Visualization of Customer Expectations from Web Text using Co-occurrence Graph and Auto-labeling in the Service Market

Ryosuke Saga[1,2], Naoaki Ohkusa[2], Takafumi Yamashita[1] and Nahomi Maki[3]
[1]*Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, Japan*
[2]*School of Knowledge and Information Systems, College of Sustainable System Sciences, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, Japan*
[3]*Dept. of Information Media, School of Informatics, Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Japan*

Keywords: Information Visualization, Service Science, Customer Expectations, Co-occurrence Graph, Clustering, Auto-labeling.

Abstract: This study describes the visualization of customer expectations using the service science domain. Customer expectations influence service quality and are considered important factors for user evaluation of services. Customer expectations are constructed from word of mouth, rumors, and user experience. Investigation using a questionnaire is useful in comprehending customer expectations, but this method is costly and time consuming. In this research, we extract customer expectations from Web text consisting of massive word-of-mouth data and visualize them using a co-occurrence graph. In addition, we apply clustering and auto-labeling methods to easily understand the results of the co-occurrence graph. In the case study of a coffee service, we are able extract topics related to customer expectations, but labeling methods are still subject to improvement.

1 INTRODUCTION

Recently, the service industry has become dominant in the world market and occupied more than 70% of the gross domestic product in Japan. Additionally, each real product has shifted toward the service industry, thereby increasing the importance of service.

Service science is a domain that systematically understands services (Vargo et al., 2004; Maglio et al., 2010; Maglio et al., 2006). Service science examines concept of service beyond several domains. It is well-known that service has the feature called heterogeneity, which indicates that quality and productivity of service are not stable. Parasuraman et al. proposed a conceptual model related to the quality of service (Parasuraman, 1985). This conceptual model shows the gaps between customer and service provider.

A significant concept called *customer expectation* is included in this model. Customer expectation refers to the perceived benefits a customer expects before having the actual service experience. Customer expectation consists of several elements, including word of mouth, personal needs, and past experience. The evaluation of the quality of service comes under the influence of the gap between customer expectation and feeling of perceived service. If perceived service meets customer expectations then customers give a high evaluation; otherwise, customers' evaluation is low.

Customer expectation leads to customer satisfaction, and identifying customer expectation is important to provide proper quality of service that fills the gap between customer expectation and perceived service. Investigation using a questionnaire is useful in comprehending customer expectations, but this method is costly and time consuming.

The Web can now be used to check for a service or a product from Google and Amazon.com (Chakrabarti, 2003; Liu, 2008). Much useful word-of-mouth and product/service information that can build up customer expectations is found in these Web pages and e-commerce sites. These Web pages can be accessed by future customers who come under the influence of word of mouth.

Saga R., Ohkusa N., Yamashita T. and Maki N.
Visualization of Customer Expectations from Web Text using Co-Occurrence Graph and Auto-labeling in the Service Market.
DOI: 10.5220/0006238303170321
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 317-321
ISBN: 978-989-758-228-8

Therefore, based on the assumption that indications of customer expectations can be found on the Web, the research question *"Can we visualize customer expectations from the Web?"* is formulated. We attempt to visualize customer expectations from the Web to answer this question. In this visualization, we use co-occurrence graphs in which nodes show the keywords in Web pages and edges indicate the relationship among keywords. In this study, we target coffee service, which is one of the representative service products because of its heterogeneity that each person feels the different taste for each coffee. Then, the co-occurrence graph is utilized to show information, such as coffee brands and tastes, to represent customer expectations. Several topics can be found in the co-occurrence graph, but customer expectation is difficut to understand from the graph when the number of nodes is large. We label each topic automatically to use the machine learning method called *auto-labeling*.

The contributions of this paper are as follows:

- The first approach is visualization of the service science area. This approach provides visualization results and helps the service provider to easily comprehend the situation of service usage and evaluation.
- Clustering and auto-labeling are combined to understand the results of the co-occurrence graphs. This approach is valuable to this visualization domain and can be applied in other co-occurrence visualizations.

2 VISUALIZATION PROCESS

The visualization process is explained in this section. First, we extract keywords from Web pages and utilize texts without illustrations or visualized elements such as pictures and videos. However, customer expectations are difficult to discover and understand if the nodes appear in co-occurrence. Clusters with dense edges show topics for clustering the co-occurrence graph and adding labels automatically.

2.1 Keyword Extraction and Edge Extraction

We extract texts from Web pages related to services. We use the term frequency-inverse document frequency (TF-IDF) algorithm to extract the keywords obtained from the top ranked words. The TF-IDF algorithm is expressed as follows:

$$tf_i = \frac{n_i}{\sum_k n_k}. \quad (1)$$

$$idf_i = \log \frac{|D|}{\{d : d \ni t_i\}}. \quad (2)$$

$$tf_i idf_i = tf_i * idf_i. \quad (3)$$

In these formulas, n_i is the appearance of word t_i, $\Sigma_k n_k$ is the sum of all of words, $|D|$ is the number of documents, and d is the number of documents including the word t_i.

We then extract the edges. This study presents the edges using the Jaccard coefficient over a threshold. The Jaccard coefficient is expressed as

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \quad (4)$$

where $|A|$ is the number of documents with a word A and $|A \cap B|$ is the number of documents with words both A and B.

2.2 Clustering

Clustering is a method used to understand the causality and topics hidden in the data and to summarize, classify, and separate data. Several clustering methods, also known as community detection methods, are used for graph mining. In this study, we use the method proposed by Newman (Newman et al. 2004) that aggregates the nodes based on modularity, which has a value close to 1. Modularity is defined as follows:

$$Q = \Sigma_i (e_{ii} - a_i^2), \quad (5)$$

where e_{ij} is the percentage of the number of edges from cluster i to cluster j to the total edges, e_{ii} is the percentage of the number of internal edges in cluster i to the total number of edges, and a_i is the percentage of the number of edges connected to cluster i to the total number of edges. The Newman method focuses on ΔQ, as shown in formula (6):

$$\Delta Q = 2(e_{ij} - a_i a_j). \quad (6)$$

The process of the Newman method is as follows:

1. Allocate all nodes as clusters.
2. Calculate ΔQ for all pairs of clusters.
3. Aggregate two clusters with the highest ΔQ.
4. Repeat steps 2–4 until ΔQ <0.

2.3 Auto-labeling for Clusters

After clustering, we attach the labels to each cluster. We refer to the study of Mei et al. (2007) for the

process. Here, we regard a label as the combination between two words.

This method generates the label showing the meaning of topics for the topic models. The topic model is a statistical representation used to infer topics based on the assumption that each document shows and contains some topics. Given that the concept of a label includes the concept of a topic, Mei et al. (2007) estimated the relationships between label and topic. The relationships are regarded as an optimization problem with the minimization of Kullback–Leibler divergence (Kullback et al., 1951) between word distributions and the maximization of mutual information between label and topic. The relationships are based on the idea that a proper label has the feature that determines the low meaning gap between word and label.

2.3.1 Meaning Score

The meaning relationship between topic θ and label l is obtained by $Score(l,\theta)$, which is calculated by the Kullback–Leibler divergence expressed as formula (7):

$$\begin{aligned} Score(l,\theta) &= -D(\theta \| l) = -\sum_w p(w|\theta)\log\frac{p(w|\theta)}{p(w|l)} \\ &= -\sum_w p(w|\theta)\log\frac{p(w|C)}{p(w|l,C)} - \sum_w p(w|\theta)\log\frac{p(w|\theta)}{p(w|C)} \\ &\quad - \sum_w p(w|\theta)\frac{p(w|l,C)}{p(w|l)} \\ &= \sum_w p(w|\theta)\log\frac{p(w,l|C)}{p(w|C)\,p(l|C)} - D(\theta\|C) - \sum_w p(w|\theta)\frac{p(w|l,C)}{p(w|l)} \\ &= \sum_w p(w|\theta)PMI(w,l|C) - D(\theta\|C) + Bias(l,C) \end{aligned} \quad (7)$$

where w is the word, l is the label, θ is the topic, and C is the text collection (or corpus), i.e., the texts from the Web pages. PMI is the pointwise mutual information, and $D(\theta||l)$ shows the Kullback–Leibler divergence between θ and l.

2.3.2 Labeling of the Cluster in the Co-occurrence Graph

Using the concept presented in Sections 2.3.1 and 2.3.2, we label the clusters in the co-occurrence graph. The labeling process is as follows:

1. Specify the documents from the extracted nodes corresponding to the keywords of each cluster in the co-occurrence graph. The specified documents are initially regarded as those belonging to a cluster based on the assumption that specified documents (i.e., C in formula (7)) express the topic θ of the cluster.
2. Generate labels $L=\{l_1,l_2, \ldots, l_m\}$ from the keywords in the documents. Each label consists of multiple words.
3. $D(\theta||C)$ and $Bias(l,C)$ are not related to the score ranks and can be ignored because l and θ are generated from the same document collection C, and $D(\theta||C)$ is a constant value. Finally, score (l,θ) is calculated using the following formula:
4. $$Score(l,\theta) = \sum_w p(w|\theta)\log\frac{p(w,l)}{p(w)p(l)}. \quad (8)$$

2.3.3 Label Disambiguation

When several topics are found, each topic should have a different label. To summarize only one topic by a label, we modify the label score in formula (8) into formula (9).

$$Score'(l,\theta_i) \approx (1+\frac{\mu}{k-1})Score(l,\theta_i) - \frac{\mu}{k-1}\sum_{j=1\ldots k} Score(l,\theta_j), \quad (9)$$

where k is the number of topics and μ is a parameter. A label is attached to the most related topic using this formula.

3 CASE STUDY

3.1 Environment and Dataset

We target the coffee market, which is one of the representative services, as a case study of service using top 200 results acquired from google.co.jp by keyword “coffee”. During keyword extraction, we regard nouns and adjectives as keywords. We extract the top 500 keywords with high TF-IDF values using the TF-IDF algorithm. Then, we extract the edges using the Jaccard coefficient of over 0.3. The label consists of two keywords, and the label candidates are generated from the combination of two keywords appearing in a cluster. We set the parameter μ to 0.7 in this labeling. Moreover, we ask seven students to label each cluster from the visualized co-occurrence graph. The labeled results are called man-label.

3.2 Result

The result of the case study is illustrated in Figure 1 and Table 1. Figure 1 is illustrated by Kamada-Kawai layout (Kamada et al. 1989) and shows the clusters classified into different colors. One large

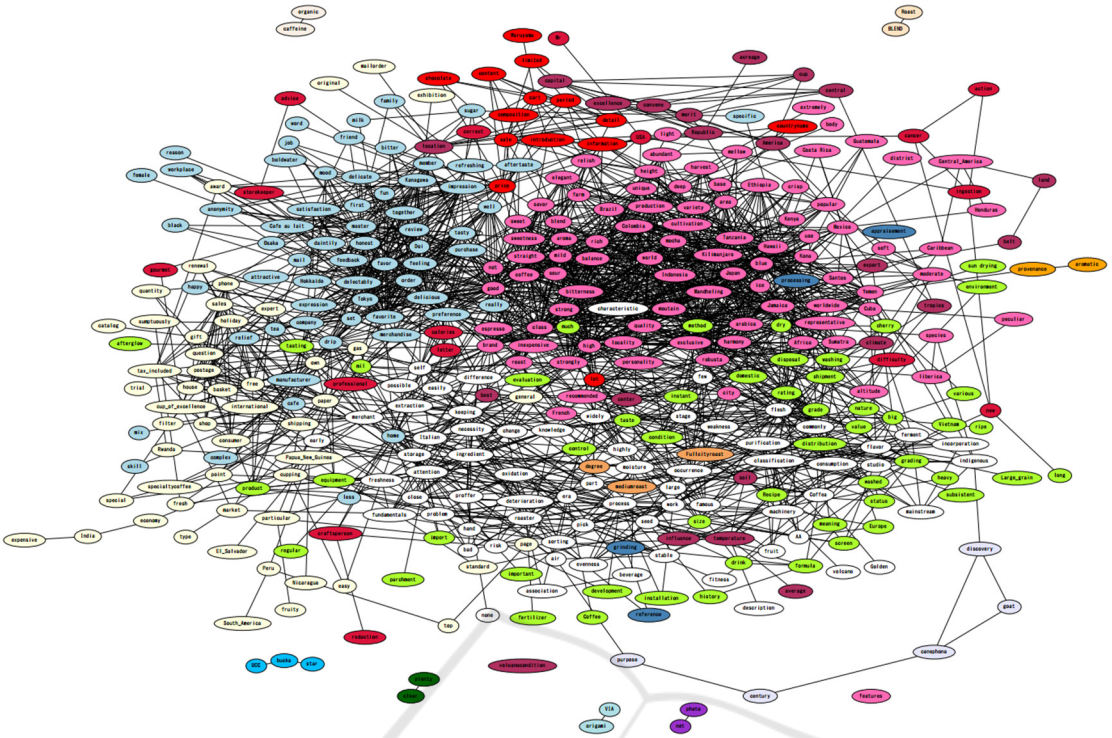

Figure 1: Visualization results from Web text about coffee service.

Table 1: System-generated labels and man-labels for each cluster.

ID	Colour	Auto Label	Man Label
1		Robusta Item	Rank, Production, Recipe
2		High Grade	Effect, Efficacy
3		Phone Selling	Store Selling, Sales
4		Customer Set	The way of Drinking, Preference, Place
5		Indonesia Robusta	Brand, Producing Area
6		Cup Weather	Country, Condition of growing, Land
7		Customer Update	Limited time offer, limited sale

component and other small components are also presented. In the large component, we can find seven clusters using the Newman method. Then, we label the cluster that includes more than six nodes because a topic of a cluster with only a few nodes can be easily predicted.

As shown in Table 1, we can extract the cluster for general customer expectation (e.g., clusters 1 and 5) indicating the coffee brand and the growing coffee production area. However, a gap can be observed between the results of the auto-label and that of the man-label because the latter is based on the co-occurrence graph and abstracted compared with the former.

Finally, we conclude the case study.

- We extract the cluster like customer expectation so that the visualization of customer expectations is succeeded by using web text and co-occurrence graph.
- Some labels by auto-labeling methods succeeded for some clusters to understand the clusters. However, there is room for improvement of this method.

4 CONCLUSION

This study described the visualization of customer expectations from Web text. For visualization, we utilized the co-occurrence graph consisting of keywords and co-occurrence relationships as nodes and edges, respectively. We conducted clustering and auto-labeling for each generated cluster to easily

comprehend the visualized graph.

In the case study, we visualized customer expectations on coffee services. For clustering, we used the Newman method and the auto-labeling method proposed by Mei et al. Consequently, we showed customer expectations using the co-occurrence graph, but the result of auto-labeling differed from that of man-labeling.

ACKNOWLEDGEMENTS

This research was supported by MEXT/JSPS KAKENHI 25420448, 25240049, and 16K01250.

REFERENCES

Vargo, S. L., Lusch, R. F., 2004. Evolving to a new dominant logic for marketing. In *Journal of Marketing*. Vol. 68, pp. 1-17.

Maglio, P. P., Kieliszewski, C. A., Spohrer, J. C., 2010. *Handbook of Service Science*, Springer.

Maglio, P. P., Srinivasan, S., Kreulen, J. T., Spohrer, J., 2006. Service systems, service scientists, SSME, and innovation. In *Communications of the ACM*. Vol 49, pp. 81-85.

Parasuraman, A., Zeithaml, V. A., Berry, L. L., 1985. A conceptual model of service quality and its implications for future research. In *Journal of Marketing*. Vol. 49, No. 4, pp. 41-50.

Chakrabarti, S., 2003. *Mining the web: discovering knowledge from hypertext data*, Morgan Kaufmann Publishers, Massachusetts, 1st edition.

Liu, B., 2008. *Web data mining: exploring hyperlinks, contents, and usage data,* Springer, Berlin, 2nd edition.

Newman, M. E. J., Girvan. M., 2004. Finding and evaluating community structure in networks. In *Physical review E.* Vol. 69, No. 2, 026113.

Newman, M. E. J., 2004. Fast algorithm for detecting community structure in networks. In *Physical review E*. Vol. 69, No. 6, 066133.

Mei, Q., Shen, X., Zhai. C., 2007. Automatic labelling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 490-499.

Kullback, S., Leibler. R. A., 1951, On information and sufficiency. In *the annals of mathematical statistics.* Vol. 22, No. 1, pp. 79-86.

Kamada, T., Kawai, S., 1989. An algorithm for drawing general undirected graphs. In Information Processing Letters. 32 (1), 7-15.