Automated Judgment of Document Qualities

Kwong Bor Ng
Queens College, CUNY, Kissena Boulevard, Flushing, NY 11367. E-mail: kbng@qc.edu

Paul Kantor
Rutgers University, New Brunswick, NJ 08903. E-mail: kantor@scils.rutgers.edu

Tomek Strzalkowski
SUNY Albany, Western Avenue, Albany, NY 12222. E-mail: tomek@csc.albany.edu

Nina Wacholder
Rutgers University, New Brunswick, NJ 08903. E-mail: nina@scils.rutgers.edu

Rong Tang
Catholic University of America, Washington, DC 20064. E-mail: tangr@cua.edu

Bing Bai, Robert Rittman, Peng Song, and Ying Sun
Rutgers University, New Brunswick, NJ 08903. E-mail: bbai@eden.rutgers.edu, rritt@scils.rutgers.edu, psong@eden.rutgers.edu, ysun@scils.rutgers.edu

The authors report on a series of experiments to automate the assessment of document qualities such as depth and objectivity. The primary purpose is to develop a quality-sensitive functionality, orthogonal to relevance, to select documents for an interactive question-answering system. The study consisted of two stages. In the classifier construction stage, nine document qualities deemed important by information professionals were identified and classifiers were developed to predict their values. In the confirmative evaluation stage, the performance of the developed methods was checked using a different document collection. The quality prediction methods worked well in the second stage. The results strongly suggest that the best way to predict document qualities automatically is to construct classifiers on a person-by-person basis.

Overview

We report the results of experiments to automate the assessment of various document qualities such as depth and objectivity. The primary purpose of the study is to develop a quality-sensitive functionality, orthogonal to relevance, to identify documents for use in an automatic question answering system (i.e., HITIQA, see Small & Strzalkowski, 2004). We note that in this article: (a) a "quality" is a property of a document, not its goodness, and (b) variables representing this kind of document properties are not necessary to be categorical (i.e., present or absent). The study consisted of two stages: the classifier construction stage and the confirmative evaluation stage. In the first stage, through focus group studies, quality judgment experiments, textual feature extraction and analysis, we generated nine document qualities and developed classifiers to predict human judgments of these qualities. In the second stage, we tested the methods developed using a different set of documents.

Some of the early findings of the first stage of the study were reported in Ng et al., 2003 and Tang, Ng, Strzalkowski, and Kantor, 2003. In this article, we report all our findings from both stages. First, we summarize the results reported in the previous two articles, fill in some missing technical details, and report on the late findings of the first stage of the study. We then report the results of the second stage (the confirmative evaluation stage) of the study.

In our study, we ask the following questions: (a) What document qualities are deemed important by information professionals? (b) Is it possible to develop a method, based on machine computable features of a document, to predict a document's qualities? (c) Which machine computable features have strong predictive power? (d) Which machine

Received September 27, 2004; revised April 21, 2005; accepted June 16, 2005

 • Published online 2 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20393

computable features form a minimal set for prediction with a detection rate much better than the false alarm rate?

We find that:

1. Different individuals may have different conceptions of the relative importance of different document qualities. Empirically, we can use focus group studies to solicit and identify the qualities that are important to a particular information professional community.
2. If we treat document qualities as having an *objective existence* and try to predict them as such, we can build workable classifiers to separate documents of high-quality rating from low-quality rating, based on textual features and part of speech tags, as discussed below. We had various success rates for different document qualities.
3. If we personalize the model so that predictions are tailored to individuals without attempting to maintain consistency in judgments, the predictive power will increase about 25–30%, and the precision of prediction can be as high as 95% in the top 20 documents.
4. Different document qualities have different strong predictors. In general, predictors constructed based on human understanding of the language and content have higher predictive power.
5. Using stepwise methods, we can eliminate redundant predictive variables. Usually, no more than 10 predictive variables will result in an receiver operating characteristic (ROC) curve that is as strong as that which results from using all the predictive variables.

The results of the study changed our understanding of the nature of document quality judgments. It led us to explore personalization of classifiers. We conclude that document quality judgments should be treated as a product of the interaction between judge and document; to automate these assessments, one must take into consideration the personal dimension.

Relevance Versus Qualities

In information retrieval studies, the primary concern is the match between the query and topical "about-ness" of documents in a collection. Other document properties, such as level of difficulty (i.e., intellectual access), reliability of information sources, authoritativeness of content, etc., are not considered because those qualities usually have no direct relationship with topical relevance (with exceptions of course, e.g., see Naumann, 2002). We believe some document properties do contribute to the usefulness of a document. In some analytical settings, it is of interest to estimate these properties in addition to the primary estimate of their relevance to the topic. We are especially interested in two broad dimensions of document qualities: presentation style and content. The presentation style dimension relates to whether the information, descriptions, explanations, and judgments contained in a document are presented in a particular manner. Different genres may have different desired styles of presentation. For example, for a news article, the desired presentation style may emphasize clarity and comprehensiveness, and the presentation structure may be some kind of inverted triangle with the conclusion of the reported event summarized in the first paragraph. The content dimension relates to whether the information, description, explanation, and judgments contained in a document are reliable, complete, objective, and unbiased. Different genres may also differ in desirable content. For example, for an advertisement, diversity of opinions or multiple viewpoints may not be desirable. In our study, we investigate those quality aspects that make some documents more desirable to information professionals than others.

Method

Identification of Document Qualities

In the first stage of our study, we identified document qualities that were deemed important by a particular community of information professional (i.e., journalists), developed a corpus of judged documents (1000 news articles each judged twice in terms of the identified document qualities), and constructed classifiers to automatically predict those qualities (Figure 1).

To identify document qualities deemed important by information professionals, we held two focus group studies. Ideally, the focus groups would consist of "perfect representatives" of information professionals. We used people associated with journalism. In our study, we learned that there are many kinds of information users and many kinds of tasks, so the concept of perfect representation of information professionals seemed a bit naïve. We will return to this issue in our analysis.

In our first focus group, we interviewed seven news professionals, a majority of whom were from a local newspaper, the *Albany Times Union*. In our second focus group, we interviewed four professional news editors. Both sessions lasted 90 minutes. The first session was a free-form discussion, and the second session was a task-oriented meeting. The important document qualities mentioned by participants in the first session included *source reliability*, *objectivity*,

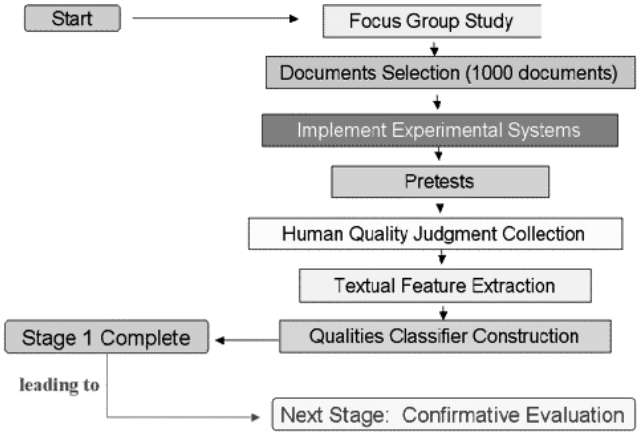


FIG. 1. Classifier construction stage.

DOI: 10.1002/asi

TABLE 1. Nine document qualities identified in the two focus group studies.

Document quality	Definition
Accuracy	The extent to which information is precise and free from known errors.
Source reliability	The extent to which you believe that the indicated sources in the text (e.g., interviewees, eye-witnesses, etc.) provide truthful account of the story.
Objectivity	The extent to which the document includes facts without distortion by personal or organizational biases.
Depth	The extent to which the coverage and analysis of information is detailed.
Author Credibility	The extent to which you believe that the author of the writing is trustworthy.
Readability	The extent to which information is presented with clarity and is easily understood.
Verbose → Conciseness	The extent to which information is well-structured and compactly represented.
Grammatical Correctness	The extent to which the text is free from syntactic problems.
One-sided → Multiviews	The extent to which information reported contains a variety of data sources and view points.

completeness/context, *wording/nuances*, *accuracy*, and *preciseness/veracity*.

Participants in the second session first reported the three quality aspects that they consider most important in their work. Each criterion was listed on the board, including *accuracy*, *author*, *variety*, *sources*, *objectivity*, *conciseness*, *depth/detail/analysis*, *easy to read*, and *grammar/syntax*. Participants were then asked to read sample documents, evaluate the documents based on the criteria listed on the board, and highlight segments of text as evidence in support of their assessments. In the end, participants were asked to rank the quality aspects in order of importance. There was much overlap between document qualities mentioned in the first and second focus groups. As a result, nine quality aspects were generated (Table 1).

Corpus Development for Stage 1

Concurrent with our focus group studies, we built a corpus of 1000 documents by selecting medium-sized news articles (100–2500 words) from the TREC collection (text retrieval conferences, http://trec.nist.gov/; Voorhees, 2001), including articles from the *LA Times*, *Wall Street Journal*, *Financial Times of London*, and the Associated Press. Documents were selected based on five topics, retrieved using the SMART system (Salton, 1971).

We recruited two kinds of participants to perform judgments: expert and student. Expert judgments were completed first. Documents judged by experts were used to train student judges to perform quality assessments at the same level as expert judges (Figure 2). The expert judges has

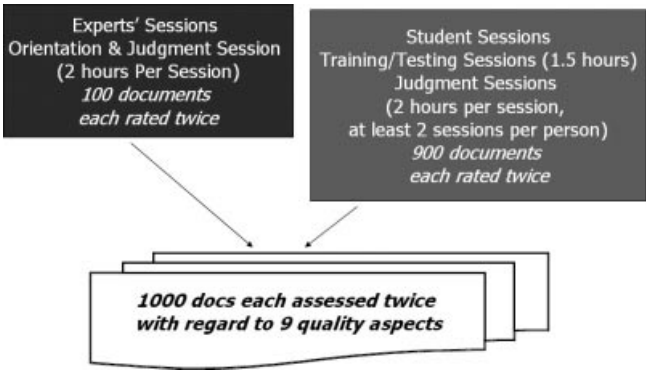


FIG. 2. Procedures for developing a corpus of judged documents.

demonstrated expertise in the five topics. They were experienced news professionals and researchers in the areas of journalism, communication, political science, and information science. Each expert judge started with a practice document, and then proceeded to evaluate 10 documents in 2 hours. In the end, 100 documents were judged according to the nine document qualities. Each document was rated twice by two different experts.

Documents were judged using a quality judgment system that we created. As shown in Figure 3, the qualities were represented in the system by a 10-point Likert scale. The right column displays the document. In the left column, a slider is used to assign scores to each of the nine qualities (minimum = 1 and maximum = 10). A score of 0 indicates there is no basis to make a judgment. We also asked judges to highlight and save portions of the text as evidence of their judgments (for each of the nine scores assigned).

Documents judged by experts were then used as training and examination materials for student judges. After an orientation session and practicing with five training documents, student judges then rated five examination documents. We compared the students' judgments with the experts' judgments, and selected students whose judgments were most similar to those of the experts, using the following method. We used the sum of squares of normalized differences between expert and student to test whether a student's judgments could have been drawn from the same distribution as the expert judgments.

$$\text{Sum of Normalized Scores} = \sum \frac{\left(\frac{x_1^e + x_2^e}{2} - x^s\right)^2}{(x_1^e - x_2^e)^2 + 1}$$

where x_1^e and x_2^e are the scores of quality variable x assigned by the two experts respectively and x^s is the corresponding score assigned by the student. For the determination of the threshold, we used Monte Carlo simulation to identify the 95% point in the observed distribution, based on the observed mean and variance. If a student's sum of squares of normalized differences exceeded the 95% point, we did not accept that student for further work. Qualified student judges were then invited back for formal judgment sessions.

DOI: 10.1002/asi

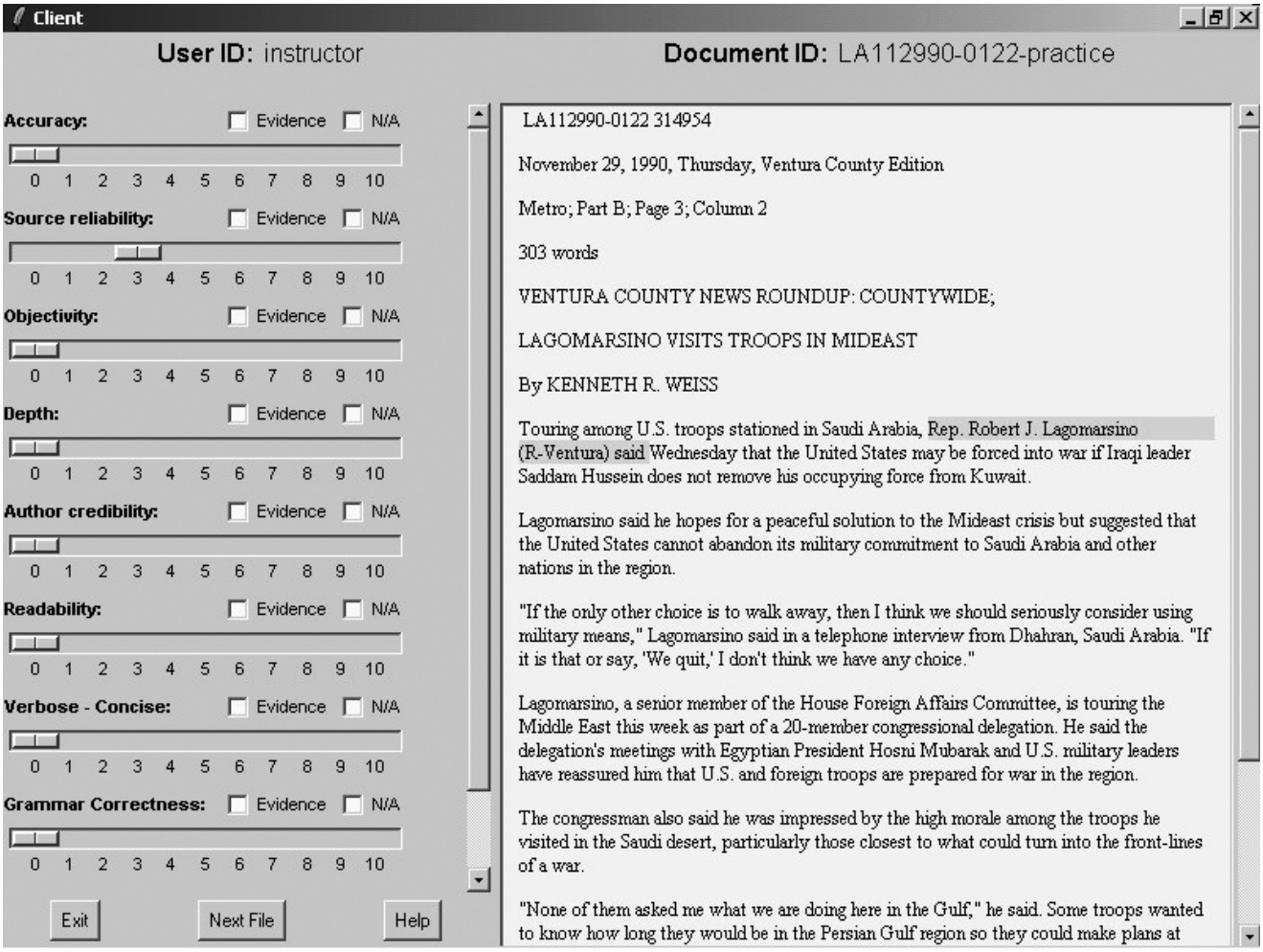


FIG. 3. Quality experiment interface.

Each session lasted 2 hours in which participants evaluated 10 documents. In the end, student participants evaluated 900 documents and each document was evaluated by two student judges.

By applying principal component analysis (Reyment & Joreskog, 1993) on the 1000 document vectors (each consisting of the nine quality variables with values equal to the average of two quality scores assigned by two judges), we identified two major components (Figure 4). The horizontal component corresponds to the content dimension. It includes credibility, reliability, accuracy, multiview, depth, and objectivity. The vertical component corresponds to the presentation style dimension. It consists of grammar, readability, and verbosity, and conciseness. Together they explain 58% of the variance of the nine quality variables.

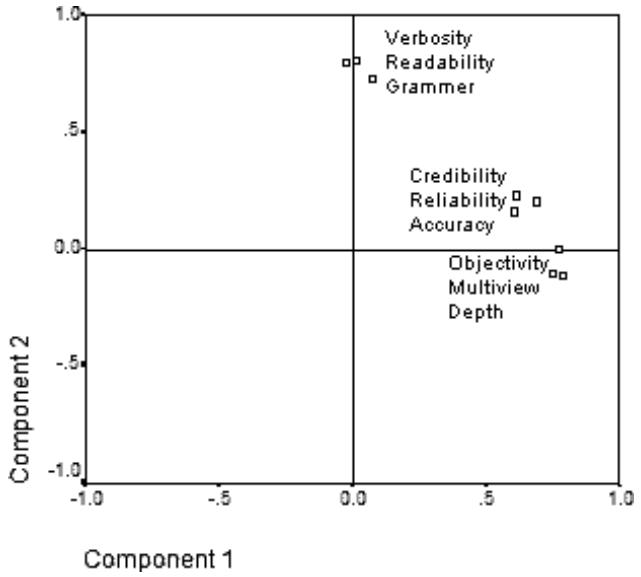
Linear Combination of Textual Features

We began by investigating whether a statistical relationship exists between the quality judgments and textual features of documents. Our objective was to identify textual features that are highly correlated with the quality variables for our collection of judged documents (It's important to note that at this point in our work, we were imagining that a quality aspect of a document is something that has a concrete reality, and that with suitable training, individuals could recognize these aspects with pretty high consistency. We now question this assumption. The results indicate to us that the best way to do this is to construct quality classifiers on a person-by-person basis.) We used existing grammatical part-of-speech "taggers" and information extraction tools developed for GATE (General Architecture for Text Engineering; Cunningham et al., 2000) and AWB (Alembic Workbench; see Day et al., 1997) (Table 2).

Based on our results, using textual features as predictive variables of document qualities seemed promising. For each of the nine quality variables, we were always able to find some highly correlated textual features, for example, see Table 3.

To pursue this direction, we explored the following model: $Q_j = \beta_0 + ? \beta_i x_i$ where Q_j is the j^{th} quality variable and x_i is the i^{th} textual feature. The result was not very good. With the textual features as the independent variables, the total variance explained was only 16.0–28.5% (Table 4).

DOI: 10.1002/asi



Pattern Matrix	Component 1	Component 2
ACCURA	.689	.206
SOURCE	.604	.158
OBJECT	.776	5.127E-04
DEPTH	.787	−.109
CREDIT	.615	−.227
READAB	1.342E-02	.811
VERBOS	−2.475E-02	.804
GRAMMA	7.313E-02	.729
MULTIV	.753	−.106

FIG. 4. Component plot and pattern matrix of the nine quality variables using principal component analysis, in rotated space. (Rotation method: Oblimin with Kaiser Normalization. Rotation converged in five iterations.)

TABLE 2. Categories of textual features used in the preliminary stage of data analysis.

Punctuation	Number of periods, question marks, exclamation marks, commas, semicolons, colons, dash, ellipsis, parentheses, brackets, quotation marks, forward slides, apostrophes, hyphens
Symbol	Number of dollar signs, percent signs, plus signs, > marks, ampersands
Length	Average length of words in characters, sentence in words, paragraph in words. Length of title, subtitle, leading paragraph, and document
Upper Case	Number of all upper-case words; number of words with first character upper case
Quotation	Average quotation length
Key Terms	Number of occurrence of the words "say," "seem," and "expert"
Unique words	Number of unique words; number of unique words excluding stop words
Parts of speech (POS)	Number of tokens, proper nouns, personal pronouns, possessive pronouns, determiners, preposition, verbs in base form, verbs in past-tense, verbs in present participle, verbs in past-participle, verbs in present tense, verbs in "ing" form
Entities	Number of persons, locations, organizations, and dates

TABLE 3. High correlation (two-tails) between quality variables and textual features.

Quality variable	Textual feature	Pearson correlation	Significance (two-tailed)
Accuracy	Personal pronoun	−0.202	0.0002
Reliability	Distinct organization	0.154	0.0048
Objectivity	Possessive pronoun	−0.219	0.0001
Depth	Distinct organization	0.236	0.0000
Credibility	Date unit, e.g., day, week	0.235	0.0000
Readability	Closing parenthesis	−0.141	0.0099
Verbose and conciseness	Subordinating preposition or conjunction	−0.197	0.0003
Multiview	Past-form verb	0.238	0.0000
Grammatical correctness	Average length of paragraph in words	−0.172	0.0016

TABLE 4. Goodness of fit of multiple linear regression.

Quality variable	R^2 of regression
Accuracy	0.181
Reliability of source	0.167
Objectivity	0.242
Depth	0.280
Credibility	0.237
Readability	0.262
Verbose and conciseness	0.210
Multiview	0.285
Grammatical correctness	0.160

We speculated that we might be more successful if we divided the quality scores into two ranges, high and low, where low = 0.5–5 and high = 5.5–10, and predicted the class instead of the actual score. To test this idea, we used discriminant analysis to combine textual features linearly. There are different ways to construct the discriminant function. We chose the classical approach (Klecka, 1980). Consider each document as a vector of textual features. We constructed the function by maximizing the difference (measured by sum of squares) between the centroids of the high group and the low group, and at the same time minimizing the difference between each vector and its centroid.

We randomly split our collection into two halves, one as training data and the other as testing data. In the training data

DOI: 10.1002/asi

set, we sought a linear combination of the frequencies of textual features (denoted by x_i in the following equation) as the basis for assigning cases into the high group and low group: $S_j = \alpha_0 + ? \alpha_i x_i$ where DS_j is the discriminant score of the quality variable j. Alpha was chosen in such a way that the ratio of between-group sum of squares to the within-group sum of squares would be at maximum:

$$\frac{\left(\frac{\sum_{i\in high} S_i}{N_{High}} - \frac{\sum_{k=1}^{N_{all}} S_k}{N_{all}}\right)^2 + \left(\frac{\sum_{j\in Low} S_j}{N_{Low}} - \frac{\sum_{k=1}^{N_{all}} S_k}{N_{all}}\right)^2}{\sum_{i\in High}\left(S_i - \frac{\sum_{i\in High} S_i}{N_{High}}\right)^2 + \sum_{j\in Low}\left(S_j - \frac{\sum_{j\in Low} S_j}{N_{Low}}\right)^2}$$

Then, we applied the discriminant function to the testing data set to classify documents. The correct classification rate could be used as a measure of performance of this method. We repeated the experiment for each quality variable. The results were good, with the correct classification rate better than chance, but not "impressive." For example, Tables 5 and 6 show the summaries of the two quality variables: depth and objectivity. The percentage in each cell is the correct classification rate of corresponding prediction. As we see, the power of the discriminant functions decreased more than 10% from the training data set to the testing data set, with only a little more than 60% correct classification rate.

Our predictive variables were POS (part of speech) tags and other features identified by the GATE and AWB systems. Selection of these variables does not consider the judges' perception of document qualities. In addition to asking judges to assign quality scores, we also asked judges to highlight and save a portion of text as supportive evidence of their judgment of each of the nine qualities. In the body of the evidence, we noticed that judges included many (a) explicit declarations, (b) implicit assertions, and (c) references to people or organizations that made the declarations and assertions. We added these three new categories of predictive variables, and gained a slight improvement (1.3–2.6%) across all qualities.

In our machine learning algorithm, scores with a large variance were treated the same as those with a small variance if they had the same mean. We suspected that we should give more attention to those judgments that agree with each other, and less attention to those judgments that disagree. In addition, if a machine could not correctly predict the membership (high or low for a particular quality) of a document about which two judges disagreed, this should not be counted as a total failure. Therefore, we added a weight factor to our machine learning and testing. Each document had nine weights (corresponding to nine quality variables), each weight was equal to $1/[(q_{i1} - q_{i2})^2 + 1]$, where q_{i1} was the score of the i^{th} quality variable assigned by one judge and q_{i2} was the score of the same quality variable assigned by the other judge. If two judges agreed, the weight would be equal to 1; the more two judges disagreed, the lower the weight. If two judges totally disagreed (i.e., one assigned the value 1 and the other assigned the value 10), the weight would be 0.012, and the case would become negligible. For the document quality "depth," using the weight mechanism, the correct classification rate improved by 5.5% in the training data set, but only by 0.4% in the testing data set.

Evaluating this result, we wondered if we might not have enough cases in the training data set. To compensate for the possibility that we might not have a large enough corpus, we used 800 cases for training and 200 cases for testing. The results were better. For example, for the quality depth, the correct classification rate improved by 2.6% in the testing data set and by 4.7% in the testing data set.

TABLE 5. Classification result of quality variable *depth* using linear discriminant analysis. Overall, 74.5% of training cases can be correctly classified, 61.60% of testing cases correctly classified.

			Predicted group membership	
Depth			Low	High
Training cases	Averaged human judgment	Low	67.7%	32.3%
		High	21.0%	79.0%
Testing cases	Averaged human judgment	Low	54.4%	45.6%
		High	33.9%	66.1%

TABLE 6. Classification result of quality variable *objectivity* using linear discriminant analysis. Overall, 75.5% of training cases correctly classified, 63.5% of testing cases correctly classified.

			Predicted group membership	
Objectivity			Low	High
Training cases	Averaged human judgment	Low	58.7%	41.3%
		High	12.7%	87.3%
Testing cases	Averaged human judgment	Low	45.5%	54.5%
		High	23.5%	76.5%

Reducing the Set of Features

We have used discriminant analysis as an exploratory tool. To arrive at a good model, all the potentially useful variables were included in the data set. We did not know in advance which of these variables are important for predicting the high–low distinction, and which variables are, more or less, extraneous. Because one of the desired products of our analysis is the identification of good predictor variables, we used a stepwise variable selection algorithm (Huberty, 1994) to see if we could reduce the number of predictive variables.

In our stepwise approach, the first variable included has the largest value for the selection criterion, then the value of the criterion is reevaluated for all variables not in the model. The remaining variable with the largest criterion value is entered next. At this point, the variable that was entered first is reevaluated to determine whether it meets the removal criterion. If it does, it is removed from the model. Next, all

DOI: 10.1002/asi

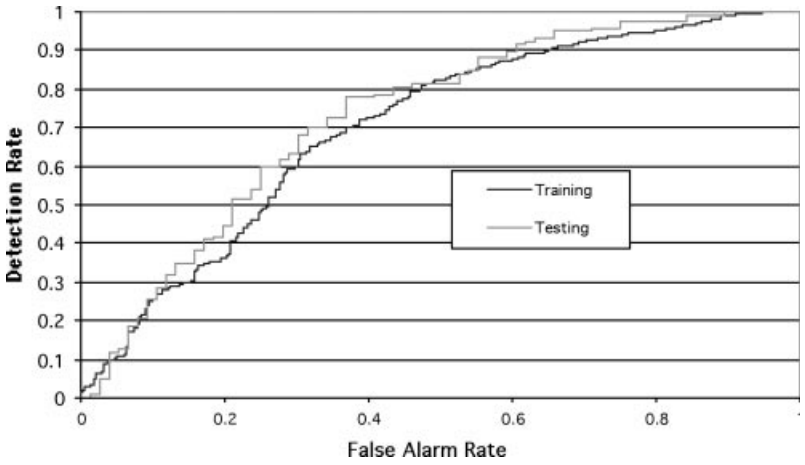


FIG. 5. Receiver operating characteristic (ROC) curves of training and testing (quality variable: depth) using stepwise discriminant analysis.

variables not in the equation are examined for entry, followed by an examination of the variables in the equation for removal. Variables were removed until none remained that met the removal criterion. Variable selection terminates when no more variables meet entry or removal criteria. Using this approach, we reduced the number of predictive variables from more than 100 to only a few without degrading the predictive power. For example, in predicting depth, we reduced the number of predictive variables to four: (a) log of sum of nouns (i.e., the sum of the frequencies of proper and common nouns, normalized by natural logarithm), (b) date unit (e.g., day, week), (c) determiner, and (d) number of all upper case words.

If we consider classifying high-depth documents correctly as "detection" and incorrectly classifying low-depth documents (as high-depth documents) as "false alarm," we can plot ROC (see Egan, 1975) curves by sorting discriminant scores numerically (Figure 5). Every point along the curves represents a possible cut-off point to discriminate between high-depth and low-depth documents. The associated detection rate of a point is the ratio of the number of high-depth documents that would be correctly classified using that point as a cutoff to the total number of high-depth documents. The associated false alarm rate of a point is the ratio of the number of low-depth documents that would be incorrectly classified as high-depth using that point as a cut-off to the total number of low-depth documents.

The two curves are very similar. The training curve is above the testing curve in some ranges, and the testing curve above the training curve in other ranges. Because the two curves are so similar, it is not surprising that there are points in the testing curve with higher detection rates and lower false alarm rates than the training curve. We observed similar results for other quality aspects. For example, using stepwise discriminant analysis for "objectivity" reduced the predictive variables from more than 100 to only 8: (a) average length of paragraph in words, (b) number of forward situations (i.e., the subject that made the declaration was found when searched forward from the declarative verb), (c) distinct organizations, (d) possessive pronouns, (e) comparative adjectives, (f) plural proper nouns, (g) distinct persons, and (h) length of leading paragraph. Overall, the discriminant equation correctly classified 73.3% of the training cases, and 71.3% of the testing cases.

One way of assessing the power of classifiers is to concentrate on the accuracy of classification for those top *N* documents for which the system makes the strongest prediction. With this in mind, Figure 6 shows results for predicting human judgments of depth. To review, we used 500 documents to train a classifier, and that classifier ranked documents according to the estimated probability to be high depth. Then, we applied the stepwise method to reduce the number of predictive variables without degrading performance. We used the resulting equation to rank the 500 testing documents in decreasing order of the probability of being a high-depth document. In Figure 6, the vertical axis is the precision rate of prediction at different cut-off points. The baseline is equal to the accuracy we would expect from random guessing, or the percentage of documents with high-depth scores in our testing data collection. Here the base line is about 62%.

When using the top 50 documents, we can be assured that about 80% of them would have high depth. On average, our prediction of depth is about 30% better than chance. This is good, but not spectacular. Depth is not the easiest document quality to predict. Our best prediction is for multiview; the worst is for objectivity. Using precision at the top 50th, top 100th, top 200th, and top 300th document as cut-off points, we can, on average, predict multiview 60% better than chance, depth 30% better than chance, and objectivity only 15% better than chance.

Stage 2: New Corpus and Interface

In the confirmative evaluation stage, we applied the algorithms developed to a new corpus to test the stability of performance. We built a larger, more diversified collection of 2200 documents (Table 7). We selected 1100 documents from the Center for Nonproliferation Studies (Monterey, CA), averaging about 400 words in length. We also selected 600 news articles from the Associated Press, the *New York Times*, and *Xinghua* (from the TREC collection), with an

DOI: 10.1002/asi

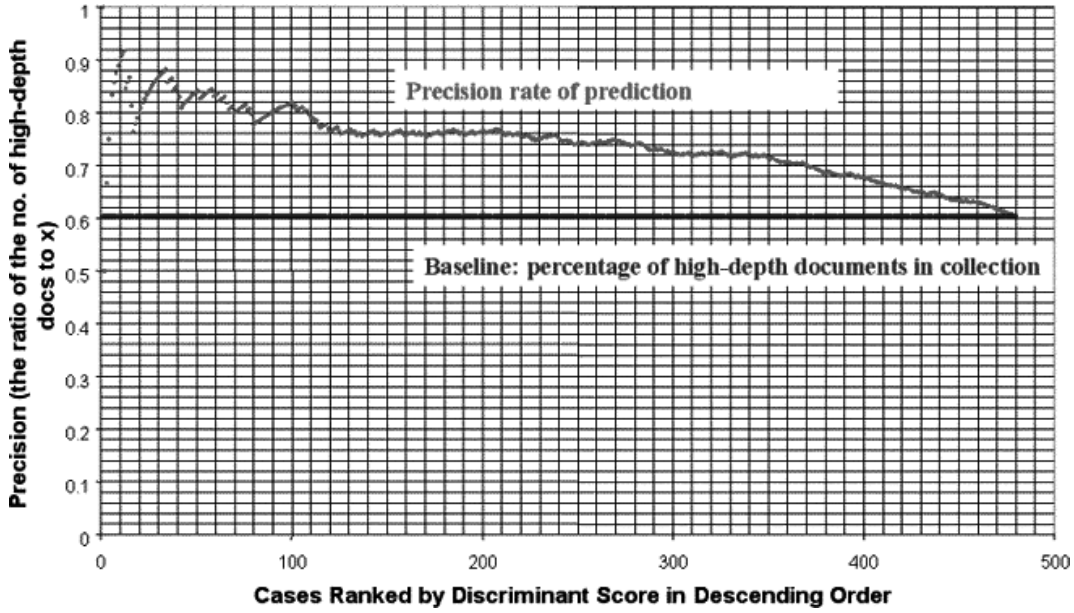


FIG. 6. Precision curve of predicting depth (split half for training and testing, stepwise method).

expert004	NYT20000424.0129	Source Reliability	7	Many of those interviewed are residents who have a stake in the iss
expert004	NYT20000424.0129	Author/Producer Credibility	5	The New York Times is balanced, but still biased
expert004	XIE19970805.0031	Author/Producer Credibility	0	Not familiar with producer, no author given
expert004	APW20000405.0141	Source Reliability	3	all the sources are government officials, no comment from environme
expert004	APW20000405.0141	Objectivity	3	Associated Press enough said. Presents a mainstream point of view
expert004	APW20000405.0141	One-Sided -- Multi-views	3	All government sources of information
expert004	WEB07D11	Accuracy	8	Hard to tell since I am unfamiliar with many of the facts
expert004	WEB07D11	Source Reliability	7	A government report may or may not be truthful, hard to tell

FIG. 7. Example of comments (last column) made by judges (first column) for the scores (fourth column) of a quality (the third column) assigned to a document (second column).

TABLE 7. Documents statistics of the second stage experiment.

Doc type	Number	Average length in words	Source
CNS	1,100	402	Nonproliferation data from the Center for Nonproliferation Studies (CNS)
AQUAINT	600	541	Articles from the Associated Press Worldstream (APW), the *NY Times* (*NYT*), and Xinghua English (XIE)
WEB	500	795	Google search

average length of about 550 words. Finally, we collected 500 Web documents using a Google search with keywords extracted from the selected news articles.

In stage 1, when we examined the evidence saved by judges, we had a hard time trying to figure out exactly how the evidence was associated with judgments. Therefore, in stage 2, we revised our interface to provide two additional functionalities: (a) users could select multiple portions of text as justification for their judgments, and (b) users could add comments to explain their scores and evidence (Figure 7). In this way, we tried to close the gap between a judge's score and the evidence she or he selected.

To recruit judges for the second stage, we followed a similar procedure as the first stage. Two qualified students judged all documents twice. By applying principal component analysis on the 2200 document vectors, we identified two similar major dimensions as in stage 1.

Performance of Stage 2

The purpose of stage 2 was to investigate the result of using the method developed in stage 1, and to see if the same approach can be used with a new corpus. The results are quite good. The approach was directly transferred from stage 1 to stage 2. For example, Figure 8 shows the result for the document quality depth. Here, we used 1100 documents for training and 1100 documents for testing. Compared to stage 1, our performance is better. The precision curve is similar, but the baseline is lower than the previous stage (about 48%). The improvement for the top 50 docs is $(0.75 - 0.48)/0.48 = 56.25\%$.

The predictive method we used was primarily based on linear discriminant analysis. There are some inherent limitations associated with discriminant analysis. For example, it assumes a multivariate Gaussian distribution of the predictive variables, and that may not be valid for the array of variables based on textual characteristics. Therefore, we supplemented it with logistic regression (Menard, 1995) to see if we could improve the performance. We ran logistic

DOI: 10.1002/asi

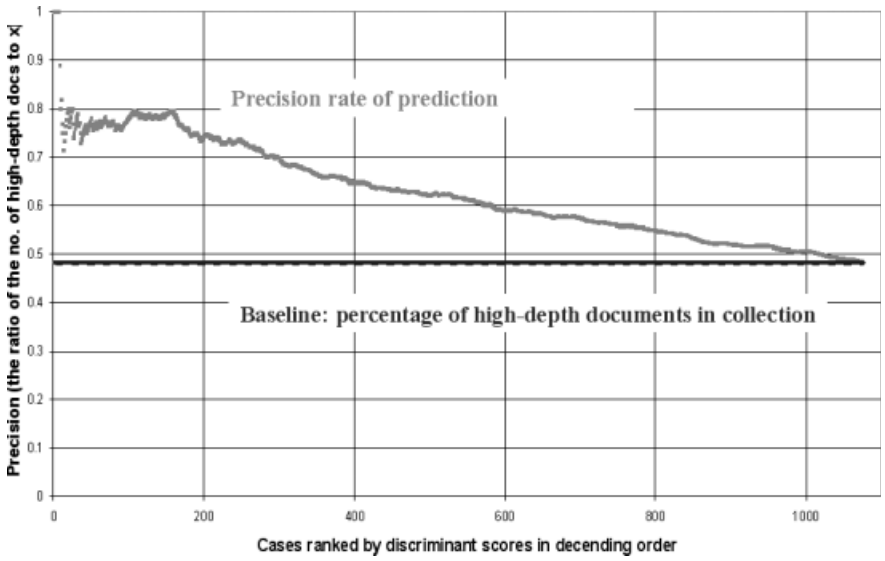


FIG. 8. Result of applying the same approach developed in stage 1 to stage 2 documents to predict document quality depth.

regression on all the quality variables; each gave us correct classification rates similar to discriminant analysis.

We also used a decision tree method, local weighted regression, and a support vector machine to see if these nonlinear methods could give us better performance, but they all had similar performance. It seems that there is not much difference between broad classes of induction methods. Perhaps it is not very surprising, as Langley and Simon (1995) have pointed out, that in machine learning, it is not uncommon for the particular induction method one employs to have little effect on the outcome.

Another strategy we tried was training on extreme cases. That is, in the training document set, we selected only documents that have a quality score higher than 7, or lower than 3, to estimate the parameters of the discriminant function. We then applied the function to the testing document set. This did not result in any improvement, for either stage 1 or stage 2 documents.

Discussion

Personalization

Although in stage 2 we could use the same approach we developed in stage 1 without modification, we were not satisfied. It seemed to us that, no matter what method we used, there was some internal barrier to improving the performance of the prediction if we just used textual features as predictive variables. We speculated that we might need to include semantic elements in our equation. However, when we looked at the evidence saved by judges, we simply could not identify discernable patterns behind the evidence. We found that two judges might give similar scores to the same document but select very different portions of text as evidence to support their judgments. Their comments did not offer much help. Different judges were simply judging quite differently, following different criteria as they saw fit.

At this point, we began to realize that different judges might have their own idiosyncratic ways of judging, and we needed to take that into consideration. Document qualities are neither physically nor textually embedded in documents. They are the result of the interaction between the mental structure of an individual judge and the textual and linguistic structures of documents at various levels. We understand that there are many interlocking problems: (a) the understanding of the meaning of the nine document qualities may vary among individuals; (b) the judgment criteria may vary among individuals; (c) the interpretation of the meaning of a document may vary among individuals; (d) the effect of presentational elements on content elements and vice versa might vary among individuals, etc. All of these problems pointed to the same solution: personalization of prediction.

In stage 1 and stage 2, we employed many judges to read and rate 20–40 documents. The resulting corpus is a mixed product of many different mental models, and cannot be used for personalization. To pursue in the direction of personalization of classifiers, we needed a personalized corpus. Therefore, we asked four of our experienced judges to judge about 500 additional documents. This was barely enough for us to train and test a qualities classifier on a personal base, but the results are impressive. For example, Figure 9 shows the result for one judge for document depth. We used stepwise discriminant analysis for split half training and testing, with only four predictive variables in the final equation. For the top 50 documents, the precision rate is always higher than 90%. Compared to the baseline, the improvement is more than 100%.

Conclusion and Future Work

From our experiments and analyses, we have come to the following conclusions: (a) The nine document quality aspects were predefined by journalists. Other information professionals may have different preferences and criteria of

DOI: 10.1002/asi

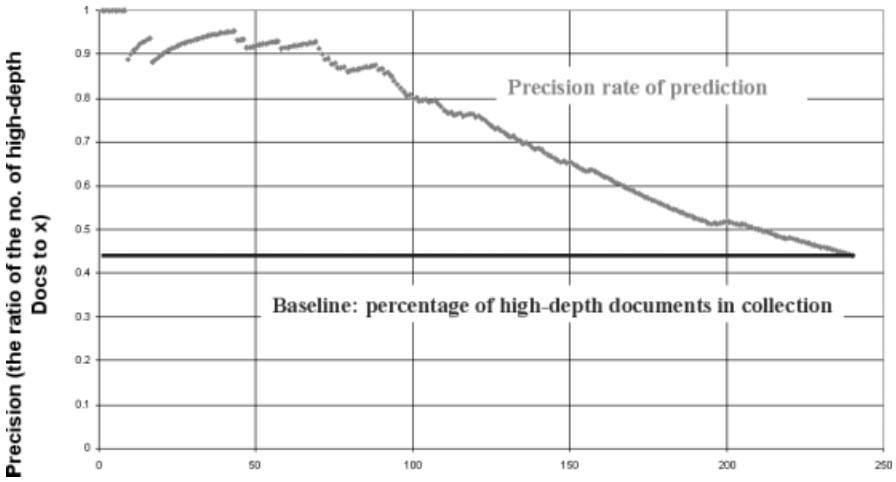


FIG. 9. Result of personalization of training (240 documents) and testing (240 documents) for the document quality depth.

document qualities; and (b) individual judges may have different interpretations and personal rating scales. We are not sure whether there is a community (common) base underlying a community, but it seems to us that instead of approaching the problem from a global perspective, a local perspective may be more promising. We are confident that our successful result is transferable to different information communities, but we need to redefine important quality aspects tailored to the needs of other communities (either by the same method, or by system adaptation). As demonstrated here, we find that (1) Empirically, document qualities can be identified from either groups of people or individuals, using social science methods like focus groups or surveys; (2) statistical analysis and machine learning methods can be used to construct document quality classifiers (based on textual or linguistic features) that can, in good cases, stay quite close to the optimum performance for quite a while. We are now working on transferring the experimental design and methodologies to the intelligence community to verify our findings.

Acknowledgment

This research was supported in part by the Advanced Research Development Activity of the Intelligence Community, under Contract # 2002-H790400-000 and NBCHC040028 to SUNY Albany, with Rutgers University as a subcontractor.

References

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., & Wilks, Y. (2000). Experience of using GATE for NLP R&D. Paper presented at the Workshop on Using Toolsets and Architectures to Build NLP Systems at COLING-2000, Luxembourg. Retrieved from ftp://ftp.dcs.shef.ac.uk/home/hamish/auto-papers/Cun00c.ps

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., & Vilain, M. (1997). Mixed-initiative development of language processing systems. In Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics. Retrieved on December 12, 2002, from http://www.mitre.org/technology/alembic-workbench/ANLP97- bigger. html

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Huberty, C.J. (1994). Applied discriminant analysis. New York: Wiley.

Klecka, W.R. (1980). Discriminant analysis. Sage University paper series on quantitative applications in the social sciences (Series no. 07-019). Thousand Oaks, CA: Sage.

Langley, P., & Simon, H.A. (1995). Applications of machine learning and rule induction. Communications of ACM, 38, 66–64.

Menard, S. (1995). Applied logistic regression analysis. Sage University paper series on quantitative applications in the social sciences (Series no. 07-106). Thousand Oaks, CA: Sage.

Naumann, F. (2002). Quality-driven query answering for integrated information systems. Berlin: Springer-Verlag. Retrieved November 9, 2002, from http://link.springer.de/link/service/series/0558/tocs/t2261.htm

Ng, K.B., Kantor, P., Tang, R, Rittman, R., Small, S., Song, P., et al. (2003) Identification of effective predictive variables for document qualities. In R.J. Todd (Ed.), Proceedings of 2003 Annual Meeting of American Society for Information Science and Technology (pp. 221–229). Medford, NJ: Information Today, Inc.

Reyment, R., & Joreskog, K.G. (1993) Applied factor analysis in the natural science. New York/Cambridge: Cambridge University Press.

Salton, G. (Ed.). (1971). The SMART retrieval system—Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall.

Small, S., & Strzalkowski, T. (2004, May). HITIQA: A data driven approach to interactive analytical question answering. Paper presented at HLT 2004, Boston, MA.

Tang, R., Ng, K.B., Strzalkowski, T., & Kantor, P. (2003) Toward machine understanding of information quality. In R.J. Todd (Ed.), Proceedings of the American Society for Information Science and Technology 2003 Annual Meeting (pp. 213–220). Silver Spring, MD: American Society for Information Science and Technology.

Voorhees, E. (2001). Overview of TREC 2001. In E. Voorhees (Ed.), Proceedings of the Tenth Text REtrieval Conference (pp. 1–15; NIST Special Publication 500-250). Gaithersburg, MD: National Institute of Standards and Technology.

DOI: 10.1002/asi