Measuring the Descriptiveness of Web Comments

Martin Potthast
Faculty of Media, Media Systems
Bauhaus University Weimar
99421 Weimar, Germany
martin.potthast@webis.de

ABSTRACT

This paper investigates whether Web comments are of descriptive nature, that is, whether the combined text of a set of comments is similar in topic to the commented object. If so, comments may be used in place of the respective object in all kinds of cross-media retrieval tasks. Our experiments reveal that comments on *textual* objects are indeed descriptive: 10 comments suffice to expect a high similarity between the comments and the commented text; 100-500 comments suffice to replace the commented text in a ranking task, and to measure the contribution of the commenters beyond the commented text.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Abstracting methods*; H.4.3 [Information Systems Applications]: Communications Applications—*Bulletin boards*

General Terms: Experimentation

Keywords: Comment Descriptiveness, Cross-media Retrieval

1. INTRODUCTION

Comments are among the oldest kinds of user-generated content on the Web and virtually all types of objects are being commented, be it texts, images, songs, videos, products, and personal profiles. Commenting may be another means to harness the wisdom of the crowds, which—unlike tagging, blogging, and wikiing—is not perceived as labor. However, comment boards are often flooded with all kinds of junk and spam, which may be a reason why research has widely neglected comments as a source of information.

Related Work. The only work which attempts to assess the intrinsic value of comments can be found in [8]. The authors investigate the impact of comments on blog search and report that the recall per query increases up to 15% if comments are included in a keyword search. However, their evaluation covers only a single retrieval task that is based on 40 queries—in terms of both scale and significance our analysis goes beyond this work. Other researchers use comments of a special type, namely product and movie reviews, in order to facilitate online shopping (e.g. [10]), or to evaluate sentiment analysis models [9]. Some use comments to extract sentences from blog posts for summarization purposes [2, 5]. Note that the relation of comments to the commented object is not analyzed in particular.

Cross-media retrieval is a subproblem of multimedia information retrieval in which the text surrounding a non-textual object has always been used to extract annotations [3, 6, 7]. Interestingly, comments have not been considered in this respect.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. COMMENT DESCRIPTIVENESS

Let D denote the set of comments related to an object x, and let d denote the concatenated text of all comments from D. To measure how much d describes the topic of x we need a retrieval model, which consists of a representation function and a relevance function. The former captures features of d and x to represent them as $\mathbf{d}$ and $\mathbf{x}$ respectively. The latter function maps $\mathbf{d}$ and $\mathbf{x}$ onto the interval $[0, 1]$, which indicates the range from no to maximum relevance. The reliability of a relevance value depends on the choice and quality of the two functions. Often, vector-based representations are employed and relevance is measured with the cosine similarity between vectors.

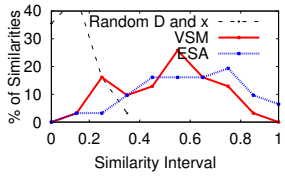
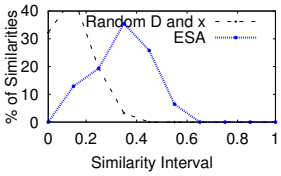
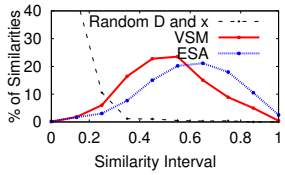
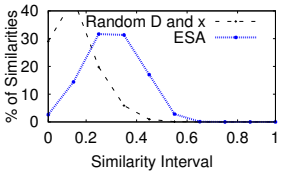
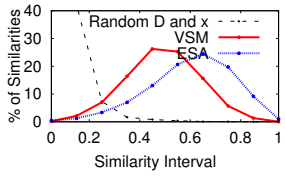
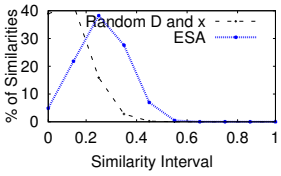
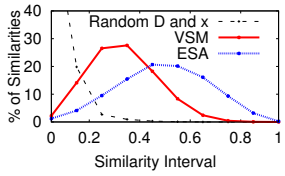
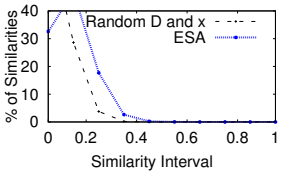
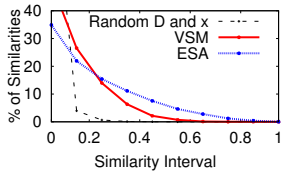
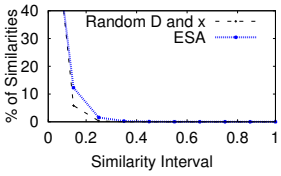
A small number of retrieval models exist which are capable of representing arbitrary objects in a cross-media feature space. They are trained on human-annotated corpora, but none of them contain any comments. However, commenting itself happens to be a cross-media phenomenon, and, a considerable amount of comments can be found on texts, such as blog posts or news articles. Under the assumption that the activity of commenting on text is not fundamentally different from that of commenting on non-textual objects we restrict our experiments to the text domain. If the assumption holds and if comments on texts prove to be descriptive, it follows that comments on non-textual objects are descriptive as well. Note that the validity of our assumption is a research question for cognitive information retrieval since commenting is a cognitive task. That said, our intuition defines well-intentioned commenting on an object x as describing x partially, up to a point, at which something new is contributed or an opinion is expressed. We expect that, compared with comments on texts, comments on images and videos are more often opinion exclamations, so that larger amounts of comments are needed to reach certain degrees of descriptiveness.

3. EXPERIMENTS

Evaluation Corpus. A rich resource for comments on text documents is the Slashdot news Web site, where news articles are published and commented in a community-driven process. The community is very active so that each article gets a considerable number of comments. We have downloaded all 17 948 articles published between January 2006 and June 2008, including the Web pages linked from each article and about 3.8 million comments.

Methodology. Two retrieval models are employed in the experiments: the well-known vector space model (VSM) and the explicit semantic analysis model (ESA), a collection-relative generalized VSM [1, 4]. In short, the latter represents a document x as vector of x's similarities to the documents of an index collection. Our index collection contains 10 000 randomly selected Wikipedia documents, and the similarity of x to each index document is computed

Table 1: Landscape of comment descriptiveness.

Experiment 1 Comment Similarity Distribution	**Experiment 2** Comment Rank Correlation	**Experiment 3** Commenter Contribution	**Comments** $\lvert D_i \rvert$ (Corpus Subset)
	VSM: 0.74 ESA: 0.84		1 000 (31)
	VSM: 0.61 ESA: 0.70		500 (951)
	VSM: 0.45 ESA: 0.42		100 (13 621)
	VSM: 0.35 ESA: 0.34		10 (17 748)
	VSM: 0.10 ESA: 0.17		1 (17 770)

Experiment Descriptions

Experiment 1: To determine the descriptiveness of comments, each document x is compared with the combined text d of its comments D. As a baseline, each x is compared once with the comments of another randomly selected document. The obtained similarity values are depicted in Column 1 of Table 1 as similarity distributions, i.e., the ratio of all similarities per similarity interval of 0.1 range.

Experiment 2: To determine if the combined text d from D can replace x in a ranking task, the remaining corpus documents are ranked twice: (*i*) wrt. their similarity to x, and (*ii*) wrt. their similarity to d. The top 100 ranks of the two rankings are compared using the rank correlation coefficient Spearman's ρ, which measures their (dis-)agreement as a value from $[-1, 1]$. The experiment has been repeated with randomly selected documents x until the averaged correlation value converged (cf. Column 2 of Table 1).

Experiment 3: To determine whether or not the observed similarities between d and x depend only on text which has been copied from x into one of D's comments, we (*i*) remove all terms from d which also occur in x, i.e., $\mathbf{d} \cap \mathbf{x} = \emptyset$, and (*ii*) exploit the fact that ESA, unlike the VSM, has the capability to measure more than just the overlap similarity between d and x. Column 3 of Table 1 shows the obtained similarity distributions.

using the VSM. To compare document representations both models use the cosine similarity. Note that we choose *basic* retrieval models to ascertain whether descriptiveness can be measured reliably. We have conducted three experiments on the evaluation corpus whose results are shown in Table 1. A detailed description of each experiment is given next to the table. In particular, it is our goal to determine the amount of comments on a document x necessary to reach a certain degree of descriptiveness. Hence, we use 5 subsets of the evaluation corpus which comprise only the documents which got at least $|D| \geq i \in \{1, 10, 100, 500, 1000\}$ comments. The experiments were repeated for each subset (= table rows), and, each experiment was repeated for each x in a subset. If an x had $|D| > i$ comments a random subset $D_i \subset D$, $|D_i| = i$, was chosen for the respective experiment.

Conclusion. Experiment 1 reveals that 10 comments are sufficient to reach a considerable similarity between a document and its comments compared to the baseline, Experiment 2 reveals that 100 to 500 comments are sufficient to reach a moderate rank correlation, and Experiment 3 reveals that 100-500 comments contain a measurable commenter contribution that is not contained in the original document. Particularly, Experiment 3 demonstrates ESA's capability to measure more than just overlap similarity and shows that the similarities measured in Experiment 1 cannot be attributed solely to duplicated text. Further, Experiment 3 may be interpreted as an indicator of the amount of comments necessary to capture the topics of non-textual objects. In all experiments the retrieval quality increases with the number of comments per document, $|D_i|$. Hence, comments on text documents can be called descriptive and it remains to be investigated whether our initial assumption holds that commenting is not entirely media-dependent.

4. REFERENCES

[1] M. Anderka and B. Stein. The ESA retrieval model revisited. In *Proc. of SIGIR'09*.

[2] J.-Y. Delort. Identifying commented passages of documents using implicit hyperlinks. In *Proc. of HYPERTEXT'06*.

[3] K. Deschacht and M.-F. Moens. Finding the best picture: cross-media retrieval of content. In *Proc. of ECIR'08*.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. IJCAI'07*.

[5] M. Hu, A. Sun, and E.-P. Lim. Comments-oriented blog summarization by sentence extraction. In *Proc. of CIKM'07*.

[6] M. Inoue. On the need for annotation-based image retrieval. In *Proc. of SIGIR'04 Workshop IRiX*.

[7] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: state of the art and challenges. *ACM TOMCCAP*, 2(1):1–19, 2006.

[8] G. Mishne and N. Glance. Leave a reply: an analysis of weblog comments. In *Proc. of WWW'06*.

[9] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in IR*, 2(1-2):1–135, 2008.

[10] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *Proc. of CIKM'06*.