Information Retrieval in the Commentsphere

MARTIN POTTHAST and BENNO STEIN and FABIAN LOOSE and STEFFEN BECKER

Web Technology and Information Systems, Bauhaus-Universität Weimar *

This paper studies information retrieval tasks related to web comments. Prerequisite of such a study and a main contribution of the paper is a unifying survey of the research field. We identify the most important retrieval tasks related to comments, namely filtering, ranking, and summarization. Within these tasks, we distinguish two paradigms according to which comments are utilized and which we designate as *comment-targeting* and *comment-exploiting*. Within the first paradigm, the comments themselves form the retrieval targets. Within the second paradigm, the commented *items* form the retrieval targets (i.e. comments are used as an additional information source to improve the retrieval performance for the commented items). We report on four case studies to demonstrate the exploration of the commentsphere under information retrieval aspects: comment filtering, comment ranking, comment summarization and cross-media retrieval. The first three studies deal primarily with comment-targeting retrieval, while the last one deals with comment-exploiting retrieval. Throughout the paper, connections to information retrieval research are pointed out.

Categories and Subject Descriptors: A.1 [**Introductory and Survey**]; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*abstracting methods, dictionaries, linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval models*; H.4.3 [**Information Systems Applications**]: Communications Applications—*bulletin boards*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Web Comments, Survey, Comment-based Retrieval, Commentsphere

1. INTRODUCTION

Numerous websites invite visitors to comment on their content. To this end, comment boards are provided at the bottom of web pages where submitted comments are shown in chronological order. Comments are one of the first kinds of user-generated web content, and virtually all types of items are commented, be them texts, images, songs, videos, products, ideas or personal profiles. Comment boards have not changed much since their debut in web-based guest books. Their very purpose is to collect user feedback, but they provide practical value to visitors as well. Hence we consider comment boards as a form of social software which create a community centered around the commented item on the respective web page. Comment boards serve as a paradigm to exploit the wisdom of the crowds since, ideally, commenters share their opinion, their criticism or extraneous information. Unlike tagging, blogging and "wiki-ing", commenting may not be considered work. In practice, however, comment boards appear less useful to the naked eye: popular web pages get flooded with up to thousands of comments, an amount impossible to be browsed by an individual user. Moreover, many comments are utterly irrelevant, spam or replications, which is why comments are often neglected as a source of useful information. It is all of these observations that form the starting point of our research.

The paper presents a study of the commentsphere, beginning with a discussion of the peculiarities of comment retrieval in Subsection 1.1. A survey of existing comment retrieval research is given in Section 2: we identify filtering, ranking, and summarization as important comment retrieval tasks,

*E-Mail: <first name>.<last name>@uni-weimar.de

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

which in turn can be classified as comment-targeting or comment-exploiting. Within this framework, Sections 3-6 report on four case studies:

Retrieval task	Usage paradigm	
	Comment-targeting	Comment-exploiting
Filtering	Case study 1, Section 3	
Ranking	Case study 2, Section 4	Case study 4, Section 6
Summarization	Case study 3, Section 5	

The original contributions of the case studies include: a comment filtering model based on writing style features, a comment ranking model based on novelty detection by similarity reduction, a comment summarization model based on a sentiment word cloud visualization, and a retrieval model for measuring cross-media similarity using comments.

1.1. Comment Retrieval Rationale

In the following, comment retrieval is compared to information retrieval. Let d_q be a web page or an item (e.g., a text, an image, a video) about a particular topic.

— *Information Need.* In comment retrieval, a user's interest in d_q is understood as an information need q that targets comments on d_q and d_q's topic. It is not assumed that d_q covers a topic exhaustively.
— *Query.* Formulating a keyword query that targets comments (in the sense of *"Retrieve all comments that contain more information on d_q's topic."*) is hardly possible. However, a good characterization of d_q's topic—and hence the information need—is d_q itself, which hence should be used as query document.
— *Relevance.* A relevant comment on d_q is a comment that *complements* the information of d_q in some respect. It is an information retrieval challenge to develop retrieval models that capture this kind of relevance.

Why is it unreasonable to formulate keyword queries against comment sets, just as we do against the web? The answer is not straightforward: a well-known principle of information retrieval states that nothing can be retrieved about an interesting topic without having a-priori knowledge about it [Belkin et al. 1982]. Within web search tasks, users formulate queries based on their incomplete knowledge about the topic for which they seek documents. The same holds true for retrieving comments, but at a much finer granularity. Since comments are short, the amount of knowledge in a comment is limited to a few facts. For example, to retrieve a comment, the user requires at least partial knowledge of the facts within. There is a (fuzzy) bound for the amount of a-priori knowledge about a fact x to be exceeded before users will start to search for x, and comments may be considered well below this bound. Another practical aspect prevents users from searching comments manually: why bother to search for a fact in a small, arbitrary set of comments when a web search engine is only one click away?

A good description of what users actually want to find in a set of comments on d_q is some kind of "surprise." This can be complementary information but also a joke. Either way the invariant is that the comments to be retrieved are on the same topic as d_q. Classical retrieval models (e.g. algebraic models like the vector space model, latent semantic indexing, explicit semantic analysis or probabilistic models like binary independence, the unigram language model or latent Dirichlet allocation) are not well-suited to measure such connections: when computing the relevance between d_q and some comment d, term overlap or concept overlap is measured in first place, which is obviously inappropriate for our purpose. In its extreme form, the classical models consider a comment as most relevant which duplicates d_q, although it contributes nothing.

1.2. Notes on Terminology

An information retrieval system organizes a set of real-world documents for efficient retrieval and provides a user interface to specify information needs. The paper in hand denotes a real-world document as d, an information need is denoted as q, and sets of documents and information needs are denoted as D and Q respectively. The heart of an information retrieval system is a retrieval model $\mathcal{R}$, which provides both the means and the rationale to compute document representations $\mathbf{D}$ for the real-world documents D, formalized information needs $\mathbf{Q}$ for their real-world counterparts Q, and a relevance function $\rho : \mathbf{Q} \times \mathbf{D} \rightarrow [0, 1]$. ρ estimates the relevance of some d with respect to some q as $\rho(\mathbf{d}, \mathbf{q})$.

The term "commentsphere" is derived from the term "blogosphere", i.e., the commentsphere is made up of all user-generated comments on web items. The term was coined in a comment on a blog post asking how blogs may be improved, in which one commenter suggested "*permalinks in the commentsphere*" [Phil Wolff 2004]. To the best of our knowledge, the first scientific paper mentioning the term is [Mishne and Glance 2006]; a variant is the term "commentosphere" [Schuth et al. 2007].

2. SURVEY OF EXISTING RESEARCH

This section presents a survey of research related to retrieval tasks in the commentsphere. Based on an analysis of 59 relevant papers, we identify three main retrieval tasks where comments are successfully utilized: filtering, ranking and summarization. Other tasks that have attracted less attention include discourse extraction and popularity prediction. We distinguish the mentioned tasks with respect to their retrieval targets, which are either the comments themselves or the commented items. We term the underlying paradigms as comment-targeting and comment-exploiting.

2.1. Comment-targeting Retrieval

Research which targets comments is organized in the Tables I and II. Special emphasis is placed on the underlying retrieval models. Two-thirds of the surveyed papers address only one specific *comment type*, namely reviews of products or movies. Reviews play an important role in decision-making when buying something online, which renders this comment type particularly interesting. Also, reviews are rated by other users, which means that they can be used for evaluation purposes. Research on reviews is known as opinion mining, and it relies on technologies developed for sentiment analysis. We emphasize the following distinction: while the very purpose of sentiment analysis is the identification of subjectivity and polarity, opinion mining employs sentiment as a single feature among many for the analysis of reviews.

All comment retrieval models rely on a feature-based vector representation, where a feature is a possibly elaborate function that quantifies a certain aspect of a comment. Table I organizes the variety of the found features according to nine research fields. Table II overviews the analyzed papers, whereas the retrieval model (multi-)column illustrates the feature usage by referring to the nine research fields of Table I: the •-symbol indicates the most important feature group of a model, which can be extended by features from other feature groups (indicated by the ◦-symbol). We also analyzed the related work with respect to the different approaches for relevance quantification:

— In comment filtering, the quality of a comment is quantified, where "good quality" refers to good writing style, the absence of vandalism and the absence of extreme sentiments. Also, the reputation of a commenter is taken into account, presuming that reputable commenters will not write low-quality comments in the future.
— In comment ranking, the relevance of a comment is put on a level commensurate with its helpfulness as it is perceived by other users. The existing retrieval models try to capture the concept of helpfulness from the human-labeled product reviews in order to predict the degree of helpfulness of a new review. This approach is related to the learning-to-rank paradigm.

Table I. Overview of features that are used within comment retrieval models.

Group	Features
IR	*(Information Retrieval)*
	– [sum of] word [n-gram] weights (Boolean, tf, $tf \cdot idf$)
	– character n-grams ($n \in [1, 5]$)
	– comment-article similartiy
	– word entropy
	– unigram language model
NLP	*(Natural Language Processing)*
	– number of dependent tokens
	– part-of-speech tokens (percentage of verbs, nouns, etc.)
	– number of verb / noun phrases
Style	*(Writing Style)*
	– word length (avg. number of syllables)
	– sentence length (avg. number of words)
	– paragraph length (avg. number of words)
	– comment length (number of words, sentences, or paragraphs)
	– highlightings (e.g. bold, italic)
	– punctuation (questions, exclamations)
	– reader's grade level (Dale Chall Formula, Power's Summer Kearl Grade, Bormuth Grade Level, Flesch Kincaid Grade, Gunning Fog Index, Coleman Liau Grade Level, Fry Readability Formula, Autom. Readability Index, SMOG)
Vand	*(Vandalism Detection)*
	– non-words, or gobbledygook
	– letter repetition
	– upper case words or spelling
Sent	*(Sentiment Analysis)*
	– word polarity (positive, neutral, negative)
	– sentence polarity (percentage of positive / negative words)
	– document polarity (percentage of positive / negative sentences)
	– word subjectivity (percentage of subjective / objective relations)
	– sentence subjectivity (percentage of subjective / objective words)
Com	*(Comment-specific Analysis)*
	– time of posting (absolute, relative)
	– user rating of the commented object
	– common debate phrases
	– user-feedbacks on helpfulness, comment quality (absolute, relative)
	– appearance of product-/movie-properties (in the comment, in sentences)
	– product information (price, sales rank, rating)
	– appearance of brand/product names (in the comment [title, body])
User	*(User Modeling)*
	– serial sharing commenter quotients (quickness, amount)
	– behaviors (percentages in rating others good or bad)
	– popularity (percentages of being rated good or bad, other's clicks/replies)
	– user identification (nickname, join date, role)
	– user relationships (friends, foes)
	– user expertise (acticvities on the same topic)
	– user success (top-rated contributions)
Other	
	– blog title string
	– use of genre core vocabulary (measured as mutual information)

Table II. Overview of comment-targeting IR technology. The tasks addressed within our studies are marked with an asterisk.

Retrieval task	Comment retrieval model								Classifier	Evaluation corpus		Reference
	IR	NLP	Style	Vand	Sent	Com	User	Other	Relevance basis		Type	Source(s)	
*	○	●	○						quality	Naïve Bayes	generic	Slashdot	**cf. Section 3**
	○						●		quality, reputation	association rule learning	generic	Slashdot	[Veloso et al. 2007]
	●						○	○	misuse, spam	Naïve Bayes, perceptron, logistic regression, SVM	generic	Rediff.com	[Sculley 2008a; 2008b]
filtering	●								spam	KL-divergence threshold	generic	–	[Mishne et al. 2005]
	○		○	○	○	●	○		quality, reputation	logistic regression	reviews	Amazon	[Jindal and Liu 2007; 2008]
			○		●	○			quality	SVM	reviews	Amazon	[Liu et al. 2007]
*	●	○							novelty	–	generic	Slashdot	**cf. Section 4**
	○	○	○	○		○	●		user preference	support vector regression	generic	Digg.com	[Hsu et al. 2009; Khabiri et al. 2009]
		●				○	○		helpfulness	radial basis function	reviews	Amazon, CNET	[Liu et al. 2008]
			●		○	○			helpfulness	multiple regression	reviews	Amazon	[Ghose and Ipeirotis 2007]
	○	○	●		○	○			helpfulness	SVM	reviews	Amazon	[Kim et al. 2006]
ranking	○	○			●				helpfulness	linear regression, SVM	reviews	Amazon	[Zhang and Varadarajan 2006]
					●				increasing positivity	custom	reviews	Ebert's Movies, ConsumerReports.org	[Lacey 2005]
					●	○			–	–	reviews	–	[Popescu and Etzioni 2005]
							●		–	stochastic simulation	reviews	eBay	[Huang et al. 2009]
*					●				word frequency	–	generic	YouTube	**cf. Section 5**
		○			●	○			sentence importance	–	reviews	Rotten Tomatoes, Amazon, CNET, IMDB	[Beineke et al. 2006; Hu and Liu 2004; Liu et al. 2005; Zhuang et al. 2006]
					●	○			sentence importance	–	reviews	CNET, Epinions, PriceGrabber	[Lerman et al. 2009; Lerman and McDonald 2009]
	○				●	○			sentence importance	SVM	reviews	Env. Prot. Agency	[Kwon et al. 2006; Kwon et al. 2007]
(opinion) summarization		○			●				sentence importance	–	reviews	NHK BS debate, ewoman.co.jp	[Fujii and Ishikawa 2006]
	○	●				○			property frequency	–	reviews	eBay	[Lu et al. 2009]
					●	○			property frequency	–	reviews	eOpinion.com	[Yang et al. 2009]

— In comment summarization, the prevalent approach is the extraction of sentences that express an opinion. The relevance of a full comment is not considered, but the importance of sentences in describing the content of a comment or all comments is.

It can be observed that the commenters on comment boards begin to discuss or even argue about the commented item d_q. Most comment boards, however, do not support discussion threading or record the reply-to structure. In this regard, Mishne and Glance [2006] tried to analyze whether a dispute is taking place on a comment board D. They represent the comment set D as a feature vector using features from almost all groups mentioned in Table I, train a decision tree classifier based on 500 annotated comment boards and achieve a classification accuracy of 0.88. Dit et al. [2008] and Schuth et al. [2007] go one step further and attempt to extract the discussion hierarchy from D.

Remarks. The wide range of features that are employed within the comment retrieval models shows the different views that can be taken on the data and makes comment-retrieval an interdisciplinary research field. A special case is comment filtering where many authors employ user-centered features and domain knowledge about reviews. These features are not applicable to other comment types since comments are often posted anonymously and do not necessarily review something.

Most of the comment ranking models use a comment's helpfulness as relevance measure, which is a reasonable choice for comment ranking. It must be considered, though, that helpfulness is hardly ever defined but derived from the ground truth of the evaluation corpora. If, for example, a model is trained on Amazon reviews it is an open question whether a "domain transfer" to other comment types will work. Moreover, it stands to reason that the initiating (query) document d_q for a set of comments D introduces a bias in the relevance assessments, which may not be crucial for reviews but for comments in general. The existing models measure the relevance of comments based on a static, predefined information need, since they do not consider d_q.

Comment summarization is done by extracting the "important" sentences from comments. Again, this can only be done reliably for reviews but not for comments in general: unlike reviews, comments tend to be short and messy, rendering sentence extraction and the quantification of their importance difficult. Altogether we observe that, while being an active research field, comment-targeting retrieval currently focuses too much on reviews. I.e., most likely the proposed models are not adequate for the wider commentsphere.

2.2. Comment-exploiting Retrieval

When retrieving web items for which a sufficient number of comments are available, the comments can be used to raise the retrieval recall. This was first observed by Mishne and Glance [2006], who performed a large-scale analysis on the importance of comments within the blogosphere. They found that comments account for up to 30% of its size and that the use of comments improves the recall of blog search by 5%-15%, indicating that comments are a vital part of the blogosphere. This work is also the first to assess the intrinsic value of comments. In blog retrieval, in order for a blog post to be relevant to a query, it suffices if one of its comments is relevant to the query, which was also the modus operandi of the TREC blog track [Ounis et al. 2008]. Recently, the same idea has been applied to video retrieval, in which comments provide a rich text resource as well—significantly larger than user-supplied tags or video titles [Cunningham and Nichols 2008; Yamamoto et al. 2008; Yee et al. 2009]. There is no doubt that this idea can be applied to image retrieval, music retrieval or other types of web items as well. In Section 6, we show that comments can be used to compare web items across media.

Comments are also exploited for the summarization of web items [Delort 2006; Hu et al. 2007; 2008; Park et al. 2008]: given an item d_q and comments D on that item, the task is to generate a summary of d_q. The comments on d_q are evaluated to find the often referred to parts of d_q. These parts are then used for the summary, circumventing the problem of identifying them solely based on d_q.

Similarly, filtering by comment exploitation is a viable monitoring and maintenance technology. Given a web item d_q about which commenters are outraged, which may be detected using the

Table III. Characterization of comment categories on Slashdot.

Category	Description of a comment d	Frequency	Avg. score
Negative Categories			
offtopic	d does not discuss d_q's topic	4%	-0.6
flamebait	d is meant to pick a fight	3%	-0.5
troll	d is a prank to waste other people's time and effort in responding	5%	-0.6
redundant	d repeats what has been said before	2%	-0.3
Positive Categories			
insightful	d makes d_q more accessible with new understandings, analogies, or examples	32%	3.1
informative	d adds new information or a new angle	14%	3.1
interesting	d does not fit in any other category	24%	2.9
funny	d is humorous respecting d_q's topic	16%	3.2

aforementioned dispute classification approach of [Mishne and Glance 2006], the web item could be automatically selected for reexamination by a site administrator. Though we have not found research addressing this, it is very likely that such measures are already being taken on sites such as YouTube, for example.

Another important task in this regard is the prediction of the popularity of a web item based on its comments [Jamali and Rangwala 2009; Kaltenbrunner et al. 2007; Mishne and Glance 2006; Szabó and Huberman 2008; Tsagkias et al. 2009; Yano et al. 2009]. Here, the comments are treated like time series data, using features, such as the increase of comments per time frame, to predict whether the commented item will become popular.

Remarks. Research on comment-exploiting retrieval is more diverse compared to the research targeting comments. This is in the nature of things as the ways in which comments can be exploited for different purposes cannot be enumerated. Approaches to comment-exploiting retrieval cannot be compared across different retrieval tasks. Within each comment-exploiting task, however, the literature is few and far between, which indicates that comments have not yet been adopted as a source of valuable information about the commented item. By highlighting comment-exploiting retrieval as a paradigm, we hope to foster research in this direction.

2.3. Evaluation Corpora

The fifth column in Table II shows places where comments can be found for evaluation purposes. Since most of the research is about reviews, Amazon is used most often as a corpus. However, there are plenty of other websites which may be useful in this respect (e.g. YouTube, Flickr, Last.fm, Digg, Picasa or news paper sites). Although there is no lack of comments in general, comments with human annotations are rare; exceptions include Amazon, Digg and Slashdot. For our studies, we have compiled two evaluation corpora based on comments from Slashdot and YouTube; both are available to other researchers upon request.

Slashdot Corpus. Slashdot is a news website for publishing and commenting technology-related news articles. The publishing process is based on a moderation system in which users can submit an article d_q, and Slashdot's editors decide whether or not d_q will be published. For each published article a comment board D is available, many of whose comments are categorized by Slashdot's comment moderators. Eight predefined comment categories are used: four of which are considered "positive" and four "negative" (see Table III for a short characterization). Based on the categories assigned by different moderators, an integer score is computed for each comment. The accounting of all assessments is mapped onto a range from -1 (negative) to +5 (maximum positive). Unfortunately, the scores do not reflect how many moderators are involved in an assessment.

We have downloaded all Slashdot articles from January 2006 to June 2008, including all comments. In total, 17 948 articles were published during this period, and about 3.8 million comments were posted. Comments are organized as discussion threads, which means that a large fraction of the comments are not direct responses to an article, but responses to other comments. Only a small fraction of all comments has been categorized by moderators. Our experiments are based on the 311 167 categorized, direct responses. Together, the second and third quartile of the articles get between 16 to 41 direct comments, while the second and third quartile of the comment lengths range from 1 to 45 words. With respect to the distribution of the comments on the categories, there seem to be only very few low-quality comments on Slashdot (see Table III). However, one should be careful to consider this result as an accurate picture considering most comments are not categorized and Slashdot policies encourage moderators to categorize positive rather than negative (i.e. moderators may spend time finding good comments instead of wasting time reading bad ones).

YouTube Corpus. YouTube is a video sharing website for homemade videos. The comments on videos are typically very short, and quite often thousands of comments per single video can be found. Since only a single video is associated per YouTube page, and since most comments are very short, we assume that most of them are some kind of opinion expression regarding the respective video. Explanations or discussions are less frequently observed than on Slashdot, for example. This makes YouTube comments especially interesting for opinion summarization. We downloaded 9.8 million comments from YouTube which were posted on 64 830 videos that appeared on several YouTube feeds at the end of 2008. Due to limitations of the YouTube API, only up to 1 000 comments per video could be retrieved, and it was not possible to adjust the time frame in which a comment or a video has been posted.

The following sections present four exploratory studies on the corpora which relate to the retrieval tasks discussed above: the filtering of low-quality comments on Slashdot (Section 3), the ranking of comments on Slashdot (Section 4), the summarization of mass opinion in YouTube comments (Section 5) and the retrieval of web items across media between Slashdot and YouTube by exploiting the associated comments (Section 6).

3. CASE STUDY: COMMENT FILTERING

Given a set of comments D, the task is to filter all comments of extremely low quality, particularly comments from spammers and vandals. The case study investigates whether comment filtering on Slashdot can be done on the basis of a writing style analysis. This analysis is interesting since existing retrieval models for this task depend primarily on user modeling [Veloso et al. 2007].

3.1. Retrieval Model

We assess a comment's quality by its readability which, in turn, depends much on its writing style. User-generated content on the web particularly lacks in this respect since users tend to use common speech, they do not revise their writing for grammatical and spelling errors and they often neglect punctuation and capitalization. On the contrary, many users seem to prefer good writing over bad writing, since comments with better style achieve consistently higher ratings on Slashdot. For this reason, as well as to ensure generalizability, our feature selection comprises features from linguistic stylometry and vandalism detection. The latter is an especially important feature class targeting low-quality and ill-intentioned comments and was proposed for use in the detection of vandalism on Wikipedia [Potthast et al. 2008].We use the following features (see Table I):

— *NLP.* The frequency of prepositions and interjections indicates common speech.

— *Style1.* The comment length indicates whether or not a commenter puts effort in her writing. This feature achieves a remarkable performance in discriminating high-quality Wikipedia articles [Blumenstock 2008].

— *Style2.* Readability formulas, such as the DC Formula [Dale and Chall 1948; Chall and Dale 1995], the FK Grade [Flesch 1948; Kincaid et al. 1975] and the GF Index [Gunning 1969] indicate the sophistication of language use:

$$\begin{aligned} \mathrm{DC} &= 0.1579 \cdot \mathrm{PDW} + 0.0496 \cdot \mathrm{ASL} + 3.6365 \\ \mathrm{FK} &= 0.39 \cdot \mathrm{ASL} + 11.8 \cdot \mathrm{ASW} - 15.59 \\ \mathrm{GF} &= (\mathrm{ASL} + 100 \cdot \mathrm{R3SW}) \cdot 0.4 \end{aligned}$$

where PDW = ratio of d's words a 4th-grader understands (based on a dictionary),
ASL = d's average sentence length,
ASW = d's average number of syllables per word, and
$\mathrm{R3SW}$ = ratio of d's words with at least 3 syllables.

— *Vand1.* The compression rate of a comment's text.

— *Vand2.* The deviation of a comment's letter frequency distribution from the expectation. This as well as the first vandalism feature indicate bad writing or non-writing (e.g. when a commenter hits the keyboard randomly).

— *Vand3.* The normalized frequency of vulgar words found in a comment.

3.2. Experiments

Based on a set D of categorized comments from the Slashdot corpus, a dichotomous classifier $c : \mathbf{D} \rightarrow \{0, 1\}$ is trained on the feature representations $\mathbf{D}$ of D. Naïve Bayes is used as the classification technology. We have also experimented with SVM classifiers but despite their otherwise good performance, Naïve Bayes could not be outperformed. The performance is measured as precision and recall with respect to each class $\in \{0, 1\}$, indicating the negative and the positive comment category on Slashdot. The training is based on a ten-fold cross-validation; Experiment 1 in Table IV shows the achieved performance results. At the bottom of the table, a baseline is given in which all comments are classified as positive.

Two issues render our classification approach particularly difficult: the class imbalance and the short length of the comments. Keeping these problems in mind, the results of the first experiment are promising but not overwhelming: only a small portion of negative comments are classified as such. Another issue which needs to be addressed in this regard is the category "funny." Funny comments are a vital part of Slashdot. We presume that neither of our features nor any of those we analyzed for our survey is capable of capturing funniness. One of the few publications targeting humor retrieval is [Mihalcea and Pulman 2007], wherein the authors try to distinguish humorous texts from other texts. But the particular case where a humorous text is a response to another not necessarily humorous text has not been studied. The Slashdot corpus appears as a valuable resource for humor retrieval.

We have conducted three additional experiments in which the "funny" category was either swapped from positive to negative, dropped or considered as a third class altogether. The results are also shown in Table IV. As is evident, when considering funny comments as negative the classification performance is significantly improved, which may be an indication that funny comments look similar to negative comments. Note that dropping funny comments results in a better performance as well. Considering funny comments as a third class does not work since they cannot be significantly separated from negative comments using our retrieval model. In [Reyes et al. 2010] we investigate this issue further.

Compared to the results of Veloso et al. [2007] who study the same classification task, we achieve a similar classification accuracy. Note however, that we employ an entirely different feature set: the retrieval model of Veloso *et al.* is specific to Slashdot since it is based primarily on a user model, whereas our model can be considered as domain-independent, more robust and applicable to anonymous comments. Finally, Experiment 5 measures the classification performance of single features. In contrast to the findings of [Blumenstock 2008], the comment length feature Style1 does not improve over the baseline, which is also the case for the vandalism feature Vand1.

Table IV. Filtering performance of the comment quality model.

Experiment	Feature(s)	Class (Categories)	Precision	Recall	F-Measure
1	All	positive	0.86	0.90	0.88
		negative	0.43	0.33	0.37
2: Swapping "funny"	All	positive exclusive funny	0.74	0.88	0.80
		negative inclusive funny	0.74	0.51	0.60
3: Dropping "funny"	All	positive exclusive funny	0.85	0.91	0.88
		negative	0.61	0.45	0.52
4: Third class "funny"	All	positive exclusive funny	0.76	0.87	0.81
		funny	0.41	0.46	0.43
		negative	0.52	0.17	0.26
5: Individual features	NLP	positive	0.82	0.98	0.89
		negative	0.35	0.06	0.10
	Style1	positive	0.81	1.00	0.90
		negative	0.00	0.00	0.00
	Style2	positive	0.84	0.94	0.89
		negative	0.43	0.19	0.26
	Vandalism1	positive	0.82	1.00	0.90
		negative	0.00	0.00	0.00
	Vandalism2	positive	0.84	0.96	0.90
		negative	0.50	0.18	0.26
	Vandalism3	positive	0.83	0.97	0.90
		negative	0.50	0.12	0.19
Baseline: All positive	_	positive	0.81	1.00	0.90
		negative	0.00	0.00	0.00

4. CASE STUDY: COMMENT RANKING

Given a document d_q and a set of comments D, the task is to rank the comments according to their novelty with respect to d_q. Ordinary novelty detection identifies sentences in a document stream which complement facts already known to the user [Soboroff and Harman 2005]. Analogously, d_q encodes a user's a-priori knowledge before exploring D. The case study investigates the applicability of novelty to comment ranking. This analysis is interesting since none of the existing approaches include d_q in their retrieval models, which may be acceptable for review ranking but not for general comment ranking.

4.1. Retrieval Model

We analyze whether the well-known maximal marginal relevance (MMR) model works for Slashdot comments. Moreover, we propose a new model, called ESA_Δ, and compare it to MRR. Both MMR and ESA_Δ are meta-models, since they employ a generic retrieval model in a sophisticated fashion in order to boost their performance. Here, the generic retrieval model is a tf-weighted vector space model, which hence is also a good baseline for comparison.

Maximal Marginal Relevance. Under the MMR model, the most relevant comment which complements a given query document d_q is computed iteratively from the comments D on d_q, based on a subset $S \subset D$ that the user already knows [Carbonell and Goldstein 1998]. In the i-th step, the comment d_i at rank i is computed as follows:

$$d_i = \arg\max_{d_x \in D \setminus S} [\lambda \cdot \varphi(\mathbf{d}_x, \mathbf{d}_q) - (1 - \lambda) \cdot \max_{d_y \in S} [\varphi(\mathbf{d}_x, \mathbf{d}_y)]],$$

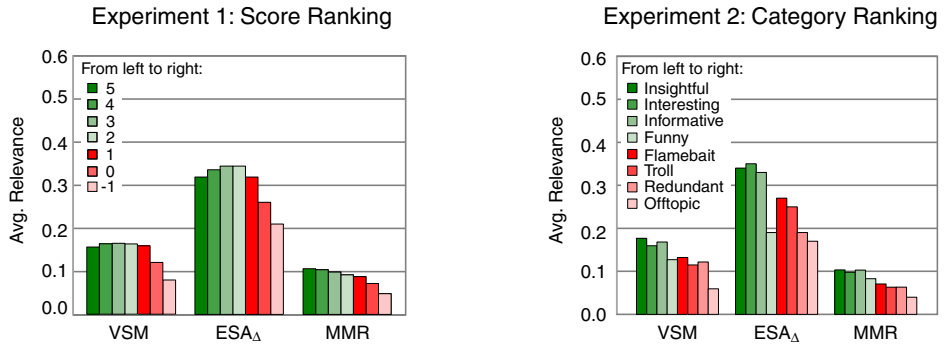

Fig. 1. Evaluation results for the ranking study: the plots show the average relevance of a comment d to d_q per score (left) and per category (right). VSM, ESA$_\Delta$, and MMR denote the vector space model, the new similarity-reduced explicit semantic analysis, and maximal marginal relevance.

where $S = \{d_q, d_1, \ldots, d_{i-1}\}$ are the a-priori known comments, φ is the cosine similarity, and λ adjusts the trade-off between d_i's similarity to d_q and the novelty of d_i in D. Initially, S contains only d_q. Note that the relevance of d_i to d_q is quantified as the value that maximizes the right-hand side of the equation. In accordance with the literature, we chose $\lambda = 0.8$.

Similarity-reduced Explicit Semantic Analysis. ESA$_\Delta$ is based on the explicit semantic analysis paradigm [Gabrilovich 2006; Gabrilovich and Markovitch 2006; 2007]. The original ESA model represents a document d as a vector $\mathbf{d}_{|D_I}$ that comprises the similarities of d to documents from a collection D_I, referred to as index collection. The similarities of d to D_I are computed using the vector space model. Each document from D_I is considered as the description of a particular concept, and documents from Wikipedia have been successfully applied in this respect. The supposed rationale of ESA is to represent d in a concept space that is defined by the index collection. Within ESA, two documents d_q and d are compared by computing the cosine similarity between the concept vectors $\mathbf{d}_{q|D_I}$ and $\mathbf{d}_{|D_I}$.

The ESA model introduces a level of indirection to the similarity computation in the form of the index collection D_I. This way, connections between d_q and d may become apparent which are not obvious when looking only at vocabulary overlap. In ESA$_\Delta$ we intend to extract exactly this portion of similarity by first reducing the overlap between d_q and d: all terms from d that appear in d_q are removed. What remains in the reduced comment $d_\Delta = d \backslash d_q$ is considered as the comment's "visible novelty." To quantify the relatedness of d_q with regard to d_Δ, the ESA vectors $\mathbf{d}_{q|D_I}$ and $\mathbf{d}_{\Delta|D_I}$ are compared.

4.2. Experiments

Two experiments were conducted on the Slashdot corpus.

Experiment 1: Score Ranking. The comment scores on Slashdot define a ranking that can be used to evaluate a retrieval model with respect to its ability to capture relevance. For the comments D of each article d_q, the average relevance value of all comments $D_i \subset D$ with score $i \in \{-1, 0, \ldots, 5\}$ to d_q was computed. The left plot in Figure 1 shows the results. The standard deviations for the models are as follows: $\sigma_{\mathrm{VSM}} = 0.13$, $\sigma_{\mathrm{ESA}_\Delta} = 0.16$, and $\sigma_{\mathrm{MMR}} = 0.08$. The vector space model and the ESA model show a comparable similarity distribution in which medium scores are ranked highest on average, while MMR places the comments in a natural order. However, the ESA$_\Delta$ model achieves significantly higher similarity values than does the vector space model, while the relevance values computed with MMR appear to be rather small. The latter is not a problem as long as the desired ordering of the comments is achieved. To determine whether this is indeed the case, we have also computed the graded relevance scores NDCG, ERR, and Kendall's τ, both on the entire rank-

Table V. Ranking performance of the comment retrieval models. Standard deviation is denoted in brackets.

Measure	Retrieval Models						Baseline	
	VSM		ESA_Δ		MMR		Random	
NDCG	0.71	(0.08)	0.71	(0.08)	0.70	(0.08)	0.70	(0.08)
ERR	0.54	(0.24)	0.54	(0.24)	0.55	(0.26)	0.53	(0.25)
Kendall's τ	0.09	(0.13)	0.08	(0.12)	0.06	(0.13)	0.01	(0.11)
NDCG@10	0.72	(0.13)	0.73	(0.13)	0.73	(0.13)	0.70	(0.12)
ERR@10	0.58	(0.24)	0.58	(0.24)	0.58	(0.26)	0.55	(0.25)
Kendall's τ@10	0.09	(0.12)	0.08	(0.12)	0.06	(0.13)	0.01	(0.11)

ings produced by the three models and restricted to the respective top 10 comments only. In addition, these scores have also been computed for a random ranking as a second baseline. Table V shows the results. As can be seen for the complete rankings, neither of the models clearly outperforms the other, and what is more, neither of the models outperforms the random baseline. On the top 10 comments, at least the latter is achieved, while the three models remain almost indistinguishable.

Experiment 2: Category Ranking. This experiment follows the design of Experiment 1, but computes the average relevance value of a comment to d_q per category. Hence, based only on positive and negative judgments, a less fine-grained ranking is demonstrated (e.g. if a user wants negative comments to be presented after the positive comment). The right plot in Figure 1 shows the results. Observe the difference in the distributions of the vector space model and MMR to the ESA_Δ model: the latter achieves significantly higher similarities for the positive categories (except for "funny") than for the negative categories, whereas the vector space model and MMR show almost no discriminative behavior.

Remarks. The results of this study should be interpreted with caution: the models induce a sensible ranking only in terms of averaged values while their graded relevance scores are inconclusive. Undesired rankings occur with a non-negligible probability. We conclude that the task of ranking comments is in its infancy and should be subject to further research.

5. CASE STUDY: COMMENT SUMMARIZATION

Given a set of comments D, the task is to generate a short text or a visualization that overviews the contents or the opinions of D: users will quickly get an idea about the comments without having to read everything. The case study investigates the use of word clouds for opinion summarization of comments on YouTube.[1] This analysis is interesting since existing comment summarization approaches rely on sentence extraction and are prevalently applied to distill customer product reviews. The respective technology cannot directly be applied to comments, which are significantly shorter than reviews—most of the comments found on media sharing sites do not even contain one sentence. It is unlikely that relevant information can be found in such comments except the opinion of the commenters. Short comments in particular are tedious to read which is why a suitable summarization for them is desired. While a single opinion may not be very useful (especially if no argument is provided), the fact that popular items inspire thousands of people to share their opinions allows us to generate a representative opinion summary.

5.1. Retrieval Model

The summarization of a comment set D divides into an offline step and an online step. Suppose that two dictionaries V^+ and V^- are given, comprising human-annotated terms that are commonly used to express positive and negative opinions respectively [Stone 1966].

[1]Our technology has been operationalized as an add-on for Firefox and Chrome which has gained considerable attention: http://www.webis.de/research/projects/opinioncloud. We have published a demo paper about the add-on at the European Conference on Information Retrieval [Potthast and Becker 2010].

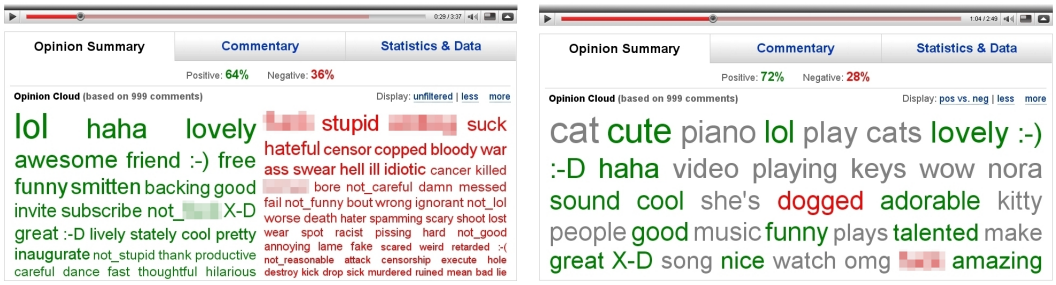

Fig. 2. Opinion summaries which contrast positive and negative words (left) and all kinds of words (right).

In the offline step, the well-known sentiment analysis approach described by [Turney and Littman 2003] is used to extend V^+ and V^- to the application domain. The extension is necessary in order to learn terms that are not covered by the dictionaries, and it is not feasible to do this manually. The semantic orientation SO of an unknown word w is measured by the degree of its association with known words from V^+ and V^-:

$$\mathrm{SO}(w) = \sum_{w^+ \in V^+} \mathrm{assoc}(w, w^+) - \sum_{w^- \in V^-} \mathrm{assoc}(w, w^-),$$

where $\mathrm{assoc}(w, w')$ maps two words to a real number that indicates their association strength. If $\mathrm{SO}(w)$ is greater than a threshold ε (less than $-\varepsilon$), w is added to V^+ (V^-); otherwise w is considered as neutral. The point-wise mutual information statistic is applied as an association measure:

$$\mathrm{assoc}(w, w') = \mathrm{PMI}(w, w') = \log_2 \frac{p(w \wedge w')}{p(w) \cdot p(w')},$$

where $p(w \wedge w')$ denotes the probability of observing w together with w', and $p(w)$ is the probability of observing w.

In the online step, when a set of comments D is observed, two summary term vectors $\mathbf{s}_D^+$ and $\mathbf{s}_D^-$ are constructed in which the absolute frequencies of the positive and negative terms occurring in D are counted, based on the dictionaries V^+ and V^-. If a word does not occur in the dictionaries, it is considered neutral.

5.2. Visualization

We visualize $\mathbf{s}_D^+$ and $\mathbf{s}_D^-$ as word clouds. Figure 2 shows examples where $\mathbf{s}_D^+$ and $\mathbf{s}_D^-$ are contrasted. Word clouds are arrangements of words in 2 or 3 dimensions in which important words are highlighted [Seifert et al. 2008]. Here, the words are colored according to their sentiment polarity, and they are scaled according to their term frequency. The font size of a word is computed as follows:

$$\mathrm{size}(w) = \mathrm{max\_size} \cdot \frac{tf(w)}{\max_{w^* \in V} (tf(w^*))},$$

where $\mathrm{max\_size}$ is a predefined maximum font size and $tf(w)$ is the term frequency of w as counted in $\mathbf{s}_D^+$ or $\mathbf{s}_D^-$, and $V = V^+ \cup V^-$. For smoothing purposes, the logarithm of the numerator and the denominator may be taken.

A few more issues need to be addressed in practice: emoticons and exclamations (such as ":-)" or "lol") require an additional set of detection rules since commenters often vary their spelling, spelling errors (which are abundant in comments) need to be corrected on the fly, and negations in front of opinion words need to be detected as a heuristic to determine the orientation of a word in context.

An important part of the visualization is the percentages of positive and negative words found on top of the word cloud (see Figure 2). For a quick overview these numbers are sufficient; however, when a user wants to know more about what other commenters thought, a click on any word in the cloud produces a list of comments containing it. As mentioned above, we have implemented a browser add-on that summarizes YouTube comments and Flickr comments on-the-fly. A lot of user feedback was obtained this way; from which it became clear that users find the summary interesting and useful, yet they criticize that sometimes words from comments are considered negative (positive) although they have been used in a positive (negative) way. For future developments, it is planned to incorporate sentiment classification of whole comments.

Finally, it is noteworthy that the word cloud shown right in Figure 2 inspired us to investigate cross-media retrieval by exploiting comments (see the next section). In this particular case, the top 3 neutral words perfectly explain what the video is all about: a cat playing the piano.

6. CASE STUDY: CROSS-MEDIA RETRIEVAL

Our final case study investigates whether a set of comments D can be used for retrieval purposes (i.e. whether the combined knowledge of D tells us something about the commented item d_q), thus allowing for the comparison of items across media. To the best of our knowledge, we are the first to analyze this possibility.[2]

Cross-media retrieval is a sub-problem of multimedia information retrieval, which again divides into various sub-tasks. Here, we consider the following task: given a set of items of different media types, the task is to pair those items which are similar with respect to their topic, regardless of their media type. A primary goal of cross-media retrieval is the construction of retrieval models that bridge the gap between different media types by means of identifying correlations between low-level features and semantic annotations. We approach this problem from a different perspective through the use of comments in lieu of the commented item. This way, model construction is not an issue since well-known text retrieval models can be directly applied. Although the text surrounding a non-textual item has always been used to extract annotations in multimedia information retrieval [Deschacht and Moens 2008; Inoue 2004; Lew et al. 2006], comments in particular have not been considered in this respect.

A premise of our approach is that comments have to describe the commented item to some extent, which is analyzed in [Potthast 2009]. In short, we found that comments on *text* are descriptive: 10 comments are sufficient to reach a considerable similarity between a text and its comments which is not rooted in duplication, while 100-500 comments contain a significant contribution of the commenters beyond the commented text. This case study proceeds in this direction.

6.1. Retrieval Model

A standard vector space model with $tf \cdot idf$ term weighting is used as our retrieval model. Given a web item d_q and its associated set of comments D, d_q is represented as a term vector $\mathbf{d}_q$ based on the index terms found in D, while applying stop word reduction and stemming. In the case that d_q is a text document, as in the Slashdot corpus, the index terms found in d_q are also included in $\mathbf{d}_q$. The representations of two items, $\mathbf{d}_q$ and $\mathbf{d}'_q$, are compared using the cosine similarity. Though nearly every retrieval model can be employed for this task, we resort to a simple vector space model in order to determine how robust a cross-media similarity assessment can be accomplished.

6.2. Experiments

Given the evaluation corpora described above, 6 000 videos from the YouTube corpus were sampled and compared to each of the 17 948 Slashdot articles. This resulted in about 107.7 million similarities being computed. Slashdot and YouTube are similar in that both are community-driven websites, so that at least some topical overlaps can be expected. However, since both corpora have

[2] We have published a poster paper about this study at the World Wide Web Conference [Potthast et al. 2010].

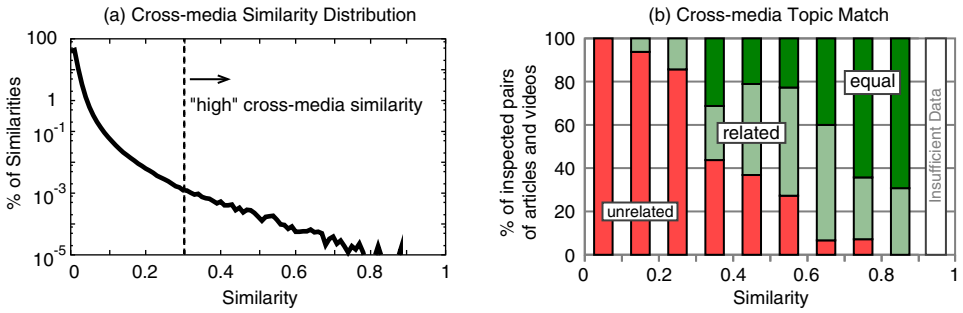

Fig. 3. (a) Distribution of cross-media similarities between YouTube videos and Slashdot articles. (b) Percentage of pairs of articles and videos with unrelated topics, related topics, and equals topics per similarity interval of range 0.1, based on a stratified sample of 150 manually inspected item pairs.

been compiled independently, we were not aware of existing overlaps. Figure 3a shows the obtained similarity distribution as a percentage of similarities over similarity intervals with an interval resolution of 0.01.

From all compared pairs of articles and videos, we have sampled a total of 150 for manual inspection by means of stratified sampling from similarity intervals of range 0.1. The sampled pairs were then classified into categories of topical match, namely pairs with equal topic, related topics, and unrelated topics. Figure 3b shows the obtained results. Topic overlaps start appearing at a similarity of 0.3, which may already be considered a "high" cross-media similarity for its considerable positive deviation from the expectation. As the comment-based cross-media similarity increases, more and more items with related or equal topics can be observed. Within high similarity ranges, pairs of articles and videos with equal topic appear most often. In addition to this manual inspection, the top 100 most similar pairs were evaluated with respect to their topical match. Table VI gives a detailed overview of these item pairs, and Table VII shows a selection of matching item pairs: 91% of the top item pairs have in fact equal or related topics. The similarity values in the table give an idea about the measured similarities and their standard deviation (stdev). Few false positives achieve high similarities, but are based on a lot more comments on the side of Slashdot. Observe that the number of comments appears to correlate with the similarity, and that more comments possibly result in a "topic drift." Often, the title of a YouTube video is descriptive, and hence we have determined the percentage of pairs where the video title overlaps with the Slashdot article. On average, this is the case in 60% of the examined item pairs, which means in turn that in this experiment 40% of the top 100 matching item pairs would not have been identified based only on their titles.

7. CONCLUSION AND FUTURE WORK

This paper studies the commentsphere from an information retrieval perspective. We present a comprehensive survey of related work and, based on this survey, identify the most relevant retrieval tasks with respect to research effort and user impact: the filtering, the ranking and the summarization of

Table VI. Overview of the top 100 inspected cross-media similarities.

Topic Match	Share	Similarity				Avg. # Comments		Title Match
		min	avg.	max	stdev	Slashdot	YouTube	
equal	36%	0.71	0.78	0.91	0.06	53	927	72%
related	55%	0.71	0.76	0.91	0.04	81	683	62%
unrelated	9%	0.72	0.78	0.87	0.05	104	872	–
Σ	100%	0.71	0.77	0.91	0.05	74	790	60%

Table VII. Selection of matching Web items found with comment-based cross-media retrieval.

Similarity	Slashdot Comments	YouTube Comments	URLs (In Acrobat the URLs are clickable) Slashdot	YouTube
0.91	83	950	http://slashdot.org/story/07/03/15/2056210	http://www.youtube.com/watch?v=RuWVMB7OxbM
0.82	69	950	http://slashdot.org/story/08/02/05/1511225	http://www.youtube.com/watch?v=Z_gKOCb4QBA
0.81	102	950	http://slashdot.org/story/08/01/02/1611240	http://www.youtube.com/watch?v=tLlHibrFATg
0.76	41	950	http://slashdot.org/story/07/10/16/1526257	http://www.youtube.com/watch?v=TluRVBhmf8w
0.74	40	950	http://slashdot.org/story/07/07/11/1246250	http://www.youtube.com/watch?v=DLxq90xmYUs
0.74	79	766	http://slashdot.org/story/07/08/13/1347253	http://www.youtube.com/watch?v=BWQ5ZMnz25I
0.74	66	78	http://slashdot.org/story/06/02/02/0024235	http://www.youtube.com/watch?v=F0uq21xjMCw
0.73	75	950	http://slashdot.org/story/08/06/04/1159207	http://www.youtube.com/watch?v=adc3MSS5Ydc

comments as well as their exploitation for the same tasks on web items. In addition, there are a number of secondary retrieval tasks that are no less exciting, including the prediction of a web item's popularity based on comments. We conducted a case study for four of the tasks mentioned above. For this purpose we compiled adequate corpora which are available to other researchers in order to foster the research activities in this field. Within the case studies, special attention was paid to the used retrieval models: our objective was to point out differences to existing retrieval models and to provide a better understanding of the challenges for retrieval tasks in the commentsphere. Moreover, we developed new retrieval model variants to address part of these challenges. Our achievements from a retrieval model perspective are:

— *Feature Overview.* We compile an overview of features used to represent comments.
— *Retrieval Models for Filtering and Ranking.* We propose a domain-independent model to filter low-quality comments which competes with other models. With ESA_Δ we present a new retrieval model to measure novelty for comment ranking.
— *Cross-media Retrieval.* We introduce a new cross-media similarity analysis paradigm.

From a research perspective, we consider the following directions as promising:

— *Retrieval Models.* Current research focuses on product and movie reviews, but web comments in general are much more diverse and require new tailored retrieval models.
— *Humor Retrieval.* If surprises are what people expect from comments, they may be funny at the same time: humor retrieval has not recognized comments as a research subject.
— *Multimedia Annotation.* If comments capture a commented item's topic well, it should be possible to extract tags and annotations for the commented item from its comments.
— *Ranking and Summarization.* Comment boards show comments in chronological order, but with increasing comment number the overview gets lost. Hence, ranking comments by their relevance and novelty, as well as summarizing comments will continue to be an important direction for research.

REFERENCES

BEINEKE, P., HASTIE, T., MANNING, C., AND VAITHYANATHAN, S. 2006. An exploration of sentiment summarization. In *Proceedings of AAAI 2003*. 12–15.

BELKIN, N., ODDY, R., AND BROOKS, H. 1982. ASK for Information Retrieval. *Journal of Documentation 33,* 2, 61–71.

BLUMENSTOCK, J. E. 2008. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. ACM, 1095–1096.

CARBONELL, J. AND GOLDSTEIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 335–336.

CHALL, J. AND DALE, E. 1995. *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books.

CUNNINGHAM, S. J. AND NICHOLS, D. M. 2008. How people find videos. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, New York, NY, USA, 201–210.

DALE, E. AND CHALL, J. 1948. A formula for predicting readability. *Educational Research Bulletin 27*.

DELORT, J.-Y. 2006. Identifying commented passages of documents using implicit hyperlinks. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM, New York, NY, USA, 89–98.

DESCHACHT, K. AND MOENS, M.-F. 2008. Finding the Best Picture: Cross-Media Retrieval of Content. In *30th European Conference on IR Research, ECIR 2008, Glasgow*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, Eds. Lecture Notes in Computer Science, vol. 4956 LNCS. Springer, Berlin Heidelberg New York, 539–546.

DIT, B., POSHYVANYK, D., AND MARCUS, A. 2008. Measuring the semantic similarity of comments in bug reports. In *Proceedings of 1st International ICPC2008 Workshop on Semantic Technologies in System Maintenance (STSM2008)*.

FLESCH, R. 1948. A new readability yardstick. *Journal of Applied Psychology 32*, 221–233.

FUJII, A. AND ISHIKAWA, T. 2006. A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. Association for Computational Linguistics, Sydney, Australia, 15–22.

GABRILOVICH, E. 2006. Phd thesis. Ph.D. thesis, Israel Institute of Technology.

GABRILOVICH, E. AND MARKOVITCH, S. 2006. Computing Semantic Relatedness of Words and Texts in Wikipedia-derived Semantic Space. Technical report cis-2006-04, Computer Science Department, Technion, Haifa, Israel.

GABRILOVICH, E. AND MARKOVITCH, S. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, M. M. Veloso, Ed. 1606–1611.

GHOSE, A. AND IPEIROTIS, P. 2007. Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality. In *Proceedings of the 2007 9th International Conference on Decision Support Systems (ICDSS 2007)*. Kolkata, India.

GUNNING, R. 1969. The fog index after twenty years. *Journal of Business Communication 6,* 2, 3–13.

HSU, C.-F., KHABIRI, E., AND CAVERLEE, J. 2009. Ranking comments on the social web. In *IEEE International Conference on Social Computing (SocialCom)*.

HU, M. AND LIU, B. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 168–177.

HU, M., SUN, A., AND LIM, E.-P. 2007. Comments-oriented blog summarization by sentence extraction. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, New York, NY, USA, 901–904.

HU, M., SUN, A., AND LIM, E.-P. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 291–298.

HUANG, S., SHEN, D., FENG, W., ZHANG, Y., AND BAUDIN, C. 2009. Discovering clues for review quality from author's behaviors on e-commerce sites. In *ICEC '09: Proceedings of the 11th International Conference on Electronic Commerce*. ACM, New York, NY, USA, 133–141.

INOUE, M. 2004. On the need for annotation-based image retrieval. In *Proceedings of the SIGIR'04 Workshop Information Retrieval in Context*. 44–46.

JAMALI, S. AND RANGWALA, H. 2009. Digging digg: Comment mining, popularity prediction, and social network analysis. Technical report gmu-cs-tr-2009-7, George Mason University, USA.

JINDAL, N. AND LIU, B. 2007. Analyzing and detecting review spam. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007*. IEEE, 547–552.

JINDAL, N. AND LIU, B. 2008. Opinion spam and analysis. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*. ACM, New York, NY, USA, 219–230.

KALTENBRUNNER, A., GÓMEZ, V., AND LÓPEZ, V. 2007. Description and prediction of slashdot activity. In *Latin American Web Conference (LA-WEB 2007)*. IEEE Computer Society, Los Alamitos, CA, USA, 57–66.

KHABIRI, E., HSU, C.-F., AND CAVERLEE, J. 2009. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *International AAAI Conference on Weblogs and Social Media*.

KIM, S.-M., PANTEL, P., CHKLOVSKI, T., AND PENNEACCHIOTTI, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sydney, Australia, 423–430.

KINCAID, J., FISHBURNE, R., ROGERS, R., AND CHISSOM, B. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75 Millington TN: Naval Technical Training US Naval Air Station.

KWON, N., SHULMAN, S. W., AND HOVY, E. 2006. Multidimensional text analysis for erulemaking. In *dg.o '06: Proceedings of the 2006 international conference on Digital government research*. ACM, New York, NY, USA, 157–166.

KWON, N., ZHOU, L., HOVY, E., AND SHULMAN, S. W. 2007. Identifying and classifying subjective claims. In *dg.o '07: Proceedings of the 8th annual international conference on Digital government research*. Digital Government Society of North America, 76–81.

LACEY, A. 2005. A simple probabilistic approach to ranking documents by sentiment. In *Proceedings of the Class of 2005 Senior Conference on Natural Language Processing*. Computer Science Department, Swarthmore College, Swarthmore, Pennsylvania, USA, 1–7.

LERMAN, K., BLAIR-GOLDENSOHN, S., AND MCDONALD, R. 2009. Sentiment summarization: evaluating and learning user preferences. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 514–522.

LERMAN, K. AND MCDONALD, R. 2009. Contrastive summarization: an experiment with consumer reviews. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Morristown, NJ, USA, 113–116.

LEW, M., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2,* 1 (Feb.), 1–19.

LIU, B., HU, M., AND CHENG, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*. ACM, New York, NY, USA, 342–351.

LIU, J., CAO, Y., LIN, C.-Y., HUANG, Y., AND ZHOU, M. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 334–342.

LIU, Y., HUANG, X., AN, A., AND YU, X. 2008. Reviews are not equally important: Predicting the helpfulness of online reviews. Technical report cse-2008-05, York University.

LU, Y., ZHAI, C., AND SUNDARESAN, N. 2009. Rated aspect summarization of short comments. In *18th International World Wide Web Conference*. 131–131.

MIHALCEA, R. AND PULMAN, S. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Springer, Mexico City, February 2007*. (2nd) best paper award!

MISHNE, G., CARMEL, D., AND LEMPEL, R. 2005. Blocking blog spam with language model disagreement. In *AIRWeb*. 1–6.

MISHNE, G. AND GLANCE, N. 2006. Leave a reply: An analysis of weblog comments. In *Third Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (WWW'06)*.

OUNIS, I., MACDONALD, C., AND SOBOROFF, I. 2008. On the TREC blog track. In *Proceedings of International Conference on Weblogs and Social Media*.

PARK, J., FUKUHARA, T., OHMUKAI, I., TAKEDA, H., AND GOO LEE, S. 2008. Web content summarization using social bookmarks: a new approach for social summarization. In *WIDM '08: Proceeding of the 10th ACM workshop on Web information and data management*. ACM, New York, NY, USA, 103–110.

PHIL WOLFF. 2004. Comment on the blog post "A vision for the next generation of blogging tools?" by David Winer. http://web.archive.org/web/20040312054524/↵ http://blogs.law.harvard.edu/bloggerCon/2004/02/24. A copy of Wolff's comment was preserved in another blog post that summarizes these comments (found at http://www.cadence90.com/wp/?p=2515); the comment says: *Permalinks in the commentsphere. Cross-posting of comments I post to my side-blog, preserving my blog as the central place to read what I write throughout the web. Notify me when someone comments in my blog. Be specific, better yet: Notify via my choice of email and IM.*

POPESCU, A.-M. AND ETZIONI, O. 2005. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 339–346.

POTTHAST, M. 2009. Measuring the Descriptiveness of Web Comments. In *32th Annual International ACM SIGIR Conference*, M. Sanderson, C. Zhai, J. Zobel, J. Allan, and J. Aslam, Eds. ACM, 724–725.

POTTHAST, M. AND BECKER, S. 2010. Opinion Summarization of Web Comments. In *Advances in Information Retrieval, Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010*, C. G. et al., Ed. Lecture Notes in Computer Science, vol. 5993. Springer, Heidelberg, 668–669.

POTTHAST, M., STEIN, B., AND BECKER, S. 2010. Towards Comment-based Cross-Media Retrieval. In *Proceedings of the 19th International Conference on World Wide Web (WWW 10)*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds. ACM, 1169–1170.

POTTHAST, M., STEIN, B., AND GERLING, R. 2008. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research (ECIR 2008)*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, Eds. Lecture Notes in Computer Science, vol. 4956 LNCS. Springer, Berlin Heidelberg New York, 663–668.

REYES, A., POTTHAST, M., ROSSO, P., AND STEIN, B. 2010. Evaluating Humor Features on Web Comments. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. European Language Resources Association (ELRA).

SCHUTH, A., MARX, M., AND DE RIJKE, M. 2007. Extracting the discussion structure in comments on news-articles. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*. ACM, New York, NY, USA, 97–104.

SCULLEY, D. 2008a. Advances in online learning-based spam filtering. Ph.D. thesis, Tufts University, USA. Carla E. Brodley.

SCULLEY, D. 2008b. On free speech and civil discourse: Filtering abuse in blog comments. In *CEAS 2008 - The Fifth Conference on Email and Anti-Spam, 21-22 August 2008, Mountain View, California, USA*.

SEIFERT, C., KUMP, B., KIENREICH, W., GRANITZER, G., AND GRANITZER, M. 2008. On the beauty and usability of tag clouds. *iv 0*, 17–25.

SOBOROFF, I. AND HARMAN, D. 2005. Novelty detection: the TREC experience. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Morristown, NJ, USA, 105–112.

STONE, P. J. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

SZABÓ, G. AND HUBERMAN, B. A. 2008. Predicting the popularity of online content. *CoRR*.

TSAGKIAS, E., DE RIJKE, M., AND WEERKAMP, W. 2009. Predicting the volume of comments on online news stories. In *ACM 18th Conference on Information and Knowledge Managment (CIKM 2009)*. ACM, ACM, Hong Kong.

TURNEY, P. AND LITTMAN, M. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst. 21,* 4, 315–346.

VELOSO, A., MEIRA, W., MACAMBIRA, T., GUEDES, D., AND ALMEIDA, H. 2007. Automatic moderation of comments in a large online journalistic environment. In *Proceedings of the 2007 International Conference on Weblogs and Social Media (ICWSM 2007)*. Boulder, Colorado, U.S.A.

YAMAMOTO, D., MASUDA, T., OHIRA, S., AND NAGAO, K. 2008. Collaborative video scene annotation based on tag cloud. In *PCM '08: Proceedings of the 9th Pacific Rim Conference on Multimedia*. Springer-Verlag, Berlin, Heidelberg, 397–406.

YANG, J.-Y., MYUNG, J., AND GOO LEE, S. 2009. The method for a summarization of product reviews using the user's opinion. *Information, Process, and Knowledge Management, International Conference on 0*, 84–89.

YANO, T., COHEN, W. W., AND SMITH, N. A. 2009. Predicting response to political blog posts with topic models. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on ZZZ*. Association for Computational Linguistics, Morristown, NJ, USA, 477–485.

YEE, W. G., YATES, A., LIU, S., AND FRIEDER, O. 2009. Are Web User Comments Useful for Search? In *Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval, co-located with ACM SIGIR 2009*, C. Lucchese, G. Skobeltsyn, and W. G. Yee, Eds. CEUR-WS, 61–68.

ZHANG, Z. AND VARADARAJAN, B. 2006. Utility scoring of product reviews. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 51–57.

ZHUANG, L., JING, F., AND ZHU, X.-Y. 2006. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, New York, NY, USA, 43–50.

Received December 2010; revised March 2011; accepted May 2011