Proceedings of HLT-NAACL 2003
Short Papers , pp. 97-99
Edmonton, May-June 2003

Automatically Predicting Information Quality in News Documents

Rong Tang
School of Information Science and Policy
University at Albany
135 Western Avenue
Albany, NY 12222
tangr@albany.edu

Kwong Bor Ng
Graduate School of Library and Information Studies, Queens College, CUNY.
New York, NY 11367
kbng@qc.edu

Tomek Strzalkowski
ILS Institute
Univerity at Albany
1400 Washington Ave
Albany, NY 12222
tomek@albany.edu

Paul B. Kantor
School of Communication Information and Library Studies
Rutgers University
New Brunswick, NJ 08901
kantor@scils.rutgers.edu

Abstract

We report here empirical results of a series of studies aimed at automatically predicting information quality in news documents. Multiple research methods and data analysis techniques enabled a good level of machine prediction of information quality. Procedures regarding user experiments and statistical analysis are described.

1 Introduction

As a part of a large-scale multi-institutional project HITIQA (High-quality Interactive Question Answering), we worked on developing an extended model for classifying information by quality, in addition to, and as an extension of the traditional notion of relevance. The project involves Computer and Information Science researchers from University at Albany and Rutgers University. Our serving clientele are intelligent analysts, and the documents that we targeted were news articles.

2 Research Approach

The term "Quality" is defined by International Organization of Standards (1986) as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied need" (Standard 8402, 3.1). Among numerous study on classification of information quality, Wang and Strong (1996) proposed four dimensions of qualities as detailed in Table 1: intrinsic, contextual, representational, and accessibility.

Categories	Elements
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Accessibility, Security
Contextual IQ	Relevancy, Value-added, Timeliness, Completeness, Amount of Information
Representational IQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation

Table 1. Information Quality Dimensions (Source: Strong, Lee, Wang, 1997, p.39)

Empirical attempts to assess quality have primarily focused on counting hyperlinks in a networked environment. Representative studies include the work by Amento and his colleagues (Amento, Terveen, & Hills, 2000), Price and Hersh (1999), and Zhu and Gauch (2000). However, as a whole, previous studies were only able to produce algorithmic measures for Web documents based on link counts and with a limited number of quality aspects such as popularity. Our approach is to record actual users' quality assessments of news articles and conduct advanced statistical models of association between users' quality scoring and occurrence and prevalence of certain textual features.

3 Methodology and Results

Multiple research methods were used. Firstly, we conducted focus-group sessions to elicit key quality aspects from news analysts. Secondly, we performed experts and students quality judgment experimental sessions. Thirdly, we identified a set of textual features, ran programs to generate counts of the features, and performed statistical analysis to establish the correlation between features and users' quality ratings.

Two focus group sessions were conducted during March and April of 2002. Participants included journalism faculty members, professional editors, and a number of journalists from a local newspaper Albany Times Union. Nine information quality criteria were considered to be salient to the context of news analysis: *Accuracy*, *Source reliability*, *Objectivity*, *Depth*, *Author credibility*, *Readability*, *Conciseness*, *Grammatically Correctness*, and *Multiple Viewpoints*.

A computerized quality judgment system that incorporated the nine quality aspects was developed. One thousand medium-sized (100 to 2500 words) news articles were selected from the TREC collection (Voorhees, 2001) with 25 relevant documents each from five TREC Q&A topics.

We recruited expert and student participants for judgment experiments. Expert sessions were performed first and ten documents judged by experts were selected and used as the training and testing material for the student participants. The entire judgment experiment period ran from May to August of 2002. As a result, each of the 1,000 documents was rated twice, by two different judges, one at Albany, and one at Rutgers.

There were high inter-judge agreements between Albany and Rutgers. Figure 1 is the normality plot of the difference between scores assigned by Rutgers' judges and Albany's judges on the variable of "accuracy," with a mean almost equals to zero (with range from – 9 to + 9). The curves of the other eight quality variables are similar to the one below, indicating a very insignificant disagreement in judgments.

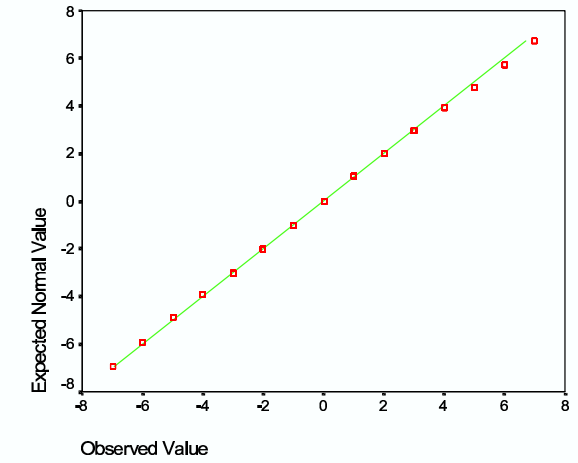

Figure 1. Normality Plot of differences in quality judgments on the aspect of "Accuracy"

Principle component analysis (PCA) revealed the same two components from Albany data as from Rutgers data. As shown in Figure 2, one component (the lower one) consists of "credibility", "source reliability", "accuracy", "multi-view", "depth", and "objectivity." The second component (the upper one) consists of "grammar", "readability", and "verbose and conciseness". Together they explain 58% of the variance.

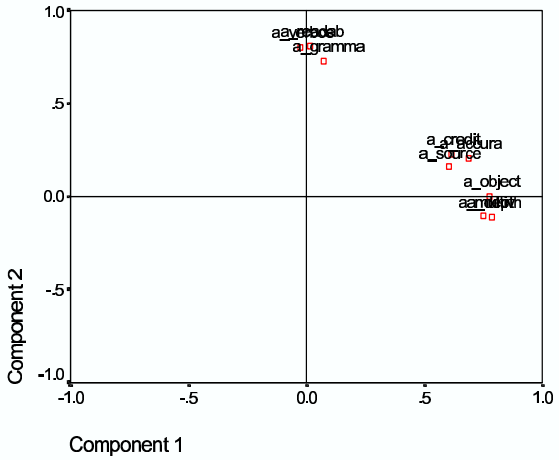

Figure 2. PCA of Judgment data, in rotated space. Rotation method: Oblimin with Kaiser Normalization. Rotation converged in 5 iterations.

We recoded users' scores 1 to 5 as low and scores 6 to 10 as high. We split the 1,000 documents into two halves by random selection. In our training round the first half was used to estimate the parameters that would give best discriminant and logistic regression functions. In our testing round, we applied the functions to the other half to predict the quality criteria of the documents.

	Discriminant Analysis Correct-Rate	Logistic Regression Correct-Rate
Accuracy	75.8%	75.9%
Source Reliability	67.8%	68.5%
Objectivity	70.6%	73.8%
Depth	77.4%	77.9%
Author Credibility	69.3%	71.7%
Readability	81.3%	83.0%
Conciseness	70.5%	70.9%
Grammar	74.9%	75.1%
Multi-view	82.1%	82.2%

Table 2. Performance of prediction (based on split-half training and testing) by two methods

We then employed stepwise discriminant analysis to select the dominant predictive variables from a range of 104 textual features. These features included elements of punctuations, special symbols, length of document segments, upper case, quotations, key terms, POS, and entities. Our further analysis suggested that certain text features are highly correlated with each of the nine aspects.

Quality Aspects	Textual Feature	Pearson correlation (2 tails)
Accuracy	Personal Pronoun	0.0002
Source	Distinct organization	0.0048
Objectivity	Pronoun	0.0001
Depth	Document length	0.0000
Author Credibility	Date unit, e.g. day, week	0.0000
Readability	Closing parenthesis	0.0099
Conciseness	Subordinating preposition or conjunction	0.0003
Multi-view	Past tense verb	0.0000
Grammatical correctness	Average length of paragraph in words	0.0016

Table 3. Highly correlated textual features and quality aspects

At this point, we are able to produce good prediction of several aspects of information quality, including Depth, Objectivity, Multi-view, and Readability. The prediction testing and training for the remaining quality aspects are currently in progress. Tables 4 and 5 illustrate the results of training versus testing classification for the criteria of "objectivity" and "depth," with ratings grouped into high and low categories.

Objectivity			Predicted Group Membership	
			Low	High
Training Cases	Original	Low	58.7%	41.3%
		High	12.7%	87.3%
Testing Cases	Original	Low	45.5%	54.5%
		High	23.5%	76.5%

Table 4. Classification result of "objectivity." 75.5% of training cases correctly classified, 63.5% of testing cases correctly classified

Depth			Predicted Group Membership	
			Low	High
Training Cases	Original	Low	64.5%	35.5%
		High	11.9%	88.1%
Testing Cases	Original	Low	51.0%	49.0%
		High	22.6%	75.4%

Table 5. Classification result of "depth." 74.5% of training cases correctly classified, 61.6% of testing cases correctly classified

4 Summary

In this study, we were able to identify important quality criteria relevant to intelligent analysts' work and we were also able to generate automatic quality metrics of news documents using users' quality judgments. Our next step is to apply our machine prediction method to produce measures of a new set of documents and have users to verify and modify machines' scoring. We hope that through this, we can collect new data to test our quality metrics and to further improve its' performance.

Acknowledgement

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000.

References

Amendo, B., Terveen, L., & Hill, W. (2000). Does "authority" mean quality? Predicting expert quality ratings of Web documents. *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 296-303.

Price, S. L., & Hersh, W. R. (1999). Filtering Web pages for quality indicators: An empirical approach to finding high quality consumer health information on the World Wide Web. *Proceedings of the AMIA 1999 Annual Symposium.* 911-915.

Voorhees, E. (2001). Overview of TREC 2001. In E. Voorhees (ed.) *NIST Special Publication 500-250: The Tenth Text REtrieval Conference*, pp. 1 – 15. Washington, D.C.

Strong, D., Lee, Y., & Wang, R. Y. (1997). 10 potholes in the road to information quality. *IEEE Computer, 30*(8), 38-46.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-34.

Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 288-295.