Journal of Library and Information Studies 16:1 (June 2018) pp.23-43
doi: 10.6182/jlis.201806_16(1).023

Identifying Food-related Word Association and Topic Model Processing using LDA

Yu-Chin Li[1], Tsung-Chih Hu[2], Kuo-En Chang[3]

Abstract

This paper presents an interdisciplinary study that combines natural language processing and psycholinguistics research. The latent Dirichlet allocation (LDA) model was used for semantic relatedness computation to enable an understanding of the mechanisms and processes through which humans encode and retrieve lexical units. To test the similarity of the output of the topic model and human word association, the "Time-limited Multiple Divergent Thinking Test of Word Associative Strategy" (TLM-DTTWAS) was used to collect data and conduct tests with three food-related stimulus words. A total of 101 subjects took the tests, producing 4,251 words. The empirical results were analyzed on two levels: (1) by the expert word association classification: taxonomic and script proposed by Ross and Murphy (1999); (2) followed by the associative hierarchy theory of Mednick (1962), to sort the vocabulary test results into two associative hierarchies, "steep" and "flat." The analysis indicated that human word association displays randomness, as well as generalization and continuity. After the experimental text was passed through the LDA latent semantic model which demonstrated highly significant correlation. This was a whole new attempt to train a data science model to make inference and prediction of human concept association which could be very useful in teaching as well as commercial applications.

Keywords: LDA (latent Dirichlet allocation); Mandarin Vocabulary Study; Semantic Priming; Time-limited Multiple Divergent Thinking Test of Word Associative Strategy (TLM-DTTWAS); Word Association

1. Introduction

In this study, the extent to which Data Science (DS) and Natural Language Processing (NLP) can make inferences and predictions of human cognition was tested. It has been shown that human cognition manifests certain generalizations which transcend culture, gender, age, and language. For example, the Bouba-Kiki effect, proposed by Ramachandran and Hubbard (2001), proved that a special connection exists between languages and objects or concepts. No matter the culture, gender, age, or language of the test subjects, the majority of their interpretations are almost identical. These mutual inter-connections or "word association" (Meara, 2009) and the aggregative effect "categorization" (Squire & Kandel, 2000), are both extremely important cognitive activities.

Most previous studies of DS and NLP focus on testing accuracy, efficiency and effectiveness

[1] Department of Teaching Chinese as a Second Language, Chung Yuan Christian University, Taoyuan, Taiwan
[2,3] Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan
* Corresponding Author: Yu-Chin Li, E-mail: yuchin.li@cycu.edu.tw

臺灣大學學術期刊資料庫

when a subject deals with tasks involving text, graphics, sound or a mixture of these (Hassabis, Kumaran, Summerfield, & Botvinick, 2017). In this study attempts have been made to focus from a different perspective. This was done by using a specifically designed word association test to categorize human cognition patterns of word association and determine if DS and NLP can simulate the patterns of human word association. Such a new approach where DS and NLP is used to make inferences and predictions of human concept association, could be of considerable value in teaching as well as commerce.

2. Literature Review

2.1 Human memory encoding and retrieval

It has been shown that human memory and cognition display clear organization and hierarchy. When humans encounter external information, cognitive mechanisms transform the experience into representations and construct "schema" or "frameworks." This creates different forms of memory which may be working, short-term, or long-term. According to Anderson's (1977) "Schema theory," schema are knowledge representation structures used by humans for the generalization and abstraction of objects, events and fields. They sketch out basic frameworks and fundamental structures. For example, the schema of "education" includes the concepts "teacher," "student," "classroom," etc. They display generalization and flexibility and the details can change according to specific circumstances and different "definitions."

On the other hand, many previous studies have resulted in construction grammar and cognitive linguists forming a premise that language is the projection and extension of the human mind. This is supported by experiment (embodiment), as well as memory storage, prototype and family resemblance theories (Rosch & Mervis, 1975). These authors recognize the tolerance of a certain vague zone, and suggest that objects and concepts can be sorted and classified by the degree of similarity between certain characteristics and features. In terms of the organization of knowledge, Langacker (1987) proposed the theories of encyclopedic knowledge and cognitive domain and suggest that knowledge (including grammar and lexical knowledge) is more like an interconnected network. The "frame semantics" proposed by Fillmore and Atkins (1992) divides the links between words, words and humans, and internal and external worlds, into five levels: domain, frame, sub-frame, synonymous phrase, and vocabulary item. Ellis (2002), on the other hand, observed that L2 acquisition involves a "process of construction and reconstruction," and demonstrated that the frequency, form and function of constructions interact at the process of learners' L2 acquisition. Lastly, from the perspective of neuroscience, the effective encoding and repeated retrieval of memory is the key factor in the transformation of short-term memory to long-term memory. Studies have shown that the pathways and patterns of memory encoding and retrieval are quite similar (Kosslyn & Smith, 2006). This is a significant statement for this study, because it means that by understanding how we retrieve the memory, we might be able to comprehend and efficiently help to encode and store memory, or specifically assist what we work on here—vocabulary acquisition.

臺灣大學學術期刊資料庫

2.2 Context and word association network

The important studies previously mentioned about human memory encoding and retrieval, lead to two key factors which have been especially selected for discussion: (1) context and (2) organization in the form of a network.

Most second language acquisition researchers agree that context is the most important factor for vocabulary acquisition, and vocabulary should be learnt in context (Gu, 2003; Nation, 2013; Schmitt, 1997). Several key factors are contextual-related, such as the salience of the word in context, the richness of contextual clues, the capacity to infer word meaning from context as well as the learner's existing repertoire of vocabulary, etc. All these play key roles in vocabulary acquisition (Nation, 2013). Important breakthrough has been made when Baddeley (Baddeley, 1982; Gathercole & Baddeley, 2014) found an interesting relationship among human memory and change of context, which he called "encoding specificity."

On the other hand, a thorny unsolved problem is how the mental lexicon is organized. From the second language acquisition perspective, the concept of word association has been introduced by Meara (Meara, 2009; Schmitt & Meara, 1997). Recent studies of the mental lexicon, proved this point even further. Drum and Konopak (1987) assert that vocabulary is learned as a nodal network, which is supported by a structural representation of domain knowledge. The mental lexicon or the development of the semantic field is presented and organized more like a network or map in which words are interrelated and connected. According to these studies, a person's vocabulary can be compared to a net in which each word is a nodal point; higher interconnection of nodes indicates a better grasp of the vocabulary in that field. The efficiency of memory retrieval is related to the distance, volume and strength of the links between nodes. The self-reference effect explains the distance index: When the cues provided are more closely related to the individual's life or experience, the memory is more easily retrieved (Symons & Johnson, 1997). Engle, Nations and Cantor (1990) have shown that when a student understands more about something, they can more easily organize and absorb new information, which indicates the volume index. The level-of-processing effects theory explains how the probability of memory storage and retrieval is greatly increased when input is related to semantics or even preferences, compared to merely aural input (Roediger, Weldon, Stadler, & Riegler, 1992), indicating the strength index of cue retrieval. These indices can provide a convincing logical basis for the interpretation of research results.

As previously mentioned, recent studies combined with new technology have developed new representation and applications about vocabulary and the mental lexicon, such as the automatic production of lexical networks or lexical graphs using data-science model building dictionaries, as well as automatic idexing or lexical ontology (Polguère, 2014; Walter, Unger, & Cimiano, 2014; Zock & Tesfaye, 2012).

2.3 Topic model processing: Latent Dirichlet Allocation (LDA)

New advances in Data Science (DS) and Natural Language Processing (NLP), allow us to teach computers to decode the relationship or "semantic relatedness" (SR) between words,

臺灣大學學術期刊資料庫

lexical units, sentences, and text. The traditional method of lexicon or corpus establishment has generally been association and marking of vocabulary by humans. This not only takes time and much hard work, but can easily result in biased word association. This paper aims to find an automatic topic modeling process that can approach the associative tendencies of humans.

Applications of these technologies in the field of Chinese are scarce. The most outstanding example was developed by Chen, Wang, and Ko (2009). They used the Balanced Corpus of Modern Chinese published by Academia Sinica in 2006, and applied Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) to build a space to represent semantic links between Chinese words (http://www.lsa.url.tw/modules/lsa/). This website uses vectors to compare the relatedness of pairs of words, sentences and texts. It can also be used to search for and rank keywords that are most closely correlated semantically to specific words or sentences. It can automatically distinguish words in sentences and texts, and calculate the frequency of words and phrases. It uses statistical vector analysis to illustrate the semantic correlation of words, sentences or documents, and reveals the semantic space of human psychological representations (Landauer, 2002).

The LSA calculation method is primarily used to collect all data from the original dataset, convert it into a word to text large-scale matrix, and perform Singular Value Decomposition (SVD) analysis on this matrix. LSA manages to use matrix dimensionality reduction to express the latent semantic space of the entire dataset (Dumais, 2004). However, this method has some limitations:

A. Polysemes cannot be interpreted: LSA uses SVD, the results of which are linear combinations of words. These combinations cannot be logically used to analyze ambiguous words in the dataset (Chen & Xie, 2005).
B. It does not yet express latent semantic space using probability: LSA predicts whether a word distribution corresponds to a normal distribution or not (Rosario, 2000), and uses matrix dimensionality reduction to search for answers. However, it cannot consider possibilities outside the normal distribution, or other possible statistical distribution (Chen & Xie, 2005).
C. Computation and storage is not conducive to big datasets: LSA uses matrix computation to find latent semantic models, but if the dataset is too big, the resultant matrices are difficult to work with and store (Rosario, 2000).

Taking into account the LSA's limitation, this paper applies LDA model methods (Blei, Ng, & Jordan, 2003) for its calculations. LDA is a statistical analysis model for random variable sets. It takes into consideration the sequential relevance of the context of data points. To find the common latent semantic properties of the entire dataset, it is assumed that the words in the dataset obey Dirichlet distribution and Bayes' theorem (Vapnik, 1998); statistical processing with words in the dataset is carried out and the results are presented as probability. Through LDA processing, the words in the dataset are converted into a latent semantic space distribution and an aggregate set. Semantically similar words will be aggregated and common topic extracted, similar topic will be aggregated and common context will also be extracted. This model can associate words, topics and context beautifully.

臺灣大學學術期刊資料庫

Traditionally, most researchers looking at the efficiency of semantic relatedness (SR) have compared it directly with human output corpus (Altınel & Ganiz, 2016; De Boom, Van Canneyt, Demeester, & Dhoedt, 2016; Szymański & Rzeniewicz, 2016).

In this study, human word association data was collected and analyzed. The results were compared to the LDA model and it was shown that DS can effectively simulate human associative patterns.

3. Research Questions

A. Does human word association, in this case related to food, display randomness or generalization?
B. Is it possible to use multiple free word association tests to identify similarities and differences in human free word association?
C. Is it possible to use data science and NLP to effectively approach a simulation of the mechanisms and results of human word association?

4. Research Hypotheses

Three hypotheses about word recognition are proposed:

A. Human word association simultaneously displays both randomness and generalization. Thus, infers two main categories of association. The first is more connected to personal experience and knowledge, and is relatively random. The second relates to experience and knowledge shared by the majority of people who use the same language, or belong to the same culture, and is relatively communal or general.
B. Test subjects will display different levels of extension in their word association, and this makes associative tendencies more diverse.
C. We can use the DS model and random variable sets to simulate human word association patterns and use simulation to predict human associative thinking (Vapnik, 1998).

5. Research Methods

In the preparation of the analysis and prediction study presented here using human associative models, reference has been made to the research methods employed by Chen et al. (2009) in their paper "The Construction and Validation of Chinese Semantic Space by Using Latent Semantic Analysis." They used the Chinese polyseme free association norm, formed of 600 polysemes produced by 300 university students collected by Hu, Chen, Chang, and Sung (1996). Chen et al. (2009) saw these 600 polysemes as representative of an internal psychological lexical semantic relationship between Chinese readers. They used the LSA model to analyze latent semantics, investigating whether or not the Chinese semantic space could reasonably reflect polysemous representations in Chinese readers' internal psychological structure (Chen et al., 2009).

However, the procedures used in this study are different in two respects:

A. The majority of word association tests use a single stimulus word to produce a single associated word (Guilford, 1967; Hu et al., 1996; Mednick, 1962). This makes it impossible to detect the continuity of human word association and so a test for multiple associated words was used in this study.
B. Based on the relative merits of LSA and LDA discussed in the papers mentioned above, LDA was used to carry out computation and analysis in this study.

臺灣大學學術期刊資料庫

Our research methods fall into three main sections: (1) word association tests, (2) topic model processing-LDA model, and (3) expert classification.

5.1 Association test

The construction of most word association analytical schemes has been based on word association tests. The earliest included the "divergent thinking model" (Guilford, 1967) and the "spreading activation model and associative theory" (Mednick, 1962). Guilford (1967) defined divergent thinking in three dimensions: computation, content and product. The three dimensions have also been used to propose a so-called structure of intelligence (Huang, Chen, Huang, & Liu, 2009), and to indicate that there are 180 factors in the structure of intelligence. Research has shown that when an individual is in the process of divergent thinking, they will first think of concepts with relatively strong and close associations, before considering weaker and more distant associations. It is quite rare for a completely new concept to be introduced at the beginning, and few people produce such associations.

Mednick (1962) proposed an Associative Hierarchy theory, and divided his test results into two types: "steep associative hierarchies" and "flat associative hierarchies." People with *steep associative hierarchies* can only produce a small number of close associations, it is harder for them to produce distant associations. Those with *flat associative hierarchies*, also think of the most closely linked words, but not as strongly as those with a steep hierarchy and their associations are spread over more distant concepts. Therefore, it is easier for them to produce more different associated concepts. In this study, the aim was to collect multiple word association and observe the randomness, continuity and generalization of the data, rather than the more usual single word association. To achieve this end, we adopt the Divergent Thinking Test of Word Associative Strategy (DTTWAS) of Huang et al. (2009), but the single word association was changed to a time-limited multiple word association test. The details of the experiment were as follows:

Participants: Participants for the study were recruited using a Google web form circulated by email to previous survey respondents, to members of social network groups, and to persons on selected mailing lists. Participants were given clear instructions and took part in an online test.

Experimental Process: Three food-related stimulus words were selected at random: "cafeteria," "apple juice" and "radish cake," all very common words in Taiwanese daily life. Food-related words were selected because they relate to a most common everyday aspect of life shared by people in the same society. The participants were asked to produce associations based on one of the stimulus words for five minutes, and then repeat the exercise for the next word. Each participant produced three sets of associations, one for each of the three words.

5.2 Topic model processing: Using the LDA model to process program data

To analyze and predict patterns of human word association, this research used the latent Dirichlet allocation (LDA) model for processing data. LDA is a statistical analysis model that uses random variable sets. The randomness of human word association is somewhat limited, and there

臺灣大學學術期刊資料庫

is a semantic contingency between words that follow each other, in accordance with Bayes' theorem. This also fits the characteristics of the LDA model and influenced our decision to use LD for the semantic correlation computation. The data processing involved two steps: (A) data cleansing, and (B) latent Dirichlet allocation model processing.

A. *Data cleansing*: Filtration was carried out before processing to remove invalid data. The conditions for filtering were: (a) tests carried out for less than, or more than, five minutes (according to the written log); (b) results that produced fewer than five, or more than thirty, associated words for a single stimulus word. After filtering, 42 valid sets of results, composed of 3,338 words or phrases, remained. After repeated words or phrases were filtered out, 1,850 words remained.
B. *Latent Dirichlet allocation model processing*: Python was used for the LDA processing, primarily with the LDA model from the Gensim library. The random variable sets statistical analysis model, was used to generate potential parameters. We selected 30 sets, each containing 5 randomly selected words, and calculated their semantic correlation. This could only generate the semantic relatedness values of a maximum of 150 words, which was within 25% of those generated by each stimulus word. The LDA model is based not only on the word frequency, but takes probability as the main operating concept. This means that if the words selected by the LDA model (~150 words) can be shown as representative of all the human subject word association, then the predictions have a reasonable degree of significance.

There are 19 parameters in the Gensim Python library LDA model. To satisfy the purpose of this study, all the parameters were set to the default except for: corpus, id2word, num_topics and eval_every. The first two parameters were introduced from our own data set, and the last two, control value num_topics and eval_every were set to 20 and 30 (underlined) to obtain a wider range. The details are shown as Figure 1.

5.3 Expert classification

Expert classification was also used to analyze the test results (Budanitsky & Hirst, 2006) and compare them with the SR values calculated by the LDA model. Two experts were professionally trained, each spending about 85 hours conducting two types of classification. to a total of around 170 hours. The first type used an index presented by Ross and Murphy (1999) who proposed two ways to classify "concepts," by taxonomic or by script category. The taxonomic category is based

Class gensim.models.ldamodel.LdaModel (*corpus = None, num_topics = 20, id2word = None, distributed = False, chunksize = 2000, passes = 1, update_every = 1, alpha = 'symmetric', eta = None, decay = 0.5, offset = 1.0, eval_every = 30, iterations = 50, gamma_threshold = 0.001, minimum_probability = 0.01, random_state = None, ns_conf = {}, minimum_phi_value = 0.01, per_word_topics = False)*

Figure 1. 19 Parameters of the Gensim Python Library LDA Model Used in Our Study

臺灣大學學術期刊資料庫

on the characteristics of items and the correlation between them, and normally uses collective nouns for the demonstration of hierarchies. For example, "beagle" can be classed as: dog, or animal. Script category is based on the roles played in events, activities or behavior (Nguyen, 2007). This type of classification tends to cross categories and is less easy to organize into hierarchies. Prediction, planning, explanation, communication and decision-making functionality is strong and related to the changes in activities and situations. This makes it an objective-oriented classification which can assist inference-making. In this case, "beagle" could be classed as: pet, sniffer dog, working dog (Ross & Murphy, 1999).

The data was sorted into three types according to these classification indices: taxonomic (consumer), taxonomic (provider), and script. Because all the three stimulus words were dining and consumption related, there were two types of associative perspective: that of the consumer and that of the provider. The taxonomic category was also divided into two based on taxonomic consumer and provider. Responses more closely related to personal experience were placed in the script category.

To differentiate the steep and flat associative hierarchies mentioned previously, data from the original results in the script category were divided into (1) directly related to the stimulus word, such as coffee–cafeteria, and (2) not directly related to the stimulus word, such as apple juice–curry. These were then sorted into five hierarchical categories with different gray scales which represented the first to the fifth associative level respectively. In the opinion of the experts, words in the second level were defined as being associated with first level words by analogy. Thus, for the stimulus word "apple juice," associated words such as "weight control" appeared, presumably because of an association produced by the previous word "diet food." Colors were used for both categories. Category (2) is easy to understand and (1) can also be placed in different levels of association while still being classed as "directly related to the stimulus word." This is true despite the expert's belief that a word may result from association with a previous word, because of an equal relationship to the stimulus word. For example, the stimulus word cafeteria produced this string of associations: "takeaway cup," "flask," "plastic lid," "mug," "glass." These words may be the result of analogy starting with "takeaway cup" leading to "flask," but because these five words are all in themselves related to cafeterias, despite being color coded otherwise, they are still categorized as "directly related to the stimulus word."

It is worth mentioning that associations on level five and above have absolutely no apparent relevance to the stimulus word (e.g., apple juice–Newton), and are categorized as blue. Manual categorization of associated words is quite a problem because it is difficult to determine how an association has been produced. The order of the words, as set down by the participants, must not be changed to ensure the stream of word association remains unaltered.

6. Results and Discussion

A total of 101 data sets were collected and after the data had been cleaned 42 valid results remained. See Table 1 for the details. These

臺灣大學學術期刊資料庫

Table 1. Participants' Information

Gender	Female		Male	
n (participants)	33		9	
Ages	18-25	26-35	36-45	46-55
n (participants)	17	15	9	1

contained a total of 3,338 words or phrases and after repeated words and phrases had been filtered out, 1,850 discrete words and phrases remained. They were divided into categories as shown Table 2.

Details of each category as Table 3.

The Table 4 shows a high degree of randomness and 71% to 75% of the word association (WA) only appear once. This demonstrates the tendency of human word association to be scattered.

On the other hand, this study also shows that human WA exhibits a high level of generalization which, in this paper, is defined as association and experience shared by the majority of people. The first 25% most frequently appearing words cover more than 50% (56% to 69%) of the overall in the taxonomic category (both consumer and provider sub-categories). This demonstrates that human taxonomic word association exhibit a high level of generalization. However, the script category associations were more scattered (44% to 52%), of which those for cafeteria were the most universalized (52%). This may be due to script category associations being more related to personal experience. It can also demonstrate that the wide diversity of personal experience can lead to the script category associations of a particular concept to be relatively scattered.

6.1 Incidence of flat and steep associative hierarchies

To differentiate participants with steep and flat associative hierarchies, the test results were not only categorized by direct and indirect relevance to the stimulus word, but also coded by different gray scale. This made it very easy to visually distinguish participants with relatively steep associative hierarchies (see Figure 2 participant a22 (Note 1)) or flat associative hierarchies (see Figure 2 participant a42). The word association of participant a22 were usually closely connected to the stimulus word. In contrast, those of participant a42 seemed more likely to be the result of so-called "jump thinking." For example, in the associations for "apple juice," as well those for many other kinds of fruit, they jumped from "apple" to Bible references, and then back to words related to fruit.

Some word association are clearly scattered and unrelated. The results from subject a42 is an example of this: the association between "7-11" and "cafeteria," "Java Island" and "apple juice," or "soup spoon" and "radish cake" are remote. Interestingly enough, the Family Resemblance Theory can be applied here to explain this phenomenon. Essentially the theory proposed that some associated concepts are like family members who may, or may not, look alike. Certain features may be inherited by only by some descendants, and there may sometimes be throwbacks of

臺灣大學學術期刊資料庫

Table 2. Total Word Association Produced (Repeated Words Have Been Deleted)

Cafeteria	Apple juice	Radish cake	Total
693	622	535	1,850

Table 3. Types of Word Association for the 3 Stimulus Words

		$\bar{\chi}$	*SD*	*N*
Cafeteria	Taxonomic (consumer)	9.3	3.89	148
	Taxonomic (provider)	12.6	2.05	289
	Script	9.4	1.73	256
Apple juice	Taxonomic (consumer)	12.4	2.31	267
	Taxonomic (provider)	3.8	2.97	79
	Script	8.8	1.23	278
Radish cake	Taxonomic (consumer)	12.3	2.93	239
	Taxonomic (provider)	4.4	2.86	97
	Script	12.5	1.59	201

Table 4. Single Appearance Words

	Single appearance words	Percent.
Cafeteria	491	71
Apple juice	467	75
Radish cake	403	75

Table 5. Percentage of Total Represented by Most Frequent 25%

	Taxonomic (consumer)	Taxonomic (provider)	Script
Cafeteria	66%	69%	52%
Apple juice	56%	63%	44%
Radish cake	62%	62%	45%

臺灣大學學術期刊資料庫

			a22		
Cafeteria(Dir)	Cafeteria(Indir)	Apple juice(Dir)	Apple juice(Indir)	Radish Cake(Dir)	Radish Cake(Indir)
coffee		apple		radish	
cup		fruit juice		Chinese New Year	
table		fruit		breakfast	
sugar		cup		Pan-fried	
milk		blander	Coasters	Soy source	
chair		drink		chopsticks	
book		thirst		eat	
Price List		sweet		Chinese Style	
menu		health		dessert	
Waiter		nutrition			
cabinet					
light					
cake					
sandwich					
water					
toilet paper					
hipster					
quiet					
flower					

			a42		
Cafeteria(Dir)	Cafeteria(Indir)	Apple juice(Dir)	Apple juice(Indir)	Radish Cake(Dir)	Radish Cake(Indir)
coffee		Orannge juice		Breakfast shop	
latte		Lemon juice		Milk tea	
Capucchino		Guava juice		Soy sauce	
desert		fruit tea		Fried	Side dish
pudding		apple	Duriian	Omelette	
cake			banana	egg	
money			wax apple	Sunny-side-up egg	
				chopsticks	Soup spoon
hipster		Apple Tree	Adan and Eve	knife	
book			snake	fork	
light			Tree of Knowledge	Pancake	
Display board			Still Life	Pan fried dumplings	
Coffee beans			Apple musele	Street vendors	
Waiter			Apple light	Afternnon Tea	
Wood tables			peach	Taiwanese cusine	
Milk tea			red	Oyster omelets	
Drink		Apple leaf		Stinky tofu	
water		Fuji apple	Mount Fuji	tofu	
paper napkin		Java Apple	Java Island	Soy milk	
food		Japanese Apple		black tea	
sandwich			Orange	Dim Sum	
Startbucks		Watermelon juice		Rice rolls	
Left Bank Coffee		Apple Cake		Steamer	
appointment		Apple cider		Steamed dumplings	
Viennese coffee		Aloe juice		frying pan	
American coffee		Apple seeds		xiaolongbao	
Italian coffee		The Wedding Banquet		Sauce	
Coffee mugs		fruit juice		chili sauce	
Tumbler		Misellaneous juice		chili powder	
friend		Juice Shop		seasoning	obesity
Green tea latte		Convenient Store		oil	
Soy latte				Crispy rice	
Panini					
Pancake					
ice cream					
Croissants					
French coffee	7-11				
	Family				
	Hi-Life				

Figure 2. Screenshot of Word Association Test Results for Participants a22 and a42

臺灣大學學術期刊資料庫

recessive genes. But an heir still belongs to the same family because it is still possible to find shared features and connections. In our data, some word association stayed close to the stimulus word, as shown in Figure 3. Number 1 is closely related to Numbers 2 to 5, which is shown with a transparent relationship. However, some word association are more like those of numbers 6 to 10 in Figure 3. The connection between 1 and 10 may be blurred and remote, but if we take all the connections (explicit or implicit) into account, the connections between them becomes clearer. This kind of thinking mode is referred to as "*continual association*" (Mednick, Mednick, & Jung, 1964).

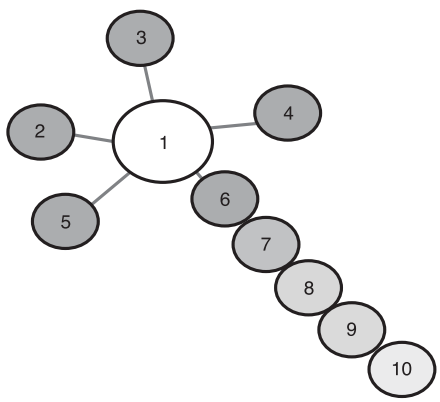


Figure 3. Diagram of Different Types of Word Association from the Stimulus Word

Table 6 lists the words produced by 42 participants (n = 3,338 words, M = 79, SD = 29.6). The participants were divided into three groups on the basis of the number of words presented: 70-90 words was taken as (1) the Medium WA group because the mean number of word association was around 80 words. 80 words + or - 10 words indicatives a neutral tendency. In contrast, group (2) participants in the Steep associative hierarchy group each produced 30 to 69 words in total. Group (3) participants in the Flat associative hierarchy group produced 91 to 150 words each. It is interesting to see there is an almost even distribution between the Steep (40%) and Flat groups (36%), but the total word output was almost double.

By extension, taking all continual WA productions in the direct and indirect categories (n = 1,116, M = 27, SD = 21.4), can also be divided into 3 groups as shown in Table 7: Here, 17-37 words was taken as (1) the Medium WA group because the mean number of continual word association was around 27 words. 27 words + or - 10 words indicates neutral tendencies. In contrast, group (2) participants in the Steep associative hierarchy group each produced 0-16 words in total. Group (3) participants in the Flat associative hierarchy group produced 40-80 words each.

Table 6. Number of Word Association by Participants with Different Associative Tendencies

		Total number of word association			*n* (participants)	Percent. (participants)
		n (word in total)	*M*	*SD*		
Steep associative	(30-69)	856	50	9.9	17	40
Medium WA	(70-90)	785	79	5.7	10	24
Flat associative	(91-150)	1,697	113	14.1	15	36

臺灣大學學術期刊資料庫

Table 7. Number of Continual Word Association by Participants with Different Associative Tendencies

Continual word association				*n* (participants)	Percent. (participants)
	n (word in total)	*M*	*SD*		
Steep associative (0-16)	144	9	4.6	16	38
Medium WA (17-37)	336	22	6.6	15	36
Flat associative (40-80)	636	58	14.0	11	26

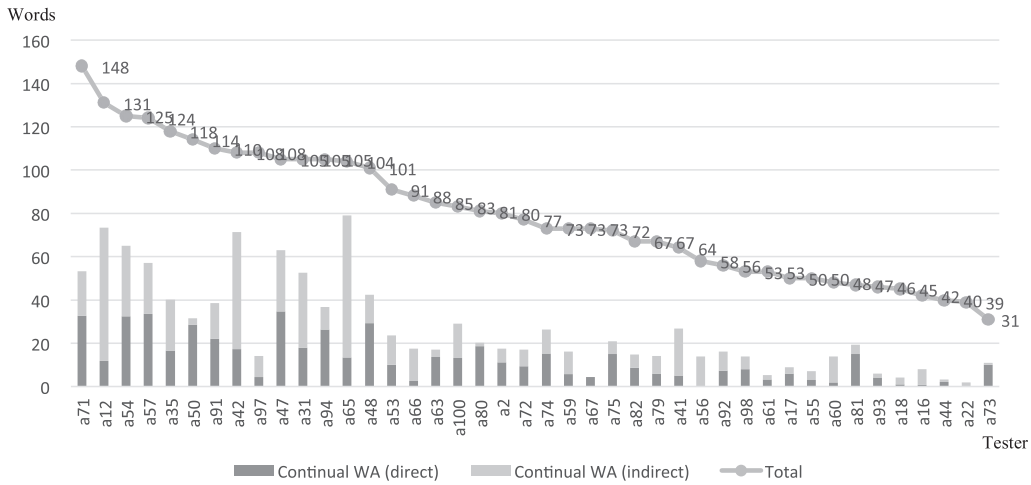


Figure 4. Total Word Association and Different Kinds of Word Association

The Spearman correlation coefficient of *continual WA* production and overall production is 0.84, which means they are highly correlated. From Table 7 it can be seen that there is a clear tendency for the more productive subject to make more continual WA (Figure 4), which also means their associations will be more creative. This result corresponds with the findings of Huang, Chen, and Liu (2012).

6.2 Results of the LDA model processing

The method of random selection, as described in the process of data cleansing, was used to select at most 150 associations for each stimulus word (Note 2) (30*5), approximately 25% of the total for each stimulus word. These results were then run through the LDA model to calculate their semantic relatedness. The Table 8 shows the null rates of the most frequent 10% of the total results of the word association test to examine the percentage that were not picked out

臺灣大學學術期刊資料庫

Table 8. The Null Rate of the Most Frequent 10% of Associations

	Taxonomic (consumer)	Taxonomic (provider)	Script
Cafeteria	0.07	0.32	0.37
Apple juice	0	0.16	0.2
Radish cake	0.12	0	0.45

by the LDA model. As explained earlier, LDA is a model for calculating semantic relatedness based on probability. The results showed that the probability of the most frequent 10% of taxonomic associations not picked up was between 0 and 0.32 (MD = 0.19, SD = 0.16). By comparison, the script category results are the highest in the three categories (0.2 to 0.45), with an average of 0.34. The script result is higher, but both are below 0.45, and the result for the stimulus word "apple juice" is only 0.2, indicating that a certain level of validity remains. This result coincides with our hypothesis: That taxonomic word association are more related to world knowledge, and the calculability of their semantic relatedness is easier to predict. On the other hand, the Script associated words are objective-oriented and more related to personal experience and situation, and semantic relatedness is more difficult to calculate. These results (Table 9) become even clearer in the correlation coefficient rate (r_s) test.

6.3 Expert-LDA correlation coefficient rate(r_s)

After the three categories had been sorted by two experts, the percentage that each word occupied in its own category was calculated. Results were compared using the Spearman correlation coefficient to compare the percentage of the words in each category with the LDA semantic relatedness values. The result, shown in Table 10 below, was termed: the "Expert-LDA correlation coefficient rate." The correlation coefficient rate of the taxonomic (consumer) category was between 0.77 and 0.94. The high Expert-LDA correlation coefficient rate of the taxonomic (consumer) category demonstrates that the LDA model has a certain ability to make inferences about human word association and LDA programs have an outstanding ability to make inferences about human word association in a specific scope (taxonomic related in this case).

However, in the category of taxonomic (provider), only the "radish cake" category was relatively significant. This might have been caused by differences in the properties of the stimulus words. It can be seen from the volume of associations that cafeteria is the word most closely related to the everyday life of Taiwanese people. This caused the number of associated words (n = 693) to be the highest for the stimulus words. The word "Starbucks" was offered by 43% of the participants (18 out of 42) in the taxonomic (provider) category for "cafeteria." The majority of words offered in the taxonomic (provider) category came from consumer experience (e.g., music, lamp, computer, waiter, etc.). The number of words reflecting the observation of details, personal experience and feelings (script category) was especially high as well, with a correlation of 0.73. This also indicated the

臺灣大學學術期刊資料庫

Table 9. Top 15 Word Association of the 3 Food-Related Words: Cafeteria, Radish Cake and Apple Juice

Ranking	Cafeteria	Radish cake	Apple juice
1	Coffee	Thick Soy Sauce	Apple
2	Cake	Breakfast	Ice
3	Coffee bean	Oil	Fruit
4	Latte	Fried	Juice
5	Latte Art	Radish	Orange Juice
6	Cappuccino	Small shrimp	Banana
7	Milk	Omelet	Apple cider
8	Waffle	Soy sauce	Apple pie
9	Foamed milk	Shredded radish	Pineapple
10	Sugar	Milk tea	Flesh (fruit)
11	Dessert	HK style	Beverage
12	Tea	Soy milk	Apple tree
13	Cookie	Sweet and sour sauce	Lemon juice
14	Set meal	White radish	Red
15	Beverage	Taro cake	Apple sauce

Table 10. Expert-LDA Spearman Correlation Coefficient Rate (R_s)

	Taxonomic (consumer)	Taxonomic (provider)	Script
Cafeteria	0.80	0.04	0.73
Apple juice	0.77	0.5	0.51
Radish cake	0.94	0.82	0.54

generalization of such associations and close links to everyday experience. This is the so-called "self-reference effect." The situation of "apple juice" is quite similar. The results show relatively few associations in the taxonomic (provider) category, and most associations being related to the word "apple" itself, at more than 50%. In contrast, results of "radish cake" were more widespread (e.g., breakfast stalls and dim sum restaurants), and there is no single vendor with a monopoly on the market (like Starbucks for "cafeteria"). Dim sum is offered in specific types of establishment (breakfast stalls, dim sum restaurants), which caused its presence in the taxonomic (provider) category to be relatively significant.

臺灣大學學術期刊資料庫

If we just focus on the output quantity (Table 3), the WA of Taxonomic (consumer) and the Script category are both abundant. However, compared with the taxonomic (consumer) category, the correlation of the script category was relatively insignificant (between 0.51 and 0.73). We think this is due to the fact that script associations are objective-oriented and more closely related to personal experience and situation, and thus their semantic correlation is relatively hard to calculate. Even so, we can still see a significant difference between "cafeteria" in contrast with "apple juice" and "radish cake" in Script category word association (20% higher).

There are two reasons for our assumption: (1) "cafeteria" is a part of foreign culture, which became popular just a few years ago in Taiwan. Thus, the word association of Taiwanese people tend to be prevalent and stable with this term. (2) Another significant difference is "apple juice" and "radish cake" are concrete and small scope entities compare to "cafeteria," a location, which covers larger scope. In this case, the scope of word association tends to be more concentrated. To the contrast, the word association of small entity tends to diffuse easily so far as personal experience is concerned. As such, we infer that these may be the main reasons even though the Expert-LDA correlation coefficient rate of the 3 stimulus words in script category is not as good as the taxonomic consumer category, but inside of the script category the "cafeteria" WA is still significantly higher than "apple juice" and "radish cake" WA.

Making a bold assumption, with our understanding from the literature, even though the inference of LDA under the script category is not as good, still, if the sample size is big enough and with the similarity of culture and social experience, we may be able to find patterns and make inference in script category as high as the WA in taxonomic consumer category. However, this is an unanswered question which is yet to be verified.

We have not yet found a useful calculation technique for steep and flat associative hierarchies and this suggests a direction for future research.

7. Conclusion

Neuroscience offers initial guidance towards architectural and algorithmic constraints for successful neural network applications (Hassabis et al., 2017). An understanding of human cognition and neural networks could be the key to the advancement of work to improve the effectiveness and efficiency of NLP and DS performance. Traditionally, most researchers looking at the efficiency of semantic relatedness have compared it directly with the data of human output (Altınel & Ganiz, 2016; De Boom et al., 2016; Szymański & Rzeniewicz, 2016).

The innovative aspect of this study shows that DS can effectively simulate human associative patterns. To improve the traditional word association test (Huang et al., 2009), and to meet the needs of this study a data collecting tool, the "Time-limited Multiple Divergent Thinking Test of Word Associative Strategy" (TLM-DTTWAS) was used for data collection. This test allowed the observation of a diversity of the human word association patterns. The main aim of this study was to find answers to three main questions and the data set collected by TLM-DTTWAS, and the classification of the empirical results by experts in two ways, gave answers to two of them.

臺灣大學學術期刊資料庫

(1) Using the classification indices for word association tests proposed by Ross and Murphy (1999) (taxonomic and script categories), the data were divided according to their properties into taxonomic (consumer), taxonomic (provider) and script categories. The results revealed that human association displays not only randomness, but also generalization.

(2) Using the Associative Hierarchies theory of Mednick (1962), the results were divided into two categories: steep associative hierarchies and flat associative hierarchies. In addition to demonstrating the validity of the proposition that humans display different associative tendencies, the analytical results also highlighted a high continuity of human word association. The results also correlated with the family resemblance theory of cognitive linguistics (Cuenca & Hilferty, 1999).

With respect to the third question. This study accumulated a data set using human word association tests and used the LDA model to calculate its SR. Three categories were then created using expert classification. A review was then conducted of the frequency of words as a proportion of their own category and their correlation with the SR value of the LDA model, using the Spearman correlation coefficient. The findings indicated that the Expert-LDA correlation coefficient rate of categories were closely related to everyday life (e.g., the taxonomic (consumer) category) and demonstrated a particularly high positive correlation (0.77 to 0.94), which clearly demonstrates a self-reference effect. This also showed that the LDA programs have an outstanding ability to make inferences about human word association.

In the future, a more comprehensive model may not only make inferences, but also predictions about human concept associations. In a future study consideration will be given to calculations with different models of DS, to see if better results can be obtained, and if steep and flat associative hierarchies can be simulated. Methods for data collection will be improved by expanding the range to include participants from different countries and cultural backgrounds. The field of stimulus words will also be expanded in a step towards a cross-linguistic, cross-cultural research project. Other types of topic model in addition to LDA will be considered, such as Word2Vec, to yield further interesting findings and allow the potential of DS to predict human word association with greater regularity.

Notes

Note 1 The answers were originally written in Chinese and have been translated.

Note 2 The answers were originally written in Chinese and have been translated.

References

Altınel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems, 108*, 50-64. doi: 10.1016/j.knosys.2016.06.021

Anderson, R. C. (1977). *Schema-directed processes in language comprehension* (Technical report No. 50). Urbana, IL: University of Illinois at Urbana-Champaign.

臺灣大學學術期刊資料庫

Baddeley, A. D. (1982). Domains of recollection. *Psychological Review, 89*(6), 708-729. doi: 10.1037//0033-295X.89.6.708

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(2003), 993-1022.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics, 32*(1), 13-47. doi: 10.1162/089120106776173093

Chen, M.-L., Wang, H.-C., & Ko, H.-W. (2009). The construction and validation of chinese semantic space by usinglatent semantic analysis. *Chinese Journal of Psychology, 51*(4), 415-435. doi: 10.6129/CJP.2009.5104.02

Chen, T., & Xie, Y.-Q. (2005). Literature review of feature dimension reduction in text categorization. *Journal of the China Society for Scientific and Technical Information, 24*(6), 690-695.

Cuenca, M. J., & Hilferty, J. (1999). *Introducción a la lingüística cognitiva* [Introduction to cognitive linguistics]. Barcelona, Spain: Editorial Ariel.

De Boom, C., Van Canneyt, S., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters, 80*, 150-156. doi: 10.1016/j.patrec.2016.06.012

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Drum, P. A., & Konopak, B. C. (1987). Learning word meanings from written context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 73-87). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230. doi: 10.1002/aris.1440380105

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143-188. doi: 10.1017/S0272263102002140

Engle, R. W., Nations, J. K., & Cantor, J. (1990). Is “working memory capacity” just another name for word knowledge? *Journal of Educational Psychology, 82*(4), 799-804. doi: 10.1037/0022-0663.82.4.799

Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer & E. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 75-102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gathercole, S. E., & Baddeley, A. D. (2014). *Working memory and language*. New York, NY: Psychology Press.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context

臺灣大學學術期刊資料庫

and strategies. *The Electronic Journal for English as a Second Language, 7*(2), 1-25.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245-258. doi: 10.1016/j.neuron.2017.06.011

Hu, Z.-W., Chen, Y.-Z., Chang, S.-H., & Sung, Y.-C. (1996). Chinese polyseme free association norm. *Chinese Journal of Psychology, 38*(2), 67-168.

Huang, P.-S., Chen, H.-C., Huang, H.-C., & Liu, C.-H. (2009). The Development of Divergent Thinking Test of Word Associative Strategy (DTTWAS). *Psychological Testing, 56*(2), 153-177.

Huang, P.-S., Chen, H.-C., & Liu, C.-H. (2012). The development of Chinese word remote associates test for college students. *Psychological Testing, 59*(4), 581-607.

Kosslyn, S. M., & Smith, E. E. (2006). *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Prentice-Hall.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of Learning and Motivation, 41*, 43-84. doi: 10.1016/S0079-7421(02)80004-4

Langacker, R. W. (1987). *Foundations of cognitive grammar. Volume I: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Philadelphia, PA: John Benjamins.

Mednick, M. T., Mednick, S. A., & Jung, C. C. (1964). Continual association as a function of level of creativity and type of verbal stimulus. *The Journal of Abnormal and Social Psychology, 69*(5), 511-515. doi: 10.1037/h0041086

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232. doi: 10.1037/h0048850

Nation, I. S. P. (2013). *Teaching and learning vocabulary*. Boston, MA: Heinle Cengage Learning.

Nguyen, S. P. (2007). Cross-classification and category representation in children's concepts. *Developmental Psychology, 43*(3), 719-731. doi: 10.1037/0012-1649.43.3.719

Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography, 27*(4), 396-418. doi: 10.1093/ijl/ecu017

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies, 8*(12), 3-34.

Roediger, H. L., Weldon, M. S., Stadler, M. L., & Riegler, G. L. (1992). Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1251-1269. doi: 10.1037//0278-7393.18.6.1251

臺灣大學學術期刊資料庫

Rosario, B. (2000). *Latent semantic indexing: An overview* (Final paper INFOSYS 240). Berkeley, CA: University of Berkeley.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605. doi: 10.1016/0010-0285(75)90024-9

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*(4), 495-553. doi: 10.1006/cogp.1998.0712

Schmitt, N. (1997). Vocabulary learning strategies. In D. N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199-227). Cambridge, England: Cambridge University Press. doi: 10.1017/S0272263197001022

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition, 19*(1), 17-36.

Squire, L. R., & Kandel, E. R. (2000). *Memory: From mind to molecules*. New York, NY: Holt Paperbacks.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371-394. doi: 10.1037//0033-2909.121.3.371

Szymański, J., & Rzeniewicz, J. (2016). Identification of category associations using a multilabel classifier. *Expert Systems with Applications, 61*, 327-342. doi: 10.1016/j.eswa.2016.05.039

Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: Wiley.

Walter, S., Unger, C., & Cimiano, P. (2014). ATOLL—A framework for the automatic induction of ontology lexica. *Data and Knowledge Engineering, 94*, 148-162. doi: 10.1016/j.datak.2014.09.003

Zock, M., & Tesfaye, D. (2012). Automatic index creation to support navigation in lexical graphs encoding part_of relations. In M. Zock & R. Rapp (Eds.), *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon* (pp. 33-52). Mumbai, India: The COLING 2012 Organizing Committee.

(Received: 2018/1/10; Accepted: 2018/4/24)

臺灣大學學術期刊資料庫

圖書資訊學刊　第16卷第1期　（2018.6）　頁23-43
doi: 10.6182/jlis.201806_16(1).023

「食」類相關的詞彙聯想識別和主題模型處理：以LDA為例

Identifying Food-related Word Association and Topic Model Processing using LDA

李郁錦[1]　胡宗智[2]　張國恩[3]

Yu-Chin Li[1], Tsung-Chih Hu[2], Kuo-En Chang[3]

摘　要

本研究結合自然語言處理及心理語言學二者，屬一跨領域研究。為理解人類對詞彙認知與習得的機制與過程，試圖以主題模型中的潛在語意模型LDA（latent Dirichlet allocation)，進行詞彙語意相關度的運算。為測試潛在語意模型的輸出與人類詞彙聯想的相似度，本研究藉由大規模的多重限時「詞彙聯想策略擴散性思考測驗」的資料搜集，以三項刺激詞進行測驗，共101位受試者參與受試，輸出共4,251項獨立詞。實驗結果透過二個層次的分析：(1)以專家分類（expert classification）的方式，透過二名專家，一方面以Ross與Murphy（1999）所提出的詞彙聯想結果的分類指標（知識及腳本分類）分類。另一方面，以Mednick（1962）的連結層級理論，將詞彙測驗結果分為二類：陡峭式與平緩式連結。分析結果指出人類聯想不僅具有隨機性，更具有普遍性及延展性。(2)實驗文本經由潛在語意模型LDA運算，二者的結果交叉比對後，證實具高度顯著相關。輸出結果符合人類學習和聯想的機制。本研究所進行的是一個全新的嘗試—資料處理科學對人類的詞彙及概念的聯想進行推理和預測。此一結果，未來在教學和商業上可提供改善及應用。

關鍵字：LDA（latent Dirichlet allocation）、華語詞彙學習、語義啟動、多重限時「詞彙聯想策略擴散性思考測驗」、詞彙聯想

[1] 中原大學應用華語文學系
Department of Teaching Chinese as a Second Language, Chung Yuan Christian University, Taoyuan, Taiwan
[2,3] 國立臺灣師範大學資訊教育研究所
Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan
* 通訊作者Corresponding Author: 李郁錦Yu-Chin Li, E-mail: yuchin.li@cycu.edu.tw

註：本中文摘要由作者提供。
以APA格式引用本文：Li, Y.-C., Hu, T.-C., & Chang, K.-E. (2018). Identifying food-related word association and topic model processing using LDA. *Journal of Library and Information Studies, 16*(1), 23-43. doi: 10.6182/jlis.201806_16(1).023
以Chicago格式引用本文：Yu-Chin Li, Tsung-Chih Hu, and Kuo-En Chang. "Identifying food-related word association and topic model processing using LDA." *Journal of Library and Information Studies* 16, no. 1 (2018): 23-43. doi: 10.6182/jlis.201806_16(1).023

臺灣大學學術期刊資料庫