tracked click-throughs of 178 people to news stories from those sites. |
| Park et al. [2009] | D | Diversity | **Contribution:** Providing multiple aspects of news.<br>**Dataset:** Conducted 3 types of user studies to evaluate the effectiveness of the proposed framework: clickstream analysis, a questionnaire, and a focus group interview. |
| Lampe et al. [2007] | H | Community Based and Personalized | **Contribution:** Filtering comments on Slashdot.<br>**Dataset:** Assembled a dataset from Slashdot logs, which contained factors that affected how comments were displayed, general user information, and information related to requests of a user. |
| Hong et al. [2012] | H | Community Based and Personalized | **Contribution:** Ranking social feeds.<br>**Dataset:** Created a dataset from the structural data and posts on 99 groups from June 2003 to February 2005 from Usenet. |
| Giannopoulos et al. [2012] | H | Personalized and Designer Based | **Contribution:** Diversifying user comments on news articles.<br>**Dataset:** Collected abstract of a news article talking about the elections of U.S. presidential candidates of 2 U.S. parties and the 11 top diverse comments |
| Kriplean et al. [2012a] | H | Interactive & Adaptive and Designer Based | **Contribution:** Diversifying user comments on news articles. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| | | | **Dataset:** Developed system that was launched on September 21, 2010, to a crowd of 150 at a Seattle City Club event. Articles were selected from the Seattle Times (9/27), KIRO News (10/5), the UW Daily (10/20), and the Yakima Herald (10/27). |
| Kriplean et al. [2012b] | H | Interactive & Adaptive and Designer Based | **Contribution:** Study on promotion of listening to diverse content.<br>**Dataset:** Conducted an evaluation on 3 topics on Slashdot. |
| Diakopoulos and Naaman [2011] | CS | Personalized | **Contribution:** Study on quality of discourse in online news comments.<br>**Dataset:** Conducted an interview with 18 people (including editors, reporters, and moderators). |
| Diakopoulos [2015] | CS | Relevance | **Contribution:** Study on relevancy of comments on the New York Times Web site.<br>**Dataset:** Full text of 2,815 news articles and 331,785 comments on these articles were collected and analyzed for the New York Times Web site (nytimes.com). |
| Liao and Fu [2013] | CS | Diversity | **Contribution:** Study on interactive effects of perceived threat and topic involvement on selective exposure to information.<br>**Dataset:** Conducted a user study by recruiting 28 participants from the Illinois Champaign-Urbana community. |
| Hullman et al. [2015] | CS | Critique | **Contribution:** Studying various commenting types on a data visualization blog.<br>**Dataset:** Collected all comments from 168 posts on the Economist's news articles resulting in a dataset of 4,468 comments across 118 posts containing 1 or more comments. |
| Questions and Answers in Q&A Platforms | | | |
| Bian et al. [2009] | C | Credibility | **Contribution:** Recognizing reliable users and content in Q&A platforms.<br>**Dataset:** Used the TREC Q&A queries; searched for these queries on Yahoo! Answers; and crawled questions, answers, and related user information. |
| Agichtein et al. [2008] | C | Quality | **Contribution:** Finding high-quality content on Q&A platforms.<br>**Dataset:** Created a dataset containing 8,366 Q&A pairs and 6,665 questions from Yahoo! Answers. Acquired basic usage features from a question thread (page views or clicks). |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| Li et al. [2008] | C | Objectivity | **Contribution:** Predicting question subjectivity orientation.<br>**Dataset:** Created a dataset with 1,000 questions from Yahoo! Answers by crawling more than 30,000 questions and gathered labeled for training using the Amazon Mechanical Turk coders. |
| Liu et al. [2008] | C | Quality | **Contribution:** Predicting information seeker satisfaction in a community.<br>**Dataset:** Collected a dataset using a snapshot of Yahoo! Answers in 2008 that contained 216,170 questions. |
| Jeon et al. [2006] | C | Quality | **Contribution:** Predicting the quality of answers with nontextual features.<br>**Dataset:** Assembled a dataset by crawling 6.8 million Q&A pairs from the Naver Q&A. Randomly chose 894 Q&A pairs from the Naver collection and judged the quality of the answers using human coders. |
| Bian et al. [2008] | C | Relevance | **Contribution:** Finding the fact answers from crowd.<br>**Dataset:** Obtained 1,250 TREC factoid questions that included at least 1 similar question from Yahoo! Answers archived from 7 years of the TREC Q&A track evaluations and labeled the data in 2 steps: (1) obtaining the TREC factoid answer patterns and (2) independently and manually labeled to validate the automatic labels obtained from TREC factoid answer patterns. |
| Yu and Hatzivassiloglou [2003] | C | Fact | **Contribution:** Separating facts from opinions.<br>**Dataset:** Used the TREC 2, 8, 9, and 11 collections that included 6 different newswire sources. |
| Harper et al. [2008] | CS | Quality | **Contribution:** Predictors of answer quality on online Q&A sites.<br>**Dataset:** Conducted controlled field study of questions and answers from 3 Q&A sites (Yahoo, Answerbag, and Metafilter). |
| Harper et al. [2009] | CS | Conversational | **Contribution:** Distinguishing informational and conversational questions on social Q&A sites.<br>**Dataset:** Compiled a dataset from 3 Q&A sites (Yahoo, Answerbag, and Metafilter) and developed an online coding tool making use of available volunteers for manual coding. |
| Mamykina et al. [2011] | CS | | **Contribution:** Study of success factors of the StackOverflow Q&A platform. |

(Continued)

Table I. Continued

| Citation | Framework | Value/Method | Experimental Dataset/ Evaluation Method |
|---|---|---|---|
| | | | **Dataset:** Conducted case study on the StackOverflow Q&A sites on a total of 300 thousand registered users who asked 833 thousand questions, provided 2.2 million answers, and posted 2.9 million comments. |
| | | User-Generated Tags | |
| Weinberger et al. [2008] | C | Quality | **Contribution:** Resolving tag ambiguity.<br>**Dataset:** Collected tags on 102 million Flickr photos that were uploaded between February 2004 and December 2007, each photo including at least 1 tag. |
| Sen et al. [2007] | C | Quality | **Contribution:** Identifying quality of tags.<br>**Dataset:** Collected 52,814 tags in 9,055 distinct tag sets from the MovieLens3 movie recommendation system. |
| Hall and Zarro [2011] | CS | Quality | **Contribution:** Study on a comparison of library-created and user-created tags.<br>**Dataset:** Compared the metadata created by 2 different communities: the ipl2 digital library and the social tagging system Delicious. |
| Nov et al. [2008] | CS | Quality | **Contribution:** Study on drives content tagging.<br>**Dataset:** Conducted a quantitative study for examining what motivation factors correlated with tagging levels using Flickr tags. |
| Sigurbjörnsson and van Zwol [2008] | CS | Quality | **Contribution:** Tag recommendation based on collective knowledge.<br>**Dataset:** Used a random set (52 million) from Flickr photos uploaded between February 2004 and June 2007. |
| Krestel et al. [2009] | CS | Quality | **Contribution:** Tag recommendation based on latent Dirichlet allocation.<br>**Dataset:** Used available dataset from Delicious, which consisted of 75,000 users, 500,000 tags, and 3,200,000 resources connected via 17,000,000 tag assignments of users. |
| Li et al. [2009] | CS | Quality | **Contribution:** Analyzing social tag relevance by neighbor voting.<br>**Dataset:** Collected 573,115 unique tags and 272,368 user IDs from Flickr. The number of distinct tags per image varied from 1 to 1,230, with an average value of 5.4. |
| Liu et al. [2009] | CS | Quality | **Contribution:** Tag ranking.<br>**Dataset:** Compiled a dataset composed of 50,000 images and 13,330 unique tags from Flickr. |

## ACKNOWLEDGMENTS

The authors would like to thank Arpita Ghosh from Cornell University, Paul Resnick from the University of Michigan, and the anonymous reviewers for insightful discussions, comments, and advice.

## REFERENCES

Fabian Abel, Ilknur Celik, Geert-Jan Houben, and Patrick Siehndel. 2011. Leveraging the semantics of tweets for adaptive faceted search on Twitter. In *Proceedings of the 10th International Conference on The Semantic Web, Volume Part I (ISWC'11)*.

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, Gilad Mishne, Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM'08)*.

Omar Alonso, Catherine C. Marshall, and Marc Najork. 2013. Are some tweets more interesting than others? #Hardquestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval (HCIR'13)*. ACM, New York, NY.

Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. 2012. Visualizing media bias through Twitter. In *Proceedings of the 6th International AAAI Conference on Weblog and Social Media (ICWSM'12)*.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*.

Marlene Asselin, Teresa Dobson, Eric M. Meyers, Cristina Teixiera, and Linda Ham. 2011. Learning from YouTube: An analysis of information literacy in user discourse. In *Proceedings of the 2011 iConference (iConference'11)*. ACM, New York, NY.

Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. 2012. Identifying content for planned events across social media sites. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*. ACM, New York, NY.

Hila Becker, Mor Naaman, and Luis Gravano. 2011a. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

Hila Becker, Mor Naaman, and Luis Gravano. 2011b. Selecting quality Twitter content for events. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. 2004. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW'04)*. ACM, New York, NY.

Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. 2010. Eddi: Interactive topic-based browsing of social status streams. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST'10)*. ACM, New York, NY.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, NY.

Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. 2009. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, New York, NY.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*. ACM, New York, NY.

Matthew Burgess, Alessandra Mazzia, Eytan Adar, and Michael Cafarella. 2013. Leveraging noisy lists for social feed ranking. In *Proceedings of the 7th International AAAI Conference on Weblog and Social Media (ICWSM'13)*.

Moira Burke, Elisabeth Joyce, Tackjin Kim, Vivek An, and Robert Kraut. 2007. Introductions and requests: Rhetorical strategies that elicit response in online communities. In *Proceedings of the 3rd International Conference on Communities and Technologies (C&T'07)*.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the 2011 IEEE 3rd International Conference on Privacy, Security, Risk, and Trust and the 2011 IEEE 3rd International Conference on Social Computing*. IEEE, Los Alamitos, CA.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM, New York, NY.

Ruggiero Cavallo and Shaili Jain. 2012. Efficient crowdsourcing contests. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Volume 2 (AAMAS'12)*.

Jilin Chen, Rowan Nairn, and Ed Chi. 2011. Speak little and well: Recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY.

Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media.*

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*.

Nicholas Diakopoulos. 2015. The editor's eye: Curation and comment relevance on the New York Times. In *Proceedings of the ACM 2015 Conference on Computer Supported Cooperative Work (CSCW'15)*.

Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM Annual conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY.

Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW'11)*. ACM, New York, NY.

Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: A scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY.

Deen G. Freelon, Travis Kriplean, Jonathan Morgan, W. Lance Bennett, and Alan Borning. 2012. Facilitating diverse political engagement with the Living Voters Guide. *Journal of Information Technology and Politics* 9, 3, 279–297.

Anindya Ghose and Panagiotis G. Ipeirotis. 2007. Designing novel review ranking systems: Predicting the usefulness and impact of reviews. In *Proceedings of the 9th International Conference on Electronic Commerce (ICEC'07)*.

Anindya Ghose and Panagiotis G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23, 10, 1498–1512.

Arpita Ghosh. 2012. Social computing and user-generated content: A game-theoretic approach. In *Proceedings of ACM SIGecom Exchanges (SIGecom'12)*. ACM, New York, NY.

Arpita Ghosh and Patrick Hummel. 2011. A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC'11)*. ACM, New York, NY.

Arpita Ghosh and Preston McAfee. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM, New York, NY.

Arpita Ghosh and Preston McAfee. 2012. Crowdsourcing with endogenous entry. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, New York, NY.

Giorgos Giannopoulos, Ingmar Weber, Alejandro Jaimes, and Timos Sellis. 2012. Diversifying user comments on news articles. In *Web Information Systems Engineering—WISE 2012*. Lecture Notes in Computer Science, Vol. 7651. Springer, 100–113.

Eric Gilbert and Karrie Karahalios. 2010. Understanding deja reviewers. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'10)*. ACM, New York, NY.

Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter* 12, 1, 58–72.

Catherine E. Hall and Michael A. Zarro. 2011. What do you call it? A comparison of library-created and user-created tags. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*. ACM, New York, NY.

F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. 2009. Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY.

F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY.

Liangjie Hong, Ron Bekkerman, Joseph Adler, and Brian D. Davison. 2012. Learning to rank social update streams. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. ACM, New York, NY.

Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW'11)*. ACM, New York, NY.

Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking comments on the social Web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering, Volume 4 (CSE'09)*. IEEE, Los Alamitos, CA.

Yuheng Hu, Shelly D. Farnham, and Andrés Monroy-Hernández. 2013. Whoo.Ly: Facilitating information seeking for hyperlocal communities using social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Crowdsourcing, attention and productivity. *Journal of Information Science* 35, 6, 758–765.

Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. 2015. Content, context, and critique: Commenting on a data visualization blog. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'15)*. ACM, New York, NY.

Shaili Jain, Yiling Chen, and David C. Parkes. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC'09)*. ACM, New York, NY.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY.

Nitin Jindal and Bing Liu. 2007. Analyzing and detecting review spam. In *Proceedings of the 2007 7th IEEE International Conference on Data Mining (ICDM'07)*.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*. ACM, New York, NY.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*.

Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Technical Report TR/SE-0401. Keele University, Keele, England.

Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent Dirichlet allocation for tag recommendation. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*. ACM, New York, NY.

Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012a. Supporting reflective public thought with ConsiderIt. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*. ACM, New York, NY.

Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012b. Is this what you meant? Promoting listening on the Web with Reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY.

Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*. ACM, New York, NY.

Cliff A.C. Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the reader: Filtering comments on Slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, New York, NY.

David Laniado and Peter Mika. 2010. Making sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference on the Semantic Web, Volume Part I (ISWC'10)*.

Baoli Li, Yandong Liu, and Eugene Agichtein. 2008. CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia* 11, 7, 1310–1322.

Q. Vera Liao and Wai-Tat Fu. 2013. Beyond the filter bubble: Interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, New York, NY.

Jingjing Liu, Yunbo Cao, Chin Y. Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*.

Yandong Liu, Jiang Bian, and Eugene Agichtein. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*.

Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest Q&A site in the West. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY.

Elaheh Momeni. 2014. *Adaptive Moderation of User-Generated Content on Web*. Ph.D. Dissertation. Retrieved November 16, 2015, from http://eprints.cs.univie.ac.at/4317/.

Elaheh Momeni, Simon Braendle, and Eytan Adar. 2015. Adaptive faceted ranking for social media comments. In *Advances in Information Retrieval (ECIR'15)*. Springer.

Elaheh Momeni, Claire Cardie, and Myle Ott. 2013a. Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects. In *Proceedings of the 7th International AAAI Conference on Weblog and Social Media (ICWSM'13)*.

Elaheh Momeni, Ke Tao, Bernhard Haslhofer, and Geert-Jan Houben. 2013b. Identification of useful user comments in social media: A case study on Flickr commons. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'13)*. ACM, New York, NY.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the 2012 ACM Conference on Computer Supported Cooperative Work (CSCW'12)*. ACM, New York, NY.

Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146, 647–651.

Sean A. Munson, Stephanie Y. Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th International AAAI Conference on Weblog and Social Media (ICWSM'13)*.

Sean A. Munson and Paul Resnick. 2010. Presenting diverse political opinions: How and how much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY.

Kevin Kyung Nam, Mark S. Ackerman, and Lada A. Adamic. 2009. Questions in, knowledge in? A study of Naver's question answering community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY.

Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. 2012. The design space of opinion measurement interfaces: Exploring recall support for rating and ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY.

Oded Nov, Mor Naaman, and Chen Ye. 2008. What drives content tagging: The case of photos on Flickr. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY.

Michael P. O'Mahony and Barry Smyth. 2009. Learning to recommend helpful hotel reviews. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*. ACM, New York, NY.

Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, New York, NY.

Myle Ott, Yejin Choi, Claire Cardie, and Jeff Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (ACL'11)*.

Tim Paek, Michael Gamon, Scott Counts, David Maxwell Chickering, and Aman Dhesi. 2010. Predicting the importance of newsfeed posts and social network friends. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*.

Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY.

Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22, 1, 457–479.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblog and Social Media (ICWSM'10)*.

Huzefa Rangwala and Salman Jamali. 2010. Defining a coparticipation network using comments on Digg. *IEEE Intelligent Systems* 25, 4, 36–45.

Stephen Robertson, Milan Vojnovic, and Ingmar Weber. 2009. Rethinking the ESP game. In *Extended Abstracts on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY.

Dana Rotman, Jennifer Golbeck, and Jennifer Preece. 2009. The community is where the rapport is—on sense and structure in the YouTube community. In *Proceedings of the 4th International Conference on Communities and Technologies (C&T'09)*. ACM, New York, NY.

Matthew Rowe, Sofia Angeletou, and Harith Alani. 2011. Predicting discussions on the social Semantic Web. In *Proceedings of the 8th Extended Semantic Web Conference on the Semantic Web: Research and Applications, Volume Part II (ESWC'11)*. 405–420.

Shilad Sen, F. Maxwell Harper, Adam LaPitz, and John Riedl. 2007. The quest for quality tags. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP'07)*. ACM, New York, NY.

Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY.

Börkur Sigurbjörnsson and Roelof van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, NY.

Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. 2014. Was this review helpful to you? It depends! Context and voting patterns in online content. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. ACM, New York, NY.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY.

Gabor Szabo and Bernardo A. Huberman. 2010. Predicting the popularity of online content. *Communications of the ACM* 53, 8, 80–88.

Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. 2012. What makes a tweet relevant for a topic? In *Making Sense of Microposts (#MSM2012)*.

Oren Tsur and Ari Rappoport. 2009. RevRank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'09)*.

Don Turnbull. 2007. Rating, voting & ranking: Designing for collaboration & consensus. In *Proceedings of CHI'07 Extended Abstracts on Human Factors in Computing Systems (CHI EA'07)*. ACM, New York, NY.

Ibrahim Uysal and W. Bruce Croft. 2011. User oriented tweet ranking: A filtering approach to microblogs.. In *Proceedings of the Conference on Information and Knowledge Management (CIKM'11)*. ACM, New York, NY.

Adriano Veloso, Wagner Meira Jr., Tiago Macambira, Dorgival Guedes, and Hélio Almeida. 2007. Automatic moderation of comments in a large on-line journalistic environment. In *Proceedings of the International AAAI Conference on Weblog and Social Media (ICWSM'07)*.

Claudia Wagner, Matthew Rowe, Markus Strohmaier, and Harith Alani. 2012a. Ignorance isn't bliss: An empirical analysis of attention patterns in online communities. In *Proceedings of the IEEE International Conference on Social Computing*.

Claudia Wagner, Matthew Rowe, Markus Strohmaier, and Harith Alani. 2012b. What catches your attention? An empirical study of attention patterns in community forums. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*.

Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys* 44, 4, Article No. 25.

Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. 2007. Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07)*.

Kilian Quirin Weinberger, Malcolm Slaney, and Roelof van Zwol. 2008. Resolving tag ambiguity. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, New York, NY.

Robert West, Ashwin Paranjape, and Jure Leskovec. 2015. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*.

Jiang Yang, Mark S. Ackerman, and Lada A. Adamic. 2011. Virtual gifts and guanxi: Supporting social exchange in a Chinese online community. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW'11)*. ACM, New York, NY.

Kyung Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. In *Proceedings of Information and Communication Technologies in Tourism (ENTER'09)*. 37–47.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*.

Zhu Zhang and Balaji Varadarajan. 2006. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY.

Received March 2014; revised August 2015; accepted August 2015