ChinaRxiv 中X

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201805.00233

Analysis of Structural Features of High-Frequency Word Co-occurrence Networks in Online Shopping Reviews (Postprint)

Authors: Li Taoying, Lu Xiaoning, Li Feng

Date: 2018-05-20T00:00:00+00:00

Abstract

Online shopping reviews constitute direct feedback from consumers on purchased goods. Mining valuable knowledge from these reviews can provide a basis for merchants to conduct precision marketing and personalized recommendation services, as well as for consumers to make purchase decisions. In view of this, this study takes clothing online shopping reviews from domestic large-scale comprehensive e-commerce platforms as the research object, performs word segmentation on the reviews, filters high-frequency words, analyzes the co-occurrence relationships among high-frequency words, constructs a high-frequency word co-occurrence network, and reveals multiple structural characteristics of hot words in online reviews along with the finding that a few nodes in the review network play a dominant role in the network' s operation, thereby providing a research approach for the field of online shopping review mining to investigate the interactive influence of the structural characteristics of high-frequency word co-occurrence networks on sales volume.

Full Text

Analysis of Structural Characteristics of High-Frequency Word Co-occurrence Networks in Online Shopping Reviews

Li Taoying, Lyu Xiaoning, Li Feng
(Transportation Management College, Dalian Maritime University, Dalian, Liaoning 116026, China)

Abstract: Online shopping reviews represent direct consumer feedback on purchased products. Mining valuable knowledge from these reviews provides crucial support for businesses to implement precision marketing and personalized recommendation services, as well as for consumers to make informed purchase decisions. This study examines clothing reviews from a major domestic e-commerce platform, performing word segmentation and filtering high-frequency terms to analyze co-occurrence relationships among them. By constructing a high-frequency word co-occurrence network, we identify multiple structural features of hotspot terms in review networks and demonstrate that a small number of nodes play a dominant role in network dynamics. This research offers a novel approach to investigating the interactive effects between the structural characteristics of high-frequency word co-occurrence networks and sales performance in the domain of online review mining.

Keywords: online shopping reviews; high-frequency words; co-occurrence network; sentiment analysis

0 Introduction

The rapid development of the Internet and widespread adoption of digital technologies have made online shopping an efficient and convenient purchasing method that profoundly influences daily life and work. As the number of online shoppers continues to grow, the volume of product reviews has increased exponentially, making it increasingly difficult for potential consumers to identify useful information for purchase decisions. Moreover, due to variations in linguistic habits, reviews tend to be unstructured and disorganized. Each review may address different product aspects, creating significant challenges for customers seeking to make purchase decisions, businesses aiming to improve sales performance, and regulatory agencies attempting to monitor market activities. Directly browsing through review content is not only time-consuming but also fails to provide an objective, holistic impression.

Existing literature on online review mining has concentrated on three primary areas:

a) **Review authenticity and usefulness** [2–8]. For instance, Liao et al. [2] employed empirical methods to investigate factors influencing the usefulness of online product reviews and their underlying mechanisms. Jiang et al. [4] argued that the varying quality of reviews significantly compromises the accuracy and credibility of demand mining, proposing a complex network-based approach for review usefulness analysis [5,6]. Other studies have examined behavioral factors influencing consumers' willingness to post reviews, such as motivational incentives [7], including small financial rewards that encourage review posting but may lead to biased or inauthentic content [9], with some reviews being entirely fabricated [10]. This stream of research primarily addresses whether reviews genuinely reflect consumer opinions and their overall quality and effectiveness.

b) **Impact of reviews on purchase decisions or sales** [13–23]. Wang and Yan [13] investigated whether reviews of popular brand products influence sales of less popular brands. Li et al. [14] examined how negative review quality, consumer involvement, and gender affect consumer satisfaction and purchase choices. Gong et al. [21] conducted an empirical analysis using book review data from Dangdang.com, finding that online consumer reviews significantly impact book sales. Other researchers have analyzed differential effects of initial versus follow-up reviews on purchase decisions [22]. These studies primarily explore whether and how review content influences product sales across various dimensions.

c) **Sentiment analysis of reviews** [24–29]. Li and Ye [24,25] analyzed relationships between sentiment orientation in product reviews and sales revenue, noting that inadequate review content quality leads to significant errors in sentiment analysis results [27]. Wang et al. [28] proposed a paragraph-level sentiment polarity classification method based on sentence-level sentiment, accounting for expression habits and corpus granularity. Wang and Wang [29] suggested employing coarse-grained sentiment polarity identification for similar product recommendations while using fine-grained identification for individual product features. This research stream focuses on subjective content identification, sentiment classification, and economic value extraction from online reviews.

Prior research has primarily examined how external factors such as review timeliness, incentive mechanisms, multiple reviews, and linguistic features influence consumers, or how reviews affect product sales and sentiment orientation. However, few studies have investigated consumer focus areas, relationships between reviews, and co-occurrence characteristics. Building upon existing online review mining research, this paper employs complex network theory to identify consumer attention hotspots by constructing a high-frequency word co-occurrence network from online shopping reviews and analyzing its structural properties.

---

1 High-Frequency Word Co-occurrence Network for Chinese Online Shopping Reviews

The review data used in this study were collected from a major Chinese integrated e-commerce platform—one of the world' s top ten Internet companies—with nearly 600 million yuan in cumulative sales during 2014–2015 and approximately 31.5 million product varieties. The platform generated a massive volume of online shopping reviews. This analysis focuses specifically on 59,730 reviews of 543 clothing items sold on the platform during 2014–2015, using this dataset to construct a high-frequency word co-occurrence network for online shopping reviews.

1.1 Chinese Word Segmentation and High-Frequency Word Extraction

Word segmentation is a critical preprocessing step that occurs after removing stop words such as "的" (de), "得" (de), and various punctuation marks. Common segmentation methods include string-matching-based approaches, statistical methods, and knowledge-understanding-based methods. Given the large dataset, this study employs a statistical segmentation method that analyzes the co-occurrence frequency of adjacent characters in context—higher co-occurrence frequencies indicate greater likelihood of forming a valid word. This approach effectively reflects word formation credibility [30,31]. Existing Chinese analysis software such as the Institute of Computing Technology Chinese Lexical Analysis System (ICTCLAS) and Wuhan University Professor Shenyang' s ROST CM can also be utilized for segmentation and word frequency statistics.

After segmentation, we obtain review words c_i with corresponding frequencies n_i. The probability of word c_i appearing across all N reviews is $p_i = n_i/N$. Words are then ranked by frequency (or probability) in descending order, ensuring $n_i \geq n_j$ (equivalently $p_i \geq p_j$) for all $i < j$. By setting a threshold K for the number of high-frequency words, we select the top K words as our high-frequency vocabulary.

1.2 High-Frequency Word Co-occurrence Matrix and Network Construction

Among the K high-frequency words from clothing reviews, we count the co-occurrence frequency e_{ij} of any two high-frequency words c_i and c_j across all N reviews. The co-occurrence relationship between any two high-frequency words is measured using mutual information from information theory [32], which quantifies the degree of association between words. The calculation formula is:

$$I_{i,j} = \log_2 \frac{P_{i,j}}{P_i P_j}$$

where $P_{i,j}$ represents the probability of c_i and c_j co-occurring, P_i represents the probability of c_i appearing, and P_j represents the probability of c_j appearing. Larger $I_{i,j}$ values indicate stronger co-occurrence relationships between c_i and c_j. The matrix formed by $I_{i,j}$ constitutes the high-frequency word co-occurrence matrix (which can be represented as an upper or lower triangular matrix due to symmetry $I_{i,j} = I_{j,i}$).

The rationale for selecting mutual information over raw co-occurrence counts can be illustrated through an example. Suppose we have 10,000 reviews, with high-frequency words c_1 and c_2 having frequencies $n_1 = 8000$ ($P_1 = 0.8$) and $n_2 = 7000$ ($P_2 = 0.7$), and co-occurring 5000 times ($P_{1,2} = 0.5$), yielding mutual information $I_{1,2} = -0.36$. In contrast, high-frequency words c_3 and c_4 have frequencies $n_3 = 5000$ ($P_3 = 0.5$) and $n_4 = 5000$ ($P_4 = 0.5$), co-occurring

4500 times ($P_{3,4} = 0.45$), with mutual information $I_{3,4} = 0.85$. Although c_3 and c_4 have lower absolute co-occurrence counts, they appear together in nearly all instances relative to their individual frequencies, making their co-occurrence relationship more significant. Alternative measures such as association rule confidence could also replace mutual information.

Based on the co-occurrence matrix, we rank relationships by mutual information in descending order and set a threshold E for the number of high-frequency word co-occurrence relationships. Selecting the top E relationships as edges and the high-frequency words involved in these edges as nodes forms the high-frequency word co-occurrence network.

For example, setting $K = 200$ high-frequency words and $E = 100$ co-occurrence relationships yields the network shown in [Figure 1: see original paper], where node size represents degree. The network contains exactly E edges, with nodes drawn from the 200 high-frequency words—only node pairs with mutual information rankings in the top 100 are included as edges in the network.

---

2 Structural Characteristics of Online Shopping Review High-Frequency Word Co-occurrence Networks

2.1 Network Topology

The topological structure of the online shopping review high-frequency word co-occurrence network is illustrated in [Figure 1: see original paper]. In this network, each node represents a high-frequency word from clothing reviews, and each edge represents a co-occurrence relationship where two high-frequency words appear together in a single review. Node size indicates degree—the number of neighboring nodes connected to it. Node degree serves as a key metric for measuring node importance in the network, with high-degree nodes often referred to as hub nodes.

Without loss of generality, assuming the network contains N nodes, we construct a binary adjacency matrix $A(N, N)$ based on co-occurrence relationships. Matrix element $a_{i,j}$ equals 1 if a co-occurrence relationship exists between high-frequency words i and j, and 0 otherwise. This symmetric matrix $A(N, N)$ enables calculation of structural properties including shortest path length, network density, degree distribution, clustering coefficient, community structure, rich-club phenomenon, and matching patterns.

2.2 Small-World Characteristics

Many real-world networks exhibit small-world properties, characterized by average path lengths comparable to those of random networks of the same size but with significantly higher clustering coefficients [33]. Network average path length is the mean of shortest path lengths between all node pairs:

$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \neq j} d_{ij}$$

where d_{ij} represents the number of edges between high-frequency word nodes i and j. The clustering coefficient is the average of individual node clustering coefficients:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i = \frac{1}{N} \sum_{i=1}^{N} \frac{2N_i}{k_i(k_i-1)}$$

where k_i is the degree of node i and N_i is the number of edges actually existing among its k_i neighbors.

For the network in [Figure 1: see original paper] with $N = 18$ nodes, calculations yield an average path length of 1.34 and a clustering coefficient of 0.84. Compared to a corresponding random network, the online shopping review high-frequency word co-occurrence network exhibits a comparable average path length but substantially higher clustering coefficient, demonstrating clear small-world characteristics. These results indicate that any two high-frequency words in the network are connected through co-occurrence relationships requiring at most one intermediate step on average, with over half of high-frequency word pairs having direct co-occurrence relationships.

These small-world properties have practical implications for consumer decision-making. For example, if a consumer searches for reviews containing the high-frequency word "comfortable," the platform should automatically recommend co-occurring terms such as "good-looking" and "well-fitting" to provide personalized suggestions.

2.3 Degree Distribution Characteristics

Node degree distribution is a crucial metric for describing complex network structures. Existing literature commonly employs either the degree distribution function $P(k)$ or the cumulative degree distribution function to characterize network properties. The former represents the proportion of nodes with degree k, while the latter represents the proportion of nodes with degree greater than or equal to k [34]. Empirical research indicates that most real-world complex networks exhibit one of three degree distribution patterns: (1) scale-free, (2) broad-scale, or (3) single-scale characteristics. This study uses the cumulative degree distribution function to analyze the degree distribution of the clothing review high-frequency word co-occurrence network.

The cumulative degree distribution function for the network in [Figure 1: see original paper] is shown in [Figure 2: see original paper]. To ensure representativeness, we also calculated cumulative degree distributions for networks with $K = 500, E = 200$ ([Figure 3: see original paper]) and $K = 1000, E = 500$

([Figure 4: see original paper]), with results presented in [Figure 5: see original paper].

As shown in [Figure 2: see original paper] and [Figure 5: see original paper], as the number of nodes increases and degree k grows, the cumulative degree distribution curves exhibit rapid initial decay followed by slower decay, indicating scale-free characteristics. Scale-free networks feature a small number of highly connected nodes that dominate network dynamics, while most nodes have few connections. The clothing review high-frequency word co-occurrence network conforms to this pattern, with nodes such as quality, color, style, size, and fabric serving as key hubs with numerous connections while other nodes remain sparsely connected.

2.4 Rich-Club Phenomenon

The rich-club phenomenon describes how high-degree nodes (hub nodes) form tightly interconnected cores within networks. This can be measured using the rich-club coefficient $\phi(k)$ [35]. Let $E_{>k}$ denote the number of connections among nodes with degree greater than k. The rich-club coefficient is defined as:

$$\phi(k) = \frac{E_{>k}}{N_{>k}(N_{>k}-1)/2}$$

where $N_{>k}$ is the number of nodes with degree greater than k, and $N_{>k}(N_{>k}-1)/2$ represents the maximum possible number of connections among these nodes. The rich-club coefficient for the online shopping review high-frequency word co-occurrence network is shown in [Figure 6: see original paper], with coefficients for networks from [Figure 1: see original paper] and [Figure 3: see original paper] displayed as (a) and (b) respectively. The coefficient increases with node degree k, indicating that hub words in the review network are more densely interconnected than low-degree nodes, forming a rich club. Notably, nodes with degree greater than 10 form a fully connected subgraph. This phenomenon demonstrates that reviews within the club constitute the network core, controlling the overall structure of review nodes and attracting substantial consumer attention—critical factors for driving purchases.

2.5 Matching Patterns

Matching patterns describe relationships between a node' s degree and its neighbors' degrees [34]. Statistical analysis involves two steps. First, calculate the average neighbor degree for node i:

$$k_{nn,i} = \frac{1}{k_i}\sum_{j\in N_i} k_j$$

where N_i is the neighbor set of node i. Second, compute the statistical average of neighbor degrees for nodes with identical degree k:

$$k_{nn}(k) = \frac{1}{N_k} \sum_{i:k_i=k} k_{nn,i}$$

where N_k is the number of nodes with degree k. If $k_{nn}(k)$ increases with k, high-degree nodes preferentially connect with other high-degree nodes, indicating assortative mixing. Conversely, if $k_{nn}(k)$ decreases with k, the network exhibits disassortative mixing. [Figure 8: see original paper] illustrates the relationship between node degree and neighbor degree in the review network, showing that $k_{nn}(k)$ decreases as k increases, confirming that the network is disassortative. This indicates that high-degree nodes preferentially connect with low-degree nodes, suggesting that during network evolution, newly entering reviews tend to establish connections with already well-connected reviews, and co-occurrence relationships preferentially form between high-connectivity and low-connectivity reviews.

2.6 Community Structure

Community structure refers to the presence of groups where intra-group connections are dense while inter-group connections are sparse. This study employs the modularity function defined by Girvan and Newman to identify community structures in the clothing review high-frequency word co-occurrence network [36]. Using the GN algorithm' s modularity metric Q, we partition the network into c subnets and define a symmetric matrix $E = (e_{ij})_{c\times c}$, where e_{ij} represents the proportion of edges connecting subnets i and j relative to total network edges. The trace of this matrix, $\sum_i e_{ii}$, describes the proportion of edges connecting nodes within the same subnet. Additionally, $b_i = \sum_j e_{ij}$ describes the proportion of edges connected to nodes in subnet i. The modularity function is:

$$Q = \sum_i (e_{ii} - b_i^2) = \mathrm{Tr}(E) - \|E^2\|$$

where $\mathrm{Tr}(E)$ denotes the sum of diagonal elements. Better partition quality yields larger Q values, though Q never exceeds 1. Values closer to 1 indicate more clearly distinguishable subnet structures, with typical real-world networks exhibiting Q values between 0.3 and 0.7. Notably, Q can be negative when intra-subnet edge proportions fall below random connection expectations.

For our clothing review high-frequency word co-occurrence network, $Q = -0.082$, with the network partitioned into 4 subnets. This maximum modularity value is significantly below 0.3, indicating that the four-subnet partition is not meaningful and community structure is not prominent. The negative Q value suggests all high-frequency words form tightly interconnected relationships without clear community boundaries.

---

3 Conclusion

Based on complex network theory, this study abstracts online shopping review relationships into network form—using high-frequency review words as nodes and co-occurrence relationships as edges—to construct and analyze the structural characteristics of a clothing review high-frequency word co-occurrence network. The main findings are:

a) The network exhibits small-world properties, where any two review words are connected through at most one intermediate step on average, with over half of high-frequency word pairs having direct co-occurrence relationships.

b) The degree distribution follows a scale-free pattern, where a minority of highly connected nodes dominate network dynamics while most nodes have few connections, indicating that online shoppers focus on a limited set of key factors.

c) The network displays a rich-club phenomenon, where hub words form the core of the review network and represent highly focused consumer attention, significantly influencing purchase decisions.

d) The network is degree-disassortative, with high-degree nodes preferentially connecting to low-degree nodes.

e) No significant community structure exists, as evidenced by a negative modularity index $Q = -0.082$, indicating that all high-frequency words form tightly integrated connections.

Furthermore, this methodology provides e-commerce retailers with comprehensive insights into their products and consumer expectations, while also supporting consumers in developing purchase strategies and regulatory agencies in formulating supervision policies. Future research analyzing how these structural characteristics interact with consumer behavior, product sales, and product popularity to inform precision marketing strategies represents a highly promising direction.

References

[1] Yang Ming, Qi Wei, Yan Xiangbin, et al. Utility analysis of online product reviews [J]. Journal of Management Sciences, 2012, 15(5): 65-75.

[2] Liao Chenglin, Cai Chunjiang, Li Yi. Empirical study on factors influencing the usefulness of online reviews in e-commerce [J]. Soft Science, 2013, 27(5): 46-50.

[3] Nie Hui. Evaluation and prediction of user review quality based on content analysis [J]. Library and Information Service, 2014, 58(13): 83-89.

[4] Jiang Wei, Zhang Li, Dai Yi, et al. Online review usefulness analysis for user requirement acquisition [J]. Chinese Journal of Computers, 2013, 36(1): 119-131.

[5] Titov I, Mcdonald R. Modeling online reviews with multigrain topic models [C]// Proc of the 17th International Conference on World Wide Web. 2008: 111-120.

[6] Yang L, Xiangji H, Aijun A, et al. Modeling and predicting the helpfulness of online reviews [C]// Proc of the 8th IEEE International Conference on Data Mining. 2008: 443-452.

[7] Cheung C M, Lee M K. What drives consumers to spread electronic word of mouth in online consumer-opinion platforms [J]. Decision Support Systems, 2012, 53(1): 218-225.

[8] Yan Qiang, Meng Yue. Factors influencing perceived usefulness of online reviews—An empirical study based on online movie reviews [J]. Chinese Journal of Management Science, 2013, 21: 126-131.

[9] Fu Dongpu, Wang Kanliang. Impact of review rewards on online product reviews—A social relationship perspective [J]. Journal of Management Sciences in China, 2015, 18(11): 1-12.

[10] Meng Meiren, Ding Shengchun. Analysis of behavioral motivations for fake review posters [J]. Information Science, 2013, 31(10): 100-104.

[11] Wang Changzheng, He Shan, Wang Kui. Research on usefulness perception of follow-up reviews in online word-of-mouth [J]. Management Science, 2015, 28(3): 102-114.

[12] Yin Guopeng. What makes online reviews more useful in consumers' eyes? —The effect of social factors [J]. Management World, 2012, 12: 115-124.

[13] Wang Junjun, Yan Qiang. Empirical study on the impact of online reviews on sales across product categories with different popularity [J]. Chinese Journal of Management Science, 2013, 21: 406-411.

[14] Li Hong, Yu Kui, Xia Jingbo. Impact of negative online reviews on consumers' online purchase decisions: An experimental study [J]. Journal of Intelligence, 2011, 30(5): 202-207.

[15] Feng Jiao, Yao Zhong. Research on the impact of online review information on purchase decisions based on social learning theory [J]. Chinese Journal of Management Science, 2016, 24(9): 106-114.

[16] Li Xinxin, Hitt L M. Self-selection and information role of online product reviews [J]. Information Systems Research, 2008, 19(4): 456-474.

[17] Mudambi S M, Schuff D. What makes a helpful online review? A study of customer reviews on amazon.com [J]. MIS Quarterly, 2010, 34(1): 185-200.

[18] Kwark Y, Chen Jianqing, Raghunathan S. Online product reviews: Implications for retailers and competing manufacturers [J]. Information Systems Research, 2014, 25(1): 93-110.

[19] Zhang Xuan, Jing Fengjie. Impact of product types on online impulse buying behavior [J]. Management Science, 2012, 25(3): 69-77.

[20] Zhang Li. Relationship between consumers' traditional offline shopping habits and online shopping behavior and its practical significance [J]. Management Science, 2006, 19(2): 13-21.

[21] Gong Shiyang, Liu Xia, Zhao Ping. How do online consumer reviews affect product sales?—An empirical study based on online book reviews [J]. China Soft Science, 2013, 6: 171-173.

[22] Shi Wenhua, Gong Xue, Zhang Qi, et al. Comparative study of initial online reviews and follow-up reviews [J]. Management Science, 2016, 29(4): 45-58.

[23] Du Xuemei, Ding Jingyu, Xie Zhihong, et al. Impact of online reviews on consumers' purchase intention [J]. Management Review, 2016, 28(3): 173-183.

[24] Zhang Ziqiong, Ye Qiang, Li Yijun. Review of sentiment analysis research for Internet product reviews [J]. Journal of Management Sciences in China, 2010, 13(6): 84-96.

[25] Hao Yuanyuan, Zou Peng, Li Yijun, et al. Empirical study on the impact of online review sentiment orientation on sales revenue based on movie panel data [J]. Management Review, 2009, 21(20): 95-103.

[26] Li Yongzhong, Hu Siqi. Sentiment analysis of online shopping reviews based on HowNet and PAT trees [J]. Library and Information Studies, 2016, 3: 66-70.

[27] Hu M, Liu B. Mining and summarizing customer reviews [C]// Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 168-171.

[28] Wang Hongwei, Zheng Lijuan, Yin Pei, et al. Sentiment polarity classification of Chinese online reviews based on sentence-level sentiment [J]. Journal of Management Sciences in China, 2013, 16(9): 64-74.

[29] Wang Wei, Wang Hongwei, Meng Yuan. Collaborative filtering recommendation algorithm research: Considering online review sentiment orientation [J]. Systems Engineering—Theory & Practice, 2014, 34(12): 3238-3249.

[30] Jiang Jianhong, Zhao Songzheng, Luo Mei. Research and application of Chinese word segmentation model combining dictionary and statistical methods [J]. Computer Engineering and Design, 2012, 33(1): 387-391.

[31] Zhang Min, Wang Chunhong. Research on web new word segmentation method based on statistical methods [J]. Computer Engineering & Science, 2010,

32(5): 133-135.

[32] Huang Degen, Ma Yuxia, Yang Yuansheng. Chinese name recognition method based on mutual information [J]. Journal of Dalian University of Technology, 2004, 44(5): 744-748.

[33] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393(4): 440-442.

[34] Newman M. E. J. The structure and function of complex networks [J]. SIAM Review, 2003, 45(2): 167-256.

[35] Zhou S, Mondragon R J. The rich-club phenomenon in the Internet topology [J]. IEEE Communications Letters, 2004, 8(3): 180-182.

[36] Girvan M, Newman M E J. Community structure in social and biological networks [C]// Proc of National Academy of Sciences of the USA, 2002, 99(12): 7821-7826.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.