REVIEW DRIVEN CUSTOMER SEGMENTATION FOR IMPROVED E-SHOPPING EXPERIENCE

Silviu Homoceanu, Sergej Dechand, Wolf-Tilo Balke
Institute of Information Systems
Technische Universität Braunschweig
Braunschweig, Germany
{silviu, balke}@ifis.cs.tu-bs.de, s.dechand@tu-bs.de

ABSTRACT

Despite intriguing commercial possibility, product search on the Web and e-shopping applications still strive to offer satisfying customer experience. The major challenge probably is to harness the power of user generated content in the form of reviews. Using the example of cell phones this paper demonstrates that user reviews, opinions, and product ratings may actually severely differ with respect to the intended product usage of individual customers or groups. Investigating individual task-based rating behavior, we show that customer segmentation paired with intuitive interface paradigms like faceted search, promises to significantly enhance user experience by combating the information flood.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information filtering, Classification;

General Terms

Experimentation, Human Factors, Languages.

Keywords

Product reviews, product search, e-commerce, customer segmentation, usage profiles, personalized search.

1. MOTIVATION

E-commerce, and especially shopping on the Web has become a huge success. According to [1] product search toady makes for about 20% of Web search queries and shopping portals like Amazon.com[1], Buy.com[2] or Ebay.com[3] are continuously creating significant revenues. However, looking at user experience comparing online shopping and traditional face-to-face shopping a striking difference is notable when providing advice to customers. Where online offerings tend to be efficient only when customers at least roughly know what product to purchase, communicating directly with salespersons is still hard to replace where customer needs are only vaguely specified. This has also been noted by online portal providers and currently a mixture of technical specifications, expert opinions and user generated content in the form of ratings or experience reports is state of the art.

Still, working through all this information is left to customers, who will often feel overwhelmed by the sheer amount of information. But taking a closer look at human interaction in sales processes reveals that a common question of salespeople is along the lines of what features are especially desirable or what tasks a user mainly wants to perform with the product. For example, when looking for a new cell phone a user might state that typical features like calendar function or a good connectivity for checking emails are important rending him/her as a 'business' user, whereas other users may be more interested for example in MP3 player functionality or a fashionable design.

This shows that the intended usage of a product plays an essential role in customers' decision making and indeed first portals have begun to also assign scores also for non-technical features that may be important for certain customer groups. Considering for example the popular CNET[4] portal for reviews, cell phones are already scored based on different feature sets with respect to being best smart phones, best basic phones, best MP3 phones, etc. (see http://reviews.cnet.com/best-cell-phones/). This specific usage-based type of product search already inspired work in the area of personalization aiming to uncover implicit features important to individual users (or groups), see [9] or [11], or scoring a product's utility regarding such features, e.g., by creating conceptual views [2].

But even sophisticated personalized search techniques can only limit down the choice to a couple of suitable products, whose details still have to be compared when deciding for one product to purchase. And again the richness of user generated content strikes: for example popular products on CNET.com can easily have more than 300 reviews to sort through. While portals like Amazon.com try to alleviate the problem by showing histograms over the respective reviews' product ratings and then allowing to navigate reviews by positive/negative opinions, the obvious idea should be to classify user reviews again according to the intended usage and then allow users of each group to only view reviews relevant with respect to this usage.

In the course of this paper we will show that user reviews can indeed be broken down to semantically meaningful usage-centered groups using simple supervised learning techniques thus paving the way for improved user experience in terms of efficient information processing. In particular, we will show -using the example of cell phones- that the ratings of groups with respect to some product may severely differ for different intended usages and thus express valuable personalized information beyond rating histograms over all user ratings. Moreover, we present a simple faceted search interface that allows users to filter product reviews according to a set of predefined user groups and to intuitively understand the important features of each group by usage-related tag clouds automatically generated from the product classes' training sets.

[1] www.amazon.com

[2] www.buy.com

[3] www.ebay.com

[4] www.cnet.com

2. INVESTIGATING TASK-BASED RATING BEHAVIOUR

The exchange of opinions about products in the form of user reviews and ratings has become commonplace. But -as we will show in this section- relying on these ratings only, may often result in a somewhat biased perception. Consider for example a Motorola Droid cell phone with CNET.com ratings in Figure 1: the product seems to perfectly satisfy more than half of the customers. But should a customer e.g., with typical business usage in mind buy it? Again customers have to dive into the reviews and weigh them individually with respect to their intended usage. However, neither a good review favoring features like MP3 player functionality or a fashionable design, nor a bad review warning about the quality of the camera will be helpful for making the decision about suitability for business tasks. Moreover, also looking at representative positive/negative user reviews -like usually done in opinion mining- might not clarify matters. In contrast, business users will rather be interested in organizer functionality, large screens for reading emails, portability, etc.

Figure 1: Rating Histogram and Average Rating for the Motorola Droid from CNET.com.

Following this idea we argue that product reviews can indeed be classified with respect to intended usages. While there always will be many rather general reviews, especially for certain product segments there will be a significant amount of reviews considering the suitability of a product for a specific task. These reviews will generally use different terms in their vocabulary making the intended tasks distinguishable. Moreover, following psychological results from cognitive economy [3], there are only a handful of such typical tasks that can usually be anticipated.

Task-Based Classification of Reviews: we decided to use a supervised learning technique for the document classification task. We manually tagged 150 cell phone reviews from CNET.com as training set for typical tasks performed with cell phones, e.g., business, multimedia, smartphone, basic use, fashion, social networking, navigation, etc. After cleaning review vocabularies by eliminating stopwords and Porter stemming, for each task we trained a support vector machine (SVM) [4], [5] on the high-dimensional term space given by the review terms and thus for each review derived a binary decision whether it focuses on the task or not. The results of this classification are quite promising: using the leave-one-out cross-validation method the classification for all tasks achieved a precision of above 85% for 50% recall values degrading gracefully for higher recalls (see Figure 2 for the three prominent tasks: business, multimedia and smartphone).

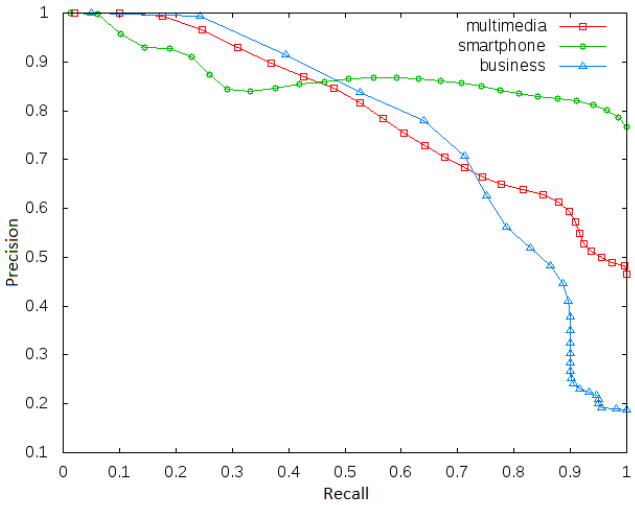

Figure 2: Precision Recall Curve for Business, Multimedia, and Smartphone tasks on 150 Test Reviews.

It is also interesting to note that by looking at the number of task-specific reviews relative to general reviews, we are also able to distinguish between different product lines developed for some task. For example looking at business tasks we found that for typical business cell phones like Blackberry cellphones or the Nokia e-series there is a significantly higher percentage of reviews focused on (and classified by our SVM as) business usage (cf. Figure 3). This reflects the intuition that business users rate more of the products known for their business capabilities, whereas task-specific reviews are rare for all purpose products.

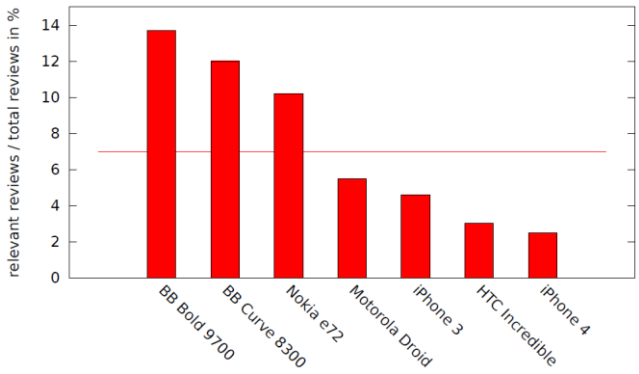

Figure 3: Relevant Reviews on Business Usage.

Rating Behavior in Task-based Review Classes: We are now ready to investigate the actual rating behavior of task-centered customers with respect to different products. Let's have a closer look at a general purpose cell phone like the Motorola Droid. Figure 4 shows rating histograms taking all (a), only business reviews (b), and only multimedia reviews (c). It's easy to see that that there is a significant difference in the shapes of the three histograms. While the histogram for the multimedia reviews (c) is slightly skewed towards the maximal rating value, the histogram for business reviews (b) shows a large plateau. What also already becomes apparent are vastly different means and variances with

respect to the same products depending on the respective customers' intended use.

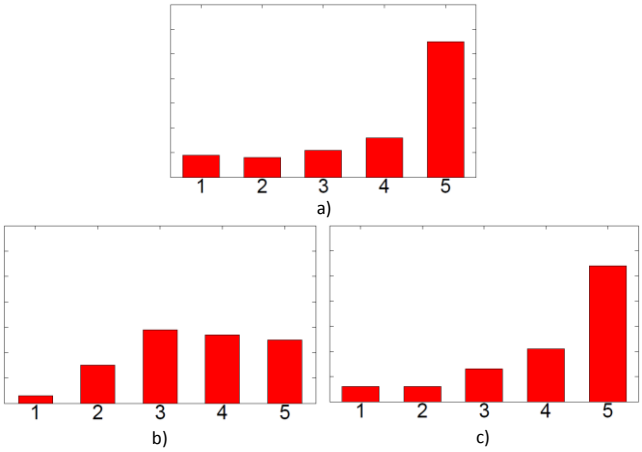

Figure 4: Rating Histograms (ranging from 1-5 stars) for the Motorola Droid – a) all reviews, b) business reviews, c) multimedia reviews.

Although we thoroughly evaluated several tasks, in the following we will only report our results on 'business' tasks for brevity reasons. The remaining tasks yielded similar results.

Table 1 shows statistical characteristics of ratings on a 1 to 5 stars scale for classified business reviews in contrast to general, i.e. non-business reviews in terms of average and variance. Moreover, following the distinguished product lines in Figure 3 we separated typical business devices (BlackBerries, Nokia e-line) from typical allrounders (like the iPhones and android devices). There is a recognizable difference in rating behavior between business and non-business reviews. But whereas for the business cell phones the means only slightly differ (0.29 on average), for the allrounders opinions tend to differ more (0.62 on average). Clearly the allrounders are punished harder for not being able to perform business tasks, respectively also rewarded higher like in the case of the iPhone4 that mainly differs in features related to business tasks from the iPhone3. Also looking at the variances reveals effects along the same line: while business reviews show rather low variances (on average 0.79 for business cell phones and 1.66 for allrounders), opinions in general reviews tend to differ more (on average 1.59 for business cell phones and 2.23 for allrounders). This clearly shows the different interest profiles of more or less homogeneous customer groups expressed by user reviews; usually more focused interests with respect to both usage and product, will result in smaller variances in the ratings.

Table 1: Rating Behavior in Reviews (wrt. business tasks)

Products \ Reviews	Business μ (σ^2)	Non Business μ (σ^2)
BB Bold 9700	4.30 (0.76)	4.02 (1.40)
Nokia E72	3.66 (0.08)	3.47 (2.30)
BB Curve 8300	3.61 (1.54)	4.00 (1.07)
Motorola Droid	3.32 (1.41)	3.84 (1.97)
iPhone 3	2.80 (2.00)	3.47 (2.57)
iPhone 4	3.88 (1.26)	3.43 (2.72)
HTC Incredible	3.19 (1.96)	4.02 (1.65)

3. FACETED SEARCH INTERFACE

In brief, the intended usage or tasks plays an important role in the way customers rate the products and thus, should also play an important role in advising customers. For instance, of the 124 reviews on CNET for the BlackBerry Bold 9700, only 17 are really focused on business tasks. Nevertheless, even if a customer can be clearly assigned to the 'business' segment, more than 100 possibly irrelevant reviews are hampering the information gathering process. Therefore, in this section we provide a simple faceted search interface allowing users to filter product reviews according to a set of predefined task-based groups.

Figure 5 shows our simple, yet useful interface for the Motorola Droid: once a user state that he/she is specifically interested in certain tasks like business usage, smartphone capabilities, or multimedia features, reviews are automatically filtered according to the classification presented in Section 2. In order to provide users also with a basic understanding of what is meant by the usage classes, each group reveals on mouse click a tag cloud of the salient terms from the combined vocabulary of all reviews in the class. Salient terms here are those terms that show highest discriminative power with respect to the SVM classifier.

Figure 5: Task-Based Faceted Search Interface.

4. RELATED WORK

Dealing with the rising information flood in the field of user generated content, especially for product search has often been addressed with opinion mining techniques. In [6] the authors propose a classification of the reviews not by topics, but by the overall sentiment or sentiment ratio. In a nutshell this means determining whether a review expresses a dominantly positive or negative opinion regarding a product. This approach, however, delivers even less information than rating histograms do. Providing the customer with the number of positive and negative reviews for a product is useless, especially if sentiments strongly vary. After all, the customer is still left with the burden of manually investigating which reviews actually address his/her needs.

Another classical approach based on opinion mining is to produce a feature-based summary of the product reviews, see for example [7], [8], [9], [10]. Product features are first identified and then adjectives in the proximity of the features are used to establish their polarity. Further improvements of this approach [11] are even able to cope with various terms people may use to refer to a single product feature (picture, photo and image all refer to the camera). But feature-based summarizations of opinions have a major drawback: the necessity of features is usually evaluated with respect to some intended task. For example not needing constant Internet access a basic phone user may find the typical battery life of 1-2 days for smartphones rather unsatisfying, while smartphone users usually find the same life span entirely sufficient.

Although the idea of actually aggregating a spectrum of opinions, or different points of view is enticing, it still poses a severe challenge: to some degree bias and diversity in opinions has recently been investigated in [12] on news data. Moreover, in [13], [14] the authors propose a solution for identifying the political orientation based on the opinion expressed in political texts. The problem is, however, that the proposed method needs a manually created "dictionary" assigning each word appearing in the texts to an ideological score, heavily challenging the applicability.

5. CONCLUSIONS

Deciding for products in online shopping applications is a major task in successful e-commerce portals. However, until very recently these processes were to a large degree not supported beyond simple SQL-style database queries. With the advent of massive user generated content on the Web, companies begin to recognize the new chances for (pro-) actively advising customers by exploiting user reviews for possibly interesting products. But the ever growing amount of available information and the diversity of opinions and ratings also turn out to be a curse in terms of putting more and more cognitive effort on customers trying to grasp the full information about a product.

Following ideas from cognitive economy in the course of this paper we have demonstrated that this curse can indeed be alleviated: when connecting user reviews to certain intended tasks or specific kinds of product usage and then filtering irrelevant reviews for specific customer segments, the information flood is (at least to some degree) contained. We have shown on the example of cell phones that simple supervised machine learning techniques and small training sets already suffice to effectively segment real world reviews taken from CNET.com. Moreover, investigating the segment-specific rating behavior for individual usage profiles, we claim that also the more consistent ratings and the focus on task-specific problems or advantages of some product will generally benefit users and enhance the overall e-shopping experience.

Of course the preliminary results presented in the paper can only be a start to a thorough investigation of the topic. Our future work will deal with central questions like how to detect customers' usage intentions efficiently, in particular without long and cumbersome elicitation cycles. Similarly, for more heterogeneous usage profiles also the question of how to effectively extract and represent bias and diversity information remains.

6. REFERENCES

[1] Kumar, R., and Tomkins, A."A Characterization of Online Search Behavior". *Data Engineering Bulletin*, 32(2), pp. 3-11. 2009.

[2] Selke, J., Homoceanu, S. and Balke, W. "Conceptual Views for Entity-Centric Search: Turning Data into Meaningful Concepts". *BTW*. Kaiserslautern, Germany. 2011.

[3] Rosch, E. and Lloyd., B., L. "Principles of Categorization". pp. 27–48. Lawrence Erlbaum, 1978.

[4] Cortes, C. and Vapnik, V. "Support-vector networks". Machine Learning. 20:273--297, 1995.

[5] Joachims, T. "Text categorization with Support Vector Machines: Learning with many relevant features." *ECML*, 1998.

[6] Pang, B., Lee, L., and Vaithyanathan, S. "Thumbs up? Sentiment Classication using Machine LearningTechniques". *EMNLP*, (pp. 79-86). 2002.

[7] Hu, M., and Liu, B. "Mining Opinion Features in Customer Reviews". *AAAI*, (pp. 755-760). 2004.

[8] Hu M and Liu B. "Mining and Summarizing Customer Reviews". *SIGKDD*.168-177. 2004

[9] Liu, B., Hu, M., & Cheng, J. "Opinion Observer: Analyzing and Comparing Opinions on the Web". *WWW*. 2005.

[10] Pang B. and Lee L. "Opinion Mining and Sentiment Analysis". *Foundations and Trends in IR*.1-135. 2008.

[11] Zhai, Z., Liu, B., Xu, H. and Jia, P. "Clustering product features for opinion mining" *WSDM*. 2011.

[12] Ma, Q. and Yoshikawa, M. "Topic and Viewpoint Extraction for Diversity and Bias Analysis of News Contents". *APWeb/WAIM. LNCS* 5446, pp. 150–161, 2009.

[13] Laver, M., Benoit, K., and Garry, J. "Extracting policy positions from political texts using words as data". *American Political Science Review* 97:311–31. 2003.

[14] Lanny, W., M., and Vanberg, G. "A Robust Transformation Procedure for Interpreting Political Text". *Political Analysis* 16.1. pp. 93-100. 2007.