Categorization of Cooking Actions Based on Textual/Visual Similarity

Yixin Zhang
zhangyx@dl.soc.i.kyoto-u.ac.jp
Kyoto University

Yoko Yamakata
yamakata@mi.u-tokyo.ac.jp
The University of Tokyo

Keishi Tajima
tajima@i.kyoto-u.ac.jp
Kyoto University

ABSTRACT

In this paper, we propose a method of automatically categorizing cooking actions appearing in recipe data. We extract verbs from textual descriptions of cooking procedures in recipe data, and vectorize them by using word embedding. These vectors provide a way to compute contextual similarity between verbs. We also extract images associated with each step of the procedures, and vectorize them by using a standard feature extraction method. For each verb, we collect images associated with the steps whose description includes the verb, and calculate the average of their vectors. These vectors provide a way to compute visual similarity between verbs. However, one type of action is sometimes represented by several types of images in recipe data. In such cases, the average of the associated image vectors is not appropriate representation of the action. To mitigate this problem, we propose a yet another way to vectorize verbs. We first cluster all the images in the recipe data into 20 clusters. For each verb, we calculate the ratio of each cluster within the set of images associated with the verb, and create a 20-dimensional vector representing the distribution over the 20 classes. We calculate similarity of verbs by using these three kinds of vector representations. We conducted a preliminary experiment for comparing these three ways, and the result shows that each of them are useful for categorizing cooking actions.

KEYWORDS

recipe data; text understanding; vectorization; word embedding;

ACM Reference Format:
Yixin Zhang, Yoko Yamakata, and Keishi Tajima. 2019. Categorization of Cooking Actions Based on Textual/Visual Similarity. In *5th International Workshop on Multimedia Assisted Dietary Management (MADiMa '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3347448.3357165

1 INTRODUCTION

Recently, there are many popular Web sites for posting and sharing recipe data, such as Allrecipes in North America and UK, Haodou in China, and Cookpad in Japan. Now several millions of recipe data are shared on these sites. The structure of recipe data posted on these sites are relatively uniform, and primarily consists of step-by-step descriptions of cooking procedures, and sometimes they also have images associated with specific steps in the procedures. Figure 1 shows an example of such recipe data.

(https://www.haodou.com/recipe/7064617/)

Figure 1: Example of recipe data posted on Haodou.

The data accumulated on these sites are now huge and relatively uniform, and thus have become a valuable resource for information on foods and recipes. As a result, the method of extracting useful information on foods and recipes from this data has become an important research issue. In addition, because these cooking procedure data are typical examples of data describing procedures of creating something, research on the analysis of these data is an important research issue in more general sense in the filed of natural language processing. Because of that, there have been much research on automatic understanding of recipe data recently [1, 3, 8].

In this paper, we particularly focus on automatic understanding of cooking actions that appear in the cooking procedures. Such actions are usually described by action verbs in the text description, and it is not difficult to extract them by using natural language processing techniques. However, automatic understanding of the meanings of these verbs are sometimes not easy. It is because the vocabulary of verbs describing cooking actions are very rich. For example, in Chinese, there are many words that basically mean "cut" but with various additional meanings and nuances. Similar situation happens in many other languages.

When we encounter a verb which is not included in the dictionary, there are two kinds of information that are useful for automatically understanding it. First, if we can know which verbs in the dictionary have meanings similar to the newly found verb, we can

MADiMa '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6916-9/19/10…$15.00
https://doi.org/10.1145/3347448.3357165

roughly infer the meaning of the verb. The vocabulary for describing cooking actions is rich as mentioned above, but it is not because there are many types of actions. It is because there are many verbs with similar meanings but with minor differences. Therefore, even if we encounter an unknown verb in a recipe description, we can expect that it has some similar verbs in the dictionary in most cases.

The second type of information useful for understanding the meaning of a newly found verb is an image associated with it. Images are, however, associated with a step, and a step description often includes more than one action verbs. In such cases, we need to determine which action verb the image is representing. Similarity of verbs is also useful for determining it. If the image associated with a step including verbs A and B is more similar to images associated with other steps including verbs similar to A than to images associated with other steps including verbs similar to B, the image probably represents that verb A. Notice that we assume A or B is a newly found verb, so we cannot simply use images associated with other steps including A or B.

As explained above, identification of verbs similar to a newly found verb is useful for automatic understanding of the verb in two ways. In order to enable it, we propose a method of automatically categorizing action verbs appearing in recipe data. Our basic approach is to produce vector representation of verbs, and compute similarity of verbs on that representation. In this paper, we compare three ways to vectorize verbs extracted from textual descriptions of cooking procedures in recipe data.

The first method is to vectorize verbs by using word embedding. We use word2vec [7], one of the standard word embedding methods. This method produces a vector of a verb based on what context a verb appears in. Therefore, the vectors produced by this method provide a way to compute contextual similarity between verbs.

The second method uses images associated with cooking procedures. We vectorize them by using a standard feature extraction method. For each verb, we collect images associated with the steps whose description includes the verb, and calculate the average of their vectors. This vector is expected to represent visual characteristics of the verb, and therefore, the vectors produced by this method provide a way to compute visual similarity between verbs.

However, one type of action is sometimes represented by several types of images in recipe data. For example, a cooking step including a verb "cut" is often associated with an image showing the action itself, i.e., an image of a knife hold in a hand and cutting something, but it is also often associated with an image showing the result of the action, i.e., something cut into pieces. It is also sometimes associated with an image showing the preparation of the action, i.e., a knife and the material to cut side by side. As a result, the images associated with the action "cut" form three clusters. In such a case, the average of their vectors does not represent the action well.

To mitigate this problem, we propose a yet another way to vectorize verbs. We first cluster all images in the data set into 20 clusters. For each verb, we calculate the ratio of each cluster within the set of images associated with the verb, and create a 20-dimensional vector representing the distribution over the 20 classes.

We calculate similarity of verbs by using these three kinds of vector representations. We conducted a preliminary experiment for comparing these three ways, and the result shows that text-based vectorization and image-based vectorization can extract different types of similarity of verbs, and are complementary with each other.

2 RELATED WORK

There have been research on automatic recognition of actions recorded or described in multimedia data in the context of cooking [11], or in more general context [9]. They focus on video data. When we have a video data corresponding to a step in a cooking procedure, the main issue is to segment the video into a sequence of primitive actions, and recognize the action in each segment or align the sequence of video segments with the sequence of actions described in the textual data. On the other hand, we assume that an image is associated with a step in the procedure which may include more than one actions. This is the case in most recipe data Web sites. We usually cannot represent more than one actions in one image, and an image associated with a step usually corresponds to one of the actions in the step. In that case, the issue is to determine which action in the description the image represents.

There have also been research on the analysis of a set of words associated with images, such as identifying important tags from a tag set associated with an image [4–6, 12] and inferring semantic relationship between tags associated with images [2]. These studies use images and tags collected from image sharing sites, such as Flickr and Instagram. Most tags used in these sites are nouns (and occasionally adjectives). On the other hand, we focus on verbs associated with images in the recipe descriptions, and examine whether semantic relationship between verbs can also be inferred from their textual and visual features.

In summary, automatic recognition of actions have been studied focusing on video data, while image data has been used for object recognition but not action recognition in the prior research. In this paper, we focus on automatic understanding of actions based on text and image data.

3 OUR METHODS

We have explained our three methods of vectorizing verbs in recipe data in Section 1. In this section, we explain some details of these methods.

3.1 Dataset

First, we explain recipe data we used in this research. We collected 12548 recipe data posted on Haodou Recipe[1], a recipe sharing Web site in China. Each data item consists of the following components: a recipe ID, a general description, ingredients, tips, and a sequence of cooking procedure steps, each of which is a pair of a text description and an optional image. Figure 1 shows an example.

We extract text data of the procedure steps, segment Chinese sentences into words, and add the Part-of-Speech tagging (POS tagging) by using Chinese language segmentation and POS tagging tool jieba[2]. We then extract verbs that are not included in the stop word list. We extracted 3175 verbs, 341 verbs (10.7%) of which appeared 100 times or more. On the other hand, 1957 verbs (61.6%) appeared less than 10 times, and 695 verbs (21.9%) appeared only once. Table 1 shows the top 20 most frequent words.

[1] http://www.haodou.com/recipe/
[2] https://github.com/fxsjy/jieba

Table 1: Top 20 Most Frequent Verbs

rank	verb		frequency
1	put in	放入	16958
2	add in	加入	12187
3	pour in	倒入	7150
4	stir fry	翻炒	5413
5	prepare	准备	4792
6	boil	煮	4625
7	stir	搅拌	4613
8	set aside	备用	4590
9	moderate amount	适量	4341
10	wash	洗净	4291
11	add	加	3716
12	put	放	3096
13	out	出	2918
14	cut into	切成	2763
15	cut	切	2733
16	be	是	2550
17	clean	清洗	2202
18	mix well	拌匀	2171
19	ferment	发酵	2119
20	cover	盖	2089

3.2 Vectorization by Word-Embedding

The first vectorization method uses word embedding. We use one of the standard word-embedding method word2vec. We learn word-embedding by using the corpus of all text data in our recipe data, and transform each verb into a 200-dimensional vector.

Given word embedding vectors of verbs, the similarity between two verbs is computed by the cosine similarity of two vectors.

3.3 Vectorization by Visual Features of Associated Images

The second method uses visual features of associated images. We vectorize an image by using a convolutional neural network VGG16 [10], which is widely used for image recognition, trained on ImageNet data. We use the output of two fully-connected layers in VGG16, which is a 4096-dimensional vector. In VGG16, this vector is used as the input of the final layer that outputs 1000-dimensional vector representing the probability distribution over 1000 classes.

We compared the performance of two variations: directly using these 4096-dimensional vectors, and reducing them to 300-dimensional vectors by using PCA. Figure 2 shows the explained variance ratio by 1 to 400 PCA components. The explained variance ratio by 300 PCA components is 0.7746.

In either case, given feature vectors of images, we compute the average of the feature vectors of the associated images for each verb. The similarity of two verbs is computed by the cosine similarity of their average image feature vectors.

3.4 Vectorization by Ratio of Clusters within Associated Images

As explained in Section 1, we also propose a method of vectorizing verbs based on the ratios of images in each cluster within a set of all

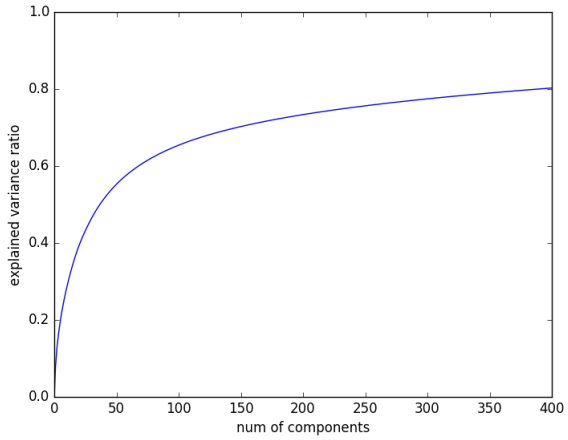

Figure 2: Explained variance ratio by 1 to 400 PCA components.

Table 2: The Number of Images in Each Class

class	#img	class	#img	class	#img	class	#img
00	775	05	3107	10	3468	15	2387
01	3244	06	2574	11	2530	16	1882
02	2561	07	3919	12	1436	17	1685
03	2696	08	2258	13	1123	18	1175
04	2664	09	3529	14	2544	19	2607

images associated with the verb. We first cluster all 48164 images in our dataset into 20 clusters by using k-mean clustering method. Table 2 shows the number of images in each class, and Figure 3 shows examples of images in each cluster. Clusters are numbered from 0 to 19. We chose images that are close to the centroids of the clusters.

We then produce a 20-dimensional vector for each verb by computing the ratio of each cluster within the set of all images associated with it. Figure 4 shows a heat map representing the ratio of each cluster for 14 example verbs. (All these verbs appear more than 100 times in the dataset.) For example, the set of images associated with the verb at the second row, which means "stir-fry", mainly consists of images in the cluster 8, and it also includes images in the cluster 17. As shown in Figure 3, both cluster 8 and 17 consists of images showing some food fried in a pan. On the other hand, the set of images associated with the verb at the top row, which means "put in", equally includes images in all the clusters. This is because the verb "put in" is used in a variety of contexts.

Table 3 shows five verbs with the highest entropy of probability distribution over 20 clusters. Low entropy means that the verb is used in some specific contexts, and high entropy means that the verb is used in a variety of contexts. All the verbs with the highest entropy shown in Table 3 are general words that are used in a variety of contexts.

As the components of a 20-dimensional vector are the ratio of 20 classes within the set of images associated with a verb, their sum

Figure 3: Example photos in the clusters 0 to 19.

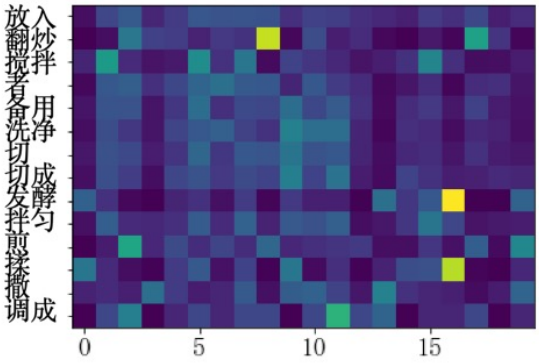

Figure 4: Cluster distribution of 14 example verbs.

Table 3: Five Words with Highest Entropy

verb		entropy
add	放	4.224
add-in	放入	4.201
dry	干	4.196
exist	有	4.185
come out	出来	4.177

Table 4: Verbs with Largest Value of Each of 20 Dimensions

	verb			verb	
0	fold	折叠	10	eat	开吃
1	break (an egg) & put in	磕入	11	eat	开吃
2	pour in	倒进去	12	cut off	切断
3	eat	开吃	13	squeeze in	挤入
4	dish arrangement	摆盘	14	press out	压出
5	whip up the egg	打蛋	15	times	回
6	make soup	煲	16	fold in	包入
7	cook porridge	煮粥	17	put in	下
8	cook	烧煮	18	fill	盛入
9	fold	折叠	19	baked	出炉

is 1. In other words, these 20-dimensional vectors are normalized by L1-norm. Given these vectors, the similarity of two verbs is computed by L1-distance between their 20-dimensional vectors.

On the other hand, Table 4 shows verbs that have the largest value of each component of the 20-dimension vectors, in other words, verbs that have the brightest cell in each column of the heat map. These verbs seem appropriate for describing images in each cluster shown in Figure 3.

4 EXPERIMENT

In this section, we show the result of the experiment that we ran in order to evaluate the three methods of vectorizing verbs explained in the previous section.

For each of the 14 example verbs we have shown in Figure 4, we computed top-10 similar verbs by using our three vectorization methods. We first compare the results of the text-based method (word-embedding vectors) and the image-based methods (image feature vectors and cluster ratio vectors) in order to examine if the results of the text-base method and the image-based methods are similar or not. We calculated how many of the top-n results given by the image-based method is also included in the top-10 results given by the text-based method for $n = 1, \ldots, 10$. Total number of the included images for 14 example verbs are shown in Table 5.

Table 5: Degree of Agreement with Text-Based Method

n	1	2	3	4	5	6	7	8	9	10
4096-vector	10	13	16	21	23	27	29	31	33	35
300-vector	10	13	17	21	24	27	29	31	35	39
20-vector	8	12	18	22	24	26	30	31	32	34

As explained before, there are two variations of the image feature vectors, one using the original 4096 dimensions, and one using the 300 dimensions extracted by PCA. The first and second rows of the table show the numbers for these two methods. The third row shows the numbers for the third method based on the ratio of the 20 clusters within the set of associated images.

When $n = 10$, if we divide the values in Table 5 by $140 = 10 \times 14$, it corresponds to the r-precision of the image-based method assuming that the text-based method is the ground truth. As shown in Table 5, the values are not very large, which means that the results of the text-based method and the image-based methods have some images in common, but are not very similar. Table 5 also shows that the degrees of agreement of the three image-based methods are not largely different from each other.

Next, we show the top-10 results of the text-based method and the method using the 20-dimensional vectors for the 14 example verbs in Table 6 and Table 7.

Although there are many verbs that are only included either in the results of the text-based method or the results of the image-based method, many of them are verbs really related to the given verb. These result shows that the text-based similarity and the image-based similarity of verbs are useful for different kind of similar actions, and are complementary with each other.

The results also show that the verbs that are ranked high only by image-based methods have more specific meanings in some cases. For example, in the image-based top-10 results for the verb "boil", "put in", "add in" and "boiled water" are verbs related to "boil". On the other hand, the result of the text-based top-10 results for "boil" include many synonyms of "boil".

5 CONCLUSION

In the prior research, automatic recognition and understanding of actions from multimedia data has been studied for video data, and image data has been used for object recognition. In this paper, we focus on the problem of automatic understanding of actions from multimedia data consisting of text data and images.

We compared three methods of vectorizing action verbs in the textual description of cooking procedures: (1) word-embedding, (2) the average of feature vectors of images associated with the verb, and (3) the ratio of images in 20 clusters within the set of images associated with the verb. To compare these three methods,

Table 6: Top-10 Similar Verbs by Text-Based and Image-Based Methods

put in 放入

	text		image	
1	pour	倒入	put in	放
2	inject	注入	pour in	倒入
3	add in	加入	let	让
4	put	放	boil	煮
5	in	入	add in	加入
6	season into	调入	fish out	捞出
7	plus	加上	continue	继续
8	season	调成	boil	开
9	fill	盛出	stew	焖
10	add	加	come out	出来

stir fry 翻炒

	text		image	
1	medium well	断生	fry	炒
2	fry	炒	heat up	烧热
3	stir well	炒匀	saute	炒香
4	fried	炒好	wok	炒锅
5	stir fry	爆炒	stir well	炒匀
6	fry to	炒至	spicy fry	爆香
7	fragrant	爆香	fry to	炒至
8	saute	炒香	raw	生
9	turn soft	变软	cook to	烧至
10	fry to	炒出	stir-fry	爆炒

stir 搅拌

	text		image	
1	whipping	搅打	mixing	翻拌
2	mix	拌成	split	分
3	stir	搅	no	无
4	stir up	搅匀	stir	搅
5	mix	拌	to	到
6	mix	翻拌	will	会
7	sift	筛入	blend	搅打
8	pump	抽	mix	拌
9	none	无	want	要
10	hit	打	be	是

boil 煮

	text		image	
1	boil	煮开	put in	放入
2	boil	煮沸	boil	煮沸
3	make soup	煲	boiling water	开水
4	cooked	煮熟	fish out	捞出
5	stew	焖	pour	倒入
6	boil	烧开	add in	加入
7	stew	炖煮	cook	煮熟
8	stew	炖	cooking wine	料酒
9	make soup	熬	stew	焖
10	turn	转	chop	切碎

standby 备用

	text		image	
1	chop	切碎	cut	切
2	standby	待用	drain	控干
3	crush	剁碎	cut into	切成
4	drain	控干	fish out	捞出
5	rub	擦	pickle	腌制
6	separate	分开	wash	洗净
7	tear into	撕成	go	去
8	wash	洗净	put in	放入
9	wok	炒锅	soak	泡
10	break	打碎	chop	切碎

wash 洗净

	text		image	
1	go	去	go	去
2	remove	去掉	cut into	切成
3	tear	撕成	remove	去掉
4	thaw	解冻	cut	切
5	remove	去除	marinated	腌制
6	crush	剁碎	spare	备用
7	remove the core	去核	wash	洗
8	soak	浸泡	soak	浸泡
9	cut into	剁成	thaw	解冻
10	cut	切	process	处理

cut 切

	text		image	
1	cut	切成	standby	备用
2	cut into	剁成	cut to	切成
3	tear	撕成	fish out	捞出
4	cut	切好	wash	洗净
5	smash	拍	drain	控干
6	smash	剁碎	go	去
7	rub	擦	pickled	腌制
8	wash	洗净	clean	洗
9	chop	切碎	soak	泡
10	chop	剁	want	要

cut into 切成

	text		image	
1	cut	切	cut	切
2	tear into	撕成	go	去
3	chop	剁成	standby	备用
4	grow	成长	wash	洗净
5	rub	擦	pickle	腌制
6	cut	切好	drain	控干
7	smash	拍	prepare	准备
8	crush	剁碎	soak	浸泡
9	chop	切碎	soak	泡
10	remove	去掉	remove	去掉

Table 7: Top-10 Similar Verbs by Text-Based and Image-Based Methods (continued)

ferment 发酵

	text		image	
1	carry	进行	pinch	捏紧
2	retract	回缩	rub	揉
3	stamp	戳	rub to	搓成
4	rub dough	揉面	rub to	揉成
5	rub	揉	roll up	卷起
6	wait	等待	conduct	进行
7	provoke	醒发	pack into	包入
8	pad	垫	rub dough	揉面
9	split	分割	divide into	分成
10	ferment	饧	press	按压

mix up 拌匀

	text		image	
1	season	调成	do	做
2	stir	搅匀	will	会
3	mix	拌成	will	要
4	mix	拌	be	是
5	mix up	匀	no	没有
6	seasoning	调味料	advance	提前
7	season	调	prepare	准备
8	add into	加入	have	有
9	filter	筛入	need	需要
10	full swing	上劲	come out	出来

fry 煎

	text		image	
1	bake	烙	fry	炸
2	fry	煎制	wait	待
3	decoct	炸	put in	放进去
4	cook to	烧至	change	变
5	stew	煸	stew	焖
6	flip	翻	put in	放入
7	add oil	加油	let	让
8	heat up	烧热	capped	加盖
9	heat	加热	flavor	入味
10	solidificate	凝固	solidificate	凝固

rub 揉

	text		image	
1	rub dough	揉面	rub to	搓成
2	rub	揉搓	rough to	揉成
3	rub to	揉成	roll up	卷起
4	rub	搓	divided into	分成
5	ferment	饧	provoke	醒
6	divided into	分成	become	成
7	provoke	醒	rub	搓
8	rub into	搓成	tightly	捏紧
9	split	分割	ferment	发酵
10	proof	醒发	roll dough	擀

spread 撒

	text		image	
1	leach	淋	put in	装
2	pour	浇	place in	摆在
3	embellish	点缀	done	做好
4	pendulum	摆	steam	蒸
5	sprinkle	撒入	come	来
6	come	刷	spread	铺上
7	put	放上	like	喜欢
8	wrap	裹	can	可
9	wipe	抹	put	放
10	fill	盛入	in	入

season 调成

	text		image	
1	mix well	拌匀	boil	汆
2	stir	搅匀	change	变
3	season	调	ribbonfish	带鱼
4	seasoned	调好	marinate	腌制
5	add	加上	put in	放入
6	mix	拌成	hot	烫
7	material	料	soak	浸泡
8	season in	调入	scoop	舀
9	drip	淋入	let	让
10	pour	倒入	put	放

we compute top-10 similar verbs by using each of these methods for 14 verbs. The result shows that text-based similarity and image-based similarity of verbs are useful for different kind of similar actions, and are complementary with each other.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR16E3, Japan.

REFERENCES

[1] Jingjing Chen and Chong-wah Ngo. 2016. Deep-based Ingredient Recognition for Cooking Recipe Retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*. 32–41. https://doi.org/10.1145/2964284.2964315

[2] Marie Katsurai, Takahiro Ogawa, and Miki Haseyama. 2014. A Cross-Modal Approach for Extracting Semantic Relationships Between Concepts Using Tagged Images. *IEEE Trans. Multimedia* 16, 4 (2014), 1059–1074.

[3] Yoshiyuki Kawano and Keiji Yanai. 2015. FoodCam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications* 74, 14 (2015), 5263–5287. https://doi.org/10.1007/s11042-014-2000-8

[4] Shangwen Li, Sanjay Purushotham, Chen Chen, Yuzhuo Ren, and C.-C. Jay Kuo. 2017. Measuring and Predicting Tag Importance for Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 12 (2017), 2423–2436. https://doi.org/10.1109/TPAMI.2017.2651818

[5] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag Ranking. In *Proceedings of the 18th International Conference on World Wide Web*. 351–360. https://doi.org/10.1145/1526709.1526757

[6] Taka Maenishi and Keishi Tajima. 2019. Identifying Tags Describing Image Contents. In *Proc. of ACM Hypertext*. https://doi.org/10.1145/3342220.3344936

Table 8: Sources of Images in Figure 3

	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step
1	7079260	3	7011659	2	7002718	10	7023990	4	7012133	6	7009031	2	7006787	8
2	7012549	8	7007718	4	7003088	4	7003589	13	7068443	4	7006082	6	7013131	7
3	7008275	7	7013153	4	7004761	5	7015787	2	7011465	3	7013879	8	7017142	7
4	7002143	6	7014095	3	7064292	10	7010708	12	7032355	2	7023307	6	7024778	1
5	7031178	7	7005793	3	7047591	10	7017542	9	7010649	3	7031834	3	7008072	9
6	7064214	9	7054320	7	7005527	8	7002900	6	7000271	11	7016705	9	7016459	6
7	7012954	10	7013467	2	7011214	7	7015240	10	7003823	16	7035566	4	7021811	1
8	7012083	7	7003989	5	7009076	14	7010883	6	7019773	4	7055258	11	7017187	6
9	7016291	1	7001324	3	7005420	3	7037091	7	7017598	1	7013128	3	7009320	1
10	7015195	11	7004091	3	7006716	12	7070894	11	7009651	22	7007400	4	7004466	4
11	7015211	9	7010991	4	7000749	9	7002580	6	7006155	5	7000177	3	7013814	3
12	7015415	4	7018293	5	7002217	7	7006074	11	7091694	7	7002635	6	7011710	12
	Cluster 7		Cluster 8		Cluster 9		Cluster 10		Cluster 11		Cluster 12		Cluster 13	
	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step
1	7003087	8	7078345	9	7066502	6	7006859	4	7013912	4	7015432	4	7003759	13
2	7019110	3	7008710	3	7000702	7	7015763	10	7010774	6	7013933	3	7004890	5
3	7016449	9	7030693	7	7008562	3	7081062	14	7015967	1	7008033	6	7006135	5
4	7059723	4	7008205	8	7013081	2	7093383	4	7023027	2	7001011	2	7066370	9
5	7080195	12	7021213	3	7031153	2	7093383	2	7012597	8	7005368	8	7017741	17
6	7009718	2	7016909	3	7011548	4	7008968	7	7080462	5	7018918	5	7001311	11
7	7005070	4	7079514	6	7055252	3	7016474	3	7011089	6	7003536	2	7015392	12
8	7004049	1	7006690	5	7000006	1	7008081	3	7014750	3	7011034	2	7027790	11
9	7020355	12	7012857	5	7017564	11	7001123	7	7015021	9	7036800	23	7012495	10
10	7013268	5	7062248	7	7000968	6	7005065	5	7078500	2	7009239	8	7018601	12
11	7017237	4	7004048	4	7090820	1	7009128	7	7007795	5	7070118	2	7035866	13
12	7003838	5	7002589	6	7024957	3	7006403	5	7004773	2	7011186	7	7014105	6
	Cluster 14		Cluster 15		Cluster 16		Cluster 17		Cluster 18		Cluster 19			
	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step	Recipe	Step		
1	7000226	7	7065141	16	7020961	5	7010343	3	7004892	2	7007724	17		
2	7008774	2	7008876	4	7048971	5	7011741	5	7003738	2	7035932	3		
3	7001393	1	7000269	3	7015868	2	7008172	8	7004665	2	7007932	11		
4	7024236	10	7020240	4	7008298	7	7013678	11	7006147	2	7080462	5		
5	7000404	1	7062335	10	7045196	4	7003091	5	7001291	14	7003492	5		
6	7016449	2	7013128	3	7015790	15	7008967	8	7008144	3	7012250	3		
7	7026540	2	7012703	4	7065220	3	7002589	4	7004273	9	7000271	11		
8	7003972	1	7012907	1	7077468	7	7015897	7	7007527	14	7009024	20		
9	7016825	1	7011020	1	7001393	1	7013509	10	7014446	6	7010167	9		
10	7091981	5	7013561	1	7007988	3	7009094	10	7014332	3	7087988	14		
11	7011219	1	7012465	2	7090780	2	7010095	4	7009255	4	7004462	14		
12	7007217	2	7089751	9	7015929	7	7007819	8	7013957	5	7011889	12		

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv e-prints* (jan 2013). arXiv:cs.CL/1301.3781

[8] Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. 2015. Named Entity Recognizer Trainable from Partially Annotated Data. In *PACLING 2015*. 148–160.

[9] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action Recognition using Visual Attention. *CoRR* abs/1511.04119 (2015). arXiv:1511.04119 http://arxiv.org/abs/1511.04119

[10] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[11] Yoko Yamakata, Hirokuni Maeta, Takuya Kadowaki, Tetsuro Sasada, Shinji Imahori, and Shinsuke Mori. 2017. Cooking Recipe Search by Pairs of Ingredient and Action—Word Sequence v.s. Flow-graph Representation—. *Transactions of the Japanese Society for Artificial Intelligence* 32, 1, WII-F (2017), 1–9.

[12] Jinfeng Zhuang and Steven C. H. Hoi. 2011. A two-view learning approach for image tag ranking. In *Proceedings of the 4th International Conference on Web Search and Web Data Mining*. 625–634. https://doi.org/10.1145/1935826.1935913

A SOURCES OF PHOTOS IN FIGURE 3

The images shown in Figure 3 were downloaded from https://www.haodou.com/recipe/. The recipe ID and the step number of each photo are shown in Table 8.