Visual Opinion Analysis of Customer Feedback Data

Daniela Oelke², Ming Hao[1], Christian Rohrdantz², Daniel A. Keim², Umeshwar Dayal[1], Lars-Erik Haug[1], Halldór Janetzko²

[1]Hewlett Packard Laboratories, Palo Alto, CA ²University of Konstanz, Germany

ABSTRACT

Today, online stores collect a lot of customer feedback in the form of surveys, reviews, and comments. This feedback is categorized and in some cases responded to, but in general it is underutilized – even though customer satisfaction is essential to the success of their business. In this paper, we introduce several new techniques to interactively analyze customer comments and ratings to determine the positive and negative opinions expressed by the customers. First, we introduce a new discrimination-based technique to automatically extract the terms that are the subject of the positive or negative opinion (such as price or customer service) and that are frequently commented on. Second, we derive a Reverse-Distance-Weighting method to map the attributes to the related positive and negative opinions in the text. Third, the resulting high-dimensional feature vectors are visualized in a new summary representation that provides a quick overview. We also cluster the reviews according to the similarity of the comments. Special thumbnails are used to provide insight into the composition of the clusters and their relationship. In addition, an interactive circular correlation map is provided to allow analysts to detect the relationships of the comments to other important attributes and the scores. We have applied these techniques to customer comments from real-world online stores and product reviews from web sites to identify the strength and problems of different products and services, and show the potential of our technique.

KEYWORDS: Visual Opinion Analysis, Visual Sentiment Analysis, Visual Document Analysis, Attribute Extraction

INDEX TERMS: I.7.5 [Document and Text Processing]: Document Capture - Document Analysis; I.5.2 [Pattern Recognition]: Design Methodology - Feature evaluation and selection

1 INTRODUCTION

1.1 Motivation

With the rapid growth of Internet technologies, there are large numbers of customer reviews on the websites. [9] reports that "81% of Internet users (or 60% of Americans) have done online research on a product at least once". Furthermore, they state that customers are willing to invest significantly more for a 5-star-rated product than a 4-star-rated product. Therefore, reviews can have a large impact on the profit margin that a company is able to realize with a specific product. While the internet users were generally satisfied with their online product research, "at the same time, 58% also report that online information was missing, impossible to find, confusing, and/or overwhelming". [9]

Over the years, manufacturers and retailers alike have collected vast amounts of reviews from their customers. This feedback is a valuable source of information for a company to improve the quality of the products, correct service failures, and guide their customers. On the other hand, as the above cited surveys show, the customers themselves are also interested in this source of information to find the product that fits best to their needs. However, the feedback is unstructured by nature; there are too many customer comments to read them all sequentially. It is time-consuming to uncover the "golden nuggets". As a result, actionable insight is not readily available and much of the feedback is ignored.

Often customers are asked to give a total score (see e.g. the webpage of amazon.com). Yet, this score does not necessarily reveal the product's true quality and may provide misleading recommendations. An attribute of a product that was important for customer A and thus had an important impact on the total score that this customer gave might be irrelevant for customer B. Thus, the latter does not mind if this feature is not available in the product or is deficient. Similarly, it is not enough for a company to know which of their products customers liked best or least. In order to learn from the feedback and be able to improve the products they need to know which attributes of the product their customers liked and disliked. Also, for marketing purposes it is interesting to see which subgroups of customers with similar opinions exist.

1.2 Our goals and Contributions

In this paper we present an approach to automatically analyze large volumes of customer reviews with respect to what was commented on positively or negatively. To achieve our goal, we developed a novel discrimination-based technique that detects automatically which product attributes were frequently commented on. We then analyze for each attribute if it was mentioned positively or negatively. In contrast to other approaches, we determine an overall score for each attribute and each review since we are not only interested in a list of sentences that comment on a specific attribute positively or negatively but want to know the specific opinion of the customer.

Furthermore, we present several novel visualization techniques that help the analyst to make use of the resulting structured but still large amount of data. Our summary report visualization provides a quick overview. In contrast to other visual summarization techniques it is scalable both with respect to the number of attributes and the number of products that are compared. In addition, we also developed a technique to analyze clusters of similar opinions in the data set. Key to the detection of meaningful clusters is our distance function that is able to group reviews with similar patterns together. The clusters themselves are visualized with special thumbnails that show what the reviews in the particular cluster have in common. Finally, our interactive circular correlation map allows analysts to detect relationships between the attributes and the user given scores.

The paper is structured as follows: In section 2 we give an overview of the whole visual analytics process. Section 3 introduces the automatic techniques that are necessary to extract

e-mail: {oelke, rohrdant, keim, janetzko}@inf.uni-konstanz.de
e-mail: {lars-erik.haug, umeshwar.dayal, ming.hao}@hp.com

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

the opinions from reviews. In section 4, we present our three visualization techniques. This section also includes an application part in which real-world data is used to demonstrate the effectiveness of our technique. An evaluation of the attribute extraction step and an analysis of the strengths and weaknesses of our approach are then presented in section 5. We end with a review of related work in section 6 and the conclusions.

2 Overview of the Process

A schematic overview of our approach is given in Figure 1. The illustrated process can be subdivided into two main tasks: (a) The fully automatic analysis of the opinions mentioned in the reviews and (b) the visual analysis of the feature vectors that were generated based on these automatically extracted opinions.

The process is designed as a pipeline. This allows us to easily exchange single steps of the complex process. Task (a) is done in three steps: the extraction of the attributes that have been frequently commented on, the detection of opinion signal words, and finally the mapping of the attributes and opinions to find out which opinion words refer to a specific attribute. Note that with “attributes” in this case we refer to the components or aspects of a specific product or type of product that are crucial when customers judge it. At the end of task (a) a feature vector is generated for each review which has one entry per attribute that can either be larger than 0 (if the attribute has been positively commented on), smaller than 0 (if it was negatively mentioned) or 0 (if no opinion was expressed on the attribute in this review). Task (b) deals with the analysis of the extracted feature vectors. As we will show, a combination of visual and automatic techniques is key to finding trends and patterns in the data.

3 Automatic Extraction of Attributes and Opinions

Given a collection of reviews our technique allows to automatically create one feature vector per review that reflects the opinion given in the review in a detailed, differentiated and comparable way. The opinion of any significant product attribute is represented by one feature dimension.

The approach includes some data preprocessing steps outlined in section 3.1 and a novel technique for attribute extraction described in section 3.2. Then, section 3.3 deals with the detection of opinion signal word and section 3.4 describes the mapping between attributes and opinion signal words.

3.1 Data Preprocessing

The attributes have to be extracted from plain text natural language reviews. As a preprocessing step we apply a base form reduction algorithm to all words in order to get e.g. singular forms for nouns and infinitive forms for verbs. In addition we use a sentence splitter and POS-tagger ([13]) as well as a NP-chunker ([11]) in order to identify noun phrases. This allows us to consider noun phrases that consist of several consecutive words (tokens). Numbers and short strings with less than 3 characters are deleted in the preprocessing step since they often correspond to punctuation marks or special characters that do not need to be considered.

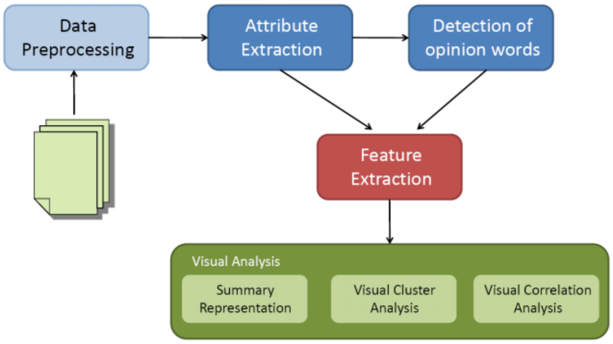
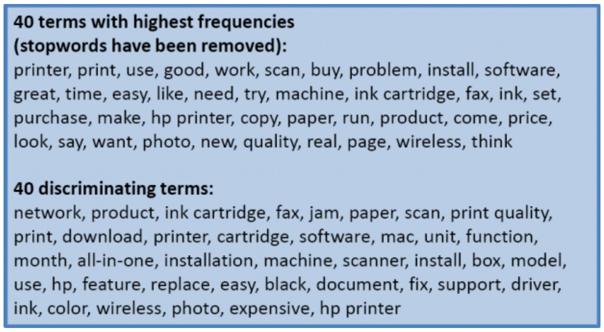


Figure 1. The consecutive steps of our visual opinion analysis process.



40 terms with highest frequencies (stopwords have been removed):
printer, print, use, good, work, scan, buy, problem, install, software, great, time, easy, like, need, try, machine, ink cartridge, fax, ink, set, purchase, make, hp printer, copy, paper, run, product, come, price, look, say, want, photo, new, quality, real, page, wireless, think

40 discriminating terms:
network, product, ink cartridge, fax, jam, paper, scan, print quality, print, download, printer, cartridge, software, mac, unit, function, month, all-in-one, installation, machine, scanner, install, box, model, use, hp, feature, replace, easy, black, document, fix, support, driver, ink, color, wireless, photo, expensive, hp printer

Figure 2. 40 most frequent terms (top) compared to the Top-40 discriminating terms. It can easily be seen that the list of discriminating terms contains more terms which are relevant with respect to the question what the customers frequently comment on whereas the list of the most frequent terms also contains many terms that are typically used in reviews but do not convey the desired information (e.g. need, like, good, etc).

3.2 Attribute Extraction

The first step of our opinion analysis method is the extraction of attributes. With attributes we refer to certain characteristics of the entity of interest which are frequently mentioned when this entity is evaluated. For a product those attributes may be components, properties or parameters that are important for customers when evaluating it. In our scenario, it is important to automatically detect a list of all relevant product attributes, since it allows us to analyze customer reviews not only with respect to their general evaluation, but look in detail at the particular attributes that customers were satisfied with or complained about.

A straightforward way to automatically extract these attributes out of textual data sources (such as reviews) would be to take the most frequent words and filter out stop words according to a given stop word list. For our printer reviews from amazon.com this results in the list of the 40 most frequent terms that is shown in the upper part of Figure 2. The problem that comes along with this approach is that not only words describing product attributes like “print” or “software” are frequent but also typical review terms like “great”, “like” or “need”.

Widely used stop word lists contain only very general terms like conjunctions, determiners, pronouns etc. and thus are not suitable to separate the printer terms from the rest. We have to apply a special term filtering that extracts the printer terms while it does not consider the review terms.

For this purpose, we developed a novel discrimination-based term extraction method: We consider the set of printer reviews to be a special class of text documents (“printer review class”) and compare it to a set of reviews from amazon.com (e.g. book reviews) which we consider to be the counter-balance class (“book review class”). Now, the aim is to find the terms that are much more important within the class of printer reviews than within the class of book reviews. We make use of the fact that both classes share the review-related vocabulary and extract the terms that discriminate the printer review class against the counter-balance class of book reviews. By our definition a term discriminates one class from another if it is much more important within this class than within the other one. In order to measure the importance of terms for a class, we weight terms according to a novel extension of the TFIDF-measure, our “Term Frequency Inverse Class Frequency” (TFICF). We determine then the set of terms that discriminate the printer review class against the counter-balance class considering the TFICF term scores.

TFICF – Our Importance Measure

The most popular approach for term scoring, TFIDF [12], is not suitable in this case. This is due to the fact that the TFIDF value determines an importance value for a certain term with respect to a document within a document collection. What we need is an importance value for a certain term with respect to a document class. Therefore, we introduce the TFICF, which is an extension of the classic TFIDF measure. The formula for TFICF is composed of two factors: a term frequency value (tf) and an inverse class frequency value (icf).

The tf value reflects the relative frequency of a term within a class as in the TFIDF measure. The icf value takes into account in how many classes the term is present. In contrast to the standard idf formula our icf formula has to operate on multiple classes of documents instead of a single class. A straightforward application of the icf formula would be to say that a term t is an element of a class c, if it occurs in at least one of the corresponding documents. However, that means that outlier documents get a high influence on the result. Therefore, we propose to define that term t is only considered element of a class c if at least X percent of the documents D (where X is a user-defined parameter) contain the term (see equation 1).

$$icf(t) = log\left(\frac{|C|}{|\{c \in C : \frac{|\{d \in c : t \in d\}|}{|\{d \in c\}|} > X\}|}\right) \qquad (1)$$

Extracting Discriminating Terms

The TFICF measure provides a term weight that is comparable among several classes. For each term we get one value per class that allows us to compare the importance of the term in the two classes. We now define that a term is discriminating for one of these classes if its score is significantly higher for this class than its scores for the other classes. To determine the discriminating terms for a class, we use a threshold called discrimination factor by which a score for one class must outnumber the scores of all other classes (see definition 1).

Definition 1: Discriminating terms

A term t is discriminating for a single class C_k if:

$$\forall i \in \{1 \dots n\} \backslash k :$$
$$tficf(t, C_k) > discrimination\text{-}factor \cdot tficf(t, C_i).$$

Extracting printer attributes

As mentioned before, for our scenario we used a counter-balance class containing book reviews and discriminated the printer review class against it. As both classes shared the review specific terms, only printer related terms were discriminating the printer class and hence got extracted. Figure 2 compares the approach that just extracts the 40 most frequent terms after filtering stop words (top) with the result of our technique using the book reviews as a counter-balance class (bottom). It is easy to see that the quality of the second list is much higher since more product-related attributes are present. Thus, our approach allows us to do domain-specific term filtering without the usage of an ontology or a specialized knowledge base, just by providing a set of documents as counter-balance class.

3.3 Detection of opinion signal words

Once the attributes that were commented on have been identified, the aim is to find out what opinions were expressed on these attributes, respectively if the attribute was mentioned negatively or positively. To this end it is crucial to determine so called opinion signal words connoted with an attribute. These opinion signal words can have a positive polarity, e.g. “wonderful” or “to like”, or they can express a negative connotation as for example “bad” or “problem”. In order to find opinion signal words we used a freely available dictionary [1] listing words that evoke positive or negative associations. All other words that were not contained in the word list were considered as neutral words. In addition, it is important to deal with the occurrence of negations. While the term “good” for instance is clearly positive, the expression “not good” is negative although it contains a positive opinion signal word. For the purpose of avoiding such misinterpretations a simple heuristic was introduced that inverts the polarity of an opinion signal word if a negation word is preceding within a short range of tokens.

3.4 Mapping of attributes and opinion signal words

Given the attribute terms and the opinion words next we determine for each attribute in a sentence which opinion word(s) refer to it. We developed a novel statistical method that detects the opinion that has been expressed on an attribute by taking the polarity of the opinion signal words (os-words) around it into account. We call our method Reverse-Distance-Weighting (RDW). The basic assumption of RDW is that the closer an os-word is to an attribute the higher is the probability that it refers to that attribute. Please note that we constrain the search for os-words to the same sentence that the attribute is in.

Within a sentence a higher influence is given to closer os-words according to the reverse-distance weight (see equation 2).

$$rd\text{-}weight(A,o) = \begin{cases} 1 & if\, dist(A,o) \leq cutoff/2, \\ 0.5 & if\, cutoff/2 < dist(A,o) \leq cutoff, \\ 0 & else. \end{cases} \qquad (2)$$

A cutoff value is used to define how many words before and after the attribute are taken into account. In our application the cutoff threshold was set to 4. Experiments showed that this cutoff value is especially important in long sentences, because it prevents errors that are caused by very distant os-words that are incorrectly mapped to the attribute.

In order to get an opinion value for an attribute A and a sentence S equation 3 is applied. The weighted polarity values for each os-word o within S are summed up. The polarity value of an os-word is either +1 or -1 depending on whether it is contained in the positive or negative word list.

$$opinion\text{-}score(A,S) = \sum_{o \in S} rd\text{-}weight(A,o) \cdot polarity(o) \qquad (3)$$

If an attribute A gets a positive opinion score, the sentence is interpreted as talking positively about the attribute. Likewise, if this sum is negative, the sentence is supposed to talk in a negative manner about the attribute. If the sum is equal to 0, the polarity of the closest os-word is decisive.

Our application is special in the sense that we do not only want to know whether an attribute was mentioned positively or negatively in a specific sentence. Instead we are interested in the overall opinion that was expressed about the attribute in the review. To get the opinion value for an attribute on the review-level the majority vote of the sentence polarities for this attribute is determined.

As a result for every review we get a feature vector that summarizes the expressed opinions on the individual attributes. For each attribute there is one feature dimension in the vector. The corresponding value of the vector for a particular attribute's

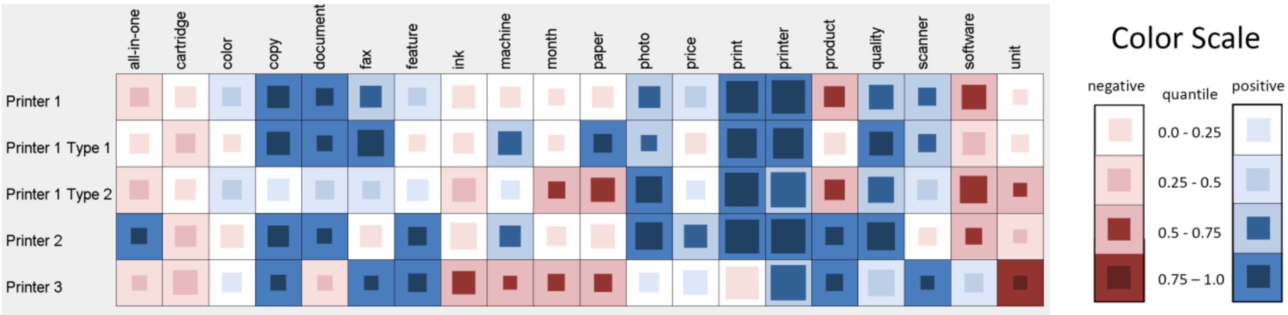


Figure 3. Summary Report of printers: Each row shows the attribute performances of a specific printer. Blue color represents comparatively positive user opinions and red color comparatively negative ones (see color scale). The size of an inner rectangle indicates the amount of customers that commented on an attribute. The larger the rectangle the more comments have been provided by the customers.

dimension indicates whether the attribute was mentioned positively (+1), negatively (-1) or neutrally / not at all (0).

4 Applications and Visual Analytics Methods

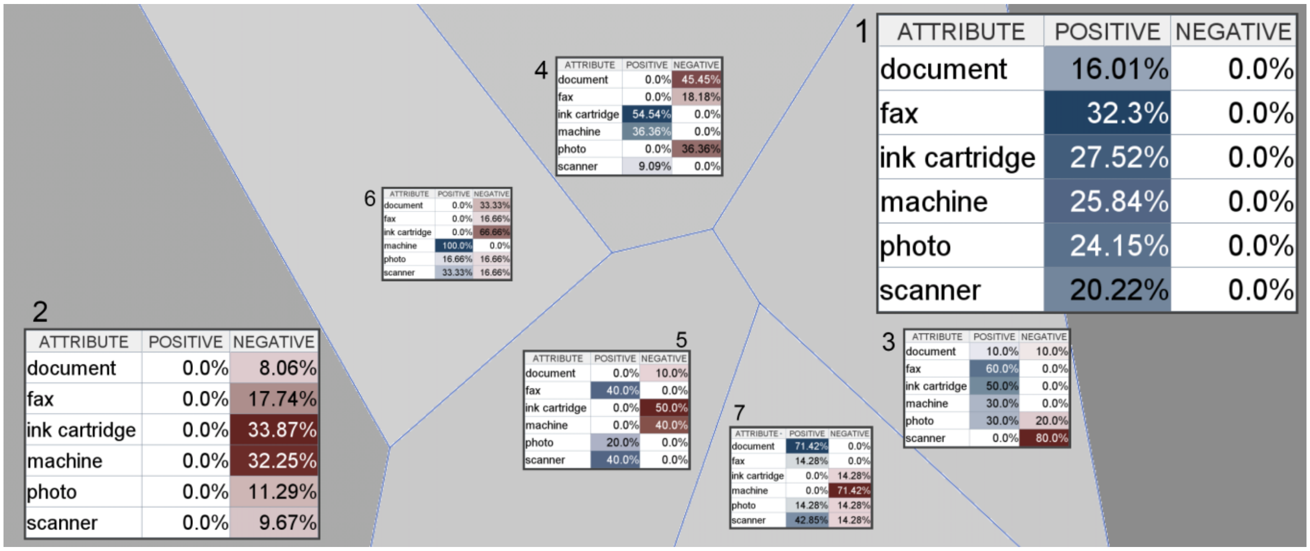
The resulting multi-dimensional feature vectors are applied to real-world customer comments from an online store and product reviews from amazon.com to identify the strength and problems of different products and services. We have introduced three new visual analytics techniques: (1) summary reports that provide a quick overview of the customer reviews without the need to read them, (2) a visualization of the clusters of reviews in which similar opinions are expressed and finally (3) circular correlation maps to uncover correlations between the user-given score and different attributes.

4.1 Visual Summary Reports

Visual summary reports provide a quick overview of the customer feedback data set. They show for each attribute extracted by our automatic algorithm whether it belongs to the category of attributes with a positive tendency (blue) or the category with a negative tendency (red). The size of the inner rectangles is determined by the percentage of reviews that commented on the attribute signaling the importance that the analyst should give to this attribute in his or her evaluation. Color is mapped to the percentage of positive or negative opinions, respectively. Using our automatic analysis method, we calculate the average percentage of positive comments per attribute and use this as a threshold. Attributes whose percentage of positive comments is above that threshold exhibit a positive tendency compared to the other attributes (color = blue), the ones that are below the threshold show a negative tendency (color = red). The stronger the positive or negative tendency is the darker the color value becomes. The intervals for the four shades of blue / red tones are determined by the quantiles of the set of positive or negative attributes.

By recalculating the threshold value for each data set instead of using a fixed one we compensate for the shift that is caused by the fact that some products are generally commented on more positively than others. Please note that the basis for the calculation of the threshold and the quantiles is always the whole set of product reviews that are displayed to ensure comparability across the different lines.

Figure 3 shows a visual summary report of reviews from amazon.com on three different printers (in total 1876 reviews were analyzed). For printer 1 we additionally show the result for two different printer types separately. This allows a detailed analysis of strength and weaknesses of specific printer families. It can be seen that there are some attributes that the customers are generally satisfied with. This is true for the general attribute


1

ATTRIBUTE	POSITIVE	NEGATIVE
document	16.01%	0.0%
fax	32.3%	0.0%
ink cartridge	27.52%	0.0%
machine	25.84%	0.0%
photo	24.15%	0.0%
scanner	20.22%	0.0%

2

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	8.06%
fax	0.0%	17.74%
ink cartridge	0.0%	33.87%
machine	0.0%	32.25%
photo	0.0%	11.29%
scanner	0.0%	9.67%

3

ATTRIBUTE	POSITIVE	NEGATIVE
document	10.0%	10.0%
fax	60.0%	0.0%
ink cartridge	50.0%	0.0%
machine	30.0%	0.0%
photo	30.0%	20.0%
scanner	0.0%	80.0%

4

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	45.45%
fax	0.0%	18.18%
ink cartridge	54.54%	0.0%
machine	36.36%	0.0%
photo	0.0%	36.36%
scanner	9.09%	0.0%

5

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	10.0%
fax	40.0%	0.0%
ink cartridge	0.0%	50.0%
machine	0.0%	40.0%
photo	20.0%	0.0%
scanner	40.0%	0.0%

6

ATTRIBUTE	POSITIVE	NEGATIVE
document	0.0%	33.33%
fax	0.0%	16.66%
ink cartridge	0.0%	66.66%
machine	100.0%	0.0%
photo	16.66%	16.66%
scanner	33.33%	16.66%

7

ATTRIBUTE	POSITIVE	NEGATIVE
document	71.42%	0.0%
fax	14.28%	0.0%
ink cartridge	0.0%	14.28%
machine	0.0%	71.42%
photo	14.28%	14.28%
scanner	42.85%	14.28%

Figure 4. Scatterplot of customer reviews on printers: Seven main opinion clusters have been identified and mapped in a 2D space, each represented by one thumbnail. The more reviews a cluster contains, the larger its thumbnail is displayed. Positive opinions are highlighted in shades blue, the negative ones accordingly in red. The color brightness is mapped to the percentage of reviews within a cluster that share a certain opinion.

"printer" but also for other attributes such as "copy, photo, print or quality". The latter ones suggest the assumption that the quality of the prints is not that much of a distinguishing factor between the different brands. On the other hand we found that there is a lot of negative feedback for the attributes "cartridge", "ink", "month", "software" and "unit".

Of course, such attribute terms can only be considered as hints that point at certain problems that the analyst should have a closer look at. This is especially true for terms such as "month" or "unit" which are not self-explanatory. In analyzing the data set, the size of the inner rectangle in combination with the color gives an idea to the analyst of how severe the problem might be. For example, in Figure 3 the "software" column sticks out because of the large size of the inner rectangles (signaling that many customers commented on it) in combination with the dark red colors (which means that a relatively large number of customers was dissatisfied with this aspect). For a customer it will be of special interest to observe the differences in the evaluation of the three printers. Strength and weaknesses of the specific printers as seen by other customers become easily visible in the summary report visualization.

4.2 Cluster Analysis

Our second visualization shows groups of customers with a similar opinion. Thus, this kind of visualization is important for companies that would like to learn about different groups of customers.

To find the different groups of customer opinions we apply a hierarchical clustering algorithm (Complete Linkage) and then project each cluster representative (on a user-selected hierarchy level) in 2D space using multi-dimensional scaling as a dimensionality reduction method. Each cluster is then visualized using a thumbnail image that depicts for each attribute the percentage of reviews in the cluster that commented on it, split up into negative and positive comments. The number of reviews that the cluster contains is mapped (non-linear) to the size of the thumbnails and to the grey tone of the corresponding voronoi cell.

Figure 4 shows an example in which clusters of customers of printer 1 are shown. The largest cluster is the one that summarizes all the reviews that did not criticize any attribute but were satisfied with the printer. On the other hand there is also a group of customers that disliked most attributes (cluster 2). More interesting however are the clusters that summarize the reviews with a rather differentiated opinion. For example cluster 3 aggregates the reviews that had an overall positive tone but in which also some critical aspects (mostly because of the scanner) are mentioned. Finally, clusters 4-6 show user groups with a clearly differentiated opinion about the product.

Key for the detection of expressive clusters is the use of a meaningful distance function that is able to measure the similarity of reviews. In the following we derive a special distance function that satisfies our needs.

Given a feature vector r = (i_1, i_2, …, i_n) for each review as described in section 3 we need a distance function that is able to discern the similarity of two reviews with respect to the opinion that is expressed in them. The similarity between two reviews is increased if they both comment on an attribute and both agree in their opinion about this attribute. Likewise, if an opposite opinion is expressed about an attribute in two reviews, the similarity value between those two reviews has to decrease. This leads us to the first part of our distance function that counts how often the two reviews state opposing opinions on an attribute (see equation 4):

$$f(r_1,r_2) = \sum_{k\in K} g(i_k^{(r_1)}, i_k^{(r_2)}), \tag{4}$$

$$\text{where } K = \{i : i^{(r_1)} \neq 0 \wedge i^{(r_2)} \neq 0\},$$

$$\text{and } g(i_k^{(r_1)}, i_k^{(r_2)})) = \begin{cases} 1 & \mathit{if\,sign}(i_k^{(r_1)}) \neq \mathit{sign}(i_k^{(r_2)}), \\ 0 & \mathit{else.} \end{cases}$$

So far the positions in the feature vectors in which at least one of the reviews does not comment on the attribute do not contribute to the distance. Can they be ignored? Consider the following two feature vectors:

```
R1:  [ +1  0 +1  0 +1  0 +1  0 +1  0 ]
R2:  [ +1 -1  0 -1  0 -1  0 -1  0 -1 ]
```

In this pair of feature vectors only the first attribute is commented on by both reviewers. As they both mention the attribute positively the above introduced distance function would consider the two reviews as stating a very similar opinion (distance = 0). However, it is obvious that those reviews do not express a similar opinion on the product. Thus, we also have to take these attributes into account that only one of the reviewers comments on. We do so by calculating for both feature vectors the percentage of positively mentioned attributes, taking only the attributes into account that the other reviewer did not comment on since the ones that both commented on already contribute to the first part of our distance function. Those two values are then subtracted from each other to measure the difference in their general opinion about the remaining attributes (see equation 5).

$$h(r_1,r_2) = \left|\frac{L_{1+}}{L_1} - \frac{L_{2+}}{L_2}\right|, \tag{5}$$

$$\begin{aligned} \text{where } L_1 &= |\{i \,:\, i^{(r_1)} \neq 0 \wedge i^{(r_2)} = 0\}| \\ L_{1+} &= |\{l \in L_1 \,:\, l > 0\}| \\ L_2 &= |\{i \,:\, i^{(r_2)} \neq 0 \wedge i^{(r_1)} = 0\}| \\ L_{2+} &= |\{l \in L_2 \,:\, l > 0\}| \end{aligned}$$

Finally, both parts of the distance function are weighted with the number of attributes that contribute to it, leading to the following distance function (equation 6):

$$\begin{aligned} dist(r_1,r_2) = {} & \frac{K}{K+L1+L2} * f(r_1,r_2) \\ & + \frac{L1+L2}{K+L1+L2} * h(r_1,r_2) * w \end{aligned} \tag{6}$$

where the factor w allows to balance the influence of the two aspects of the distance function.

4.3 Circular Correlation Map

The Circular Correlation Map offers a detailed view on the data. It was introduced in [18] and can be used to find correlations between the different aspects of the data set (such as the attributes, the total score or the assigned cluster id). Figure 5-7 shows an application example in which feedback of customers that bought a notebook in an online store was analyzed. The feedback was directly collected by a company in order to find out where improvements are necessary.

A feature vector is added to the diagram as follows: for each attribute that the customer commented on a line is drawn from the position of the document ID on the right semicircle to the respective score value of that review in the middle and from the score to the position of the attribute on the left semicircle. The color of a line is determined by the opinion that was expressed on the attribute (blue = positive, red = negative). If multiple lines are

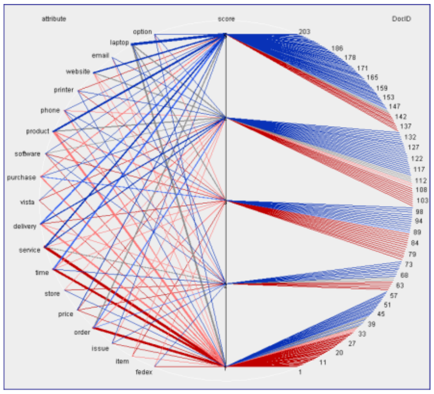


Figure 5. *All Customer Comments.* Most comments have an overall positive tendency (many blue lines).

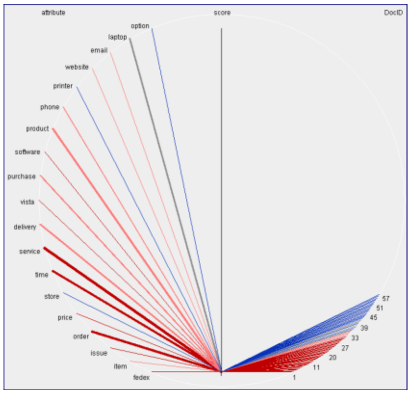
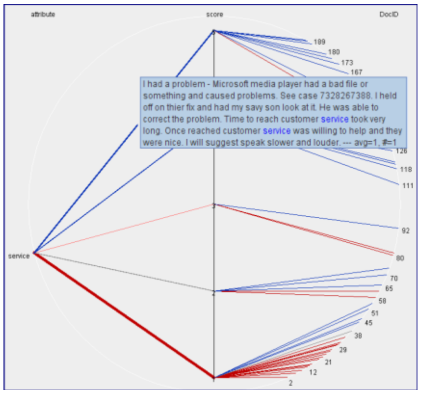


Figure 6. *Analyzing the score 1 comments:* Service is one of the attributes that is often mentioned negatively.



I had a problem - Microsoft media player had a bad file or something and caused problems. See case 7328267388. I held off on thier fix and had my savy son look at it. He was able to correct the problem. Time to reach customer service took very long. Once reached customer service was willing to help and they were nice. I will suggest speak slower and louder. --- avg=1, #=1

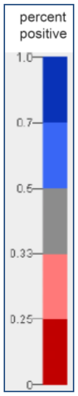


Figure 7. *Subset of comments on service.* Not all the customers are dissatisfied with the service. But this was a hot topic in the score 1 comments.

on top of each other the percentage of positive comments is calculated and the lines are colored accordingly. The width of a line represents the number of lines that are on top of each other.

The tool allows to interactively select subsets of the graph to analyze them in detail. Figure 5 gives a visual overview of the relationships between attributes, customer score, and the document ID. In Figure 6 only the lines with a user-given score of 1 are displayed. The visualization allows us to detect the main problems that led to the low scores. The thick red lines that lead to the attributes "service, time and order" reveal that those aspects troubled the customers. To find out if the service is in general perceived negatively by the customers, we next select the subset of lines that represent the comments on "service" (Figure 7). It can be seen that this is not the case as customers that gave a high score on average also commented positively on the service attribute. However, the thickness of the line from "service" to score 1 in relation to the total number of score 1 reviews signals that this was a hot topic among the reviews with score 1. Please note that in this visualization it is also possible to go to the lowest level of abstraction and read the text of the review (see Figure 7).

5 Evaluation

The evaluation of our approach consists of an evaluation of the novel attribute extraction method as well as a detailed discussion of the results with a description of opportunities for performance enhancements.

5.1 Evaluation of the attribute extraction

In order to evaluate the quality of our attribute extraction approach a small user study was conducted. For the evaluation scenario the 40 top terms according to frequency were compared to the 40 top terms extracted by our discrimination-based approach (listed in Figure 2). For each of the terms the participants of the user study had to decide whether it is a printer attribute of which they would want to know if users generally liked or disliked it before buying a particular printer, which are precisely the kind of terms that should be extracted. In order to avoid any bias the terms extracted by both approaches were merged and the terms were ordered alphabetically. Thus, the participants did not know by which method a term was originally extracted. As participants of the user study five rather experienced printer owners were recruited.

An interesting outcome of the user study was that users have quite varying preferences on attribute terms. For 31 out of the 40 terms that our method extracted at least one participant thought that they were useful printer attributes. For the 40 top-frequency terms only 21 terms were found to be useful by at least one user. The detailed comparison of the result of both methods is shown in Figure 8. Our method clearly outperforms the standard frequency-based method by a significant margin (at least 44% more relevant attributes).

5.2 In-depth evaluation of error sources and improvement potentials

As a ground truth for the performance evaluation of our method we used a manually annotated benchmark data set of product reviews which is publicly available [16]. The dataset contains customer reviews in which for every sentence the commented product attributes are listed and additionally the polarity of the attribute evaluation is given. For the evaluation only subjective sentences were considered, i.e. sentences containing positive or negative statements about the attributes. Furthermore the list of

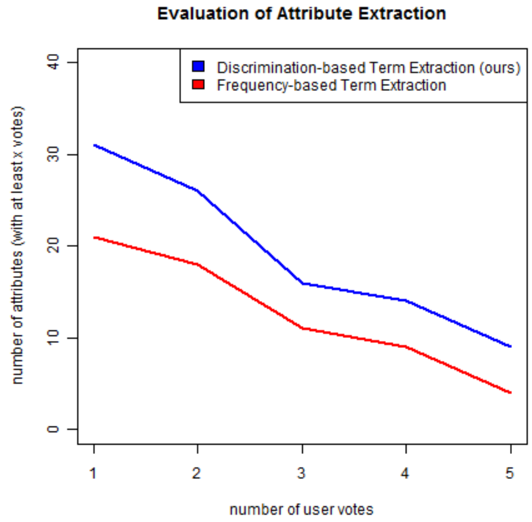


Figure 8. Results of the user-study. On the x-axis the number of users is listed that have voted for an extracted term. The y-axis indicates how many terms were identified by at least x users as useful attributes. A number of 5 users implies that it was an unanimous vote. For each individual user vote threshold our method finds at least 44% more useful attributes than the top-frequency method.

attributes was given.

The overall accuracy of our method on this benchmark dataset was 0.72. In order to discover the error sources we studied these sentences in detail for which the attribute polarity was incorrectly determined. The observed error sources can be subdivided into two main categories: Errors that have their origin in the inadequacy of the opinion word list and errors that are due to the limited possibilities to automatically derive semantics from natural language sources respectively to solve semantic ambiguities. Table 1 shows a subdivision of the 51 manually examined errors into 7 error sources.

Main Error Sources	In % of all Errors
1. Errors in opinion word list	19.6 %
2. Domain-dependency of opinion words	23.5 %
3. Attribute-dependency of opinion words	5.9 %
4. Opinion word combinations	13.7 %
5. Fixed expressions and phrases	21.6 %
6. But-clauses	7.8 %
7. Implicit attributes	2.0 %

Table 1. Overview of the main error categories

Errors due to the opinion word list:

1. The opinion signal word list that we use turned out to be erroneous which means that on the one hand it contains words that definitely are no opinion signal words and on the other hand certain opinion signal words are missing.
2. The general nature of the opinion word list implies that it does not contain context-dependent opinion signal words. For digital cameras important opinion signal words could be for example “sharp”, “light”, “blurry” or “quick” which are not included in the general list. At the same time words such as “capture” or “shot” might evoke negative emotions in general but cannot be considered as negative in the context of a digital camera.
3. Some of the opinion signal words have no generally valid polarity but depend on the nature of the attribute they describe. For example a “short” response time may be favorable for a camera but a “short” battery life is not.
4. Finally, composed opinion signal words such as “last long”, “easy to manipulate”, and “exceeded my expectations” are lacking in the list.

Errors due to the ambiguity of natural language respectively the difficulties of automatic detection of semantics:

5. Composed expressions and fixed phrases cannot be detected by just searching for single signal words (e.g. “this camera will not let you down”, “I got this camera about a month ago and I can’t put it down”, “features…are unmatched for any camera in this price range”, “8mb for a camera like this is a joke”). The same applies for poetic descriptions, metaphors or ironic statements.
6. Our mapping strategy may fail if a turn of opinion is introduced by conjunctions such as “but, however …” like in the following sentence: “The battery life seems to be on the short side but adequate for most situations.”
7. There are cases when an attribute is not explicitly but only indirectly mentioned. For example in the sentence “The camera is too small” the attribute “size” is implicit.

While natural language ambiguities (errors 5-7) are hard to resolve automatically the adaptation of an opinion word list is a onetime effort that could lead to considerable improvements. Therefore we manually edited our general opinion signal word list by deleting problematic words and inserting missing ones (see error 1). In addition, we included opinion words that are opinion-bearing in our concrete task and eliminated words from the list that in our case were not useful (see error 2). Without any further refinement we were able to increase our performance by 15% and got an overall satisfactory accuracy of 0.83. Further improvement would be possible by a separate opinion word list for each attribute (error 3) and opinion word combinations (error 4).

6 RELATED WORK

Within the context of opinion analysis three main tasks can be distinguished: Subjectivity Analysis (detecting whether a text is subjective or objective), Sentiment Analysis (detecting whether the general opinion in a text is positive, neutral or negative), and Opinion Mining (additionally analyzing what has been commented on positively or negatively). As our approach is clearly situated within the context of opinion mining we are going to review primarily these latter approaches in this related work section. A more comprehensive overview on opinion analysis techniques can be found in [9]. Furthermore, we compare our approach to existing techniques for the visualization of extracted opinions.

6.1 Attribute-based opinion mining

Attribute-based opinion mining is often made by two successive steps: First, the attributes (sometimes also called features), that have been commented on, are identified. Secondly, the respective opinion that has been expressed on them is detected.

Different approaches to extract the attributes exist. In [5, 17] the Apriori algorithm is used to find frequent features (that means sets of terms that occur frequently together in a sentence). Subsequently, two pruning steps are applied to refine the result. In contrast to this approach, Popescu et al. [10] consider all noun phrases as attributes whose frequency is above a certain threshold. The list of attributes is then further filtered by calculating the PMI score (Point-wise mutual information) between each phrase and discriminator phrases (such as "is a scanner" or "scanner has" etc. in case of reviews on scanners). The PMI scores are calculated on a set of web documents containing the product name. The paper reports a 22% increase in precision with 3% loss of recall compared to the results reported in [5]. [15] proposes and compares two different approaches. One of them is based on a mixture language model and the other one applies the likelihood-ratio test. They report better results when using the likelihood-ratio test. Titov and McDonald introduce in [19] the concept of a Multi-Aspect Sentiment model that is based on an adaption of Latent Dirichlet Allocation to extract rated attributes (here called aspects) from reviews. Finally, the approach of Kim and Hovy [6] is based on FrameNet, an online lexical database which consists of 800 semantic frames. The idea is to label each sentence that contains an opinion-bearing term with semantic roles and defer the attribute and opinion holder in the sentence from these.

Our approach is different from all the above in that we determine the attribute by our novel discrimination-based method. Our mapping of opinion words to the attributes is similar to [5, 17]. However, we use a different weighting function and a cut-off value, which both improves the performance. We also aggregate the sentence level opinion into an overall opinion value on the review level.

6.2 Visual Opinion Analysis

Visualization of opinions has not yet been a major focus of research in the area. In [7], for example, the authors suggest to use traditional bar charts to visualize how many positive respectively negative statements exist within the document corpus. The advantage of our technique is that it is much more scalable with respect to the number of attributes and the number of products that can be displayed. Furthermore, our matrix-based visualization simplifies the comparison between both different attributes and different products.

Among the other visualizations of opinion mining results is Pulse [2] which clusters reviews according to topic and then calculates the average opinion per cluster. The result is displayed in a treemap with color being mapped to the average opinion. The BLEWS system introduced in [3] analyzes blogs with respect to their political orientation. The average emotional sentiment of a set of articles is visualized as a glow around the bars that represent the number of documents that link to a specific news article. Morinaga et al. [8] display characteristic phrases for the group of positive or negative sentences in a 2D scatterplot. In [14] the development over time of RSS feeds that report on the U.S. elections is visualized. Finally, [4] has to be mentioned as the only technique that does not only display positive or negative sentiment but also other aspects such as pleasure, pain, power conflict etc. The detected emotions are visualized in an adapted rose plot.

To the best of our knowledge currently no technique exists that analyzes and visualizes customer feedback with respect to clusters of reviews in which similar opinions are expressed. In addition, we provide correlation maps to visualize the detailed distribution and correlations of attributes, opinions, and scores.

7 Conclusions and Future Work

Providing better access to the large amount of textual customer feedback data has a high impact both for the companies that need to know what the customers like or dislike about their products and for the customers who want to find the product that fits best to their needs. In this paper we address the whole opinion analysis pipeline and introduce novel techniques for the automatic feature extraction as well as for the visual analysis of the detected opinions. First, we extract the important information out of the unstructured natural language reviews and make it accessible in a structured format. Our main contributions here are a new discrimination-based attribute extraction method and a novel opinion feature construction that generates one feature vector per review breaking the overall opinion down into opinions about individual product attributes. Secondly, we develop innovative visual analytics methods in order to make the large amount of review opinions easily and quickly accessible to the analyst. Interesting opinion profiles are revealed by review summary and clustering visualizations, to allow an interactive exploration and to discover hidden patterns and relations. Our contributions are the new visual summary reports, a novel distance function for clustering review feature vectors, the cluster thumbnails and the circular correlation maps.

The presented visualization techniques have been developed in cooperation with a product manufacturer and are thus optimized for their specific needs. We believe that those visualizations also provide valuable insight for customers. However, a visualization that is specialized on potential buyers might additionally highlight the (negative) outliers in the set of product reviews as they are often considered to be the most informative ones for deciding if a product fits ones needs.

In the future, we would like to provide more detailed information by analyzing the context of the attributes (co-occurrence analysis). In addition, improving the automatic opinion extraction by applying more sophisticated linguistic methods for negation handling, anaphora resolution and the development of techniques to detect irony are challenging problems that need more research.

References

[1] V. Buvac, Internet General Inquirer, http://www.webuse.umd.edu:9090/

[2] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, pages 121–132. 2005.

[3] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König. BLEWS: Using blogs to provide context for news articles. In *Proc. of 2nd AAAI Conf. on Weblogs and Social Media*, 60-67, 2008.

[4] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Workshop on Sentiment and Subjectivity in Text,* pages 23–30, 2006.

[5] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proc. of the tenth ACM SIGKDD Conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[6] S.-M. Kim and E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proc. of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics, 2006.

[7] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[8] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *KDD '02: Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349. ACM, 2002.

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

[10] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05: Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Assoc. for Computational Linguistics, 2005.

[11] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In *Proc. of the Third ACL Workshop on Very Large Corpora*, 1995.

[12] G. Salton, A.Wong, and C. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[13] K. Toutanova, D. Klein, C. D. Manning, and Y. Singe. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, 2003.

[14] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. A. Keim. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)*, 2009.

[15] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *ICDM '03: Proc. of the Third IEEE International Conference on Data Mining*, page 427. IEEE Computer Society, 2003.

[16] M. Hu and B. Liu. Review Datasets. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[17] X. Ding, B. Liu, and P.S. Yu. A holistic lexicon-based approach to opinion mining. In *Proc. of the international Conference on Web Search and Web Data Mining (WSDM '08)*. ACM, pages 231-240, 2008.

[18] D. Keim, M. Hao, U. Dayal: "Business Process Impact Visualization and Anomaly Detection". In *Information Visualization Journal.* Palgrave Publisher, January, 2006.

[19] I. Titov and R. McDonald, A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In: *Proceedings of ACL-08: HLT, pages 308-316, Assoc. for Computational Linguistics,* 2008.