Postprint: Dietary Topic Discovery from User-Generated Content in Online Social Networks

Authors: Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】 To conduct topic detection through large-scale text clustering techniques and automatically select high-quality topics. 【Method】Using diet-related Weibo content on Sina Weibo as the data source, topic detection is performed by combining text clustering and deep learning knowledge. By matching the month of Weibo publication, posts are divided into four seasonal categories; using vector space model and text clustering methods, topic detection is conducted on Weibo posts from different seasons to obtain candidate topics; integrating deep learning knowledge, the concept of topic coverage rate is proposed to automatically evaluate topic quality and eliminate low-quality topics. 【Results】The topic filtering results based on topic coverage rate conform to manual selection expectations, successfully extracting high-quality topics with a topic coverage rate exceeding 0.5. 【Limitations】 The evaluation of topic detection quality is primarily based on qualitative assessment. 【Conclusion】 By calculating topic coverage rate to automatically select high-quality topics, this method demonstrates high efficiency and strong generalizability; the obtained topics are readily comprehensible and effectively reveal the topic distribution of diet-related Weibo posts across the four seasons.

Full Text

Preamble

Research on Identifying Food Topics from User-Generated Content in Online Social Networks
ChinaXiv Cooperative Journal

Zhang Xiaoyong[1,2], Zhou Qingqing[1,2], Zhang Chengzhi[1,2,3]
[1](School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)
[2](Alibaba Research Center for Complex Sciences, Hangzhou Normal University,

Hangzhou 311121, China)
[3](Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing 210093, China)

Abstract

[**Objective**] This study aims to detect topics through large-scale text clustering techniques and automatically select high-quality topics. [**Methods**] Using food-related microblog content from Sina Weibo as the data source, we combine text clustering with deep learning for topic detection. By matching microblogs with their publication months, we categorize them into four seasonal groups. Using the vector space model and text clustering methods, we detect topics for different seasons to obtain candidate topics. Drawing on deep learning knowledge, we propose the concept of topic coverage to automatically evaluate topic quality and eliminate low-quality topics. [**Results**] The topic selection results based on topic coverage align with manual selection expectations, successfully extracting high-quality topics with topic coverage values above 0.5. [**Limitations**] Topic detection quality evaluation is primarily based on qualitative assessment. [**Conclusions**] By calculating topic coverage to automatically select high-quality topics, this method is efficient, highly generalizable, and produces easily understandable topics that effectively reveal the distribution of food-related microblog topics across the four seasons.

Keywords: Topic detection; User-generated content; Topic coverage; Food mining

1. Introduction

The development of Web 2.0 concepts and technologies has driven the rapid growth of social media. Various social platforms provide great convenience for user communication, and increasingly more people share their views on topics through social networks. Meanwhile, with improving living standards, public attention to diet has grown significantly, with users sharing food experiences, recommending recipes, discussing dietary benefits, and searching for local specialties on social networks.

As a primary platform for information acquisition and sharing, Weibo contains substantial food-related content. According to statistics, as of December 2015, Sina Weibo had 230 million users, with 36.7% sharing information about nearby food and attractions through the platform [1]. Therefore, conducting food topic detection based on Weibo data is both feasible and reliable.

The rapid popularization of social networks and unprecedented enthusiasm for online participation have led to an explosion of online information [2]. How to efficiently and accurately locate hot topics from complex, massive, and heterogeneous social network comments has long been a research focus in public opinion monitoring and competitive intelligence [3-5].

Traditional topic detection primarily targets ordinary text, obtaining topics through large-scale text clustering [6]. In this approach, topics are typically represented by all documents within a cluster, containing only category information that is not easily understood and often requires manual review to identify high-quality topics. This paper uses Sina Weibo as the research object, combining text clustering with deep learning knowledge for topic detection to achieve automatic selection of high-quality topics. In the text representation model, we screen feature words based on microblog corpus characteristics to address data sparsity and improve clustering efficiency. In the clustering process, we use the K-means algorithm to cluster microblogs and determine the total number of clusters based on clustering evaluation results to obtain candidate topics. By calculating topic coverage, we automatically evaluate topic quality, eliminate low-quality topics, and avoid the manual selection step, thereby improving topic detection efficiency.

Corresponding author: Zhang Chengzhi, ORCID: 0000-0001-8121-4796, E-mail: zhangcz@njust.edu.cn.

This work is supported by the National Social Science Fund Project "Research on User-Based Knowledge Organization Models in Online Social Networks" (Project No.: 14BTQ033), the National Social Science Fund Key Project "Research on Social Public Opinion and Decision Support Methodology System in Big Data Environment" (Project No.: 14AZD084), and the Jiangsu Provincial Graduate Student Research Innovation (Practice) Program Project "Multi-Granularity Movie Review Mining Based on Social Media" (Project No.: SJLX15_{0166}).

2. Related Work

2.1 Topic Detection and Tracking Research

The rapid development of the Internet has led to exponential growth of information resources, making efficient retrieval of hot topics a key focus in public opinion monitoring and competitive intelligence [3]. Topic Detection and Tracking (TDT) technology emerged to address this challenge, aiming to solve information overload problems [7] by automatically summarizing information related to topics for human review [8-9]. Current TDT research primarily focuses on online news reports and blogs, with emphasis on story segmentation, topic tracking, topic discovery, and new event detection [8].

Traditional topic discovery techniques mainly use clustering methods, including K-means algorithm [10-11], hierarchical clustering [12], centroid vector method [7, 13], and Single-Pass [13-14]. These methods have achieved good results in topic detection tasks for ordinary text, such as in TDT corpora [15]. However, these techniques typically represent topics using all documents within a cluster, which is not easily understood and often requires manual review to obtain high-quality topics. Additionally, with the rise of topic models [16-18], some studies have used LDA model [17] and its extensions to obtain topics [19]. For example, reference [20] extracted topics from scientific literature based on LDA topic

models, then calculated topic intensity and influence for trend analysis; reference [21] combined LDA models with adaptive clustering algorithms based on affine propagation for topic discovery; reference [22] proposed an MB-LDA model suitable for microblog topic mining by considering contact relationships and text associations in microblogs. The disadvantage of these techniques is that the extracted topic words have poor interpretability and high time costs.

In addition to these representative techniques, there are many other topic discovery methods with unique features. These methods each have their advantages, and there is currently no unified evaluation standard. Therefore, specific needs must be considered in practical applications. This paper comprehensively compares multiple algorithms for topic extraction and combines deep learning knowledge to propose an indicator called "topic coverage" to automatically evaluate topic quality, thereby improving the efficiency of topic detection.

2.2 Food Mining Research

Current food mining research is concentrated in fields such as history [23-26], sociology [27-29], and geography [30-31], aiming to study the impact of dietary culture changes on these domains. Due to the lack of systematic data support, related research has primarily been conducted through qualitative approaches such as field investigations and historical analysis [32], with few quantitative and systematic studies.

With the increasing abundance of online recipe data, quantitative research has begun to emerge in food mining. Reference [33] analyzed 56,498 recipes from multiple countries and regions, demonstrating that Western cooking tends to use multiple spices to create mixed flavors, satisfying the so-called Food Pairing Hypothesis, while Eastern cuisine does the opposite. Reference [34] analyzed small-scale recipes and concluded that climate is the main factor affecting chefs' choice of condiments; however, reference [35] statistically analyzed 8,498 recipes from 20 Chinese cuisines and proved that geographical distance has a greater impact on dietary habits than climate.

In summary, the rich recipe data on the Internet and massive food reviews on social networks have made quantitative research in the food domain possible. Existing food mining has focused on recipe data: analyzing the impact of geographical distance and climate on food preferences, exploring ingredient pairing preferences in different regions, etc., while topic discovery research based on food reviews is relatively scarce. This paper uses the vector space model and text clustering methods to obtain relevant topics from food reviews; combines deep learning knowledge to automatically select high-quality topics by calculating topic coverage, improving topic detection efficiency. Meanwhile, the experimental results effectively reveal the distribution characteristics of food topics in microblogs, which helps further explore consumer concerns and demands in the food domain.

3. Methodology

3.1 Research Framework

To mine food topics of interest to people from massive data, this paper uses Sina Weibo content as the research object for food topic discovery. Since the distribution of food topics varies significantly across seasons, we conduct topic detection separately for different seasonal microblogs. First, we collect food-related microblogs from Sina Weibo and categorize them into four seasonal groups based on publication month; second, we obtain topics based on text representation models and text clustering; finally, we select high-quality topics based on topic coverage combined with deep learning knowledge. The specific research framework is shown in Figure 1 [Figure 1: see original paper].

3.2 Microblog Content Representation Model and Feature Selection

This paper adopts the vector space model to represent food microblog content and screens feature items based on microblog corpus characteristics.

(1) Text Preprocessing
In the preprocessing stage, we use OPENCC[1] to convert traditional Chinese to simplified Chinese in microblog text, and use Jieba Chinese word segmentation[2] to complete word segmentation and part-of-speech tagging. Since there are numerous dish names in food microblogs, we add dish name data to Jieba's custom dictionary for segmentation.

(2) Vector Space Model
The vector space model [36] (Vector Space Model, VSM), proposed by Salton et al. in 1973, represents text as vectors in document space, with each selected feature term serving as one dimension of the text. Assuming the total number of feature items in the text space is M, the i-th text d_i can be represented as:

$$V(d_i) = (f_1, w_1(d_i); f_2, w_2(d_i); \ldots; f_M, w_M(d_i))$$

where f_j is the j-th feature item; w_j is the weight of feature f_j in text d_i. This paper uses the tf-idf algorithm to obtain weights, with the formula as follows:

$$w_j(d) = \mathrm{tf}_j(d) \times \log\left(\frac{N}{n_j}\right)$$

where $\mathrm{tf}_j(d)$ is the term frequency of feature f_j in document d, n_j is the total number of documents in the corpus containing term f_j (i.e., document frequency or DF value), and N is the total number of documents in the corpus.

(3) Feature Filtering Strategy
Due to the large corpus size, this paper uses individual words as feature items in the vector space. After word segmentation and filtering of all stop words, microblog short texts still contain many high-frequency words unrelated to topic

mining, such as emoticons and modal particles. Therefore, we first filter all emoticons from microblogs; then, by calculating the document frequency (DF) of feature words, we filter out the top 100 highest DF words and low-frequency words with DF values below 100 (both thresholds determined through manual verification). Examples of feature items to be filtered are shown in Table 1 .

[1]http://opencc.byvoid.com
[2]http://www.oschina.net/p/jieba

From Table 1, we can see that these high-frequency words and emoticons appear in most microblogs and have weak discriminative power; low-frequency words are mostly meaningless user nicknames with weak topic relevance, and thus can all be filtered out.

3.3 Text Clustering

This paper deals with a large dataset requiring clustering of approximately 5 million food-related microblogs. Considering time cost and topic interpretability, and after comprehensive comparison of multiple algorithms, we ultimately selected the K-means algorithm [37], which has the fastest running speed and best topic interpretability, for text clustering.

K-means is a prototype-based clustering technique (using cluster centroids as prototypes in this paper), where the centroid is the cluster center point. The algorithm randomly selects K initial centroids, where K is the user-specified total number of clusters; calculates the Euclidean distance between each point and the centroids, assigns each point to the nearest centroid, with the set of points assigned to a centroid forming a cluster; updates each cluster's centroid based on points within the cluster; repeats the assignment and update steps until the centroids no longer change, completing the clustering.

Since the K-means algorithm requires specifying the total number of clusters, this paper specifies cluster numbers K=10, 15…45, 50 for clustering respectively, and determines the final cluster count based on clustering evaluation results.

3.4 Topic Coverage Rate

To avoid the manual selection step for determining high-quality topics in traditional methods, we propose an indicator called "topic coverage" to evaluate topic quality by combining deep learning knowledge and word similarity calculation. Below we describe the key technologies and concepts of topic coverage and word similarity calculation.

(1) Topic Coverage Calculation
To quantitatively evaluate the quality of different topics, we reference the "intra-cluster cohesion" concept proposed in reference [38] and extend it with deep learning knowledge to propose the "topic coverage" concept.

Reference [38] defines two concepts: core representative features and core articles. Core representative features refer to the 20 features with the highest DF values in a cluster; core articles refer to articles containing more than m core representative features. Intra-cluster cohesion is ultimately defined as ic/N, where ic represents the total number of core articles and N represents the total number of articles in the cluster.

From the intra-cluster cohesion concept, we know that core articles are obtained based solely on statistical features without semantic association with core representative features. Due to the short text characteristics of microblogs and data sparsity, even when m=1, the proportion of microblogs meeting the criteria is extremely small.

To obtain semantic connections between core representative features and microblogs, this paper combines deep learning knowledge to propose the "topic coverage" concept: core representative features are defined as the top n features with the highest DF values in a cluster, denoted as Top-n; core microblogs are defined as microblogs with at least m terms having word similarity greater than p with core representative features. Taking Top-n=20, m=3, p=0.9, the topic coverage calculation formula is as follows:

To calculate word similarity, this paper uses the Distribute Representation method proposed by Hinton [39] to represent word vectors based on deep learning knowledge, aiming to express terms as low-dimensional and fixed-length real vectors, where similarity in vector space represents semantic similarity in text. Since this method is more suitable for large-scale computation, it has been widely used in recent years.

To use the above method, we utilize the Skip-Gram model in Word2Vec[1] to represent text, training on the entire segmented microblog corpus to convert words into 400-dimensional real vectors. Since Cosine distance is commonly used to measure differences between individuals, we calculate Cosine distance between word vectors to measure word similarity. The value range of Cosine distance is [-1, 1], with larger values indicating greater word similarity.

4. Experiments

4.1 Experimental Dataset

Dish name data comes from the Meishijie website[2], collected by Zhu et al. [35] in April 2012. The dataset covers 20 Chinese cuisines with a total of 8,498 dish names.

The food microblog data in this paper comes from Sina Weibo. We define microblogs containing the above dish names in their main text as "food microblogs" and collected the main text and basic user information of food microblogs from Sina Weibo for the entire year of 2013, totaling 8,747,190 entries. The microblog main text includes user ID, microblog content, and publication time, as shown in Table 2 :

Table 2 Example of Microblog Main Text Content
User ID | Content | Time
1785###### | A bowl of spicy noodles + braised pork knuckle + soy sauce egg + one liang of soup dumplings + one piece of fried pork chop = stuffed | 20:04:37
1700###### | Fried pork chop with spicy soy sauce is absolutely amazing! | 19:59:24

Basic user information includes user ID, gender, and location, as shown in Table 3 :

Table 3 Example of Basic User Information
User ID | Gender | Location
1000###### | Na Er### | Urumqi, Xinjiang
1000###### | Xiao Ruiqi### | (location data)

Based on user ID, we matched microblog main text with basic user information, filtered out microblog main text missing user basic information, and finally obtained 8,737,464 microblogs. Given that food topic distribution varies significantly across seasons, we categorized microblogs into four seasonal groups based on publication month for topic distribution detection in different seasons.

4.2 Experimental Results Analysis

(1) Seasonal Division of Food Microblogs
Based on the Gregorian calendar months of the four solar terms (Beginning of Spring, Beginning of Summer, Beginning of Autumn, Beginning of Winter) in the 2013 lunar calendar, we designated February-April 2013 as spring, May-July as summer, August-October as autumn, and January plus November-December of the previous year as winter. By matching the publication months of microblogs in the dataset and excluding 19,678 microblogs missing month information, we obtained a total of 8,717,786 food-related microblogs for each season in 2013.

During the clustering process for seasonal microblogs, we found that the microblog dataset contained a large number of "spam microblogs" that were too short and contained no topic information. These microblogs were numerous and seriously affected clustering efficiency and result interpretability. After screening microblog feature words, we filtered out 3,783,652 microblogs with fewer than 10 feature words (this parameter is empirical data), which greatly improved the interpretability of clustering results. We finally obtained 4,934,134 valid food-related microblogs across seasons in 2013. The total number of seasonal food microblogs before and after filtering is shown in Figure 2 [Figure 2: see original paper].

(2) Clustering and Clustering Evaluation
This paper uses the K-means algorithm to cluster microblogs for each season separately. Considering practical needs for topic detection, we specify the num-

ber of clusters between 10-50. Since the total number of topics varies across seasons, we specify cluster numbers K=10, 15…45, 50 for K-means clustering of seasonal food microblogs respectively, and determine the final cluster count based on clustering evaluation results.

To quantitatively evaluate clustering effectiveness, we use cohesion and silhouette coefficient as validity functions. K-means is a prototype-based clustering technique (using cluster centroids as prototypes), so we define cluster cohesion (SSE) as the sum of proximities of points to the cluster prototype [40]; to measure absolute distances between points in space, proximity dist() is generally measured by Euclidean distance. The calculation formula is as follows [40]:

$$\text{Cluster SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}(c_i, x)$$

where x represents an object and c_i represents the center of cluster C_i. Lower cohesion indicates smaller average distance between objects within a cluster and better intra-cluster cohesion.

The silhouette coefficient combines the advantages of cohesion and separation. The silhouette coefficient for an individual point is calculated as follows [40]:

For the i-th object, calculate the average Euclidean distance from i to all other objects in the same cluster, denoted as a_i;

For the i-th object and any cluster not containing it, calculate the average Euclidean distance from the object to all objects in the given cluster and find the minimum value, denoted as b_i;

For the i-th object, the silhouette coefficient is calculated as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

The silhouette coefficient value ranges between -1 and 1, with larger values indicating better clustering quality. By calculating the average silhouette coefficient of all objects, we can obtain a total measure of clustering quality.

Figures 3 [Figure 3: see original paper] and 4 [Figure 4: see original paper] show the cohesion and silhouette coefficients for different seasons with different numbers of clusters. Lower cohesion and higher silhouette coefficient indicate better clustering performance. We can see that both evaluation metrics show consistent trends with the number of clusters. Based on the distribution trends of cohesion and silhouette coefficient, we can determine that the optimal number of clusters is 50 for spring, 15 for summer, 45 for autumn, and 10 for winter.

(3) High-Quality Topic Selection Based on Topic Coverage

Traditional text clustering-based topic detection techniques generally require manual selection to obtain clusters with high topic significance after clustering results are obtained. This approach involves high manual participation and low

efficiency. This paper automatically scores and ranks the topic significance of each cluster by calculating its topic coverage, avoiding this problem through quantitative evaluation and greatly improving topic detection efficiency.

To demonstrate the effectiveness of this method, Table 4 uses spring clustering results as an example to rank clusters with different intra-cluster cohesion values. Each cluster is represented by the top 15 core representative features with the highest intra-cluster DF values.

Table 4 Example of Spring Clustering Results

Topic Coverage | Intra-cluster Cohesion | Topic Name | Core Representative Features

0.94 | 0.00025 | Throat swelling and pain | throat swelling#drink#water soak#throat#hair#honey dates walnut#dry#cucumber#ginger#light salt water#chapped#kiwi#lips#throat dryness#swelling and pain

0.91 | 0.00025 | New Year greetings | smooth#everything#prosperity#wealth#dragon horse spirit#smooth sailing#good health#all the best#no taboos#wealth

0.88 | 0.00025 | Recipe sharing | pour in#stir fry#a little#wash clean#cooking wine#evenly#heat up#fish out#scallion#starch#low heat#mix well#light soy sauce#soy sauce

0.85 | 0.00025 | Beauty care | skin#honey#ginger soup#goji berry#skin#improve#tea#beauty#milk#beauty care#heatiness#vinegar#pain#fire food#effect

0.82 | 0.00025 | Ingredient: Pumpkin | pumpkin#wash clean#peel#stir fry#paste#dough#pour in#pumpkin slices#pumpkin cake#slice#pumpkin porridge#small pumpkin#marinate#white sugar#appropriate amount

0.79 | 0.00025 | Beauty care | red dates#longan#goji berry#wash clean#tremella#lotus seeds#moisten lungs#emperor#yam#lily#goji berry porridge#beauty care#strengthen spleen#low heat#walnut

0.15 | 0.00025 | Unidentifiable | food#after drinking#fat#food#health#milk#diet#fruit#vegetables#vitamins

0.12 | 0.00025 | Unidentifiable | sun#breeze#shine#sunshine#weather#mood#happiness#walk#moon#aftern

0.08 | 0.00025 | Illness | illness#family#vomit#open#pitiful#restaurant#find#sad#sell#dizzy#boss#special# bites

0.05 | 0.00025 | Unidentifiable | walk#happiness#friend#cute#gift#special#time#China#bag#laugh#die#as

In Table 4, some topics are marked as “unidentifiable” because their core representative features lack strong correlation, making manual identification impossible. Experimental results show that clusters with higher topic coverage (>0.5) have higher topic significance, while clusters with too low topic coverage are difficult to interpret. This ranking based on topic coverage aligns with manual selection expectations, effectively explaining the distribution of food-related topics across the four seasons and confirming the feasibility of our method.

Table 4 compares the values of intra-cluster cohesion and topic coverage for topics of different quality. We can see that topic coverage values with added semantic association are more reasonable and more evenly distributed, such as for “recipe sharing” and “beauty care” topics. Due to data sparsity issues, the number of core microblogs obtained solely through statistical features is too low, resulting in very low intra-cluster cohesion values for most topics that cannot

effectively evaluate topic quality, such as the "illness" and other "unidentifiable" topics.

(4) Comparative Experiments

To demonstrate method effectiveness, we added two groups of comparative experiments: one based on the Doc Embedding model [41] combined with K-means clustering to obtain topics; the other based on the LDA topic generation model [17] to obtain topics. Using spring as an example, both experiments specified a total of 50 topics for topic detection on spring microblogs, with relevant result examples shown in Tables 5 and 6 respectively.

Table 5 Example of Spring Topics Obtained by Doc Embedding Technology

Core Representative Features (Top_n=10)
wash clean#pour in#stir fry#a little#cooking wine#fish out#low heat#cut into#light soy sauce#evenly
set meal#value#enjoy#group purchase#portion#sell#original price#choose#today#100
appropriate amount#50#30#20#materials#ingredients#wash clean#milk
pour in#cooking wine#stir fry#a little#low heat#starch#soy sauce#evenly#fish out#light soy sauce
food#health#honey#fire#skin#drink#effect#hair#fat#diet
squid#hot and sour noodles#potato#barbecue#hamburger#Chinese hamburger#small meatballs#fried#octopus#pizza
apple#lunch#milk#a cup#rice#banana#fruit#diet#morning#soy milk
birthday#thanks#happy birthday#gift#gift#wish#happy#dear#cute#thanks
recipe#a dish#Douguo#net#kitchen#world#collection#look#simple#complete collection
red dates#pot#stew#lily#wash clean#goji berry#yam#longan#tremella#lotus seeds

Table 6 Example of Spring Topics Obtained by LDA Model

Topic Words (Top_n=10)
1. Shandong#aunt#little#sweet and sour#crab roe#dark soy sauce#extremely#beef noodles#Ding Ding#Jinan
5. Chinese hamburger#rice noodles#stir-fried noodles#potato#gluten#fresh shrimp#crab pieces#spinach#clear stir-fry#tofu
7. Naxi#record#spare ribs#broad beans#meat slices#flavor#shredded chicken#Japanese style#lettuce#cooking skills
10. sushi#shepherd's purse#shredded meat#buckwheat#dried tofu#carp#pickled vegetables#West Lake#abalone#rose

Through the Doc Embedding technique, each microblog text is expressed as a 100-dimensional vector; topics are obtained through K-means clustering; topics are ranked based on topic coverage, with relevant topics represented by 10 core representative features. From Table 5, we can see that compared with spring topics in Table 4, topics obtained by this method have poorer interpretability and more homogeneous topic types. Additionally, topic quality ranking based on topic coverage aligns with manual selection expectations, again proving the effectiveness of this indicator.

Based on part-of-speech and statistical features, we filter out topic-unrelated terms; topic distribution is obtained through LDA topic modeling. Each topic is represented by the 10 topic words that best represent the theme. From Table 6, we can see that topics obtained by this method are difficult to interpret, and the distinction between topics is also low. The composition pattern of topic words is basically "location+person+food or ingredient", such as "Shandong+aunt+crab roe" in topic 1 and "canteen+dad+specialty" in topic 2.

Through the above comparative experiments, we can see that the topic detection method combining vector space model and text clustering technology in this paper obtains topics with stronger interpretability, and the topic distribution across seasons also matches actual conditions; the topic quality evaluation method based on topic cohesion is efficient and highly generalizable, and can replace the manual selection step for high-quality topics.

(5) Topic Distribution Difference Analysis

To measure the distribution of topics within each season, Figure 5 [Figure 5: see original paper] shows the distribution of topic coverage across seasons.

From Figure 5, we can see that spring and autumn have more topics, and the number of topics with high topic coverage is also much higher than in summer and winter. Therefore, taking representative topics with topic coverage higher than 0.4 in each season as an example, we categorize topics into several major categories to compare and analyze topic distribution differences and their causes across the four seasons. After manual review, this paper categorizes seasonal food topics into four major themes: "benefits", "festivals", "cooking", and "travel", with relevant topics in each season enumerated under each major theme. Topic examples are shown in Table 7 .

Table 7 Examples of Representative Topics Across Seasons

Spring Representative Topics	Autumn Representative Topics
Benefits	Cough relief#cough#honey#radish#wind-cold#white radish#ginger jujube soup#cold#fresh pear#phlegm removal	Fire#drink#water soak#throat#hair#honey dates walnut#dry#cucumber#ginger#light salt water
Benefits	Skin#honey#ginger soup#goji berry#skin#improve#tea#beauty#milk#beauty care	Red dates#longan#goji berry#wash clean#tremella#lotus seeds#moisten lungs#emperor#yam#lily
Benefits	Cough relief#cough#honey#radish#wind-cold#cold#ginger jujube soup#white radish#fresh pear#lung abscess	Fire#water soak#throat#hair#cucumber#chapped#kiwi#lips#salt water
Benefits	Fluid production#phlegm removal#heat relief#weight loss#health#beauty care#tremella#winter melon soup#cough relief#slimming	Red dates#autumn#health preservation#honey#food#goji berry#effect#moisten lungs#lily#nourish yin
Festivals	Smooth#everything#prosperity#wealth#dragon horse spirit#smooth sailing#good health#all the best#no taboos#wealth	Moon cake#five nuts#egg yolk#egg yolk pastry#Mid-Autumn Festival#fresh meat#lotus seed

paste#red bean paste#Cantonese style#filling
Cooking | Wash clean#cooking wine#low heat#appropriate amount#a little#fish out#mix well#light soy sauce#pour in#starch | Wash clean#appropriate amount#pour in#a little#low heat#add#cooking wine#mix well#fish out#cut into
Travel | Beijing#exploded tripe#fermented bean drink#fried liver#Zhajiang noodles#snacks#Chengang#Dan Dan noodles#tangyuan#China | Taiwan#snacks#night market#pineapple cake#bun#Taipei#braised pork rice#large intestine#nougat#small intestine
Travel | Travel#explore#journey#experience#world#trip#appreciate#dream#culture#scenery | Taiwan#night market#snacks#pineapple cake#bun#large intestine#small intestine#Shilin#Taipei#thin noodles

Summer Representative Topics	Winter Representative Topics
Benefits	Fluid production#phlegm removal#heat relief#lily#health#moisten lungs#beauty care#tremella#winter melon soup#fire
Benefits	MM#try#girl#hot recommendation#ice clear jade clean#Korea#crystal#enzyme#big S#red and tender
Benefits	Food#health#milk#red dates#nutrition#health preservation#diet#honey#apple#lily
Cooking	Wash clean#pour in#stir fry#cooking wine#fish out#low heat#appropriate amount#light soy sauce#evenly#starch

Due to space limitations, relevant topics are represented by 10 core representative features. Comparing topic distribution across seasons in the table, we can draw the following conclusions:

Topics about food "benefits" are prominent throughout the year, with specific types of benefits changing with seasonal characteristics. For example, "fire relief" and "cough relief" are distributed across all four seasons; spring and autumn specifically have "beauty care" and "nourish yin" topics; summer adds "heat relief" topics; winter adds "health preservation diet" topics.

Topics about "cooking" tutorials are distributed throughout the year with high homogeneity.

Spring and autumn have suitable temperatures and more opportunities for users to travel, so there are more food-related topics such as "local specialty snacks" and "travel food recommendations" under the "travel" category.

Spring and autumn have many important traditional festivals, such as “Spring Festival” and “Mid-Autumn Festival”, where users tend to share representative foods for specific holidays, such as “moon cakes”.

Summer and winter have more extreme climates, users travel less, lacking “travel”-related topics; there are fewer food-related holidays, lacking “holiday”-related topics.

Through the above analysis, we explain why spring and autumn have more topics than summer and winter; we also prove that the four-season topics obtained by our method align with actual conditions.

5. Summary and Outlook

Rich recipe data on the Internet and massive food reviews on social networks provide data support for food mining research. How to detect hot topics from these reviews to provide decision-making basis for consumers and marketers has become a widespread concern. Traditional topic detection tasks mainly obtain topics through large-scale text clustering. Since topics obtained by this method only contain category information that is not easily understood, manual review is often required to remove low-quality topics, resulting in low topic detection efficiency.

This paper uses Sina Weibo as the data source and combines text clustering with deep learning knowledge for topic detection. After obtaining seasonal food topics through text clustering, we automatically select high-quality topics based on topic coverage. Our method is highly generalizable and efficient, avoiding the manual topic selection step after clustering. Experimental results effectively reveal the distribution of food microblog topics across the four seasons, helping to further explore consumer focus and demand in the food domain. Future work will further consider: the application of this topic detection method in other domains; combining feature word extraction technology to achieve more accurate and in-depth topic detection tasks.

References

[1] China Internet Network Information Center. The 37th Report of Chinese Internet Development [R/OL]. (2016-01-22). [2016-05-25]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf. (in Chinese)

[2] Yin Fengjing, Xiao Weidong, Ge Bin, et al. Incremental Algorithm for Clustering Texts in Internet-oriented Topic Detection[J]. Application Research of Computers, 2011, 28(1): 54-57. (in Chinese)

[3] Wang Wei, Xu Xin. Online Public Opinion Hotspot Detection and Analysis Based on Document Clustering[J]. New Technology of Library & Information Service, 2009(3): 74-79. (in Chinese)

[4] Xu Dongliang. Research of Public Opinion Information Mining on Bulletin Board Systems Based on Cluster Analysis[D]. Harbin: Harbin Institute of Technology, 2010. (in Chinese)

[5] Zhu Hengmin, Li Qing. Public Opinion Propagation Model with Topic Derivatives in the Micro-blog Network[J]. New Technology of Library & Information Service, 2012(5): 60-64. (in Chinese)

[6] Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87. (in Chinese)

[7] Allan J, Carbonell J, Doddington G, et al. Topic Detection and Tracking Pilot Study Final Report[C]. In: Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop. 1998.

[8] Lu Rong, Xiang Liang, Liu Mingrong, et al. Discovering News Topics from Microblogs Based on Hidden Topics Analysis and Text Clustering[J]. Pattern Recognition & Artificial Intelligence, 2012, 25(3): 382-387. (in Chinese)

[9] Luo Weihua, Liu Qun, Cheng Xueqi. Development and Analysis of Technology of Topic Detection and Tracking[C]. In: Proceedings of the 7th National Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2003: 560-566. (in Chinese)

[10] Xu J, Croft W B. Cluster-based Language Models for Distributed Retrieval[C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.

[11] Wartena C, Brussee R. Topic Detection by Clustering Keywords[C]. In: Proceedings of the 19th International Conference on Database and Expert Systems Application. IEEE Computer Society, 2008: 54-58.

[12] Yang Y, Pierce T, Carbonell J. A Study on Retrospective and On-line Event Detection[C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998.

[13] Jia Z Y, Qing H E, Zhang H J, et al. A News Event Detection and Tracking Algorithm Based on Dynamic Evolution Model[J]. Journal of Computer Research & Development, 2004, 41(7): 1273-1280.

[14] Jia Ziyan, He Qing, Zhang Haijun, et al. A News Event Detection and Tracking Algorithm Based on Dynamic Evolution Model[J]. Journal of Computer Research & Development, 2004, 41(7): 1273-1280. (in Chinese)

[15] Ma Bin, Hong Yu, Lu Jianjiang, et al. A Thread-based Two-stage Clustering Method of Microblog Topic Detection[J]. Journal of Chinese Information Processing, 2012, 26(6): 121-128. (in Chinese)

[16] Hofmann T. Probabilistic Latent Semantic Indexing[C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.

[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[18] Griths T L, Steyvers M. A Probabilistic Approach to Semantic Representation[C]. In: Proceedings of the 24th Annual Conference of the Congnitive Science Society. 2002.

[19] Shan Bin, Li Fang. A Survey of Topic Evolution Based on LDA[J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49. (in Chinese)

[20] He Liang, Li Fang. Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model[J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115. (in Chinese)

[21] Wu Yonghui, Wang Xiaolong, Ding Yuxin, et al. Adaptive On-Line Web Topic Detection Method for Web News Recommendation System[J]. Acta Electronica Sinica, 2010, 38(11): 2620-2624. (in Chinese)

[22] Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research & Development, 2011, 48(10): 1795-1802. (in Chinese)

[23] Civitello L. Cuisine and Culture: A History of Food and People[M]. Wiley, 2011.

[24] Tregear A. From Stilton to Vimto: Using Food History to Re-think Typical Products in Rural Development[J]. Sociologia Ruralis, 2003, 43(2): 91-107.

[25] Wang Renxiang. Diet and Chinese Culture[M]. Beijing: People's Publishing House, 1993. (in Chinese)

[26] Zhang Jingming. Chinese Nomads Food Culture[M]. Beijing: Cultural Relics Press, 2008. (in Chinese)

[27] Mennell S, Murcott A, Otterloo A H V. The Sociology of Food: Eating, Diet and Culture[M]. London: Sage Publications, 1992.

[28] Beardsworth A, Keil E T. Sociology on the Menu: An Invitation to the Study of Food and Society[J]. British Journal of Sociology, 2002, 49(2): 327-328.

[29] Germov J, Williams L. A Sociology of Food and Nutrition: The Social Appetite[M]. The 3rd Edition. Oxford University Press, 2008.

[30] Chen Chuankang. The Culture of Chinese Diet: Regional Differentiation and Developing Trends[J]. Acta Geographica Sinica, 1994, 49(3): 226-235. (in Chinese)

[31] Cai Xiaomei, Situ Shangji. A Review on the Studies of Food Culture from Geographical Perspective[J]. Yunnan Geographic Environment Research, 2006, 18(5): 83-88. (in Chinese)

[32] Lan Yong. On The Reasons and Distrbution of Pungent Flavour in Chinese Food and Drink[J]. Geographical Research, 2001, 16(5): 229-237. (in Chinese)

[33] Ahn Y Y, Ahnert S E, Bagrow J P, et al. Flavor Network and the Principles of Food Pairing[J/OL]. Scientific Reports, 2011: Article No. 196. http://www.nature.com/articles/srep00196.

[34] Sherman P W, Billing J. Darwinian Gastronomy: Why We Use Spices[J]. Bioscience, 1999, 49(6): 453.

[35] Zhu Y X, Huang J, Zhang Z K, et al. Geography and Similarity of Regional Cuisines in China[J]. PLoS One, 2013, 8(11): e79161.

[36] Salton G, Yang C S. On the Specification of Term Values in Automatic Indexing[J]. Journal of Documentation, 1973, 29(4): 351-372.

[37] Arthur D, Vassilvitskii S. K-means++: The Advantages of Careful Seeding[C]. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. 2007.

[38] Peng Nanyun, Wang Houfeng, Ling Chentian. Event Mining in On-line News Based on Hierarchical Clustering[A]. //Advances of Computational Linguistics in China[R]. Beijing: Tsinghua University Press, 2011: 487-492. (in Chinese)

[39] Hinton G E. Learning Distributed Representations of Concepts[C]. In: Proceedings of the 8th Annual Meeting of the Cognitive Science Society. 1986.

[40] Tan P N, Steinbach M, Kumar V, et al. Introduction to Data Mining[M]. Pearson, 2010.

[41] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. ArXiv: 1301.

[42] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

Author Contributions

Zhang Xiaoyong: Literature research and organization, draft writing;
Zhou Qingqing: Assisted with experiments, paper revision;
Zhang Chengzhi: Proposed research ideas, discussed research plan, revised final version of paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: riyao95@qq.com.

[1] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. Seg_{Kmeans}.py. Seasonal Food Microblog K-means Clustering and Clustering Evaluation Algorithm.
[2] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. Inside_{Cohesion}.py. Intra-cluster Cohesion Calculation Algorithm.
[3] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. train_{{{word2vec}}*{{model}}}.py. *Word Vector Training Algorithm.*
[4] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. Weibo{{{data}}*{{initial}}}.txt. *Original Sina Weibo Crawled Data.*
[5] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. Weibo{{{data}}*{{seg}}}.txt. *Segmented Sina Weibo Data.*
[6] Zhang Xiaoyong, Zhou Qingqing, Zhang Chengzhi. Inside{{{Cohesion}}_{{sorted}}}.txt. Seasonal Food Topic Extraction Results and Intra-cluster Cohesion Ranking Data.

Received: 2016-05-26
Revised: 2016-07-18

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