What Users Prefer and Why: A User Study on Effective Presentation Styles of Opinion Summarization

Xiaojun Yuan, Ning Sa, Grace Begany, Huahai Yang

What Users Prefer and Why: A User Study on Effective Presentation Styles of Opinion Summarization

Xiaojun Yuan[1], Ning Sa[1], Grace Begany[1], Huahai Yang[2]

[1]College of Computing and Information, University at Albany,
State University of New York
{xyuan,nsa,gbegany}@albany.edu
[2]Juji, Inc.
hyang@juji-inc.com

Abstract. Opinion Summarization research addresses how to help people in making appropriate decisions in an effective way. This paper aims to help users in their decision-making by providing them effective opinion presentation styles. We carried out two phases of experiments to systematically compare usefulness of different types of opinion summarization techniques. In the first crowd-sourced study, we recruited 46 turkers to generate high quality summary information. This first phase generated four styles of summaries: Tag Clouds, Aspect Oriented Sentiments, Paragraph Summary and Group Sample. In the follow-up second phase, 34 participants tested the four styles in a card sorting experiment. Each participant was given 32 cards with 8 per presentation styles and completed the task of grouping the cards into five categories in terms of the usefulness of the cards. Results indicated that participants preferred Aspect Oriented Sentiments the most and Tag cloud the least. Implications and hypotheses are discussed.

Keywords: Text summarization● consumer decision making ●user studies ●user interface design

1 Introduction

The widespread use of the Internet in many aspects of human activities has resulted in an abundance of publicly-accessible opinions. People can find opinions on a variety of topics in venues such as Twitter, Weibo, forums, e-commerce sites and specialized opinion-hosting sites such as Yelp. While most of these opinions are intended to be helpful for others, the sheer quantity of them often makes most of the opinions underutilized, as the information overload overwhelms many potential consumers. For example, Amazon.com has more than 18,000 reviews for Kindle Fire, a single product alone. Summarizing these reviews in some concise form could bring enormous benefits to consumers and business alike. Not surprisingly, research on opinion summarization is gaining increased attention [14, 18, 19, 22, 24,]. However, most of the

adfa, p. 1, 2011.

research emphasizes technical advances in underlying algorithms, while paying less attention to the presentation of the results, which is the focus of this work. Correspondingly, evaluation of opinion summarization research is normally based on certain notions of precision and recall calculation commonly used in information retrieval [28] and data mining [29]. Studies have only begun to investigate the effectiveness of opinion summarization in term of usability (e.g. [31]). Such studies focus on testing the newly-proposed techniques. A systematic comparison of the usefulness of different types of opinion summarization is still lacking. This paper reports our effort in addressing this deficiency.

One major difficulty with studying the effectiveness of opinion summarization is a confounding effect between content effectiveness and presentation effectiveness. It is often not clear whether a technique's empirical superiority can be attributed to its superior text analytics quality or its effective information presentation style. We plan to isolate the two factors and focus on studying the effect of presentation styles. This goal is achieved by using human-generated summarization as the content, so as to ensure the content has consistent high quality regardless of the presentation styles. We can then vary the presentation styles of the summaries to investigate their effect on the usefulness ratings of the summaries. Any differences found between the usefulness ratings of the summaries can be safely attributed to the differences in presentation styles. We identified four types of presentation styles of opinion summarization through a crowd-sourcing study on Amazon Mechanical Turk, and then conducted a lab user-centered experiment to compare the effectiveness of the four styles.

2 Previous work

Although not abundant, studies investigating the effectiveness of opinion summarization from a perspective of both usability and user preference are emerging. Several recent studies explore feedback from users regarding their preferences for certain opinion summarization styles and approaches. Most recently, Qazi et al. [26] addressed a gap in existing studies examining the determination of useful opinion review types from customers and designers perspectives. Specifically, the researchers used the Technology Acceptance Model (TAM) as a lens to analyze users' perceptions toward different opinion review types and online review systems. The study, a pilot study, focused on three review types which are related to perceived usefulness, perceived ease of use, and behavioral intention: A (regular), B (comparative), and C (suggestive). Suggestive reviews, the speech acts which are used to direct someone to do something in the form of a suggestion, were newly identified by the researchers as a third innovative review type. To examine user perspectives, researchers used a closed card sorting approach to analyze reviews from Amazon, blogs, and a self-deployed website. The results of their work indicated that the review types play a significant role in developing user perception regarding a new product or system, with suggestive reviews more significant for both customers and designers to find more usefulness that ultimately improves their satisfaction level.

Further, in another work [31], researchers conducted a user study of a review summarization interface they created called "Review Spotlight." Review Spotlight is based on a tag cloud and uses adjective-noun word pairs to provide an overview of online restaurant reviews. Findings indicated that study participants could form detailed impressions about restaurants and make faster decisions between two options with Review Spotlight versus traditional review webpages. In a large-scale, comprehensive human evaluation of three opinion-based summarization models – Sentiment Match (SM), Sentiment Match + Aspect Coverage (SMAC), and Sentiment-Aspect Match (SAM) – Lerman, Blair-Goldensohn and McDonald [15] found that users have a strong preference for sentiment-informed summaries over simple, non-sentiment baselines. This finding reinforces the usefulness of modeling sentiments and aspects in opinion summarization. In another study, Lerman and McDonald [16] investigated contrastive versus single-product summarization of consumer electronics and found a significant improvement in the usefulness of contrastive summaries versus summaries generated by single-product opinion summarizers. To find out which visual properties influence people viewing tag clouds, Bateman, Gutwin and Nacenta [2] conducted an exploratory study that asked participants to select tags from clouds that manipulated nine visual properties (font size, tag area, number of characters, tag width, font weight, color, intensity, and number of pixels). Participants were asked to choose tags they felt were "visually important" and results were used to determine which visual properties most captured people's attention. Study results indicated that font size and font weight have stronger effects than intensity, number of characters or tag area. However, when several visual properties were manipulated at one time, no one visual property stood out among the others. Carenini, Ng and Pauls [4] also employed a user study as part of their wider comparison of a sentence extraction-based versus a language generation-based summarizer for summarizing evaluative text. In their quantitative data analysis, the researchers found that both approaches performed equally well. Qualitative data analysis also indicated that both approaches performed well, however, for different, complementary reasons. In a related work, Carenini, Ng and Pauls [5] examined the use of an interactive multimedia interface, called "Treemaps," for summarizing evaluative text of online reviews of consumer electronics. Treemaps presents the opinion summarizations as an interactive visualization along with a natural language summary. Results of their user study showed that participants were generally satisfied with the interface and found the Treemap summarization approach intuitive and informative.

In more recent work, [12] researchers presented a novel interactive visual text analytic system called, "OpinionBlocks." OpinionBlocks had two key design goals: (1) automated creation of an aspect-based visual summary to support users' real-world opinion and analysis tasks, and (2) support of user corrections of system text analytic errors to improve system quality over time. To demonstrate OpinionBlock's success in addressing the design goals, researchers employed two crowd-sourced studies on Amazon Mechanical Turk, According to their results, over 70% of users successfully accomplished non-trivial opinion analysis tasks using OpinionBlocks. Additionally, the study revealed that users are not only willing to use OpinionBlocks to correct text

classification mistakes, but that their corrections also produce high quality results. For example, study participants successfully identified numerous errors and their aggregated corrections achieved 89% accuracy.

Additionally, Duan et al. [7], introduced the opinion mining system, "VISA" (VIsual Sentiment Analysis), (derived from an earlier system called TIARA). The VISA system employs a novel sentiment data model to support finer-grained sentiment analysis, at the core of which is the "sentiment tuple," composed of four elements: feature, aspect, opinion, and polarity. Researchers conducted a user experiment to explore how efficiently people could learn to use VISA and demonstrate its effectiveness. Study results indicated that VISA performed significantly better than the two comparison tools (TripAdvisor and a text edit tool) due to its key features, namely mash-up visualizations and rich interaction features.

In the current study, investigation of user perspectives on opinion summarization styles is taken further with the evaluation and comparison of four distinct, popular summarization styles focused on textual opinions; namely, Tag Clouds, Aspect-Oriented Sentiments, Paragraph Summaries, and Group Samples.

Following we introduce the four presentation styles used in our study, the methodology, results of the experiment, discussion and conclusions.

3 Opinion summarization presentation styles

Some opinion hosting sites allow opinion writers to give numerical ratings in addition to the textual opinions. Since the visualization of numerical values is a well-studied problem, we focus instead on the summarization of textual opinions. Similarly, we do not compare visualization systems that emphasize statistics rather than the textual content of the text collections (e.g. [6]). Opinion summarizations studied here are of the kind that could potentially be used in place of the full documents.

Based on our survey of the literature, we have categorized the presentation styles of such opinion summarization into four major types.

3.1 Tag clouds (TAG)

Tag clouds are perhaps the most popular form of summarization on the Internet today [3]. This type of text presentation has also been used extensively in research (e.g. [25, 28]). They consist of enumerations of the most common words or phrases in a collection of opinions, laid out on a 2D canvas. The size of the words or phrases often indicates how frequently they were mentioned. The larger the word or phrase, the more frequently it received mentions. The effect of various visual features of the tag clouds on their effectiveness have been investigated [2], but the comparison with other styles of summarization has yet to be done. See Figure 1.

Fig. 1. Tag clouds

3.2 Aspect oriented sentiments (ASP)

Aspect oriented sentiment summarization is an active area of research in text mining [11, 13, 14, 19, 23]. In this approach, some important aspects or topics (also known as features) of opinions are extracted from an opinion collection. Sentiment orientation of the text snippets containing the aspects are then estimated and summary statistics reported. A typical summarization for one aspect might look like this: *for a collection of reviews on Kindle Fire, "screen" is identified as an aspect, and 100 text snippets in the collection are found to be about this aspect, 60 of them have positive sentiment, 30 of them are negative, and the rest are neutral*. Representative text snippets for each sentiment orientation may also be provided. See Figure 2.

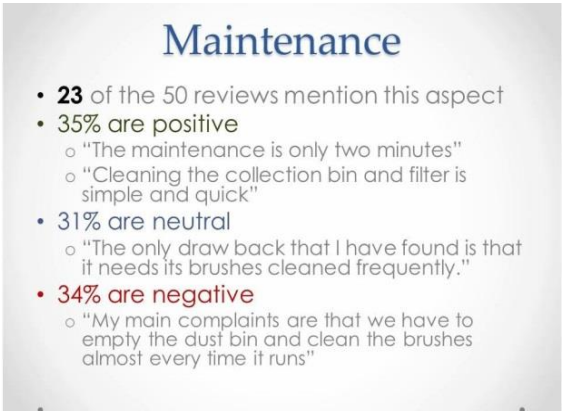

Fig. 2. Aspect Oriented Sentiments

3.3 Paragraph summaries (PRG)

Automatic text summarization systems traditionally produce short passages of text as summary [10, 27]. The summarization is called extractive when the sentences are selected from original documents; abstractive when the sentences are generated by the system [8]. Regardless of the approach, the output could be a readable abstract that resembles what humans would write for generic purposes in order to emphasize intrinsic properties such as fluency and coverage [21]. See Figure 3.

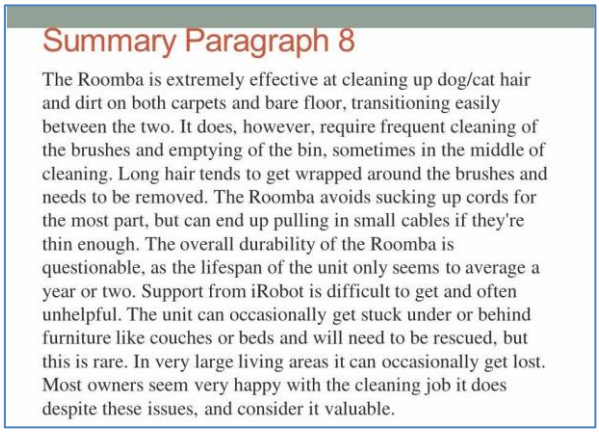

Fig. 3. Paragraph summary

3.4 Group samples (GRP)

Clustering algorithms are typically used in post-query user interfaces as a way of organizing retrieved documents [9]. It has also been used in summarization, where similar documents are grouped together and the representatives of the groups are displayed [1, 20]. This approach has shown to be effective in interactive information retrieval [17, 30]. See Figure 4.

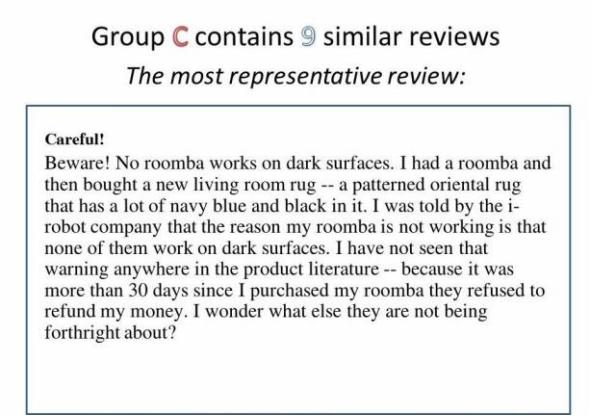

Fig. 4. Group samples

3.5 Summary of the four presentation styles

Tag clouds may perform most differently from the other three presentation styles. Because Tag clouds do not show full sentences, they lack the context needed to make accurate judgments about the value of the information. On the other hand, Tag clouds have the best support for fast perception as they package the most important words or phrases efficiently in space, and visually highlight their relative importance. Aspect oriented sentiments are similar to Tag clouds in that they lack a prose structure. On the other hand, they are also similar to Paragraph summaries because they include text snippets that are reader friendly and provide the context missing in Tag clouds. Paragraph summaries are close to Aspect oriented sentiments because they may cover similar amounts of information, as paragraphs often list pros and cons in a form that resembles aspect oriented sentiments. However, the prose structure in a well-written paragraph summary affords deep analysis and accurate assessment of the context. Group samples are similar to paragraph summaries in form. However, unlike paragraph summaries that are written anew, group samples are directly drawn from the original document collection, and retain the most amount of contextual information.

4 Research questions

We hypothesize that humans respond differently to different presentation styles of opinion summaries, and some styles would be more effective in terms of human acceptance.

In this two-phase study, we are interested in investigating the following research questions:

1. Will users prefer (or not prefer) a particular opinion summarization style in making judgments about product reviews?

2. What are the reasons that users may prefer (or not to prefer) a particular opinion summarization style in making judgments about product reviews?

4.1 Phase I: crowd sourcing opinion summarization

As mentioned earlier, in order to study the effect of presentation styles alone, we want to ensure the consistently high quality of the summaries. We achieved the goal through leveraging the wisdom of the crowd. Essentially, we elicited four styles of the opinion summaries with the help of Amazon's Mechanical Turk.

Procedure. First, we collected the top 50 reviews for one model of the iRoomba cleaning robot from Amazon. We chose this collection of opinion text because of the relative novelty of the product and the ability for the general public to relate to it. Using a within-subject design, we recruited 46 turkers located in the USA to answer a survey we developed to gather information for generating the opinion summaries from the text collection. In the survey, turkers were first directed to the raw text of the 50 reviews, and asked to read them in full. Then, questions about the reviews were asked. These questions were directly mapped from the information need for the four presentation styles of summaries. All questions were mandatory and were individually validated to ensure the quality of the answers. On average, turkers spent 56.3 minutes on the survey, and each was paid 4 US dollars.

Generating opinion summaries using Turkers. For Tag clouds, turkers were asked to list five short phrases to characterize the cleaning robot. They were also asked to estimate what percentage of the 50 reviews had opinions consistent with each phrase. The phrases turkers came up with were remarkably consistent and converged to 38 phrases (phrases with minor variations were grouped as one). All 38 phrases were used in the subsequent lab study. The average percentages of turkers' estimations were used in the subsequent lab study to determine the font size of the phrases. A total of 8 Tag clouds were drawn by hand, with each cloud containing 4 or 5 phrases.

Turkers were asked to list three important aspects of the product according to the reviews they read. For each aspect, they were asked to give an estimate of how many reviews mentioned the aspect, as well as the estimated percentage for positive, neutral and negative sentiment towards the aspect. The top 8 most-frequently listed aspects were used in the subsequent study. Again, the averages of the estimations were used in the display of the aspects.

Each turker was also asked to write a summary of all the reviews so that "consumers who read your summary can make an informed decision on the product, without having to read the reviews themselves". Among the 46 summaries, the top 8 most readable summaries, as agreed by two judges, were used in the lab study.

We asked turkers to identify similar reviews and group them together. They were required to list 3 groups of similar reviews.

When two reviews appeared in the same group once, their similarity measure increase by one. This way, we were able to generate a similarity matrix among the 50 reviews.

Using the matrix as input, we used a hierarchical clustering algorithm to cluster the 50 reviews. Four clusters produced the optimal fit, and the four cluster prototypes were used in the lab study as group samples. In addition, for each cluster, the closest summary to the prototype was also selected, so that there were 8 group samples in total.

4.2 Phase II: comparing presentation styles of summary using card sorting technique

The goal of this phase was to compare the four presentation styles of opinion summary in a consumer decision making context to respond to the two research questions.

Experiment design. The comparison of the opinion summaries was conducted as a lab card-sorting task. For each of the four presentation styles (experiment conditions), eight opinion summaries were prepared according to the procedure described in the previous section.

Each opinion summary was put on a single image of 960 x 720 pixels. Figure 1, Figure 2, Figure 3 and Figure 4 show a sample display for each of the conditions. In total, 4*8=32 image items were placed in the preparation bin in a random order.

The lab experiment was a within-subject design. Each participant's task was to take all 32 image items and place each of them in one of the five category bins. These category bins were defined as "Not at all useful", "Somewhat useful", "Useful", "Very useful" and "Extremely useful". Participants were told to ignore the card order within each category box. Essentially, we asked participants to give a usefulness rating for each opinion summary. We use such a card-sorting setting in order to record participants' thought process, as they were asked to think-aloud while placing the cards.

Participants. Thirty-four participants were recruited from University at Albany, half of them were males. All of the participants stated that they read online reviews regularly for making purchase decisions.

Procedure. Each participant was tested individually in a human computer interaction lab in a University campus in the USA. The subjects first filled out a consent form. Next, the subjects completed an entry questionnaire. The participants were then directed to http://websort.net/ to do the card sorting. After they completed the experiment, they were asked to answer several questions regarding the four presentation styles and their thoughts about the experiment. The whole experimental process was logged by Techsmith Morae 3 software.

5 Content analysis scheme

To address the earlier mentioned research question(What are the reasons that users may prefer (or not prefer) a particular opinion summarization style in making judgments about product reviews?) we employed a qualitative content analysis using an open coding approach to analyze the exit interview data of the Phase II experiment. The content analysis began with a comprehensive read-through and evaluation of all 34 participants' interview transcripts by each of the first three authors, the primary investigator and two doctoral students. In terms of the initial review and several discussions between the authors, a number of themes emerged from the interviews that pertained to the reasons of preference towards the presentation styles. Themes included: Comprehensiveness (Comprehensiveness of Information), Time (Time required to read the summary), Organization/categorization (Organization/categorization of the summary's content), Length/Amount (Length/amount of information), Appearance (Appearance of summary content), and Ease of use (Ease of use of summarization style). A coding scheme was designed according to these themes and is shown in the Table 1.

Table 1. Coding scheme

Code	Reasons
CMPP CMPNCMPU	Comprehensiveness (Comprehensiveness of Information) **(CMP)**
TMEP TMEN TMEU	Time (Time required to read the summary) **(TME)**
ORGP ORGN ORGU	Organization/categorization (Organization/categorization of the summary's content) **(ORG)**
LENP LENN LENU	Length/Amount (Length/amount of information) **(LEN)**
APPP APPN APPU	Appearance (Appearance of summary content) **(APP)**
E2UP E2UN E2UU	Ease of use (Ease of use of summarization style) **(E2U)**

The unit of analysis for the open coding was the individual interview document. Each of the 34 interview documents were independently coded by the three researchers and data collected in an Excel spreadsheet. The average pairwise percent agreement among the 3 coders is 0.81. Along with each code, snippets of supporting text were extracted from the interview data.

6 Results

6.1 User perception of the presentation styles

As initially noted, the goal of this paper is to better understand the relationship between opinion summarization styles and user decision judgment and the underlying reasons. The results are aligned with this goal and are based on a qualitative content analysis of the participant exit interview data.

In the exit interview, the participants answered questions relevant to the four presentation styles, including, (1) most helpful/least helpful, (2) easiest/most difficult, (3) fastest/slowest, (4) most informative/least informative, (5) sacrifice/keep the most, and (6) like/dislike. The six questions were the basis for the measurement of user perception of the four presentation styles. The results are displayed in Figure 5.

Participant responses also included their opinions on the usefulness of the summary in general and strategies they used to make the decision, as well as suggestions on system features they best liked, disliked, and thought could be added in the future.

Although "think-aloud" data and additional computer log data was collected using the Morae software, the current paper focuses exclusively on results from the exit interview data.

As can be seen from Figure 5, participants felt that paragraph summary was the most helpful, and the most informative presentation style while it was also the most difficult and the slowest one. Tag clouds were reported to be the easiest and the fastest to use, but accordingly they were the least informative and the least helpful style and the one disliked by most of the participants. On the other hand, though Aspect Oriented Sentiments didn't get the most votes, they were found to be generally helpful, easy to use, fast, and informative and the participants liked them the most. Group samples were relatively less helpful, more difficult, and slower.

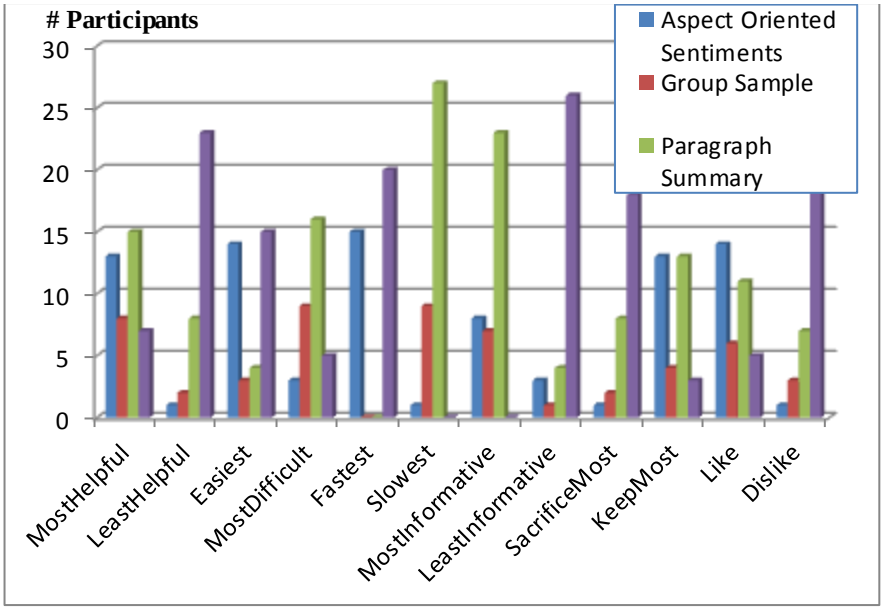

Fig. 5. User perception of the four presentation styles (y axis is the number of participants who voted for the style)

6.2 Users' opinion on the presentation styles

We were interested in finding out the major reasons influencing users' preference of the various opinion summarization styles. It appears that "Comprehensiveness", "Organization/Categorization" and "Ease of use" are equally important key reasons affecting users' preference of the presentation styles. The "Appearance", "Time", and "Length/Amount" were found to be less important reasons to the participants. Table 2 shows the distribution of reasons coded across participants. To generate the table, each unique code instance was counted for each participant.

We further investigated the distribution of the identified reasons across the four presentation styles. Table 3 shows the top 5 reasons for each style. Tag clouds received positive comments on "Appearance" and "Time" and a nearly equal amount of positive and negative comments on "ease of use." However, most of the participants agreed that they were negative regarding "comprehensiveness." This finding can explain the above-mentioned finding that participants disliked tag clouds the most. As mentioned by participants, "They don't, they don't have details." All of the top 5 reasons related to aspect oriented sentiments were positive and covered all of the main reason categories except "time." The participants liked them because they "contain negative and neutral and the positive opinions," were "very convenient or easy to read," and "very clear, brief," among other reasons. This finding correlates with the findings in Figure 5 and also explains why the participants liked aspect oriented sentiments the most. Most participants found the paragraph summary good in terms of

"comprehensiveness" and some liked its organization and found the formatting was "…what's most normal for me." But on the other hand, the paragraph summary received negative comments regarding "time" and "length." The participants found it "too long to read" and they needed to "Spend time reading it." Compared with the previous three styles, the group sample received far fewer comments. Some of the participants mentioned that it was good in terms of "comprehensiveness" and "organization," but, some didn't like it because of the "length," "ease of use," and "organization." Typical user comments can be found in Table 3.

Unsurprisingly, in both paragraph summary and the group sample styles, comprehensiveness was the predominant reason in users' preference decision. Specifically, participants claimed that the paragraph summary is "getting a lot of information in a fairly simple package" and the group sample helps them "imagine what if I had that product."

For tag clouds, though the participants liked them because they were "much faster" and "the font size was there for the words," they agreed that "They don't, they don't give details" and were negative regarding "comprehensiveness." On the contrary, the paragraph summary was long and time-consuming, but most of the participants found it provided "a lot of information in a fairly simple package" and was positive in its "comprehensiveness." Overall, aspect oriented sentiments were the best among the four styles. They were brief and, at the same time, comprehensive. The participants found them easy to use and liked their appearance and organization.

Table 2. Major reasons affecting user preference of the four presentation styles

Code	Reasons	No. Responses
CMPP; CMPN;CMPU	CMP	33
ORGP; ORGN; ORGU	ORG	32
E2UP; E2UN; E2UU	E2U	30
APPP; APPN; APPU	APP	27
TMEP; TMEN; TMEU	TME	26
LENP; LENN; LENU	LEN	26

Table 3. Reasons affecting user preference per presentation style

Styles	Code	Reasons	No.	User Comments
TAG	CMPN	CMP	27	They don't, they don't give details.
	TMEP	TME	14	It's very fast; the tag clouds was much faster
	APPP	APP	13	I can know how many people like it, because the font size was there for the words.
	E2UN	E2U	12	It was like I couldn't do anything with it. It was like it didn't seem to drive the decision, it wouldn't drive my decision making.
	E2UP	E2U	11	We can construe what the product was like in very short time.
ASP	ORGP	ORG	23	It contains negative and neutral and the positive opinions.
	APPP	APP	15	Because color coding and numbers
	E2UP	E2U	15	They're very convenient or easy to read
	CMPP	CMP	10	Percentages mean a lot for people to review all these products, and there's examples, a lot of examples on it.
	LENP	LEN	7	It was very clear, brief… concise information
PRG	CMPP	CMP	24	That's getting a lot of information in a fairly simple package
	TMEN	TME	20	Spend time reading it.
	LENN	LEN	16	Too long to read.
	ORGP	ORG	12	I enjoy the formatting and that's what's most normal for me
	E2UN & ORGN	E2U	10	… it's a lot of effort for little result sometimes. It's a tossup; It's because of so much text and even if you read it, it is not organized sometimes…
GRP	CMPP	CMP	10	Yeah the personal story with the rich information helps me to imagine what if I had that product.
	ORGP	ORG	8	I'm seeing the whole review versus just one person's interpretation of all of the reviews.
	LENN	LEN	7	In those group samples, that I think was too large is the content was too large.
	E2UN	E2U	5	because they tended to repeat the same information
	ORGN	ORG	5	It was not categorized properly. And if it isn't organized properly, it's confusing.

7 Discussion

In this paper, we were interested in discovering users' preferences and the reasons affecting their preferences of the representation styles in an opinion summarization card-sorting task.

Results demonstrate that: (1) Aspect oriented sentiments are the most preferred presentation style; (2) comprehensiveness, time, organization/categorization, length/amount, appearance, and ease of use are the major reasons impacting users' preferences of a presentation style in making decisions in a product review task.

Our results supported the finding reported in [3] in that our participants disliked the tag clouds the most. As [3] pointed out, "tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article" In our study, the participants acknowledged that tag clouds were the easiest to use because "We can construe what the product was like in very short time,," and the fastest to use because "It's very fast; the tag clouds was much faster." But, in making the decision on the usefulness of the presentation styles, their first priority was the comprehensiveness of the information in the summary presentation. As mentioned by participants, they understood people liked the tag clouds because of the "font size" and "color," but, they disliked them because they don't "give details" and "drive my decision making." This finding raised an important issue here for the design of information systems: How can the user interface balance the need for comprehensiveness of information and the need to provide key features enabling users to quickly grasp the desired information? On the other hand, [3] reported that, compared with human-assigned tags, automated tagging produces "more focused, topical clusters." The tag clouds in our study were generated by the turkers. As a result, the comprehensiveness of automatically generated tags might be our future research direction.

It was interesting to learn that participants liked the aspect oriented sentiments the most. Organization/ categorization is the most critical reason in users' decision making relevant to this presentation style. Most importantly, they liked them because they contain "negative and neutral, and the positive opinions," "color," "number" and "percentages."

Our results indicated that there may be a relationship between consumers' information needs and their preference of an opinion summary presentation style. With regards to the opinion summary of a cleaning robot, it is within expectation that consumers may want to look for information about system usability, performance, and reliability. This factor may contribute to the finding that participants prefer aspect oriented sentiments, not tag clouds.

It can also be noticed that there exist biases potentially introduced by manually generated summary presentations. Our summarizations were generated by human turkers, but, this generation could have been influenced by the instructions and contents distributed by the researchers.

Results of this study have practical implications for developers of text summarization. A few design considerations for improving usability and user experiences emerged based on participant responses and our observations. First, a deeper understanding of users' information behavior and their information needs in using information systems supporting consumer decision-making is important. In this study, we made a step towards this understanding in using turkers to generate the summary reviews. Second, after having identified key features in the consumer decision-making system, a good design should well balance the number of the features and the amount of information provided in the interface. Third, an appropriate and comprehensive organization/categorization scheme should be selected in terms of targeted user groups and task and design considerations. Many participants expressed their opinions about the importance of organization/categorization. We feel it should be given greater attention in the design process in the future experiments.

8 Conclusion

This paper reports a study comparing the effectiveness of four major styles of opinion summarization in a consumer purchase decision scenario. The study leverages the power of the crowd to bypass issues of text mining quality in order to reach more meaningful conclusions.

Our next step is to design and implement an experimental system based on the findings of this study. Such an experimental system will provide customers with a better view in the system interface. Additionally, the experimental system will be compared with our baseline system in a user-centered lab experiment to test its effectiveness and efficiency. Our goal is to contribute to improving the user experience and usability of information systems that support consumer decision-making.

As a lab-based user-centered study, limitations exist. In this experiment, the generalizability of the findings was restricted by the limited types of tasks, the number of topics, and the sample pool. Additionally, the coding scheme we generated is a simple, initial one. Deeper, more fine-tuned coding and analysis could be applied to the data in a subsequent analysis. Despite the limitations, the results of this type of research will have implications for the design of information systems that support consumer decision-making.

References

1. Ando, R., Boguraev, B., Byrd, R., and Neff, M. Multi-document summarization by visualizing topical content. In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization (2000), 79–98.
2. Bateman, S., Gutwin, C., and Nacenta, M. Seeing things in the clouds: the effect of visual features on tag cloud selections. In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, ACM (2008), 193–202.

3. Brooks, C., and Montanez, N. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In Proceedings of the 15th international conference on World Wide Web (2006), 625–632.
4. Carenini, G., Ng, R., and Pauls, A. Multi-document summarization of evaluative text. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL) (2006), 305–312.
5. Carenini, G., Ng, R., and Pauls, A. Interactive multimedia summaries of evaluative text. In Proceedings of IUI '06 of the Association for Computing Machinery (2006), 1-8.
6. Chen, C., Ibekwe-SanJuan, F., SanJuan, E., and Weaver, C. Visual analysis of conflicting opinions. In Visual Analytics Science And Technology, 2006 IEEE Symposium On (2006), 59–66.
7. Duan, D., Qian, W., Pan, S., Shi, L., and Lin, C. VISA: A Visual sentiment analysis system. In Proceedings of VINCI '12, Hangzhou, China (2012).
8. Erkan, G., and Radev, D. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. (JAIR) 22 (2004), 457–479.
9. Hearst, M., and Pedersen, J. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (1996), 76–84.
10. Hovy, E., and Lin, C. Automated text summarization and the summarist system. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (1998), 197–214.
11. Hu, M., and Liu, B. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004), 168–177.
12. Hu, M., Yang, H., Zhou, M.X., Gou, L., Li, Y., and Haber, E. OpinionBlocks: A crowd-powered, self-improving interactive visual analytic system for understanding opinion text. In Human-Computer Interaction – INTERACT 2013 (2013), 116–134.
13. Jo, Y., and Oh, A. Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining (2011), 815–824.
14. Ku, L., Liang, Y., and Chen, H. Opinion extraction, summarization and tracking in news and blog corpora. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs (2006).
15. Lerman, K., Blair-Goldensohn, S., and McDonald, R. Sentiment summarization: Evaluating and learning user preferences. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (2009), 514–522.
16. Lerman, K., and McDonald, R. Contrastive summarization: An experiment with consumer reviews. In Proceedings of NAACL HLT of the Association for Computational Linguistics (2009), 113-116.
17. Leuski, A. Evaluating document clustering for interactive information retrieval. In Proceedings of the tenth international conference on Information and knowledge management (2001), 33–40.
18. Lu, Y., and Zhai, C. Opinion integration through semi-supervised topic modeling. In Proceedings of the 17th international conference on World Wide Web (2008), 121–130.
19. Lu, Y., Zhai, C., and Sundaresan, N. Rated aspect summarization of short comments. In Proceedings of the 18th international conference on World wide web (2009), 131–140.
20. Ma˜na-L´opez, M., De Buenaga, M., and G´omez-Hidalgo, J. Multidocument summarization: An added value to clustering in interactive retrieval. ACM Transactions on Information Systems (TOIS) 22, 2 (2004), 215–241.

21. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. The tipster summac text summarization evaluation. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (1999), 77–85.
22. Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web (2007), 171–180.
23. Mukherjee, A., and Liu, B. aspect extraction through semi-supervised modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2012), 339348.
24. Pang, B., and Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1-2 (2008), 1–135.
25. Potthast, M., and Becker, S. Opinion summarization of web comments. Advances in Information Retrieval (2010), 668–669.
26. Qazi, A., Raj, R.G., Tahir, M., Waheed, M., Khan, S.U.R, and Abraham, A. A preliminary investigation of user perception and behavioral intention for different review types: Customers and designers perspective. The Scientific World Journal, (2014), 1–8.
27. Radev, D., and McKeown, K. Generating natural language summaries from multiple online sources. Computational Linguistics 24, 3 (1998), 470–500.
28. Salton, G., and McGill, M. Introduction to modern information retrieval. McGraw-Hill, Inc., 1986.
29. Witten, I., Frank, E., and Hall, M. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2011.
30. Wu, M., Fuller, M., and Wilkinson, R. Using clustering and classification approaches in interactive retrieval. Information Processing & Management 37, 3 (2001), 459–484.
31. Yatani, K., Novati, M., Trusty, A., and Truong, K. N. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In Proceedings of the 2011 annual conference on Human factors in computing systems, CHI '11 (2011), 1541–1550.