Evaluations on Classified Selection of Dense Vectors for Vegetable Geographical Origin Identification System Using Trace Elements

Nobuyoshi Sato[1], Minoru Uehara[2], Koichiro Shimomura[3], Hirobumi Yamamoto[3], and Kenichi Kamijo[3]

[1] Faculty of Software and Information Science, Iwate Prefectural University
152-52, Sugo, Takizawa, Iwate 020-0193 Japan
nobu-s@iwate-pu.ac.jp
[2] Depertment of Information and Computer Sciences, Toyo University
2100, Kujirai, Kawagoe, Saitama 350-8585 Japan
uehara@cs.toyo.ac.jp
[3] Plant Regulation Research Center, Toyo University
1-1-1, Izumino, Itakura, Gunma 374-0193 Japan
{shimomur,yamamoto,kamijo}@itakura.toyo.ac.jp

Abstract. Recently, in Japan, some farming districts established their locality as brands, and prices of agricultural products differs from their grown places. This induced some agricultural food origin forgery cases. Food traceability systems are introduced and some are now in operation to solve this problem. However, food traceability systems have vulnerabilities in their nature because they traces only artificially attached IDs. So there are possibility to forge ID and packages, and switching the vegetables in packages. So, we developed a geographical origin identification system for vegetables by using their trace element compositions. Trace element means very small quantities of elements. This system gathers trace element data of vegetables when shipping from farms, and stores them into databases located in farming districts. In case of a vegetable which has doubtful geographical origin is found in markets, their trace element compositions are measured and compared with data in databases to find its actual geographical origin. Our system judges geographical origin by whether correlation coefficient. This requires calculating correlation coefficients for identifying one and all stored data. However, this is not scalable for the number of data. In this paper, we describe a method to limit the number of data to be used to calculate correlation coefficients before calculating them, and realize scalability.

Keywords: Geographical origin identification by trace element compositions, efficient retrieval of dense vector.

1 Introduction

Recent years in Japan, some farming districts established their districts as brands; their products are traded in expensive values. This gave rise to forging geographical

T. Enokido, L. Barolli, and M. Takizawa (Eds.): NBiS 2007, LNCS 4658, pp. 385–396, 2007.

origin of agricultural products and problems to guaranty food safety. To solve these two problems, food traceability systems are now on the way to be used universally. However, since food traceability systems employ artificial materials such as barcode, RFID tags as ID attached into cardboard box and containers, it is possible to forge IDs, packages, and switching packages or contained foods.

In these situation, we developed a vegetable geographical origin identification system which distinguishes differences of trace element compositions of vegetables depend on cultivated places [1]. Here, trace element means that very small quantities of elements; and it is a technical term in chemists. Normally, vegetables are cultivated on the soil. Since trace element compositions of the soil differ from districts, even farms, absorbed trace element compositions in vegetables also differ. In our system, trace element compositions are measured at farms when shipping, they are registered in databases located in each farming district. In case of a vegetable of which cultivated place is cheated was found, its trace element composition is measured and it is compared to all registered trace element data. Correlation coefficient between doubtful one and one of stored data is calculated for comparison. If the correlation coefficient overcomes a certain threshold, it is considered that two vegetables may be cultivated the same farm. In case of a vegetable of which geographical origin seem to be cheated is found, its trace element composition is measured and compared to all registered data in databases. If geographical origin of doubtful vegetable by food traceability system is included in a set of geographical origin by our system, it is considered that geographical origin of doubtful vegetable is not cheated, otherwise, it is considered that it is cheated and its real geographical origin is a certain farm which has the highest correlation coefficient.

In our system, trace element compositions are expressed as high dimensional vectors. Our system compares query vector and stored vectors in databases by calculating correlation coefficient. This is one-to-one comparison, and this is a obstruction to make system scalable. To realize scalability, there are two possible strategies; the one is to narrow target of one-to-one comparison, the other is to develop yet another techniques rather than one-to-one comparison. However, the latter is not realistic way. Therefore, we have realized a certain extent of success to realize efficiency by using Similarity Preserve Hash (SPH)[2], presented by Moses Charikar. This method succeeded to reduce the number of one-to-one comparison and time to identify. SPH itself is jock job method and it is not enough efficient. However, we invented employment of SPH and we reduced time to retrieve data from database to about few seconds [3]. Although SPH can be used for practical generic method to narrow high dimensional dense vector, reducing retrieval time by SPH seem to reach the limit. Therefore, in this paper, we discuss on a method to reduce retrieval time to extract trace element data from database by encoding trace element compositions into some classes. This method aims at distribution of values of trace element compositions.

Organization of this paper is as following; in Section 2, we describe some related works on geographical origin identification. In Section 3, we explain our vegetable geographical identification system. In section 4, we discuss on requirements to classify trace element compositions, and describe a method to encode the classes. In section 5, we will show evaluation. Finally, we will conclude.