Knowledge Portal Construction for Powerful Knowledge Services

Researchers

A. Kawtrakul and Team
Department of Computer Engineering, Kasetsart University,Thailand
ak@vivaldi.cpe.ku.ac.th, aim@vialdi.cpe.ku.ac.th

Abstract

Knowledge is one of the most crucial success factors for both country development and business intelligence. However, sources of valuable knowledge are scattered at several locations and websites with heterogeneous formats. Moreover, the needed information was too difficult to find since there was no semantic relation and organization and, even if it was found, often overload since there was no content digestion.

This paper focuses on ontology-driven information and knowledge extraction with natural language processing techniques and a framework of user-centered design for accessing the information they need when they need it.

Keywords

Knowledge Portal, Knowledge Services Provision, Language Engineering, Knowledge Engineering, Ontology Engineering

Introduction

In order to understand a situation, an operator and decision-maker needs data, information and knowledge. The knowledge consists of data items and/or information organized and processed to convey understanding, experience, accumulated learning, and expertise that are applicable to an intended problem solving [Turban, 05]. 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 was too difficult to find since the traditional search engines return ranked retrieval lists that offer little or no information on the semantic relationships among those scattered information, and even if it was found, often overload since there was no content digestion. Accordingly, the users or knowledge workers must spend a substantial amount of their time browsing and reading to find out how information are related to one another 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-centered design for accessing the information they need when they need it. Since the web consists of a large extent of unstructured or semi-structure natural language text, several techniques both in language engineering, knowledge engineering and ontology engineering are applied to the knowledge portal construction. These include named-entity recognition [Chanlekha, 04], discourse processing [Grosz, 95], information extraction, knowledge discovery [Bloom, 56; Wah, 99; Prather, 97], and ontology maintenance [Kawtrakul, 04].

Knowledge Portal Construction

Knowledge portal, here, means an interrelated knowledge of specific domain. Knowledge portal construction, then, is a combination mode that refers to the creation of new explicit knowledge by extraction, integration and synthesizing explicit existing knowledge to offer their users high quality access to and interaction possibilities within the contents of the portal such as one-stop service of interrelated knowledge in well-organized form, disease diagnosis by using inference engine.

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 market/distributor/ exporter, harvest technology, weather forecast, disaster warning and so on (see figure 1).

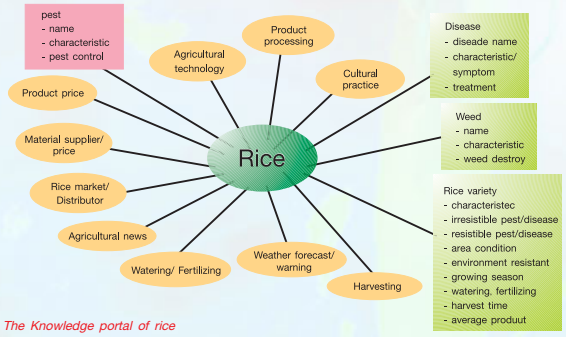

Figure 1 The Knowledge portal of rice

Three multidisciplinary areas: Language Engineering such as named entity recognition, shallow parsing, anaphora resolution and discourse processing, Knowledge Engineering such as information extraction, knowledge extraction and generalization, and Ontology Engineering such as ontology maintenance, are applied to each component of the system (see figure2).

- **Distributed information collecting by using web-crawler**. The information, which is unstructured or semi-structured documents, is gathered from many sources. Named entity recognition and domain specific parser are used to annotate the meta-data for interoperability between disparate and distributed information systems.
- **Knowledge portal processor**.
  - Ontologies are used as a key to facilitate both information extraction and integration. There are two types of integration: summary into relational database; such as <Rice breed, Disease, Pest, Yield, Harvest Period, Fertilization>, and related document hyperlink such as Product processing, Cultural practice and Fertilizing.
  - The process of knowledge extraction from text pursues the goal of extracting useful knowledge such as genralization rules from large amounts of text. It comprises of a natural language processing for extracting causal knowledge such as symptoms of plant diseases, and application of factor analysis, and C4.5 for knowledge generalization.

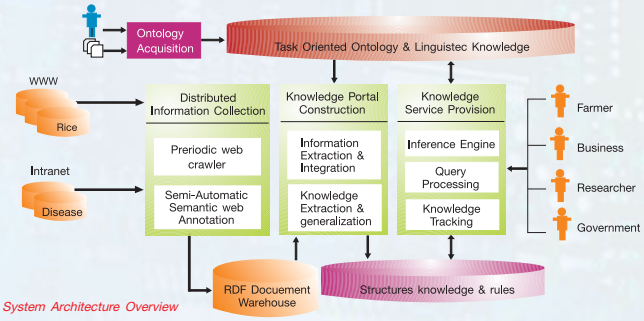

Figure 2 System Architecture Overview

Knowledge Service Provision

Four different target users group are distinguished from the view point of differences in interest, motivation, the nature of information they pursue, and background knowledge they bring (see figure 2).

- The farmers who require some useful information such as how to analyze symptom of plant diseases and how to protect plant from diseases,
- The researchers who would like to track the problem and literate the previous researches,
  Small and Medium Enterprise, who is generally more interested in business, would like to follow up, such as, the material resource planning, etc.,
- Intelligent Command Center of the Government who needs portal of Executive information, cross sector and multidiscipline analysis, and Intelligent Real-time warnings and alerts.
  Many engines have also been developed for supporting services, i.e., inference engine, query expansion and knowledge tracking.

Examples

The system use knowledge that is extracted from information extraction, together with predefined rules, to assist the farmer by suggesting the appropriate rice seed.

To use the system, the users must provide the information about the type of rice, region, characteristics and irrigation of their land. After that, the expert system will analyze the knowledge that is extracted from information extraction system and will output the rice seed that suit for their land, as well as the summary detail of rice seed, and link to the full document.

Acknowledgement

This project has been funded by NECTEC.

References

- [Bloom, 56] B. S. Bloom (Eds.): *Taxonomy of educational objectives: The classification of educational goals: Handbook* 1. cognitive domain, New York: Toronto Longmans, Green, 1956.
- [Chanlekha, 04] H. Chanlekha, and A. Kawtrakul: *Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information.* Proc. IJCNLP, Hainan, China, 2004.
- [Gomez, 94] F. Gomez, R.Hull, C. Segami: *Acquiring Knoledge From Encyclopedia Texts, Department of Computer science.* University of Central Florida, Proc. the 4th ACL conference on Applied Natural Language Processing, Struttgart, Germany, October 13-15, 1994.
- [Grosz, 95] B. Grosz, A. Joshi, and S. Weinstein: Centering: *A Framework for Modeling the Local Coherence of Discourse,* In Computational Linguistics 2(21), (1995), 203-225.
- [Kawtrakul, 04] A. Kawtrakul, M. Suktarachan and A. Imsombut,: *Automatic Thai Ontology Construction and Maintenance system,* Proc. Of Ontolex workshop on LREC (2004).
- [Prather, 97] J. C. Prather, D. F. Lobach, L. K. Goodwin, J.W. Hales, M. L. Hage, W. E. Hammond: *Medical Data Mining: Knowledge Discovery in a Clinical* Data Warehouse, Duke University Medical Center, Durham, North Carolina, Division of Medical Informatics ,1997.
- [Sakurai, 04] S. Sakurai, A. Suyama: *Rule Discovery from Textual Database on Key Phrase Patterns,* ACM Symposium on Applied Computing, 2004.
- [Wah, 99] B. W.Wah: *Generalization and Generalizability Measures,* IEEE Transactions on Knowledge and Data Engeineering, Vol. 11, No. 1, 1999.
- [Efraim Turban, 05] E. Turban, J. E. Aronson, T. P. Liang, *Decision Support Systems and Intelligent Systems,* Pearson Prentice Hall, 2005.

โครงการนี้ได้รับทุนอุดหนุนการวิจัยและพัฒนา จาก
ฟ้ายเครือข่ายการวิจัยและพัฒนา (RDD)
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
112 อุทยานวิทยาศาสตร์ประเทศไทย ถนนพหลโยธิน
ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120
โทรศัพท์ 02-564-6900 ต่อ 2501-10 โทรสาร 02-564-6901..2

http://www.nectec.or.th/