Quality-Based Automatic Classification for Presentation Slides

Seongchan Kim, Wonchul Jung, Keejun Han, Jae-Gil Lee, and Mun Y. Yi

Dept. of Knowledge Service Engineering, KAIST, Korea
{sckim,wonchul.jung,keejun.han,jaegil,munyi@kaist.ac.kr}

Abstract. Computerized presentation slides have become essential for effective business meetings, classroom discussions, and even general events and occasions. With the exploding number of online resources and materials, locating the slides of high quality is a daunting challenge. In this study, we present a new, comprehensive framework of information quality developed specifically for computerized presentation slides on the basis of a user study involving 60 university students from two universities and extensive coding analysis, and explore the possibility of automatically detecting the information quality of slides. Using the classifications made by human annotators as the golden standard, we compare and evaluate the performance of alternative information quality features and dimensions. The experimental results support the validity of the proposed approach in automatically assessing and classifying the information quality of slides.

Keywords: Information Quality (IQ), Presentation Slides, Classification.

1 Introduction

Computerized presentation slides have become a popular and valuable medium for various occasions such as business meetings, academic lectures, formal presentations, and multi-purpose talks. Online services focused on presentation slides, SlideShare[1] and CourseShare[2] to name a few, offer the ability to search and share computerized presentation slides on the Internet. Millions of slides are available on the Web and the number is growing continuously. However, most of the slide service platforms suffer from the problem of discerning the quality of available slides. This problem is acute and getting worse as the number of slides is continuously increasing. Further, on most platforms anyone can upload their slides. An automated classification approach, if effective, offers several benefits: 1) users are directed to select high quality slides among a group of similar slides, and 2) the assessed quality of a slide can be integrated into the searching and ranking strategies of slide-specialized search engines. For instance, none of the currently available slide search engines (e.g., SlideShare and CourseShare) support automated slide categorization or ranking by quality.

[1] http://www.slideshare.net
[2] http://www.courseshare.org

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 638–643, 2014.

When issuing a query to these slide-specialized search engines, end-users will get only a list of keyword-relevant slides, with no information on slide quality. The automated classification of high quality slides is an important issue for the advancement of search engines focused on presentation slides.

In the area of information retrieval, measuring information quality (IQ) for Web documents and for Wikipedia has recently been attempted by several studies. A set of quality features for Web documents for improving retrieval performance [1, 3] have been suggested, and a different set of quality indicators for better ranking and automatic quality classification of articles on Wikipedia [2, 4] have also been reported. However, these quality indicators for Web documents and Wikipedia are inappropriate for presentation slides because they overlook the importance of the representational aspects.

To the best of our knowledge, this study is the first attempt to define quality metrics for automatic classification of presentation slides. In this study, we consider only lecture slides, as they are the most popular. The key contributions of this paper are: 1) we investigate presentation slide characteristics and propose a new, comprehensive framework of information quality developed specifically for presentation slides on the basis of a user study, and 2) we assess the validity of the identified quality features and dimensions for the task of automatic quality assessment.

2 Related Work

IQ related research work has recently been receiving considerable attention in the information retrieval research community; however, no research has yet been attempted that has considered slide quality. Several studies on the classification and ranking of Web documents [1, 3] and Wikipedia articles [2, 4] have been reported.

For Web documents, Zhou and Croft [3] devised a document quality model using *collection-document distance* and *information-to-noise ratio* to promote the quality of contents in the TREC corpus. More recently, Bendersky et al. [1] utilized various quality features related with content, layout, readability, and ease of navigation, including the number of visible terms, the average length of the terms, the fraction of anchor text on the page, and the depth of the URL path. They reported a significant improvement in the retrieval performance with ClueWeb and GOV2 corpus.

For Wikipedia, Hu et al. [4] suggested the PEERREVIEW model, which considers the review behavior. The PROBREVIEW model was proposed to extend the PEERREVIEW model with partial reviewership of contributors. These models were used to determine Wikipedia article quality with features including the number of authors per article, reviewers per article, words per article, and so on. The proposed models were evaluated and found to be effective in article ranking. Dalip et al. [2] tried to classify Wikipedia articles with the following quality indicators: 1) text features (length, structure, style, readability, etc.), 2) review features (review count, reviews per day/user, article age, etc.), and 3) network features (page rank, in/out degree, etc.). They demonstrated that these structure and style features are effective in quality categorization.

We adopt several content features such as entropy and readability [1, 2] from previous studies in our experiment. However, these features are not discriminative enough to determine slide quality, because they overlook useful features about representational aspects of slides. Thus we suggest new features including font color, font size and the number of bold words for representational clarity. To the best of our knowledge, this is the first study of automatic quality classification of slides.

3 Quality Features of Presentation Slides

In this section, we present the quality features developed in order to determine presentation slide quality. Table 1 shows the 28 quality indicators derived for the five IQ dimensions, whose definitions were previously presented in [5, 6].

Table 1. Description of extracted quality features

Dimension	Indicator	Description
Informativeness (I)	numSlides	Number of slides
	numTerms	Number of terms in the slides [1, 2]
	avgNumTerms	Number of terms per slide [1, 2]
	numImgs	Number of images [2]
	avgNumImgs	Number of images per slide [2]
	preExample	Presence of example
	numExamples	Number of examples
	preTable	Presence of table [1]
	numTables	Number of tables [1]
	preLeaObj	Presence of learning objective
Cohesiveness (C)	entropy	Entropy of texts in the slide [1]
Readability (R)	numStops	Number of stopwords [1]
	fracStops	Stopword / non-stopword ratio [1]
	avgTermLen	Average term length of texts [1]
Ease of Navigation (EN)	preTableCnts	Presence of table of contents
	preSlideNums	Presence of slide numbers
Representational Clarity (RC)	numBolds	Number of bolds
	numItalics	Number of italics
	numUnderlines	Number of underlines
	numShadows	Number of shadows
	sumHighlights	Sum of bolds, italics, underlines, and shadows
	numRichTexts	Number of styled text blocks
	numFontSizes	Number of font sizes
	avgFontSize	Average size of fonts
	numFontNames	Number of font names
	numFontColors	Number of font colors
	numLineSpaces	Number of line spaces
	avgLineSpace	Average line space

We adopt some quality features from previous studies [1, 2]; the others are inspired by our own user study, which was conducted to determine the quality criteria of presentation slides. Our user study involved 60 students, recruited from two universities in order to balance their backgrounds and individual characteristics, each of which was asked to view five slides and think aloud while comparing the given slides. The verbal statements were all recorded and transcribed. Through extensive coding analysis, we identified the criteria of IQ and determined their dimensions as shown in Table 1. The entropy of document D is computed over the individual document terms as in Equation (1).

$$-\sum_{w \in D} p_D(w) log\ p_D(w), \quad \text{where} \quad p_D(w_i) = {tf_{w_i,D}}\Big/{\sum_{w_j \in D} tf_{w_j,D}} \tag{1}$$

4 Experiments

We automatically conducted a preliminary classification for high, fair, and low quality lecture slides using the proposed 28 features from the five IQ dimensions.

We randomly collected 200 MS PowerPoint presentation slides from SlideShare[1] in two courses*: data mining* and *computer network*. We manually annotated the 200 slides according to quality by hiring six graduates who had completed those courses successfully. Three annotators were assigned per course; as a result, each slide was judged by three annotators. Annotators were instructed to classify a given slide into only one class out of the three (high, fair, and low), considering all dimensions of quality such as informativeness, representational clarity, and readability etc. The inter-annotator agreement among the three annotators was $\kappa = 0.67$, which is considered to indicate substantial agreement according to Fleiss kappa [7]. Finally, we obtained 178 slides that had yielded agreement on labeling quality (high, fair, and low) from more than two annotators. These 178 slides (high: 55, fair: 83, and low: 40) were used for our classification. In order to extract the proposed quality features of the slides, shown in Table 1, we used the Apache POI[3], which is a Java API for reading and writing Microsoft Office files such as Word, Excel, and PowerPoint. We extracted the features from textual contents, file metadata, layout, etc., of PowerPoint files (ppt/pptx).

The classification in three classes: high, fair, and low was conducted using 10-fold cross validation. We used *SVM* and *Logistic Regression (LR)*, which have been widely adopted for classification, in the Weka toolkit [8]. The default parameter values given in Weka were chosen for our experiment. We report on three measures: *precision (P)*, *recall (R)*, and *F-1 score (F1)* (micro-averaged).

The results of classification are as shown in Table 2. Performance was measured by adding features of each dimension. The results clearly reveal the effectiveness of the proposed features and show that there is a little difference between SVM and LR.

[3] `http://poi.apache.org/`

Table 2. Performance of classification by adding individual dimensions

Features	SVM			LR		
	P	R	F	P	R	F
RC	0.402	0.479	0.371	0.509	0.515	0.500
RC+I	0.547	0.533	0.476	0.551	0.550	0.549
RC+I+EN	0.602	0.592	0.577	0.592	0.586	0.586
RC+I+EN+R	0.619	0.604	0.588	**0.617**	**0.615**	**0.614**
All included	**0.622**	**0.609**	**0.596**	0.600	0.598	0.597

Table 3. Top 11 features by information gain (IG)

Rank	Features	IG	Dimension
1	numTables	0.112	Informativeness
2	numFontColors	0.095	Representational Clarity
3	preSlideNums	0.093	Ease of Navigation
4	numImgs	0.093	Informativeness
5	numItalics	0.088	Representational Clarity
6	numSlides	0.087	Informativeness
7	numFontNames	0.077	Representational Clarity
8	preTable	0.034	Informativeness
9	preTableCnts	0.033	Ease of Navigation
10	preExample	0.031	Informativeness
11	preLeaObj	0.003	Informativeness

SVM with all features from all dimensions achieves the best performance of 0.596 in F1. With LR, the best performance of 0.614 in F1 is achieved when C (cohesiveness) was excluded. Performance is increased when features of each dimension are added, except C with LR.

To analyze the individual feature importance, we computed information gain (IG) for each feature. The top 11 features by IG are reported in Table 3. The results show that 10 features among the proposed features are considerably discriminative for the classification. It should be noted that there is a significant drop between the 10th and 11th. Among the top 10 features, `numTables` is the most discriminant, followed by `numFontColors and preSlideNums`. The results imply that rich tables, font colors, images, italicized fonts, slides, and font faces, and existence of slide numbers, table, table of contents, and example are engaging characteristics with which high quality slides can be discerned reliably. Remarkably, `numFontColors, preSlideNums, numItalics, numFontNames,` and `preTableCnts` are distinctive and discriminative features for slides, even though these categories are not considered and used for other documents such as Web and Wikipedia in previous studies [1-4]. Furthermore, these results are contrary to previous reports, in which it has been seen that, in terms of quality measurement, readability features with stopword fraction and coverage are effective features for Web documents [1] and structure features related to the organization of the article such as sections, images, citations and links are useful for Wikipedia articles [2]. As for features about readability, they might not be effective for slides because most slides are written in a condensed form to summarize the contents.

It is clar that newly proposed features about representation in this study are effective in slides. These results also support the necessity of our study for development a different set of features for slides.

5 Concluding Remarks

In this paper, we presented a new, comprehensive framework of information quality developed specifically for computerized presentation slides and reported the performance of automatically detecting the information quality of slides using the features captured by the framework. Although the study results may be considered not so highly strong, the study supports the validity of the proposed approach in automatically assessing and classifying the information quality of presentation slides. In our future work, we need to develop further salient features from the slide layout and content to improve the overall performance while considering other IQ dimensions such as consistency, completeness and appropriateness.

Acknowledgements. "This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0029185)." We thank the anonymous reviewers for the helpful comments.

References

1. Bendersky, M., Croft, W.B., Diao, Y.: Quality-biased ranking of web documents. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp. 95–104. ACM, Hong Kong (2011)
2. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 295–304. ACM, Austin (2009)
3. Zhou, Y., Croft, W.B.: Document quality models for web ad hoc retrieval. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 331–332. ACM, Bremen (2005)
4. Hu, M., Lim, E.-P., Sun, A., Lauw, H.W., Vuong, B.-Q.: Measuring article quality in wikipedia: models and evaluation. In: Proceedings of the 16th ACM International Conference on Information and Knowledge Management, pp. 243–252. ACM, Lisbon (2007)
5. Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C.: A framework for information quality assessment. J. Am. Soc. Inf. Sci. Tec. 58, 1720–1733 (2007)
6. Alkhattabi, M., Neagu, D., Cullen, A.: Assessing information quality of e-learning systems: a web mining approach. Computers in Human Behavior 27, 862–873 (2011)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–382 (1971)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations Newsletter 11, 10–18 (2009)