Ontology Driven K-Portal Construction and K-Service Provision

**Asanee Kawtrakul, Chaveevan Pechsiri, Trakul Permpool, Dussadee Thamvijit,
Phukao Sornprasert, Chaiyakorn Yingsaeree, Mukda Suktarachan**

Department of Computer Engineering, Kasetsart University,Thailand
ak@ku.ac.th,asanee_naist@yahoo.com

Abstract

Knowledge has been crucial for the country's development and business intelligence, where valuable knowledge is distributed over several websites with heterogeneous formats. Moreover, finding the needed information is a complex task since there has been lack of semantic relation and organization. Even if it has been found, an overload may occur because there is no content digestion.

This paper focuses on ontology-driven knowledge extraction with natural language processing techniques and a framework of user-centric design for accessing the required information based on their demands. These demands can be expressed in the form of Know-what, Know-why, Know-where, Know-when, Know-how, and Know-who for a question answering system

1. Introduction

In order to understand a situation, a decision-maker needs data, information and knowledge. The knowledge consists of data items and/or information organized and processed to convey understanding, experience, and expertise that are applicable for problem solving (Turban et. al., 2005). However, sources of these data are scattered at several locations and websites with heterogeneous formats that offer structured information to large volumes of unstructured information. Moreover, the needed knowledge has been too difficult to find since the traditional search engines return ranked retrieval lists that offer little or no information on the semantic relationships, and even if it has been found, often overload since there is no content digestion. Accordingly, the users or knowledge workers must spend time browsing and reading to find out how various information are related and where each falls into overall structure of the problem domain.

In this paper, we present a systematic attempt to construct the knowledge portal, which aims to integrate and organize the data/information resources dispersed across web resources in a manner that makes them useful, and a framework of user-centric design for accessing the requested information. Since the web consists of a large extent of unstructured or semi-structure natural language text, several techniques both in language & knowledge engineering and ontology engineering are applied to the knowledge portal construction. These include named-entity recognition (Chanlekha and Kawtrakul, 2004), discourse processing (Grosz et. al., 1995; Kongwan and Kawtrakul, 2005; Wattanamethanont et.al., 2005) information extraction, knowledge discovery (Bloom, 1956; Prather et. al., 1997; Wah, 1999; Pechsiri and Kawtrakul, 2005) and ontology maintenance (Kawtrakul et. al., 2004).

The experiment is set-up in Plant Knowledge portal such as Rice. The related information about rice dispersed across web resources consisting of varieties, disease, pest, rice exporter, harvest technology, weather forecast, etc.

2. System Overview

This paper focuses on ontology driven information extraction and integration. At extraction level, task-oriented and real-world taxonomy ontology are used to construct information schema and scenario construction. At integration level, demand-driven or pragmatic-oriented ontology is used to aggregate information from multiple heterogeneous sources. The developed system is consisted of three main components as shown in Figure 1.

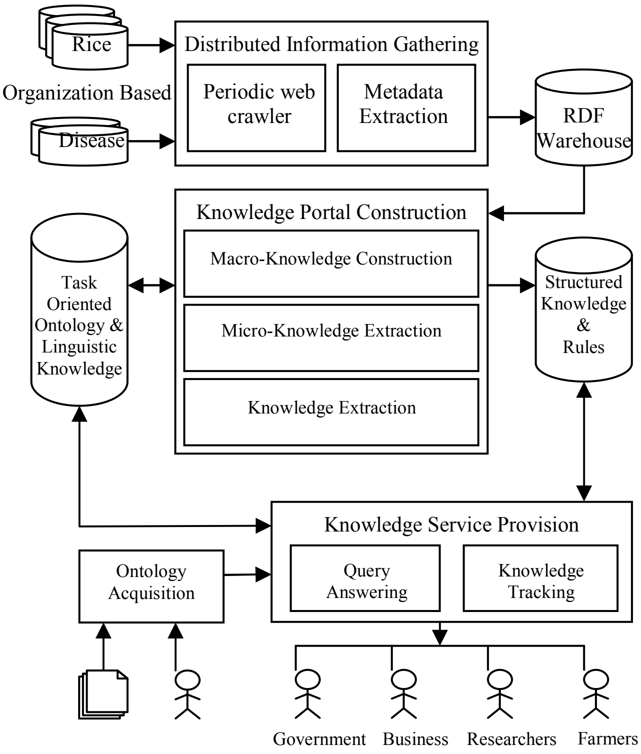

Rice
Organization Based
Disease
Distributed Information Gathering
Periodic web crawler
Metadata Extraction
RDF Warehouse
Knowledge Portal Construction
Macro-Knowledge Construction
Micro-Knowledge Extraction
Knowledge Extraction
Task Oriented Ontology & Linguistic Knowledge
Structured Knowledge & Rules
Knowledge Service Provision
Ontology Acquisition
Query Answering
Knowledge Tracking
Government
Business
Researchers
Farmers

Figure 1 System Overview

- **Distributed Information Gathering.** The information, both unstructured and semi-structured documents are gathered from many sources. Periodic web crawler and HTML parser (Thamvijit et.al, 2005) are used to collect and organize related information. Domain-specific parser (Kawtrakul and Yingsaeree, 2005) are used to extract and generate metadata for interoperability between disparate and distributed information. The output of this stage is represented in RDF format.
- **Knowledge Portal Construction**. Ontologies are used as a key to facilitate both information extraction and integration**.** There are two types of integration: summarize into relational database; such as <Rice varieties and yield**,** Disease dispersion**,** Pest characteristics **>,** and document hyperlink such as Product processing, Cultural practice and Fertilizing. Additionally, the process of knowledge extraction from text is needed to pursue the goal of generating useful knowledge, such as general symptoms of plant diseases, from the large amount of text. The output of this stage is structured knowledge and rules.
- **Knowledge Service Provision.** Four different target users groups, i.e., farmers, researchers, SME and Intelligent Command Centre, are distinguished by different viewpoints that are characterized by each user's interest. At this stage, knowledge tracking and summarization are applied for knowledge service provision. Moreover, K-service in the form of Know-what, Know-why, Know-where, Know-when, Know-how, and Know-who has been also provided.

Three multidisciplinary areas: Language Engineering such as word segmentation (Sudprasert and Kawtrakul, 2003), named entity recognition (Chanlekha and Kawtrakul, 2004), shallow parsing (Satayamas et. al., 2005), shallow anaphora resolution and discourse processing (Kongwan and Kawtrakul, 2005), Knowledge Engineering such as knowledge extraction (Pechsiri and Kawtrakul, 2005), representation and Ontology Engineering such as ontology construction and maintenance (Kawtrakul et. al., 2004; Kawtrakul et. al., 2005), are applied to each component of knowledge portal construction.

3. Ontology driven K-Portal Construction

Knowledge portal construction is not a trivial task since information is distributed in various information sources. For example, the related information about rice dispersed across web resources consisting of varieties, disease, pest, harvest technology, weather forecasting, disaster warning, etc. Hence, knowledge portal construction is a combination model that refers to the creation of new explicit knowledge by extraction, integration, reasoning and synthesizing existing explicit knowledge to help users accessing interrelated knowledge in a well-organized form. This paper presents a method for automatically constructing knowledge portal from electronic documents by using two types of ontology which are pragmatic-oriented ontology and task-oriented ontology.

3.1. Pragmatic-oriented Ontology for Macro-Knowledge Construction

Pragmatic-oriented ontology means systemic and related topic map applied to aggregate expected information from multiple heterogeneous sources. The example of rice ontology in pragmatic point of view for aggregating information for one stop accessing is shown in Figure 2.

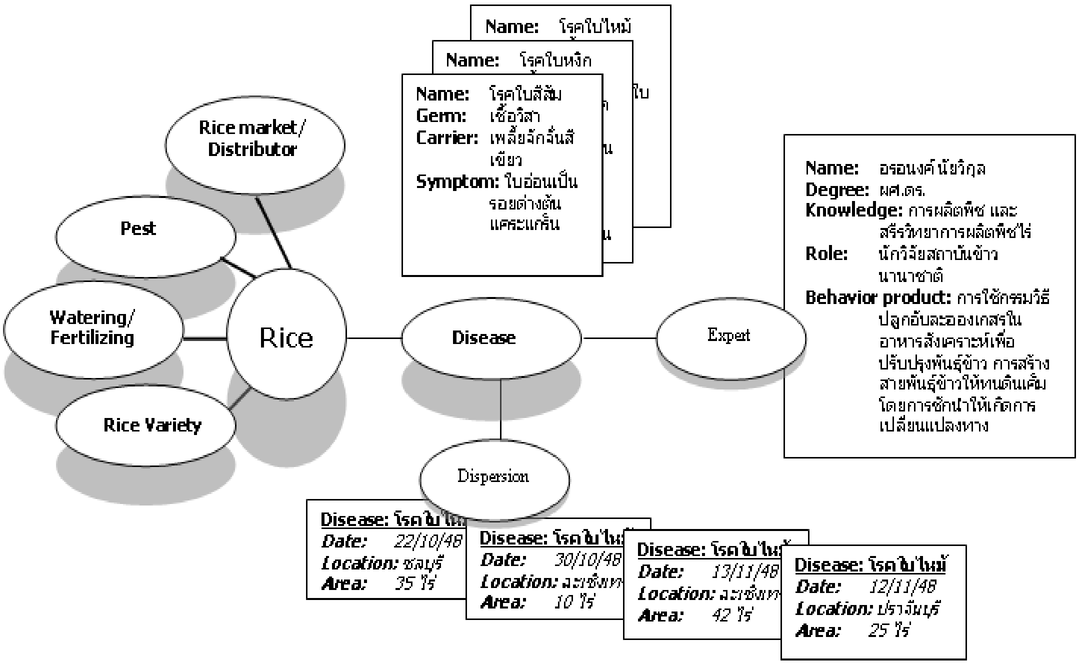


Figure 2 Ontology Driven Rice Portal

3.2. Frame-based Ontology for Micro-Knowledge Extraction

Although most of agricultural knowledge is contained in electronic document, manually reading and digesting these documents is time consuming. To solve this problem, information extraction using task-oriented ontology has been developed to extract relevance information from those documents and reorganize it into a well-defined structure format. In this system, task-oriented ontology is represented in a frame and slots scheme (Figure 3). Manually designed extraction rules are also required to guide the extraction process.

By using this predefined knowledge together with natural language processing techniques, the information extraction process contains two main steps as the following:

- **Pre-processing:** This process performs language processing consisting of word segmentation, name entity recognition, noun phrase analysis, and heuristic-based anaphora and ellipsis resolution (Kongwan, 2005).

<Rice_variety>:=

Variety_name: list value

Rice_type: {ข้าวเจ้า/Rice, ข้าวเหนียว/Sticky Rice}^

Farming_system: {ข้าวนาปรัง/off-season rice field,

ข้าวนาปี/inขseason rice field}^

Harvest_time: numeric value -

Average_yield: numeric value -

Irresistible_pest/disease: symbolic value *

Resistible_pest/disease: symbolic value *

Area_condition: symbolic value*

Note : ^ represents only one value

- represents only one value or the optional value.

* represents zero value or many values

Figure 3 An Example of Rice_Variety Frame and Slot Scheme

- **Information extraction:** The objective of this step is to extract required information based on task-oriented ontology and predefined extraction rules. The example of extracted information is shown in Figure 4.

<Rice_variety>:=

Variety_name: ข้าวญี่ปุ่น กวก.1/Khao Yipun DOA1,

ข้าวญี่ปุ่น กวก.2/Khao Yipun DOA2

Rice_type: ข้าวเจ้า/Rice

Farming_system: ข้าวนาปรัง/off-season rice field

Harvest_time: 90 วัน/day -

Average_yield: 500 กก/ไร่ kg/1600m^2 -

Irresistible_pest/disease: 0

Resistible_pest/disease: Bakanae

Area_condition: high land

Figure 4 An Example of Extraction Result

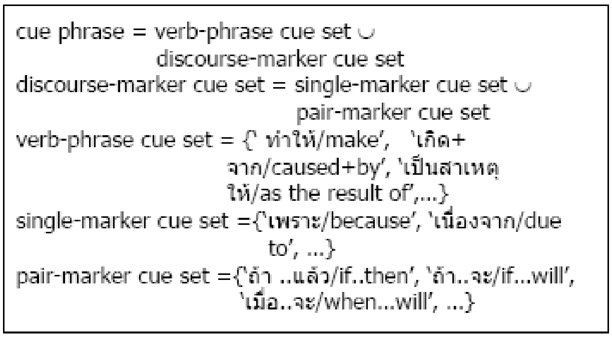
3.3. Knowledge Extraction

In this work, our knowledge extraction system will mainly focus on causality knowledge, existing between causative antecedent and effective consequent discourse units. Mining Know-Why or explanation knowledge will induce a knowledge of reasoning that is beneficial for our daily use in diagnosis.

Framework for Causality Extraction

To extract a causative unit and its effective information unit in the form of an intra-causal EDU or an inter-causal EDU, there are three main tasks as the following (Chaveevarn, et.al., SNLP 2005) :

- **Pre-processing**: This process provides word segmentation, part of speech tagging, including Named entity recognition, word-formation recognition (Pengphon et. al., 2002) to solve the boundary of Noun phrase and EDU segmentation (Chareonsuk et. al., 2005).
- **Causality Learning**: This task is to learn cause-effect expression between causative events and effective events from the corpus. This expression can be expressed by the combination of verb pairs from different EDUs or by a lexical pair from a lexico syntactic pattern within one EDU. In the verb pair expression, one event verb is the causative verb(vc) from a causative EDU. The other event verb is the effective verb or result verb(ve) from an effective EDU. For the lexical pair, the causative and the effective events are determined from the noun phrases presented in the lexico syntactic pattern of one EDU.
- **Causality recognition and extraction**: The objective of this task is to extract the cause-effect relation from an input text. There are two main steps as in the following:
  - Cause-effect Identification: This step is to identify the causative unit by searching any element in Vc from the cause-effect verb pairs (for the inter-causal EDU) and any element in the verb-phrase cue set along with lexical pairs (for the intra-causal EDU) from the causality learning step. For the basic EDU that contains an embedded EDU, paraphrasing is required before searching.
  - Cause-effect Boundary Determination: This boundary determination will be applied only in the inter-causal EDUs. The determination contains 2 main approaches, applying the cue phrase and the centering theory. Cue phrase is used for identifying a causality boundary. There are 2 kinds of the cue phrases, as the verb-phrase cue set and the discourse-marker cue set, as shown below. The centering theory will be applied when the implicit cue phrase is recognized.
  -



cue phrase = verb-phrase cue set ∪
discourse-marker cue set
discourse-marker cue set = single-marker cue set ∪
pair-marker cue set
verb-phrase cue set = {' ทำให้/make', 'เกิด+
จาก/caused+by', 'เป็นสาเหตุ
ให้/as the result of',...}
single-marker cue set ={'เพราะ/because', 'เนื่องจาก/due
to', ...}
pair-marker cue set ={'ถ้า ..แล้ว/if..then', 'ถ้า..จะ/if...will',
'เมื่อ..จะ/when...will', ...}

Figure 5 An Example of implicit cue phrase

This theory relates to the focus of attention within a discourse segmentation (Grosz et. al.,1995) to determine the smooth shift occurrence.

To evaluate the system, we use documents containing 3000 sentences from the Department of Agricultural Extension corpus. Our model of causality extraction shows the precision and recall of 86% and 70% respectively, where our evaluation is based on the expert's results.

4. K-Service Provision

Four different target users group are distinguished from the different points of views depending on their objective, interest, affectation and benefit.

- The farmers require some useful information, i.e. how to analyze the type of pests, or symptom of plant diseases and how to protect plant from diseases,
- The researchers prefer to track the problems and literate the previous researches,
- Small and Medium Enterprise requires to follow up the state of business,
- Intelligent Command Center of the Government needs portal of Executive information, cross sector analysis, and Intelligent Real-time warnings/ alerts or event tracking.

Accordingly, K-service provision consists of Question-Answering and Knowledge Tracking.

4.1 Question Answering

Unlike normal information retrieval systems or search engines, QA can supply users with the satisfy information instead of providing just a list of searching result. Moreover the QA is an effective technique to solve the task of patterns and templates processing that were driven by ontology. In this research QA is applied for the applications below.

4.1.1. Know-what

In Agricultural domain, the Object-Property knowledge, such as pest's characteristics (color, size, and etc), is a useful knowledge that serves the "What" question for a farmer or an agriculturalist, for example

"What is kind of the insect, that its body is black, size is about 2 millimeter long?"

In order to answer the question effectively, the system needs to handle the fuzzy value of the object's properties such as "approximately 2 millimeter", "light green", "dark red". To handle these problems, we use the fuzzy concept for representing the property value in Object-Property knowledge. In order to answer the query, we use the similarity function to retrieve the answer (Kongkwan et.al, 2005). Based on 440 sentences experiment, the precision of the system is 88.88% and the recall is 47.05%.

4.1.2. Know-who

Finding an appropriate person for solving some specific problems is a complicated task. For example, when a farmer wants to find an expert who knows how to control a rice blast, he/she may ask the system with the query "Who is rice blast expert?" The system will process that query and return a list of rice blast experts to him/her containing expert's name, position, affiliation, area of expert, and contacting address. The example of the output from the system may look like this:

Name: Theerayut Toojinda
Position :Research Scientist
Affiliation: Kasetsart University
Email address : Theerayut@dna.kps.ku.ac.th
Expert field : RICE BREEDING/ MARKER AIDED SELECTION

Name : Ladha, J.K.
Position :senior researcher
Work place : IRRI India Office,NASC Complex,DPS Marg, Pusa,New Delhi 110012,India
Email address : j.k.ladha@cgiar.org
Expert field : Soil fertility, plant nutrition, biological N2 fixation, rice-wheat system.

Name : Barman, Bhubaneswar
Postion : Chief Scientist
Affiliation : Assam Agricultural University ,RARS, Shillongani, Nagaon, Assam, INDIA
Email address : brrihq@bdonline.com
Expert field :cereals and grains

Figure 6 An Example of Know-who result

4.1.3. Know-why

This knowledge-question is very important for the diagnosis and preventive system because knowing the cause is relevant for problem solving. For example: the Blast disease is an awful disease for rice farming. Then, farmers will have some questions as

> "*How do we know whether it is the Blast disease? and What is the causation*?"

Our system will answer as

> "*When the disease presents, rice will show symptoms on young leaves of tillering plants initially appearing as small whitish-gray or dark, reddish-brown spots. In severe cases, entire plants may be defoliated, become stunted and die. And Rice blast is caused by the blast fungus.*"

However, our system will be beneficial to the people who look for the reasoning knowledge, with 86% precision and 70% recall.

4.2 Event Tracking

In order to fulfill knowledge service provision, we also provide tracking system called Event Tracking. Figure 5 shows the example of "dispersal" and "prevention" Frame of disease that modified from (http://framenet.icsi.berkeley.edu/).

To track the events, we need to know n-tuple information including time dimension such as the dispersal of rice blast, we need to know 3-tuples information consisting of place, time and result.

$$E = \{< p_1, t_1, r_1 >, < p_2, t_2, r_2 >, \ldots < p_n, t_n, r_n >\}$$

Where:

$p_1, \ldots, p_n$ = place 1 to n
$t_1, \ldots, t_n$ = time 1 to n
$r_1, \ldots, r_n$ = result 1 to n

5. Conclusion and Future work

This work is the research on-demand that we got the requirement from the end-users. To construct knowledge portal, two type of engineering is required as basic elements, which are language engineering, and ontology engineering. For the current prototyping, ontology is modified from Frame net and predefined manually. For the next step, ontology will be semi-automatically constructed by using a variety of terminological resources, such as raw text, dictionaries and thesauri.

Acknowledgement

The work described in this paper has been supported by the grant of NECTEC No. NT-B-22-14-12-46-06, within the project "A Development of Information and Knowledge Extraction from Unstructured Thai Document".

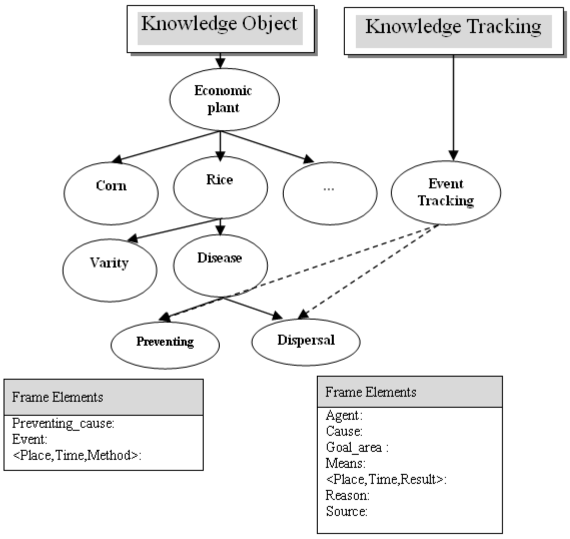


Figure 7 An example of event tracking frame

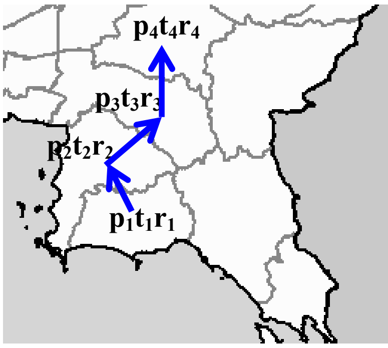


Figure 8 shows the spatial visualization event tracking of rice blast dispersal.

6. References

Kongwan, A., and Kawtrakul, A. (2005). Know-What: A Development of Object-Property Extraction from Thai Texts and Query System. In *the proceeding of The Sixth Symposium on Natural Language Processing,* Chiang Rai, Thailand.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook1 Cognitive Domain*, New York: Toronto Longmans, Green.

Chanlekha, H. and Kawtrakul, A (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. *In the proceeding of IJCNLP*, Hainan, China.

Chareonsuk, J., Sukvakree, T. & Kawtrakul, A. (2005). Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. In *the proceeding of NCSEC 2005*,Thailand.

Pechsiri, C., Kawtrakul, A., and Piriyakul, R.,(2005) Mining Causality Knowledge From Thai Textual Data. In *the proceeding of the Sixth Symposium on Natural Language Processing*, Chiang Rai, Thailand.

Grosz, B. J., Joshi, A.K , and Weinstein, S. (1995). Centering: A Framework for Modelling the Local Coherence of Discourse. *In Computational Linguistic 21(2).* pp. 203-225.

http://ai.bpa.arizona.edu/research/coplink/tkp.htm

http://framenet.icsi.berkeley.edu/

http://www.fao.org/agrovoc/

Imsombut, A., Suktarachan, M., Yingsaree, W., and Kawtrakul, A. (2004). Country Report and Activity from Thailand: Ontology Construction and Maintenance System in Agricultural Domain. In *the proceeding of AFITA/WCCA*, pp. 62-74.

Kawtrakul, A. , and Yingsaeree, C.(2005). A Unified Framework for Automatic Metadata Extraction from Electronic Document. *The International Advanced Digital Library Conference* in Nagoya.

Kawtrakul, A., et al. (2005). Automatic Term Relationship Cleaning and Refinement for AGROVOC. *In EFITA/WCCA2005*, The Sixth Agricultural Ontology Service Workshop, Vila Real, Portugal.

Kawtrakul, A., Suktarachan, M. and Imsombut, A. (2004). Automatic Thai Ontology Construction and Maintenance System. *Proc. of OntoLex Workshop on LREC,* Lisbon, Portugal.

Wattanamethanont, M., Sukvaree, T. and Kawtrakul, A. (2005). Thai Discourse Relations Recognition By Using naive Bayes Classifier", In *the proceeding of the Sixth Symposium on Natural Language Processing*, Chiang Rai, Thailand.

Pengphon, N., Kawtrakul, A. and Suktarachan, M. (2002). Word Formation Approach to Noun Phrase Analysis for Thai. In *the proceeding of SNLP2002*, Thailand.

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J.W., Hage, M. L., and Hammond, W. E.(1997). Medical Data Mining : Knowledge Discovery in a Clinical Data Warehouse, Duke University Medical Center, Durham, North Carolina, Division of Medical Informatics.

Satayamas, V., Thumkanon, C. and Kawtrakul, A.(2005). Bootstrap Cleaning and Quality Control for Thai Tree Bank Construction, In *the proceeding of the 9th NCSEC*, University of the Thai Chamber of Commerce, Bangkok, Thailand.

Sudprasert, S., Kawtrakul, A. (2003). Thai Word Segmentation based on Global and Local Unsupervised Learning, In *the proceeding of the 9th NCSEC'2003*, Chonburi, Thailand.

Thamvijit, D., Chanlekha, H., Sirigayon, C., Permpool, T., and Kawtrakul, A.(2005). Know-who: Person Information from Web Mining, In *the proceeding of the 9th NCSEC*, University of the Thai Chamber of Commerce, Bangkok, Thailand.

Turban, E., Aronson, J. E., and Liang, T. P. (2005). Decision Support Systems and Intelligent Systems, Pearson Prentice Hall.

Wah, B. W. (1999). Generalization and Generalizability Measures, IEEE Transactions on Knowledge and Data Engeineering, Vol. 11, No. 1.