TECHNICAL REPORT
YR-2007-005

FINDING HIGH-QUALITY CONTENT IN SOCIAL MEDIA WITH AN APPLICATION TO COMMUNITY-BASED QUESTION ANSWERING

Eugene Agichtein[1], Carlos Castillo[2], Debora Donato[2],
Aristides Gionis[2], Gilad Mishne[3]

[1] Emory University Atlanta, USA
[2] Yahoo! Research Barcelona Spain
[3] Search and Advertising Sciences, Yahoo!

25 September 2007

Santa Clara, California • Berkeley, California • Burbank, California
New York, New York • Barcelona, Spain • Santiago, Chile

Yahoo! Research Report No. YR-2007-005

Yahoo! Research Report No. YR-2007-005

FINDING HIGH-QUALITY CONTENT IN SOCIAL MEDIA WITH AN APPLICATION TO COMMUNITY-BASED QUESTION ANSWERING

Eugene Agichtein[1], Carlos Castillo[2], Debora Donato[2], Aristides Gionis[2], Gilad Mishne[3]

[1] Emory University Atlanta, USA
[2] Yahoo! Research Barcelona Spain
[3] Search and Advertising Sciences, Yahoo!

25 September 2007

ABSTRACT: The quality of user-generated content varies drastically from excellent to abuse and spam. As the availability of such content increases, the task of identifying high-quality content in sites based on user contributions—social media sites—becomes increasingly important. Social media in general exhibit a rich variety of information sources: in addition to the content itself, there is a wide array of non-content information available, such as links between items and explicit quality ratings from members of the community. In this paper we investigate methods for exploiting such community feedback to automatically identify high quality content. As a test case, we focus on Yahoo! Answers, a large community question answering portal that is particularly rich in the amount and types of content and social interactions available in it. We introduce a general classification framework for combining the evidence from different sources of information that can be tuned automatically for a given social media type and quality definition. In particular, for the community question answering domain, we show that our system is able to separate high-quality items from the rest with an accuracy close to that of humans.

1. Introduction

Recent years have seen a transformation in the type of content available on the web. During the first decade of the web's prominence—from the early 1990s onwards—most online content resembled traditional published material: the majority of web users were consumers of content, created by a relatively small amount of publishers. From the early 2000s, *user-generated content* has become increasingly popular on the web: more and more users participate in content creation, rather than just consumption. Popular user-generated content (or social media) domains include blogs and web forums, social bookmarking sites, photo and video sharing communities, as well as social networking platforms such as Facebook and MySpace, that offer a combination of all of these with an emphasis on the relationships between users of the community.

Community-driven question-answering portals are a particular form of user-generated content that is gaining a large audience in recent years. These portals, in which users answer questions posed by other users, provide an alternative channel of obtaining information on the web: rather than browsing results of search engines, users present detailed information needs—and get direct responses authored by humans. In some markets, this information seeking behavior is dominating over traditional web search [28].

An important difference between user-generated content and traditional content that is particularly significant for knowledge-based mediums such as question-answering portals is the variance in the quality of the content. As Anderson [3] describes, in traditional publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content. This makes the tasks of filtering and ranking in such systems more complex than in other domains. However, for information retrieval tasks social media systems present inherent advantages over traditional collections of documents: their rich structure offers more available data than in other domains. In addition to document content and link structure, social media exhibit a wide variety of user-to-document relation types, and user-to-user interactions.

In this paper we address the task of identifying high-quality content in community-driven question-answering sites, exploring the benefits of the additional sources of information in this domain. As a test case, we focus on Yahoo! Answers, a large portal that is particularly rich in the amount and types of content and social interaction available in it. We focus on the following research questions:

1. *What are the elements of social media that can be used to facilitate automated discovery of high-quality content?* In addition to the content itself, there is a wide array of non-content information available, from links between items to explicit and implicit quality rating from members of the community. What is the utility of each source of information to the task of estimating quality?

2. *How are these different factors related? Is content alone enough for identifying quality items? Can community feedback approximate judgments of specialists?*

To our knowledge, this is the first large-scale study of combining the analysis of the content with the user feedback in social media. In particular, we model all user interactions in a principled graph-based framework (Section 3 and Section 4), allowing us to effectively combine the different sources of evidence in a classification formulation. Furthermore, we investigate the utility of the different sources of feedback in a large-scale, experimental setting (Section 5) over the market leading question-answering portal. Our experimental results show that these sources of evidence are complimentary, and allow our system to exhibit accuracy close to that of humans in identifying high quality content (Section 6). We discuss our findings and directions for future work in Section 7, which concludes this paper.

2. Background and Related Work

Social media content has become indispensable to millions of users. In particular, community question answering portals are a popular destination of users looking for help with a particular situation, for entertainment, and for community interaction. Hence, in this paper we focus on one particularly important manifestation of social media – community question answering sites, specifically on Yahoo! Answers. Our work draws on significant amount of prior research on social media, and we outline the related work before introducing our framework in Section 3.

2.1. Yahoo! Answers

Yahoo! Answers[1] is a question-answering system where people ask and answer questions on any topic. What makes this system interesting is that around a seemingly trivial question-answer paradigm, users are forming a social network characterized by heterogeneous interactions. As a matter of fact, users do not only limit their activity to asking and answering questions, but they also actively participate in regulating the whole system. A user can vote for answers of other users, mark interesting questions, and even report abusive behavior. Thus, overall, each user has a threefold role: asker, answerer and evaluator.

The central element of the Yahoo! Answers system are questions. Each question has a lifecycle. It starts in an "open" state where it receives answers. Then at some point (decided by the asker, or by an automatic timeout in the system), the question is considered "closed," and can receive no further answers. At this stage, a "best answer" is selected either by the asker or through a voting procedure from other users; once a best answer is chosen, the question is "resolved."

As previously noted, the system is partially moderated by the community: any user may report another user's question or answer as violating the community guidelines (e.g., containing spam, adult-oriented content, copyrighted material, etc.). A user can also award

[1] http://answers.yahoo.com/

a question a "star," marking it as an interesting question, sometimes can vote for the best answer for a question, and can give to any answer a "thumbs up" or "thumbs down" rating, corresponding to a positive or negative vote respectively.

Yahoo! Answers is a very popular service (according to some reports, it reached a market share of close to 100% about a year after its launch [26]); as a result, it hosts a very large amount of questions and answers in a wide variety of topics, making it a particularly useful domain for examining content quality in social media. Similar existing and past services (some with a different model) include Amazon's Askville[2], Google Answers[3], and Yedda[4].

2.2. Related Work

Link analysis in social media Link-based methods have been shown to be successful for several tasks in social media [29]. In particular, link-based ranking algorithms that have been shown to be successful in the context of evaluating quality of Web pages have been applied in this context. Two of the most prominent link-analysis algorithms are PageRank [24] and HITS [20].

Consider the graph $G = (V, E)$ with vertex set V corresponding to the users of a question/answer system and having a directed edge $e = (u, v) \in E$ from a user $u \in V$ to a user $v \in V$ if user u has answered to at least one question of user v. ExpertiseRank [31] corresponds to PageRank over the transposed graph $G' = (V, E')$, that is, a score is propagated from the person receiving the answer to the person giving the answer. The recursion implies that if person u was able to provide an answer to person v, and person v was able to provide an answer to person w, then u should receive some extra points given that he/she was able to provide an answer to a person with a certain degree of expertise.

The HITS algorithm was applied over the same graph [7, 17] and it was shown to produce good results in finding experts and/or good answers. The mutual reinforcement process in this case can be interpreted as "good questions attract good answers" and "good answers are given to good questions"; we examine this assumption in Section 5.2.

Propagating reputation Guha et al. [12] study the problem of propagating trust and distrust among Epinions[5] users, who may assign positive (trust) and negative (distrust) ratings to each other. The authors study ways of combining trust and distrust and observe that, while considering trust as a transitive property makes sense, distrust can not be considered transitive.

Ziegler and Lausen [32] also study models for propagation of trust. They present a taxonomy of trust metrics and discuss ways of incorporating information about distrust into the rating scores.

[2] http://askville.amazon.com/
[3] http://answers.google.com/
[4] http://yedda.com/
[5] http://epinions.com/

Question-answering portals and forums The particular context of question-answering communities we focus on in this paper has been the object of some study in recent years. According to Su et al. [30], the quality of answers in question-answering portals is good on average, but the quality of specific answers varies significantly. In particular, in a study of the answers to a set of questions in Yahoo! Answers, the authors found that the fraction of correct answers to specific questions asked by the authors of the study, varied from 17% to 45%. The fraction of questions in their sample with at least one good answer was much higher, varying from 65% to 90%, meaning that a method for finding high-quality answers can have a significant impact in the user's satisfaction with the system.

Jeon et al. [15] extracted a set of features from a sample of answers in Naver,[6] a Korean question answering portal similar to Yahoo! Answers. They built a model for answer quality based on features derived from the particular answer being analyzed, such as answer length, number of points received, etc., as well as user features, such as fraction of best answers, number of answers given, etc. Our work expands on this by exploring a substantially larger range of features including both structural, textual, and community features, and by identifying quality of questions in addition to answer quality.

Expert finding Zhang et al. [31] analyze data from an on-line forum, seeking to identify users with high expertise. They study the user answers graph in which there is a link between users u and v if u answers a question by v, applying both ExpertiseRank and HITS to identify users with high expertise. Their results show high correlation between link-based metrics and the answer quality. The authors also develop synthetic models that capture some of the characteristics of the interactions among users in their dataset.

Jurczyk and Agichtein [18] show an application of the HITS algorithm [20] to a question-answering portal. The HITS algorithm is run on the user-answer graph. The results demonstrate that HITS is a promising approach, as the obtained authority score is better correlated with the number of votes that the items receive, than simply counting the number of answers the answerer has given in the past.

Campbell et al. [7] computed the authority score of HITS over the user-user graph in a network of e-mail exchanges, showing that it is more correlated to quality than other simpler metrics. Dom et al. [9] studied the performance of several link-based algorithms to rank people by expertise on a network of e-mail exchanges, testing on both real and synthetic data, and showing that in real data ExpertiseRank outperforms HITS.

Text analysis for content quality Most work on estimating the quality of text has been in the field of Automated Essay Grading (AES), where writings of students are graded by machines on several aspects, including compositionality, style, accuracy, and soundness. AES systems are typically built as text classification tools, and use a range of properties derived from the text as features. Some of the features employed in systems are lexical, such as word length, measures of vocabulary irregularity via repetitiveness [6] or uncharacteristic

[6] http://naver.com/

co-occurrence [8], and measures of topicality through word and phrase frequencies [27]. Other features take into account usage of punctuation and detection of common grammatical error (such as subject-verb disagreements) via predefined templates [4, 23]. Most platforms are commercial and do not disclose full details of their internal feature set; overall, AES systems have been shown to correlate very well with human judgments [5, 23].

A different area of study involving text quality is *readability*; here, the difficulty of text is analyzed to determine the minimal age group able to comprehend it. Several measures of text readability have been proposed, including the Gunning-Fog Index [13], the Flesch-Kincaid Formula [19], and SMOG Grading [21]. All measures combine the number of syllables or words in the text with the number of sentences—the first being a crude approximation of the syntactic complexity and the second of the semantic complexity. Although simplistic and controversial, these methods are widely-used and provide a rough estimation of the difficulty of text.

Implicit feedback for ranking Implicit feedback from millions of web users has been shown to be a valuable source of result quality and ranking information. In particular, clicks on results and methods for interpreting the clicks have been studied in references [1, 16, 2]. We apply the results on click interpretation on web search results from these studies, as a source of quality information in social media. As we will show, content usage statistics are valuable, but require different interpretation from the web search domain.

3. Content Quality Analysis in Social Media

We now focus on the task finding high quality content, and describe our overall approach to solving this problem. Content quality evaluation is an essential module for performing more advanced information retrieval tasks on the question/answer system. For instance, a quality score can be used as input to ranking algorithms. On a high level, our approach is to exploit features of social media that are intuitively correlated with quality, and then train a classifier to appropriately select and weight the features for each specific type of item, task, and quality definition.

In this section we identify a set of features of social media and interactions that can be applied to the content quality task. In particular, we model the intrinsic content quality (Section 3.1), the interactions between content creators and users (Section 3.2), as well as the content usage statistics (Section 3.3). All these feature types are used as an input to a classifier that can be tuned for the quality definition for the particular media type (Section 3.4). In the next section, we will expand and refine the feature set specifically to match our main application domain of community question answering portals.

3.1. Intrinsic Content Quality

In this paper we focus on social media content items that are primarily textual in nature (e.g., blogs, answers). Hence, our intrinsic quality metrics (i.e., the quality of the *content* of each item) focus on the text. For user-generated content of other types (e.g., photos or bookmarks), intrinsic quality may be modeled differently.

As a baseline, we use textual features only—with all word n-grams up to length 5 that appear in the collection more than 3 times used as features. This straightforward approach is the de-facto standard for text classification tasks, both for classifying the topic and for other facets (e.g., sentiment classification [25]).

Additionally, we use a large number of *semantic features*, organized as follows:

Visual Quality Poor quality text, and particularly of the type found in online sources, is often marked with low conformance to common writing practices. For example, capitalization rules may be ignored; excessive punctuation—particularly repeated ellipsis and question marks—may be used, or spacing may be irregular. Several of our features capture the visual quality of the text, attempting to model these irregularities; among these are features measuring punctuation, capitalization, and spacing density (percent of all characters), as well as features measuring the character-level entropy of the text. A particular form of low visual quality is misspellings and typos; additional features in our set quantify the number of out-of-vocabulary words and spelling mistakes.

Syntactic and Semantic Complexity Advancing from the visual level to more involved layers of the text, other features in this subset quantify the syntactic and semantic complexity of it. These include simple proxies for complexity such as the average number of syllables per word or the entropy of word lengths, as well as more intricate ones such as the readability measures mentioned in Section 2.2.

Grammaticality Finally, to measure the grammatical quality of the text, we use several linguistically-oriented features. We annotate the content with part-of-speech (POS) tags, and use the tag n-grams (again, up to length 5) as features. This allows us to capture, to some degree, the level of "correctness" of the grammar used. Some POS n-grams are typical of correctly-formed questions: e.g., the sequence WRB TO VB ("how to identify...") is typical of lower-quality questions, whereas the sequence WRB VBP PRP VB ("How do I remove...") is more typical of correctly-formed content.

Additional features used to represent grammatical properties of the text are its formality score [14], and the distance between its (trigram) language model and several given language models, such as the Wikipedia language model or the language model of the Yahoo! Answers corpus itself (the distance is measured with KL-divergence).

3.2. Contributor and User Relationships

A significant amount of quality information can be inferred from the relationships between users and items. For example, we could apply link-analysis algorithms for propagating quality scores in the entities of the question/answer system, e.g., we use the intuition that, say, "good" answerers write "good" answers, or vote for other "good" answerers, etc. The main challenge we have to face is that our dataset, viewed as a graph, often contains nodes of multiple types (e.g., questions, answers, users), and edges represent a set of interaction among the nodes having different semantics (e.g., "answers", "gives best answer", "votes for", "gives a star to").

These relationships are represented as edges in a graph, with content items and users as nodes. The edges are typed, i.e., labeled with the particular type of interaction (e.g., "User u answers a question q"). Besides the user-item relationship graph, we also consider the user-user graph. This is the graph $G = (V, E)$ in which the set of vertices V is composed of the set of users, and the set E represents implicit relationships between users. For example, a user-user relationship could be "User u has answered a question from user v."

The resulting user-user graph is extremely rich and heterogeneous, and is unlike traditional graphs studied in the web link analysis setting. However, we believe that (in our classification framework) traditional link analysis algorithm may provide useful *evidence* for quality classification, tuned for the particular domain. Hence, for each type of link we performed a separate computation of each link-analysis algorithm. We computed the hubs and authorities scores (as in HITS algorithm [20]), and the and PageRank scores [24]. In Section 4 we discuss the specific relationships and node types developed for community question answering.

3.3. Content Usage Statistics

Users of the content (who may or may not also be contributors) provide valuable information about the items they find interesting. In particular, usage statistics such as the number of clicks on the item and dwell time have been shown useful in the context of identifying high quality web search results, and are complementary to link-analysis based methods. Intuitively, usage statistics measures are useful for social media content, but require different interpretation from the previously studied settings.

In particular, we exploit the number of item views (clicks), as well as derived statistics such as the expected number of views for a given category, the deviation from the expected number of views, and other second-order statistics designed to normalize the values for each item type. For example, all items within a popular category such as celebrity images or popular culture topics may receive orders of magnitude more clicks than, say, Science topics. Nevertheless, when normalized by the item *category*, the deviation from expected number of clicks can be used to infer quality directly, or can be incorporated into the classification framework. The specific normalization methods are directly adapted from references [1] and [2].

3.4. Quality Classification Framework

We cast the problem of quality ranking as a binary classification problem, in which a system must learn automatically to separate high-quality content from the rest.

We experimented with several classification algorithms, including those reported to achieve good performance with text classification tasks, such as support vector machines and log-linear classifiers; the best performance among the techniques we tested was obtained with stochastic gradient boosted trees [11]. In this classification framework, a sequence of (typically simple) decision trees is constructed so that each tree minimizes the error on the residuals of the preceding sequence of trees; a stochastic element is added by randomly sampling the data repeatedly before each tree construction, to prevent overfitting. A particularly useful aspect of boosted trees for our settings is their ability to utilize combinations of sparse and dense features.

Given a set of human-labeled quality judgments, the classifier is trained on all available features, combining evidence from semantic, user relationship, and content usage sources. The judgments are tuned for the particular goal. For example, we could use this framework to classify questions by genre or asker expertise. In the case of community question answers, described next, our goal is to discover interesting, well formulated and factually accurate content.

4. Modeling Content Quality in Community Question Answering

Our goal is to automatically assess the quality of questions and answers provided by users of the system. We believe that this particular sub-problem of quality evaluation is an essential module for performing more advanced information retrieval tasks on the question answering or web search system. For example, a quality score can be used as a feature for ranking search results in this system.

Note that Yahoo! Answers is *question-centric*: the interactions of users are organized around questions, as shown in Figure 1. This screenshot shows the main user interface elements that allow the forms of interactions between contributors (asking a question, answering a question, selecting best answer, voting on an answer). These relationships are explicitly modeled in the relational features described next.

4.1. Application-Specific User Relationships

The relationships between questions, users asking and answering questions, and answers can be captured by a tripartite graph outlined in Figure 2, where an edge represents an explicit relationship between the different node types. Since a user is not allowed to answer his/her own questions, there are no triangles in the graph, so in fact all cycles in the graph have length at least 6.

Note that our dataset, viewed as a graph, contains multiple types of nodes and multiple types of interactions, as illustrated in Figure 3. This leads us to introduce multi-relational features to describe multiple classes of objects and multiple types of relationships between

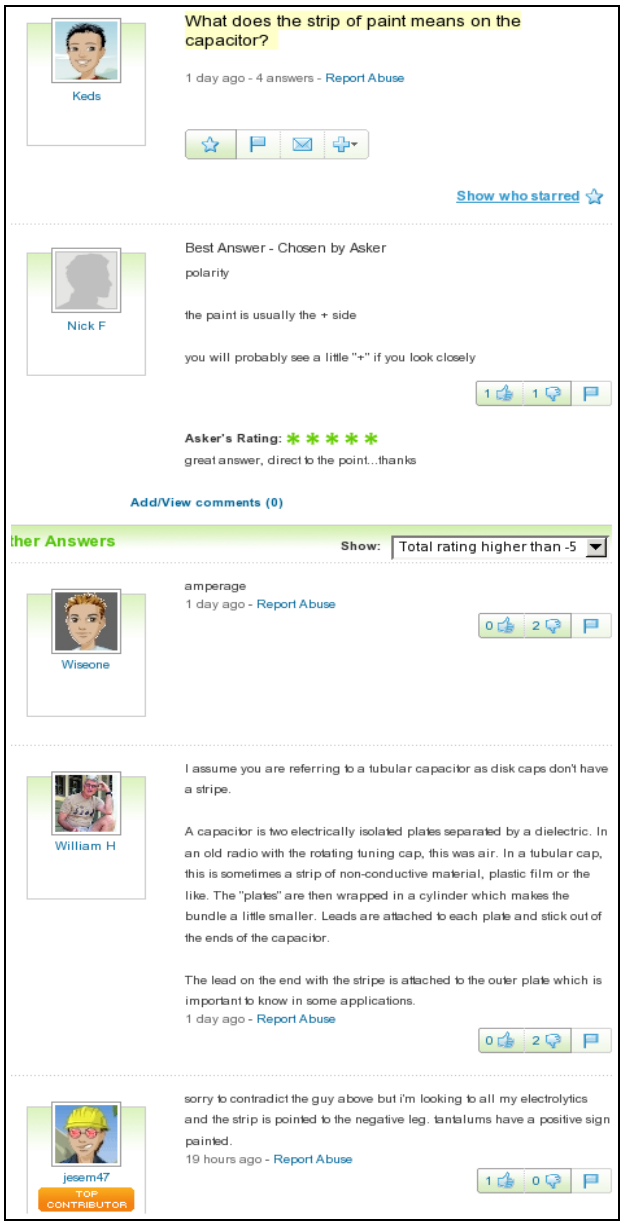

Figure 1: Example Yahoo! Answers question thread demonstrating some of the common interactions between contributors and other users.

these objects. In this section, we expand on the general user relationships ideas of the previous section to develop specificialized relational features that exploit the unique characteristics of the community question answering domain.

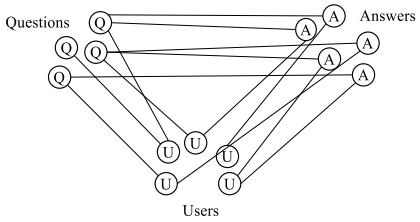

Figure 2: Depiction of the users-questions-answers interactions as a tripartite graph.

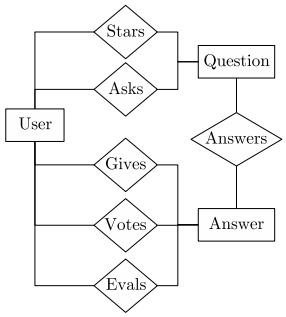

Figure 3: Partial entity-relationship diagram of answers.

Answer features In Figure 4, we depict the user relationship data that is available for a particular answer. The types of the data related to a particular answer form a *tree*, in which the type "Answer" is the root. So, an answer $a \in A$ is at the zero-th level of the tree, the question q that a answers to, and the user u who posted a are in the first level of the tree, and so on.

To streamline the process of exploring new features, we suggest naming the features with respect to their position in this tree. Each feature corresponds to a data type, which resides in a specific node in the tree, and thus, it is characterized by the path from the root of the tree to that node.

Hence, each specific feature can be represented by a path in the tree. For instance, a feature of the type "`QU`" represents the information about a question (`Q`) and the user (`U`) who asked that question. In Figure 4, we can see two subtrees starting from the answer being evaluated: one related to the question being answered, and the other related to the user contributing the answer.

The types of features on the question subtree are:

- `Q` Features from the question being answered
- `QU` Features from the asker of the question being answered
- `QA` Features from the other answers to the same question

The types of features on the user subtree are:

- `UA` Features from the answers of the user

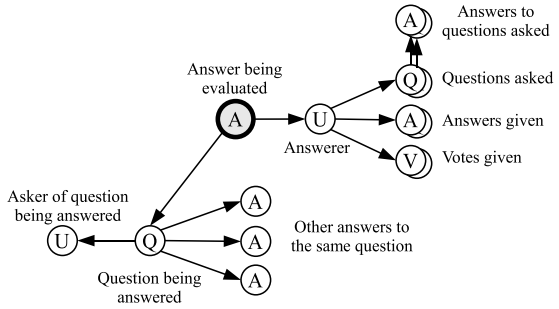

Figure 4: Types of features available for inferring the quality of an answer.

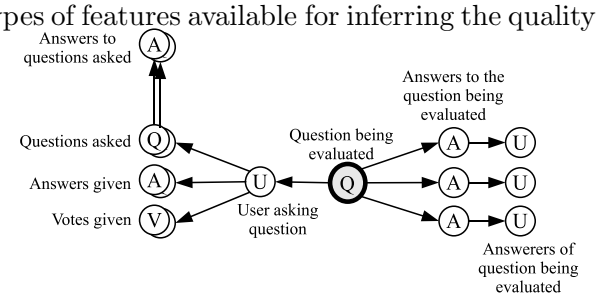

Figure 5: Types of features available for inferring the quality of a question.

`UQ` Features from the questions of the user
`UV` Features from the votes of the user
`UQA` Features from answers received to the user's questions
`U` Other user-based features

This string notation allows us to group several features into one bundle by using the wildcard characters "`?`" (one letter), and "`*`" (multiple letters). For instance, `U*` represents all the features on the user subtree, and `Q*` all the features in the question subtree.

Question features We represent user relationships around a question similarly to representing relationships around an answer. These relationships are depicted in Figure 5. Again, there are two subtrees: one related to the asker of the question, and the other related to the answers received.

The types of features on the answers subtree are:

`A` Features directly from the answers received
`AU` Features from the answerers of the question being answered

The types of features on the user subtree are the same as the ones above for evaluating answers.

Implicit user-user relations As stated in Section 3.2, besides the user-question-answer graph, we also consider the user-user graph. This is the graph $G = (V, E)$ in which the set of vertices V is composed of the set of users and the set $E = E_a \cup E_b \cup E_v \cup E_s \cup E_+ \cup E_-$ represents the relationships between users as follows:

- E_a represents the answers: $(u, v) \in E_a$ iff user u has answered at least one question asked by user v.
- E_b represents the best answers: $(u, v) \in E_b$ iff user u has provided at least one best answer to a question asked by user v.
- E_v represents the votes for best answer: $(u, v) \in E_v$ iff user u has voted for best answer at least one answer given by user v.
- E_s represents the stars given to questions: $(u, v) \in E_v$ iff user u has given a star to at least one question asked by user v.
- E_+/E_- represents the thumbs up/down: $(u, v) \in E_+/E_-$ iff user u has given a "thumbs up/down" to an answer by user v.

For each graph $G_x = (V, E_x)$, we denote by $hub(x)$ the vector of hub scores on the vertices V, by $auth(x)$ the vector of authority scores, and by $pr(x)$ the vector of PageRank scores. We also denote by $pr'(x)$ the vector of PageRank scores in the transposed graph.

To classify these features in our framework, we consider that PageRank and authority scores are related mostly to in-links, while the hub score deals mostly with out-links. For instance, let's take $hub(b)$. It is the hub score in the "best answer" graph, in which an out-link from u to v means that u gave a best answer to user v. Then, $hub(b)$ represents the answers of users, and is assigned to the answerer record (`UA`).

The assignment of these features is done in the following way:

`UQ` To the asker record of a user: $auth(a)$, $auth(b)$, $auth(s)$, $pr(a)$, $pr(b)$

`UA` To the answerer record of a user: $hub(a)$, $hub(b)$, $pr'(a)$, $pr'(b)$, $auth(v)$, $pr(v)$, $auth(+)$, $pr(+)$, $auth(-)$, $pr(-)$

`UV` To the voter record of a user: $hub(v)$, $pr'(v)$, $hub(s)$, $pr'(v)$, $hub(+)$, $pr'(+)$, $hub(-)$, $pr'(-)$

4.2. Content Features for QA

As the base content quality features for both questions and answer text individually we use directly the semantic features from Section 3.1. We rely on feature selection methods and the classifier to identify the most salient features for the specific tasks of question or answer quality classification.

Additionally, we devise a set of features specific to the QA domain that model the relationship between a question and an answer. Intuitively, a copy of a Wall Street Journal article about economy may have good quality, but would not (usually) be a good answer to a question about celebrity fashion. Hence, we explicitly model the relationship between the question and the answer. To represent this we include the KL-divergence between the language models of the two texts, their non-stopword overlap, the ratio between their

lengths, and other similar features. Interestingly, the text of answers often relates to other answers for the same question. In the example of Figure 1, the third answer actually disagrees with the answer chosen as best by the asker. While this information is difficult to capture explicitly, we believe that our semantic feature space is rich enough to allow a classifier to effectively detect quality questions (and answers).

4.3. Usage Features for QA

Recall that community QA is *question-centric*: a question thread is usually viewed as a whole, and the content usage statistics are available primarily for the complete question thread. As a base set of content usage features we use the raw page view numbers, as well as temporal statistics (i.e., how long ago the question was posted).

In addition, we exploit the rich set of metadata available for each question. In particular, the topical and genre categorization for each question, selected by the asker, is an extremely valuable resource, since clickthrough counts on a question are heavily influenced by the topical and genre category: celebrity and dating topics receive orders of magnitude more clicks than, say, Science questions. We capture this information both by directly providing category information as features to the classifier, and by computing *deviation* information (i.e., how different is usage statistics for the question thread) compared to the expected number of page views, conditioned on the question category.

In summary, while many of the item content, user relationship, and usage statistics features are designed and are applicable for many types of social media, we augment the general feature set with additional information specific to the community question answering domain. As we will show in the empirical evaluation presented in the next sections, both the generally applicable, and the domain specific features turn out to be significant for quality identification.

5. Experimental Setting

This section describes the experimental setting, datasets, and metrics used for producing our results in Section 6.

5.1. Dataset

Our dataset consists of 6,665 questions and 8,366 question-answer pairs. The base usage features (page views or clicks) were obtained from the total number of times a question thread was clicked (e.g., in response to a search result). All of the above questions were labeled for quality by human editors, who were independent from the team that conducted this research. Editors graded questions and answers for well-formedness, readability, utility, and interestingness; for answers, an additional correctness element was taken into account. Additionally, a high-level type (informational, advice, poll, etc.) was assigned to each

question. In a subset of 300 questions from this dataset, the inter-annotator agreement was $\kappa = 0.68$. The assessors were also asked to look at the type of questions. They found that roughly 1/4 of the questions were seeking for an opinion (instead of information or advice).

Following links to obtain user relationship features Starting from the questions and answers included in the evaluation dataset we considered related questions and answers as follows. Let Q_0 and A_0 be the sets of questions and answers, respectively, included in the evaluation dataset.

Now let U_1 be the set of users who have made a question in Q_0 or given an answer in A_0. Additionally we select Q_1 to be the set of all questions asked by all users in U_1. Similarly we select A_1 to be the set of answers given by users in U_1 and A_2 to be the set of all the answers to questions in Q1. Obviously $Q_0 \subseteq Q_1$ and $A_0 \subseteq A_1$. Our dataset is then defined by the nodes $(Q_1, A_1 \cup A_2, U_1)$ and the edges induced from the whole dataset.

Figure 6 depicts the process of finding related items. The relative size of the portion we used (depicted with thick lines) is exaggerated for illustration purposes: actually the data we use is a tiny fraction of the whole collection.

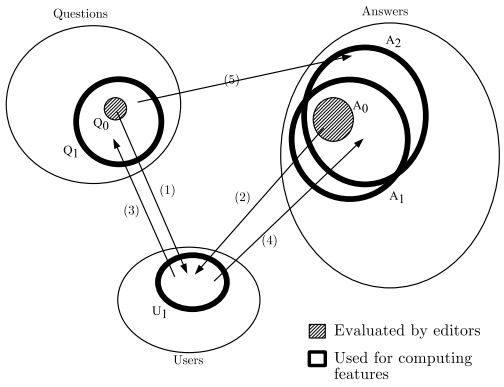

Figure 6: Sketch showing how do we find related questions and answers, depicted with thick lines in the figure. All the questions Q_0 and answers A_0 evaluated by the editors are included at the beginning, and then (1) all the askers U_0 of the questions in Q_0, (2) all the answerers U_0 of the answers in A_0, (3) all the questions Q_1 by users in U_0, (4) all the answers A_1 by users in U_0, and (5) all the answers A_2 to questions in Q_1.

This process of following links to include a subset of the data only applies to questions and answers. In contrast, for the user rating features, we included all of the votes received and given by the users in U_1 (including votes for best answers, "stars" for good questions, "thumbs up" and "thumbs down"), and all of the abuse reports written and received.

5.2. Dataset Statistics

The degree distributions of the user interaction graphs described earlier are very skewed. The (complementary) cumulative distribution of the number of answers, best answers, and votes given and received is shown in Figure 7. The distribution of the number of votes given and received by the users can be modeled accurately by Pareto distributions with exponents 1.7 and 1.9 respectively.

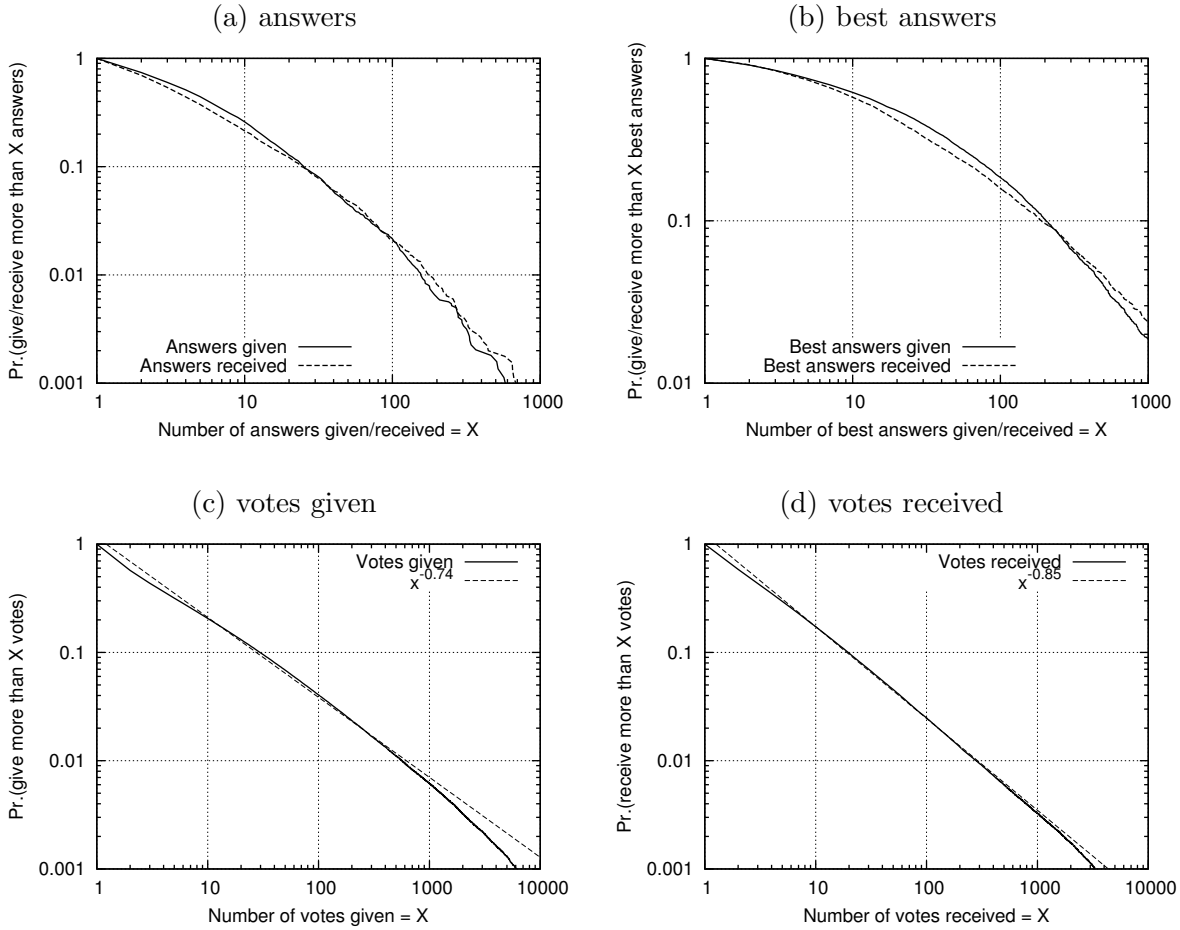

Figure 7: Distribution of degrees in the graph representing relationships between users: (a) number of answers given and received; (b) number of best answers given and received; (c) number of votes given; and (d) number of votes received. The "votes" including votes for best answer, start, "thumbs up" and "thumbs down".

In each of the graphs $G_x = (V, E_x)$, with $x \in \{a, b, v, s, +, -\}$, we computed the hubs and authorities scores (as in HITS algorithm [20]), and the and PageRank scores [24]. Note

that in all cases we execute HITS and PageRank on a subgraph of the graph induced by the whole dataset, so the results might be different than the results that one would obtain if executing those algorithms on the whole graph.

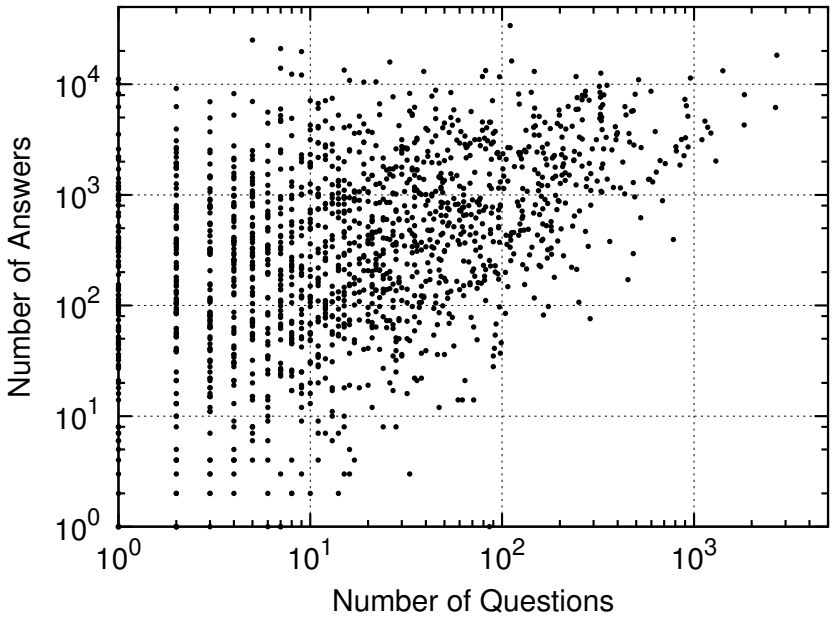

Figure 8: Number of questions and number of answers for each user in our data.

The distributions of answers given and received are very similar to each other, in contrast to [10] where there were clearly "askers" and "answerers" with different types of behaviors. Indeed, in our sample of users, most users participate as both "askers" and "answerers". From the scatter-plot in Figure 8, we observe that there are no clear roles of "asker" and "answerer" such as the ones identified by Fisher et al. [10] in USENET newsgroups. The fact that only users with many questions also have many answers is a by-product of the incentive mechanism of the system (points), where a certain number of points is required to ask a question, and points are gained mostly by answering questions.

In our evaluation dataset there is a positive correlation between question quality and answer quality. In Table 1 we can see that good answers are much more likely to be written in response to good questions, and bad questions are the ones that attract more bad answers. This observation is an important consideration for feature design.

5.3. Evaluation Metrics and Methodology

Recall that we want to automatically separate high-quality content from the rest. Since the class distribution is not balanced, we report the precision and recall for the two classes,

Table 1: Relationship Between Question Quality and Answer Quality

	Question Quality		
Answer Quality	A. High	B. Medium	C. Low
A. High	41%	15%	8%
B. Medium	53%	76%	74%
C. Low	6%	9%	18%
Total	100%	100%	100%

"high quality" and "normal or low quality" separately: both are measured when the classifier threshold is set to maximize the F1 measure. We also report the area under the ROC curve for the classifiers, as a non-parametric single estimator of their accuracy.

6. Experimental Results

In this Section we show the results for answer and question content quality. Recall that as a baseline we use only textual features for the current item (answer/question) at the level $\emptyset$ of the trees introduced in Section 4.1. In the experiments reported here, 80% of our data was used as a training set and the rest for testing.

6.1. Question Quality

Table 2 shows the classification performance of the question classifier, using different subsets of our feature set. *Text* refers to the baseline, bag-of-n-gram features; *Intrinsic* is the features derived from the text, described in Section 3.1; *Usage* refers to click-based knowledge described in Section 3.3; and *Relation* features are those involving the community behavior, described in Section 3.2.

Table 2: Precision P, Recall R, and Area Under the ROC Curve for the Task of Finding High-Quality Questions

	High qual.		Normal/low qual.		
Method	P	R	P	R	AUC
Text (Baseline)	0.654	0.481	0.762	0.867	0.523
Usage	0.594	0.470	0.755	0.836	0.508
Relation	0.694	0.603	0.806	0.861	0.614
Intrinsic	0.746	0.650	0.829	0.885	0.645
T+Usage	0.683	0.571	0.798	0.865	0.575
T+Relation	0.739	0.647	0.828	0.881	0.659
T+Intrinsic	0.757	0.650	0.830	0.891	0.648
T+Intr.+Usage	0.717	0.690	0.845	0.861	0.686
T+Relation+Usage	0.722	0.690	0.845	0.865	0.679
T+Intr.+Relation	**0.798**	0.752	0.874	**0.901**	0.749
All	0.794	**0.771**	**0.885**	0.898	**0.761**

Clearly, a standard text classification approach—used in our baseline, the first line in Table 2—does not address the task of identifying high quality content adequately; but relying exclusively on usage patterns, relations, or intrinsic quality features derived from the text (next 3 lines in the table) results in suboptimal solutions too.

In-line with intuition, we witness a consistent, gradual increase in performance as additional information is made available to the classifier, indicating that the different feature sets we use provide, to some extent, independent information.

The 10 most significant features for question quality classification, according to a chi-squared test, included features from all subsets, as follows:

- `UQ` Average number of "stars" to questions by the same asker
- ∅ The punctuation density in the question's subject
- ∅ The question's category (assigned by the asker)
- ∅ The clicks on the question, normalized by the average clicks on its category
- `UA` Average number of "Thumbs up" received by answers given to other questions asked by the asker
- ∅ Number of words per sentence
- `UA` Average number of answers with references (URLs) given to questions asked by the asker
- `UA` Fraction of questions asked by the asker in which he opens the question's answers to voting (instead of picking the best answer by hand)
- `UQ` Average length of the questions by the asker
- `UV` The number of "best answers" authored by the user
- `U` The number of days the user was active in the system

We performed a comprehensive exploration of our feature spaces, in particular focusing on user relational features and the content usage features. Due to space constraints, we discuss here only the effectiveness of different content usage, or implicit feedback, features. These features are derived from page views statistics as described in Section 3.3. A variant of the C4.5 decision tree classifier was used to predict quality based on click features alone. Table 3 breaks down the classification performance by feature type.

Table 3: Overall Precision, Recall, and F1 for the task of finding high-quality questions using only usage features

Features	Precision	Recall	F1
Page Views	0.540	0.250	0.345
+ Question category	0.600	0.410	0.510
+ Deviation from expected	0.630	0.460	0.530
All Usage features	0.594	0.470	0.530
Top 10 Usage features	**0.630**	**0.540**	**0.580**

These results support our hypothesis that topical category information is crucial for interpreting usage statistics. As we can see, normalizing the raw page view counts by question category significantly improves the accuracy, as well as modeling the deviation from the

expected page view count, which provides additional improvement. Finally, including top 10 content usage features selected according to chi-squared statistic provide some additional improvement. Interestingly, including all derived features similar to those described in [1] actually degrades performance, indicating overfitting when relying on usage statistics alone without the benefit of other forms of user feedback.

Because of the effectiveness of the relational and usage features to independently identify high-quality content, we hypothesized that a variant of co-training or co-boosting [22] would be effective to expand the training set in a partially supervised setting. However, our experiments did not result in an classification improved accuracy, and this remains an open question for future work.

6.2. Answer Quality

Table 4 shows the classification performance of the answer classifier, again examining different subsets of our feature set. In this case, we did not use the Usage subset, as there are no separate clicks on answers within Yahoo! Answers (an answer is displayed on the question page, alongside other answers to the question).

Table 4: Precision P, Recall R, and Area Under the ROC Curve for High-Quality Answer Finding

	High qual.		Normal/low qual.		
Method	P	R	P	R	AUC
Text (Baseline)	0.668	0.862	0.968	0.906	0.805
Relation	0.552	0.617	0.914	0.890	0.623
Intrinsic	0.712	0.918	0.981	0.918	0.869
T+Relation	0.688	0.851	0.965	0.915	0.821
T+Intrinsic	0.711	**0.926**	**0.982**	0.917	**0.878**
All	**0.730**	0.911	0.979	**0.926**	0.873

Once again, we observe an increase in performance attributed to both additional feature sets used; however, in this case improvement is milder. An examination of the data shows that one particular feature—the answer length—is dominating over other features, resulting in relatively high performance of the baseline.

The 10 most significant features for answer quality, according to a chi-squared test, were:

- ∅ Answer length
- ∅ The number of words in the answer with a corpus frequency larger than c
- UA The number of "thumbs up" minus "thumbs down" received by the answerer, divided by the total number of "thumbs" s/he has received
- ∅ The entropy of the trigram character-level model of the answer
- UA The fraction of answers of the answerer that have been picked as best answers (either by the askers of such questions, or by a community voting)
- ∅ The unique number of words in the answer

`U` Average number of abuse reports received by the answerer over all his/her questions and answers

`UA` Average number of abuse reports received by the answerer over his/her answers

$\emptyset$ The non-stopword word overlap between the question and the answer

$\emptyset$ The Kincaid[19] score of the answer

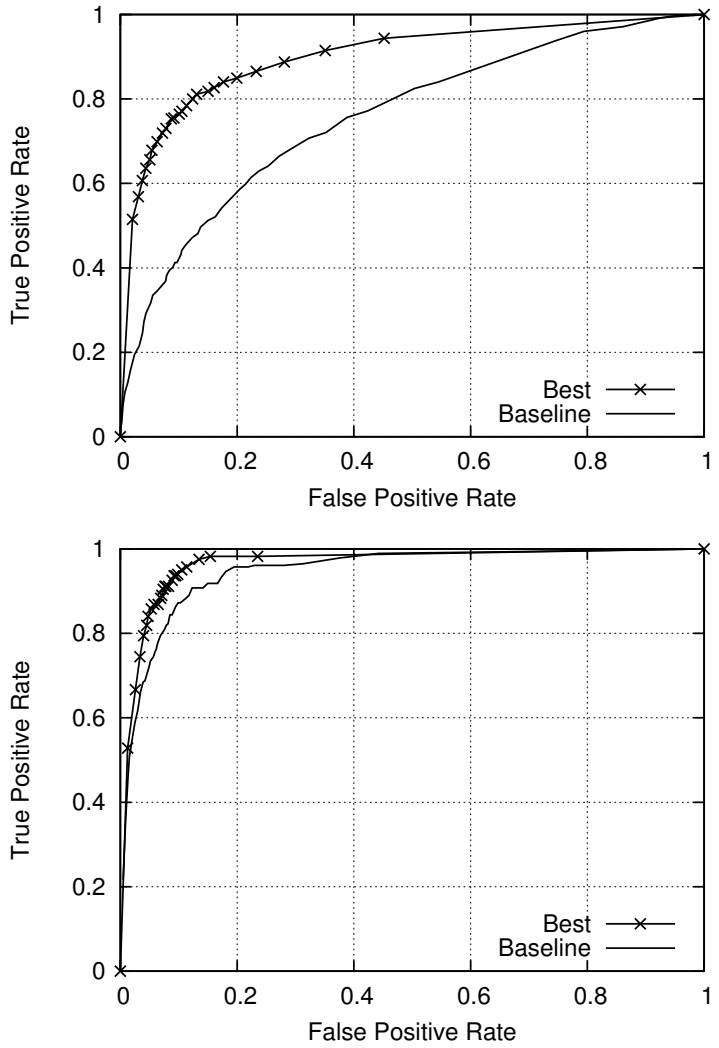

Figure 9: ROC curve for the best-performing question quality classifier (top) and answer quality classifier (bottom).

ROC curves for the baseline question and answer classifiers from Tables 2 and 4, as well as for the classifiers with the maximal area under the curve appearing in these tables, are shown in Figure 9.

7. Conclusions

We presented a general classification framework for quality estimation in social media. As part of our work we developed a comprehensive graph-based model of contributor relationships and combined it with content- and usage-based features. We have successfully applied our framework to identifying high quality items in a web-scale community question answering portal, resulting in a near-human level of accuracy on the question and answer quality classification task. Community QA is a popular information seeking paradigm that has already entered mainstream, and our results provide significant understanding of this new domain.

We investigated the contributions of the different sources of quality evidence, and have shown that some of the sources are complementary – i.e., capture the same high-quality content using the different perspectives. As one implication, the resulting classifier is likely to be more robust to spam, since the classifier automatically assigns appropriate weights to different features, including those from user relationships or usage statistics. In the future, we plan to more specifically explore the relationships and usage features to automatically identify malicious users.

We demonstrated the utility of our approach on a large-scale community QA site. However, we believe that our results and insights are applicable to other social media settings, and to other emerging domains centered around user contributed-content.

Acknowledgments: the authors wish to thank Byron Dom, Benoit Dumoulin, and Ravi Kumar for many useful discussions.

References

[1] E. Agichtein, E. Brill, S. T. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, 2006.

[2] K. Ali and M. Scarr. Robust methodologies for modeling web click distributions. In *WWW*, pages 511–520, 2007.

[3] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, July 2006.

[4] Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3), February 2006.

[5] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M. D. Harris. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics*, pages 206–210, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[6] J. Burstein and M. Wolska. Toward evaluation of writing style: finding overly repetitive word use in student essays. In *EACL '03: Proceedings of the tenth conference*

on European chapter of the Association for Computational Linguistics, pages 35–42, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[7] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of CIKM*, pages 528–531, New Orleans, LA, USA, 2003.

[8] M. Chodorow and C. Leacock. An unsupervised method for detecting grammatical errors. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 140–147, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[9] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of Workshop on Data Mining and Knowledge Discovery*, pages 42–48, San Diego, CA, USA, 2003. ACM Press.

[10] D. Fisher, M. Smith, and H. T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. volume 3, pages 59b–59b, 2006.

[11] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, 2002.

[12] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 403–412, New York, NY, USA, 2004. ACM Press.

[13] R. Gunning. *The technique of clear writing*. McGraw-Hill, 1952.

[14] F. Heylighen and J.-M. Dewaele. Variation in the contextuality of language: An empirical measure. *Context in Context. Special issue Foundations of Science*, 7(3):293–340, 2002.

[15] J. Jeon, B. W. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235, New York, NY, USA, 2006. ACM Press.

[16] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, 2005.

[17] P. Jurczyk and E. Agichtein. Discovering authorities in question-answer communities using link analysis. In *ACM Conference on Information and Knowledge Management (CIKM)*, 2007.

[18] P. Jurczyk and E. Agichtein. Hits on question answer portals: an exploration of link analysis for author ranking. In *SIGIR (posters)*. ACM, 2007.

[19] J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical Report Research Branch Report 8-75, Millington, Tenn, Naval Air Station, 1975.

[20] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[21] G. H. McLaughlin. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[22] K. Nigam and R. Ghani. Analyzing the applicability and effectiveness of co-training. pages 86–93. ACM Press, 2000.

[23] E. B. Page. Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 1994.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[25] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques, May 2002.

[26] L. Prescott. Yahoo! Answers captures 96% of Q and A market share, 2006.

[27] L. M. Rudner and T. Liang. Automated essay scoring using bayes. *Journal of Technology, Learning, and Assessment*, 1(2), June 2002.

[28] C. Sang-Hun. To outdo Google, Naver taps into Korea's collective wisdom. *International Herald Tribune*, July 4 2007.

[29] J. P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

[30] Q. Su, D. Pavlov, J.-H. Chow, and W. C. Baker. Internet-scale collection of human-reviewed data. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 231–240, New York, NY, USA, 2007. ACM Press.

[31] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM Press.

[32] C.-N. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, December 2005.