Ontology based Knowledge Map Construction for a Smart Knowledge Service

Asanee Kawtrakul1, 2, Therawat Tooumnauy1, Vasuthep Khunthong1, and Aree Thunkijjanukij3
Kanok Lohapiyaphan4, Yuphayao Pothipaki4, Werachai Narkwiboonwong4 and Anan Pusittigul4

1 Department of Computer Engineering, Kasetsart University, Bangkok, Thailand
   Email: {asanee.kawtrakul, therawat.tooumnauy, vasuthep.khunthong}@uknowcenter.org
2 National Electronics and Computer Technolgy Center, Thailand
   Email: asanee.kawtrakul@nectec.or.th
3 Agriculture Knowledge Centre, Kasetsart University, Bangkok, Thailand
   Email: libarn@ku.ac.th
4 Agricultural Land Reform Office, Bangkok, Thailand
   Email: {webmaster, yuphayao, ananp}@alro.go.th , werachai_n@yahoo.com

Abstract

Problem-solving is one important form of intelligent behavior, where the goal is to find a solution which satisfies certain criteria. Lesson learned from the past in problem solving (e.g. how to protect the disease, how to control the disease) and valuable information tracked from the previous events and the recurrence; e.g. disease outbreak, are very important for knowledge sharing. This valuable knowledge is distributed over several websites among various kinds of sources in heterogeneous and unstructured formats. In order to reduce time consumption for users to access, construction and linking information space that attached with digesting information should be developed. This paper presents a framework for constructing a knowledge map from the webs that spread throughout the Internet focusing on semantic links between Problems, problem-solving Methods and problem-solver Man (PMM map). Based on specific-task ontology as a representation of specific-domain conceptualization, a knowledge map could be constructed for smart accessing as one stop service. Based on specific-information extraction techniques, an indicative knowledge could be summarized and attached to the knowledge map for supporting multi-layers knowledge visualization. This proposed framework will be evaluated on rice applications under the project called ALRO-CyberBrain1 and used by the knowledge broker who is expected to provide knowledge services at the Community Learning Center of each district office.

Keywords: Explanation Relation Extraction, Indicative Summarization, Rice Ontology, Knowledge Map, Smart Knowledge Services

1 ALRO-Cyberbrain is ongoing project supported by Kasetsart University, NECTEC (National Electronics and Computer Technology Center) and Agricultural Land Reform Office.
1. Introduction

Knowledge portal has been developed for many purposes such as specific-domain knowledge organizing (Schneble J., 2002), or a portal for accessing specific information, e.g. Terrorism Knowledge Portal (http://ai.bpa.arizona.edu/research/coplink/tkp.htm), insurance and finance Knowledge Portal (www.cii.co.uk) and Health Knowledge Portal (www.nks.nhs.uk).

In this paper, knowledge portal (Kawtrakul A., 2007) is constructed for a smart knowledge access. Task-oriented ontology (Kawtrakul A. et al., 2006) has been applied for aggregating the related knowledge consisting of Problems, problem solving Methods and Man who solves the problems, hereafter called PMM. Since the web consists of a large amount of unstructured or semi-structured natural language text, natural language processing and information extraction techniques have been applied for supporting multi-layers knowledge visualization.

The PMM framework brings together information from different sources and different structures which scattered in the web for the construction of the knowledge map. The usage of Ontology based integration approach provides the higher degree of knowledge aggregation. A sample of rice ontology has been developed to represent the concepts/topics with specific properties as knowledge and relationships, such as Disease (P: problem), Preventive or Corrective Solutions (M: Methods), Expert (M: huMan or Solver). Man or solver refers to a person or group of persons that have either solutions or the policies/strategies about the problem. In this work, we extend Know-why (Pechsiri C., Kawtrakul A., 2007), Know-who (Thamvijit D. et al., 2005) and Know-how (ongoing project) to extract the web-based information to produce the knowledge map of “Why the problem happens”, “How to solve it” and “who has potential to solve this problem”.

Rice is one of the most important crops of Thailand, with paddy output of 23.17 million tons in 2007 for domestic consuming and exporting. Thailand is also the world's biggest rice exporter: annual shipments are worth more than $10,000 million and reached 7.5 million tons in 2007. However, rice production is facing serious problems such as global warming, oil crisis, saline and sulfate soil problems, disease and pests and also fertilizer price increasing. These problems will reduce the amount of the rice yield and increase cost of planting. Accordingly, this work intends to develop PMM by using rice domain as a case study.

The remainder of this paper is organized as follows. In section 2 we introduce background knowledge behind this framework. Section 3 describes the framework architecture of knowledge map construction system. Conclusion and future work is presented in section 4.

2. Requirements in Knowledge Aggregation related to Language processing

Lessons learned from the past in problem solving of food safety and human health and tracking valuable information from the previous events and the recurrences are very important. In order to reduce time consumption for users accessing, the construction and linking of information space with salient information should be developed (see Fig.1).

Problem-solving is one important form of intelligent behavior (Kennedy J., et.al, 2001) where the goal is to find a solution which satisfies certain criteria. It requires abductive reasoning where we apply deductive reasoning in combination with natural language processing. In classical applications as well as in expert systems, abductive inference shown in (Shohei K., et al., 2003) is a complex problem (creation and maintenance) as it is simulated by deductive procedures or rules. In addition, qualitative reasoning is concerned with modeling and inference techniques where continuous phenomena are discrete in a finite number of qualitative categories.
In order to provide knowledge service to serve a user, by following the concept of PMM, we need to collect information distributed over several websites. For example (see Fig. 2) if the farmers found that their rice has gray scar on leaves, they need to find what happened. Even though they could found this symptom, cause of disease (Problem) and method to cure (Method) on website of Rice Department of Thailand but they couldn't found who or which organization they should contact with. To get all information needed, farmer or knowledge broker who is expected to provide knowledge services may need to search and explore many websites that may consume time and require very well search skill.

Figure 1. Knowledge Integration between Problem, Solution and Solver

Figure 2. An Example of Knowledge Aggregation for a Query “brown scar on rice leaves”
Since almost information represented in unstructured text, Language Engineering is emerged to Knowledge Engineering for understanding and extracting the intended answers. This paper, will focus on Know-why (i.e., what is the cause of Problem), Know-how (i.e. how to solve the problem or Method) and Know-who (who is the right person for supporting or Man). The examples of training corpus answer extraction for each query are shown as the follows:

Example 1: A sample of annotated corpus for training the system for “Know-why” answer.

```
<disease>โรคไหม้ Rice Blast</disease>
<cue>มีอาการดังนี้ have symptom as follow</cue>
<symptom1>ใบมีแผลจุดสีน้ำตาลคล้ายรูปตา Leaves have brown scar shape like eye</symptom1> <symptom2>มัน center of scar is gray</symptom2> <symptom3>ความกว้าง of scar wide around 2-5 mm. and length around 10-15 mm.</symptom3>
```

Example 2: A Sample of annotated corpus for training the system for “Know-how” answer.

```
<disease>โรคไหม้ Rice Blast</disease>
<cue>การป้องกันกีจัด Prevent and Cure</cue>
<method1>1. ใช้พันธุ์ต้านทานโรค เช่น Suphanburi1, Suphanburi2 and Chainat1 / 1. use resistant breed as Suphanburi1, Suphanburi2 and Chainat1</method1>
<method2>ฉีดพ่นคาซูกะมัยซิน follow ratio 20 g per water 20 liter / 2. spray Kasugamycin follow ratio 20 g per water 20 liter</method2>
```

In case of “Know-who”, the system aims to find Man who will help to solve the problem (may be expert or person in charge of responsible organization) where the complete information might be distributed over several sites or sources. As the example shown below, the contact information of “Mr. Poonsak Mekwattanakan” is from the source 2.

Source1:
```
<research>การพัฒนาสายพันธุ์ข้าวต้านทานโรคไหม้ Development rice breed : improvement resistant to rice blast</research>
<author>นายพูนศักดิ์ เมฆวัฒนากาญจน์ Mr. Poonsak Mekwattanakan</author>
```

Source2:
```
ข้าราชการปฏิบัติงานด้านการวิจัย Government’s researcher
<researcher>นายพูนศักดิ์ เมฆวัฒนากาญจน์ Mr. Poonsak Mekwattanakan</researcher>
<telno>โทรศัพท์ : (045) 344103 – 4 ต่อ ext 122</telno>
```

Besides that we also need an intelligent application for serving user to browse all the knowledge. Ontology becomes an essential tool as a skeleton of knowledge which not only used for knowledge aggregation but also for query expansion. Rice Ontology is then provided by the expert in agriculture field. The ontology is tested in information retrieval process by using concepts in topic of “Rice diseases control”. We found that retrieval result from ontology delivered Recall value 3 folds as high as the Recall value delivered by the conventional method. The following examples show a set of ontology-based queries: (Thunkijjanukij, A., 2008)
Query strategy: concept[RICE] AND concept[BLAST DISEASE]
Query concept: concept[RICE] with relation “hasDisease” concept[BLAST]
concept[BLAST] with relation “causedBy” concept[BLAST PATHOGEN]
concept[BLAST] with relation “hasControlMethod” concept[CULTURAL METHOD]
concept[BLAST] with relation “hasResistantVariety” concept[RICE VARIETY]
concept[BLAST] with relation “hasChemicalControlSubstance” concept[CHEMICAL SUBSTANCE]
concept[BLAST] with relation “hasCulturalControlMethod” concept[CULTURAL METHOD]

Term definition:
- concept[RICE] hasScientificName Oryza sativa
- concept[RICE] hasCommonName rice
- concept[RICE] hasLocalName ข้าว
- local name of concept[RICE] hasSynonym ข้าวเหนียว ข้าวเจ้า
- concept[BLAST] hasMainTerm blast
- term of concept[BLAST] isTranslatedTo โรคไหม้
- concept[BLAST PATHOGEN] hasScientificName Pyricularia grisea
- Scientific name of [BLAST PATHOGEN] hasSynonym Pyricularia oryzae; Magnaporthe grisea
- concept[CHEMICAL SUBSTANCE] hasSubClass carbendazim, tricyclazole
- concept[RICE VARIETY] hasSubClass RD1, RD9, RD11, RD21, Klong Luang 1
- concept[CULTURAL METHOD] hasSubClass ..... etc.

Search query: (Oryza sativa OR rice OR ข้าว OR ข้าวเหนียว OR ข้าวเจ้า) AND (blast OR โรคไหม้ OR Pyricularia grisea OR Pyricularia oryzae OR Magnaporthe grisea OR Pyricularia oryzae OR Magnaporthe grisea) AND (carbendazim OR tricyclazole OR RD1 OR RD9 OR RD11 OR RD21 OR Klong Luang 1)

At the current state, even we have developed Thai Rice Production Ontology which comprises of 3,544 concepts, 5,518 terms and 45 relations; it is still not enough due to the lexicon growth. Language engineering, especially name entities recognition becomes one of the solutions for ontology maintenance.

3. A Framework Architecture for Constructing the Knowledge Map

Our framework is separated into 3 parts (Fig 3). First module is information collector and preprocessor, which collects information and prepare necessary data for the next process. Second one is knowledge map constructor, which extracts, merges and keeps knowledge in useful form. The third module is knowledge service provider consisting of visualization browser and question-answering processing.
3.1 Information Collector and Preprocessor

This part is composed of two components, periodic web crawler and preprocessing consisting of noise cleaning and language preprocessing. Periodic web crawler will crawl related document from Internet, seeds URLs set by expertise in their domain. In this paper, our domain of interest is agriculture field. List of seeds URLs containing the information related to human expert from government, universities, research institutes, newspaper and other high reliable web sites, are collected. After getting related document from several web sites, the preprocessing is needed for cleansing a noise like advertisement, navigation, hyperlink, etc. and for language preprocessing like word segmentation, part-of-speech recognition (Sudprasert S., et al., 2003), and Thai EDU segmentation (Chareonsuk J., et al., 2005).

3.2 Knowledge Map Constructor

This part is the core of system that will construct knowledge map by using natural language processing techniques. The main process is separated into 4 parts, ontology acquisition, information extraction & integration, knowledge extraction & generalization and inference engine.

3.2.1 Ontology Acquisition

The ontology is created and managed by experts in the community of agriculture domain. To handle this task, we use AGROVOC workbench tool (Yongyuth P., et al., 2008) that allows experts to create, verify, and manage ontology in order to create clean and useful ontology. AGROVOC can process content in multi-lingual environment and supports community of expert by authority system. Moreover, AGROVOC ontology can be exported to many other formats (TBX, SKOS, XML and OWL file), which increases the flexibility and compatibility with other components.

3.2.2 Information Extraction & Integration

This module will extract information to answer main question know-how (Method) and know-who (Man). Main NLP techniques used in this part is Name Entity Recognition (NER) and Information Extraction (IE).

Traditional name entity (NEs) follows the MUC framework, defined as name of persons, organizations, and locations, temporal information, and numerical expressions. In order to extend the system to support a specific domain like agriculture, we have to extend the NER accordingly to fit the new domain. In this work, we enrich the list of traditional NEs with the following terms from agriculture domain: plant/biomedical name, animal name, disease name, chemical name, and pathogen name. To solve NE recognition (NER) problems, we apply techniques based on (Chanlekha H., et al., 2004) and improve it by adding global feature and change machine learning model to conditional random field which providing better result than using maximum entropy model.

For NER, we can use the same class and method to cope know-why; know-how and know-who but information extraction is more specific problem and need different module to handle. In know-why, we extract causality from text by using cue word and pair of cause-effect verb following work (Pechsiri C., Kawtrakul A., 2007).
In know-how, we are interested in extracting procedural information that can be separated into three types, instruction, advice and warning. The process starts with identifying the answer units by using cue word set, such as “ดังต่อไปนี้” / “as in the following” “ดังนี้” / “as follows” “โดย” / “by” etc. In order to identify procedural information in unstructured text, we also need to analyze both syntactic and semantic in EDU level based on heuristic rules such as the set of procedural verb set, {“ใช้” / “use” “พ่น” / “spray” “ผสม” / “mix” ……}, could be used as a prompt for identifying the procedural EDU units.

In Know-who, system will extract personal data from web including their education and work experience. This process focuses on finding man who have role in agriculture domain. We use person class in NER to identify person, then do the reasoning about their expertise area based on their research papers if they were researchers or if they have position in some agriculture organization, they should have knowledge about agriculture domain with respect to their positions.

3.2.3 Inference Engine

After extraction of new knowledge, it needs consistency check with the existing knowledge. Sometimes, new knowledge can be combined with the existing one to derive more knowledge. This module will do reasoning to create new knowledge. By using reasoning, we can get more knowledge than just union of new and exist. Reasoning is done by using some heuristic rules that is defined based on some observation data.

3.3 Knowledge Service Provider

This part shows the web application (see Fig.4) that utilizes the knowledge to serve user following the concept of PMM, providing information as a one stop service. User can access the information online via browser through input keywords related to their problems. For providing knowledge service, terms in ontology can be used to extend an query that user inputs and cluster the answers according to the answer concept. For example, when user input the query as “Suphanburi3” (Rice Breed), the system will get all knowledge about “Suphanburi3” and shows the results by linking those information space consisting of the related Expert, Disease, Pest, Production, etc.

Figure4. An Example of the output of PMM application
4. Conclusion and Future Works

This paper presented our on-going research on collaborative knowledge annotation based on 'Know-why', 'Know-who' and 'Know-how' analysis tool for specific answer indications (e.g. the process of Disease Prevention or control, and the process of Rice Preparation). We introduced the conceptual framework, which is ongoing project called ALRO-CyberBrain for PMM map generation. Based on Rice Ontology, a knowledge map related to rice problem-solving such as “who do what”, “how to do” and “What is the cause of X?” will be a tool supporting knowledge service to the farmers using through the Knowledge Broker at the Community Learning Center of each district office.

Acknowledgement

This research has been partial supported by FAO under the project “AGROVOC Concept Server Workbench” and “The Development of Ontological Information for Agriculture in Thailand” project under the collaboration between University of Tokyo, AIT (Asian Institute of Technology), Kasetsart University and NECTEC (National Electronics and Computer Technology Center). We would like to give special thanks to Pacharee Varasrai and Mukda Suktarachan for their support in corpus preparation.

References


Thunkijjanukij, A. (2008) Study and development of agricultural knowledge base and ontology for decision support and question-answering system : A case study of Thai...
rice. Research Report, Kasetsart University Research and Development Institute, Kasetsart University, Bangkok.

Schneble J., Case Study: Knowledge Portal at Cisco, Published by the American Society for Training and Development (ASTD), May, 2002.

