A Regional Food's Features Extraction Algorithm and Its Application

Trung Duc Nguyen

Faculty of Environment and Information Studies

Keio University

5322 Endo Fujisawa Kanagawa 252-8520 JAPAN

Submitted in partial fulfillment of the requirements for the degree of Bachelor

Advisors:

Professor Kiyoki Yasushi
Diep Nguyen-Thi Ngoc

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter describes the background of our research and the overview of the problem.

1.1 Background

In this paper, we present a food analysis system to discover the taste of food and understand the featured ingredients in a specific geographical region.

Cooking is the art of making foods. Nowadays, together with the development of technology and the availability of equipment in cooking, many supporting systems are introduced. For example, the cooking support system utilizing built-in cameras and projectors [?], the cooking support system by using ubiquitous sensors [?], the calorie measurement system by image processing [?] or the system which helps inexperienced users in understanding non-professional recipe descriptions [?], etc. However, by using analysis job we can discover the dominant ingredients and tastes in foods and understand how to alter the taste from one to another.

We can observe that in geographical regions that are far apart from each other often have different features and tastes. For example, the Kanto region, which is located in the East of Japan, often has dense taste in its foods, while the foods in Kansai region, which lies in the southern-central region of Japan's main island Honshu, often has a diluted taste. The reason is because each region has its own special materials for foods and people in these regions have different habits in cooking food. To understand each region's featured taste we need to answer the following questions: "How can we understand the different features of each region's food?" and "What effects change the region's food taste?". Among the many factors that affect a food's taste, the combination of materials is a direct and important factor. Each recipe has a list of its own ingredients together with their amount. This leads us to the idea that we could automatically achieve the features of a region's foods by analyzing the materials. In order to make advantages of recipes in food analysis, we collected recipes in many regions to build a recipe database and propose some analysis algorithms based on text processing.

1.2 Challenges and Research Goals

We also realize that understanding the region's featured taste and the preferred materials has an application in supporting cooking activities. For example, imagine someone living in Kanto region who wants to eat some traditional foods in the Kansai region. They know the original recipe but there are some tastes in Kansai region that are not favored. They would prefer that traditional foods with replaced ingredients that are easy for Kanto people to eat. Conversely, someone living in Kanto region might want to try Kanto foods with Kansai taste. Solving this kind of problem means we can build up a system which can help people satisfy their taste. The recipes, which are made by the system, would be flexible and diverse.

In the existing cooking support systems, the methods vary such as image processing, text retrieval, sensing, etc. We use the text processing approach to directly analyze the recipes with their ingredients and amount of ingredients.

1.3 Structure of Thesis

The outline of this paper is as following. The algorithm and the experimental results are introduced in Section 2 and Section 3 respectively. Section 4 describes the web-based application using the proposed algorithm while Section 5 concludes and discusses the remaining problems and future works.

Chapter 2

Food's Feature-Ingredient Extraction Algorithm

In this section, we propose an algorithm for analyzing the dominant materials which are often used in a region. We define a material in a region to be a featured one if it appears many times with a large amount and be unique among recipes in that region. To evaluate whether it is featured or not, we suppose that the following questions should be answered: "How often, how much, and how unique the material is?". Respectively, we propose three kind of functions to answer these questions. They have the key role of the metrics for the featured ingredient's evaluation.

2.1 Ingredient Frequency

The first function named IF (Ingredient Frequency) is used to treat the question "How often does the material appear in a region?". The higher frequency an ingredient appears in a region, the higher possibility it is the region's featured ingredient. In each recipe, an ingredient only appears one time. Thus, the time that ingredient appears in the region is the number of the recipes in the region has it as ingredient. Because the database we have from the Internet are often unbalanced, there are some regions that have more recipes than others. Thus to make it indenpendent from the database, we prefer to use the ingredient's frequency rather than its appearance times. This function is formed by the number of times the ingredient appears in the region's recipes over the number of total recipes in that region. Let R be the set of all recipes (r) in a region and i be an ingredient which appears in the region. The function is formed as follows:

$$IF(i, R) = \frac{|\{i|i \in r, r \in R\}|}{|R|}$$

Becaue the IF value is the ingredient's frequency, it takes the value between 0 and 1.

2.2 Ingredient Amount

The ingredient's frequency has little meaning if there is a small amount of it in the recipes. Thus, the taste of a food not only depends on the ingredients, but also the amount of the ingredients. Even when an ingredient has a high value of IF, it might not be the region's featured ingredient. Thus, the second function, IA, is proposed for the question "How much?"

Let r be a recipe in the set of recipes S and ingredient i is in r. We define the mean function $M(i, S)$ be the mean amount of i in S as follows:

$$M(i, S) = \frac{\sum^{i \in r, r \in S} amount(i, r)}{|\{i|i \in r, r \in S\}|}$$

in which $amount(i, r)$ is the amount of ingredient i in recipe r.

We also assume that AR is the set of all recipes in the country regardless of the region it belongs to, while R is the set of all recipes just in a specific region. Thus, $M(i, R)$ calculates the mean amount of ingredient i in the region's recipes (R) while $M(i, AR)$ calculate the mean amout of ingredient i in all the country's recipes (AR). We have the IA function as follows:

$$IA(i,R) = \frac{M(i,R)}{M(i,AR)}$$

Because the IA function calculates the mean of ingredient's amount, it is independent to the frequency of that ingredient. The higher IA value is, the higher possibility it is the region's featured ingredient. Because both numerator and denominator in the formula have the same unit, the IA value is non-unit. Therefore, regardless to the variety of the ingredient's unit, we have a stable metric for evaluating the ingredient's amount.

2.3 Ingredient Unique

The IF and IA functions above might tell us how often an ingredient appears in the region, but this ingredient can often appear in many regions. To be a featured ingredient of a region, the ingredient must satisfy the condition that it appears in the region but doesn't appear in many other regions. We propose the third function IU as follows:

$$IU(i,A) = \log \frac{|A|}{|\{i|i \in a, a \in A\}|}$$

in which i is the ingredient in region a and A is the set of regions.

This function calculates the uniqueness of an ingredient among all the regions. The more often an ingredient appears in different regions the less unique it is. In other words, it is not the featured ingredient of the region. The higher IU value corresponds to higher possibility it is the region's featured ingredient. We use the log scale to make sure the IU values are not too big.

2.4 Featured Index

Featured Index, which is denoted by FI, is the index used to rank ingredients in a region in term of featured ingredient. We realize that these three functions are all proportional to the rank of the featured ingredient, thus we proposed FI to be the production of these three function's values as follows.

$$FI(i, R) = IF(i, R) \times IA(i, R) \times IU(i, A)$$

The FI function returns the featured index of ingredient i in a region which has a set of recipes R. A is the set of all regions in the country. The ingredients which have the highest FI would be the featured ingredients. On the other hands, the ingredients which have the lowest FI would be considered as the common ingredients for every region.

Chapter 3

Implementation of HUSTLE

This section describes the recipe databse and experimental studies on this database by applying the featured ingredient analysis algorithm.

3.0.1 The Recipe Database

We build a recipe database in which recipes are grouped by region. A script written in Python crawls all the recipes from a Japanese cooking website [?]. We chose this website because the recipes are typical foods grouped by regions. The website is only for Japanese recipes, thus we now only have the database for Japanese foods. Each food is characterized by its name, the region it belongs to and its recipe. Each recipe is stored as a map collection in which the ingredient is the key and the couple of amount and unit is the value. Each of the recipes we get from the website is created for various amounts of people. For example, there are recipes for 4 people but there are also recipes for 3 people. Thus we need to normalize the ingredients' amount in each recipe for one person.

There are about 200 recipes over 7 regions in Japan: Kanto, Hokkaido-Tohoku, Shikoku, Tyubu, Kyusyu-Okinawa, Kansai and Tyugoku. We calculate all the above functions for every recipe in Japan, but we only show the experimental results of Kanto and Shikoku within this paper. We chose these two regions because they lie far apart in different islands of Japan. The experimental results are discussed below.

Table 3.1: Ingredient Frequency of Ingredients in Kanto region vs Shikoku region

Kanto region		**Shikoku region**	
Ingredient	**IF**	**Ingredient**	**IF**
Soy Sauce ()	1.00	Soy Sauce ()	1.00
Miso ()	0.9	Salt ()	1.00
Sugar ()	0.83	Rice ()	0.83
Sake ()	0.83	Sake ()	0.67
Salt ()	0.67	Green onion ()	0.50
...	...	...	...
Dried bonito ()	0.08	Kelp soup ()	0.16
Pumpkin ()	0.08	Deep-fried Tofu ()	0.16
Kamaage Shirashi ()	0.08	Seared bonito ()	0.16

3.0.2 Ingredient Frequency

UTF8min

Table **??** shows that there are some common ingredients which often appear in both Kanto and Shikoku regions such as Soy Sauce (), Sake (), Salt (),... This is reasonable because we know that these ingredients are common in Japan. Because they often appear in other regions, the IF function is not enough to evaluate the region's featured ingredients. However, it helps us partially understand the habit in using materials in regions. For example, Green onion () often appears in Shikoku but not in Kanto region and Sugar () often appears in Kanto but not in Shikoku region. This leads us to the idea that typical Kanto foods are often sweeter than Shikoku foods.

3.0.3 Ingredient Amount

UTF8min

Table **??** shows the result of the IA value for Kanto and Shikoku region. We can see that

Table 3.2: Ingredient Amount of Ingredients in Kanto region vs Shikoku region

Kanto region		**Shikoku region**	
Ingredient	**IA**	**Ingredient**	**IA**
White radish ()	4.27	Shredded seaweed ()	6.00
Tempura flour ()	3.20	Carrot ()	3.95
Shredded seaweed ()	3.00	Tempura flour ()	3.20
...	...	...	...
Taro ()	0.02	Sweet potato ()	0.06
Cake flour ()	0.02	Chicken thigh ()	0.05
Field mustard ()	0.02	Sushi vinegar ()	0.05

most of the IA values are around 1, which means there is not much difference in the way of using an ingredients' amount between Shikoku region and other regions. However, there are some interesting results. For example, in Kansai region, the mean amount of pepper () is 11 times greater than the mean amount of total peper in Japan. See details in Table **??**.

3.0.4 Ingredient Uniqueness

UTF8min

Table **??** reflects the fact that the common ingredients such as Salt (), Sweet cooking wine (), Ginger (), Soy sauce () appear in almost every regions in Japan while the ingredients such as Peanut () and Chive () are not too common and mostly appear in only one region. The ingredients which have the IU value of 0 appear in every region.

3.0.5 Featured Index

UTF8min

The Featured Index (FI) is the main metric we use to elvaluate the regions' featured ingre-

Table 3.3: Ingredient Amount of Ingredients in Tyubu region vs Kansai region

Tyubu region		**Kansai region**	
Ingredient	**IA**	**Ingredient**	**IA**
Pork loin ()	25.50	Pepper ()	11.00
Seaweed ()	6.00	Sweet cooking wine ()	6.88
Green onion ()	3.72	Soy sauce ()	5.49
Onion ()	3.60	Green onion ()	3.72
...	...	...	...
Taro ()	0.03	Milk ()	0.04
Cake flour ()	0.02	Minced chicken ()	0.03
Pepper ()	0.01	Soup ()	0.01

Table 3.4: Ingredient Uniqueness of Ingredients in Japan

Ingredient	**IU**
Peanut ()	2.80
Chive ()	2.80
...	...
Salt ()	0.00
Sweet cooking wine ()	0.00
Ginger ()	0.00
Soy sauce ()	0.00

Table 3.5: Featured Index of Ingredients in Kanto region and Shikoku region

Kanto region		**Shikoku region**	
Ingredient	**FI**	**Ingredient**	**FI**
Natto ()	0.60	Kelp ()	1.40
Dried radish ()	0.47	Sea bream ()	0.94
Saury ()	0.47	Ponzu suace ()	0.90
...	...	...	...
Vineger ()	0.00	Sweet cooking wine ()	0.00
Shredded seaweed ()	0.00	Egg ()	0.00
Wine ()	0.00	Wine ()	0.00
Ginger ()	0.00	Rice ()	0.00

dients. Table **??**, which is the experimental result of FI calculation for Kanto vs Shikoku region, shows us some interesting information. For example, Natto () is the ingredient which has the highest FI value in Kanto region. This means Natto () is possibly the featured ingredient of Kanto region. In Shikoku region, Ponzu sauce () is also often used for Shikoku's foods. The FI of the same ingredient for different regions might differentiate but we figure that if an ingredient ranks high in one region, it cannot rank high in any other regions. The same thing is true for the low-rank ingredients.

Acknowledgments

Foremost, I would like to express my great appreciate to my supervisor, Professor Hideyuki Tokuda, Professor Hideyuki Tokuda who gave me advice, guidance and encouragement from the beginning when I has just joined to Hide.Tokuda Laboratory.

I would like to thank Professor Jun Murai, Osamu Nakamura and Kenji Takeda, Associate Professor Keisuke Uehara, Hiroyuki Kusumoto, Jin Mitsugi, Rodney D. Van Meter, and Kazuki Takashio, Assistant Professors Massaki Sato, Noriyuki Shigechika and Jin Nakazawa as well.

I am extremely thankful for Dr.Takuro Yonezawa, who has always supported and guided me when I has just joined to CPSF research group of Hide.Tokuda Laboratory. He also helped me analyze the meaning of my research, clarify my research goal and guided me to write technique papers.

I would like to thank my advisor, Mr.Takuya Takimoto, who has always taken care of me since I joined Hide.Tokuda Laboratory. He gave me a lot of advice and support when I am not sure about my research motivations. I would like to thank Mr.Tomotaka Ito and Mr.Ogawa Masaki for a listening and giving me many useful discussion points and advices.

I would like to thank members of CPSF research group in Hide.Tokuda Laboratory for their great friendship and support.

I would like to thank Mr.Tomotaka Ito, Mr.Ogawa Masaki, Ms.Vu Le Thao Chi, Ms.Nguyen Thi Ngoc Diep, my friend Ms.Vu Thi Thai Ha, Ms.Luu Thanh Huong, for their comments for the thesis structure and my English writings.

I send my special thank to my experiment supporters: Takuya Takimoto, Yuuki Nishiyama, Teruaki Ishiguro, Hiroki Shoji, Yutaro Kyono, Nguyen Anh Tien, Do Trung Kien, Nguyen Tien Thanh, Tran Duc Thang Nguyen Thanh Tung, Nguyen Trung Duc, Tran Ngoc Anh and Dinh Hoang Long.

And last but not least, I am heartily grateful to my family. This is the first time I have left my family to start a new life in Japan, many things happened, success and failure, sadness and happiness, tears and smiles. But they still and always are there, giving me strength to follow all goals of my life.

January 4, 2014

Trung Duc Nguyen

Publications

Long Talk

Trung Duc Nguyen, Diep Thi-Ngoc Nguyen, Yasushi Kiyoki.A Regional Food's Features Extraction Algorithm and Its Application. Workshop on Cooking and Eating Activities in conjunction with ACM Conference on Multimedia. Oct 21, Barcelona, Spain.

Poster

SFC OPEN RESEARCH FORUM 2014.