Generation of Knowledge Base for Conversation Agent of Nusantara Culinary Tour From Free Text Using Information Retrieval

Dwi Cahyono
Faculty of Engineering
Dr. Soetomo University,
Surabaya , Indonesia
dwik@unitomo.ac.id

Veronika Nugraheni Sri Lestari
Economic Department
Dr. Soetomo University
Surabaya , Indonesia
venugra@unitomo.ac.id

Abstract

Knowledge base is generally given through structured data or documents that can be given through database, XML documents. In a research developed by the author associated with nusantara culinary tour agency, knowledge base with culinary tourism domain is necessary. Culinary tour information today is high and easy to find either through printed or online media and becomes a strength for this research to be developed, to enrich knowledge base of culinary tourism conversation agent media needed to capture such abundant information by using information retrieval in accordance with the keyword "isA, location, partOF, hasA", the result of this information retrieval will then become knowledge base of the culinary tour agent presented in Relational Data Base Management System (RDBMS) PostgreSql. This research will contribute as an alternative media to enrich the knowledge base of culinary tour agent which is generated from free text using keyword for its search by using information retrieval (Information Retrieval).

Keywords
Information Retrieval, Agent, Knowledge Base, Free Text, Culinary Tour.

1 INTRODUCTION

Culinary tourism today is very popular. Changes of lifestyle in society also contribute AS People eat not only to satisfy their stomach, but also look for atmosphere and service as part of the dish of food ordered. Many regional foods serve different flavours on the tongue and many also understand the details of the typical regional food.

An agent is a person or something that has the ability to perform a particular task/job in accordance with his capacity for something or someone else (Romi Satria Wahono, 2003). Conversation agent is expected to be an alternative presentation of information needed by the community associated with culinary tourism.

Most conventional interfaces are based on direct manipulation where the user is solely responsible for monitoring and executing all tasks. However, indirect management interfaces are now being developed which allow the user to delegate some tasks to an interface agent(Beskow & Mcglashan, 1997)

In the previous research it has been concluded that Indonesian Free Texts can be used as basis to add knowledge base of agent with Information Retrieval (IR) module which generate it automatically (Cahyono et al., 2008).

Generally agent knowledge base can be given by entering directly from the database and use the knowledge of the agent or any other process that is from structured sources e.g XML and relational data.

Knowledge Base (KB) approach uses mechanism of cause and language of query ontology to retrieve source of information. Document is one of the sources in retrieving information(Ceri et al., 2013).

The most widespread applications of IR are the ones dealing with textual data. As textual IR deals with document sources and questions, both expressed in natural language, a number of textual operations take place “on top” of the classic retrieval steps.

Focus of this research is to generate an agent's knowledge base from an abundant source of information on culinary tourism in the form of free text into the knowledge base of postgreSql RDBMS-shaped agents.

Most information retrieval systems point to documents or parts of documents, giving physical access, or at maximum bibliographic access via representations(Logan, 1993).

On this research Free text provided in the process uses information retrieval module (IR) based on keyword adalah (is A), lokasi (location), bagian (part Of), memiliki (has A) related to ontology domain of culinary tour.

2 RESEARCH

In generating KB, this research agency adopted several steps that have been done in previous research by Dwi Cahyono, 2008 using 3 steps as in figure 1.

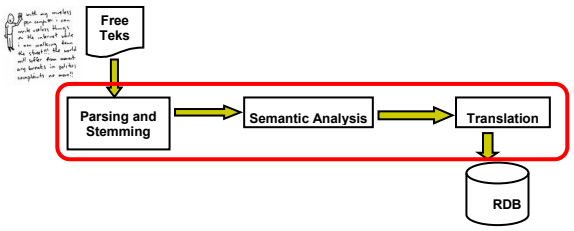

Figure 1: Block system diagram of generation of agent knowledge base (Cahyono et al., 2008).

Indonesian- based free texts given are broken down to the basic word in the process of parsing and stemming in accordance with culinary tourism domain in the culinary tourism system.

The result of parsing and stemming processes in the form of words and structure from the sentence given in free text is then processed by semantic analysis(Cahyono et al., 2008).

A semantic analysis relates the syntactic structure of phrase, sentence to paragraph level(Cahyono et al., 2008). Therefore, words, phrases, to sentences related to ontology domain of culinary tourism agent. This analysis is conducted to obtain a correct perception of the information to be presented by the agent.

Translation is a process that produces relevant information to be included in RDB (figure 1) of semantic expression generated by semantic analysis process involving semantic network of ontology domain(Cahyono et al., 2008).

2.1 Information Retrieval

Information retrieval in this study is used to collect information relevant to culinary tour from free texts of Indonesian language provided by utilizing keywords for tracking relevant information. The keywords used are presented in table 1 below:

Table 1: Lists of keywords from IR.

Keywords	Description
is (isA)	Explain something or thing
part (partOf)	Describe part of something
has (hasA)	Shows the property relationship of something
location (location)	Shows the location of something.
Composition	Point out the composition of something

The example is as follows:

“Rujak cingur adalah makanan khas berasal dari jawa timur”

It be explained from the sentence that the word "is" is a keyword that explains the word that comes before it, which is "rujak cingur" , while the word “berasal dari” indicates location.

2.2 Generation of Agent Knowledge Base

The knowledge about each data concept-not the values for a particular patient, but general domain
Knowledge such as default values for attributes, strategies for inferring them if they are not available,
etc.-is stored in a frame, and these concepts are typically organized in a frame hierarchy(Chandrasekaran, 1986).

In the process of generating responses from users, a question mark that indicates the relationship of the question asked by the user to the facts that exist in the agent's knowledge base is involved, as shown in table 2.

Table 2: List of question mark with its relation.

Question Mark	Relation	Description
What	Something/thing/ food	Asking something or form of food
Where	Location	Asking the origin of something
how many	Amount	Asking the number of things

Knowledge base of this culinary tour conversation agent is generated by following the stages in Figure 1, where each process is related to each other until the relevant information is obtained related to the culinary tour. Each stage will be explained clearly in the following sub-chapters

2.2.1 Relational Data Base Management System (RDBMS)

A database management system, or DBMS, is software designed to assist in maintaining and utilizing large collections of data, and the need for such systems, as well as their use, is growing rapidly(Raghu Ramakrishnan, 2000).
POSTGRESQL is the most advanced open source database server and open source software. The term “open source software” often confuses people. With
commercial software, a company hires programmers, develops a product, and sells it to users. With Internet communication, however, new possibilities exist(Momjian & Bruce, 2001)
In this study Relational Database System (RDBMS) using postgreSql version 8.2 is used, which is the development of author’s previous research, the database name is "RDB_Ir" with tables shows on table 3 and table 4.

Table 3: List of tables in RDB_Ir.

Table name	Function of ythe table
Things	Accommodate knowledge base of conversation agent, resulting from the generation of knowledge base
Question Mark	Accommodate Questionnaire to be used by the conversation agent
Synonyms	accommodates synonymous words
Stoplist	Contains a list of stopword
Relationship	Accommodates words with their relationships

Knowledge base of agents in this study will be stored in the table of things that possesses data dictionary shown in table 5.
Definition create thing table is follows

```
CREATE TABLE things
(
property character varying (50),
thing character varying (50),
character varying properties (50),
image character varying (20)
)
WITH (
OIDS = TRUE
);
```

Table 4: Lists the postgreSql function.

Function Name	Description
Sentence parsing	Breaks the given text into several sentences
Parsing 2	Break the sentence that has been given by the function of sentence parsing into words
urai_kalimat	Parsing the input of sentencesinto word groups in this case is called the index group each of which will become knowledge base (KB)
Stemming	Function to find the base word of a given word
Urai_kalimat_v4	The output of parsing2 is processed further into *function* urai_kalimat_v4 to find relevant information in accordance with the given keyword according to table 2 (Cahyono & Prihartono, 2013)

Table 5: dictionary data base knowledge of things (Cahyono et al., 2008).

Attribute	**Data type**
Property	character varying (50)
Thing	character varying (50)
Character	character varying (50)
picture	character varying (20)

2.2.2 Parsing and Stemming

There are two parts of parsing: *keyword parsing* and *grammar based parsing*, *keyword based parsing* is used in this study. This is simple and effective parsing model in parsing input text, *keyword based parsing* does not involve knowledge syntax .

Stemming is the process of finding the stem (root) of a word, by stripping away the affix attached to the word. In many languages words are often obtained by affixing existing words or roots (Indradjaja & Bressan, 2003). Stemming process is to search the word base of the input word given by removing the attachment attached to the input word.

Sample text input in sentence form 1 is

“Lontong Kupang is a typical food of the city of Sidoarjo”

The word "is a" in sentence 1 above describes the relationship of definition, table 1 shows the relation is a (isA).

At this stage a function created in the postgreSql function with the parsing2 function name as in table 4.

Parsing process is performed by executing parshing2 function with input parameter in the form of output sentence from this function is a list of words that have passed stemming process.

Below is syntax query for execution of parshing2 function.

"Select * from parshing2 (' Lontong Kupang is a typical food of the city of Sidoarjo ') as (varchar word)"

So the output of the above function execution is shown in table 6 below:

Table 6: Result of parshing2 function execution.

Index	**Words**
0	Lontong
1	Kupang
2	Is a
3	food
4	typical
5	city
6	Sidoarjo

In table 6 index 3 the word "food" is the result of stemming from the input word "food".

2.2.3 Semantic Analysis

In a sentence that contains the corresponding keywords also include words that precede and follow the keyword or showing the relation in KB, semantic analysis plays role to find out words that precede and relate with the keywords shown on table 1 searched in free text based on the culinary tour domain to avoid mistakes in interpretation of the given word.

The output of the "semantic analysis" process is "semantic expression" of sentences relevant to culinary tourism domain of the archipelago.

The semantic analysis of this research utilizes the RDB urai_kalimat_v4 function of the previous research by executing it to generate relevant words or sentences from the given free text.

For example if given a sentence from the free text “Clover is one of the typical foods of the city of Surabaya” then execution of its function is as follows:

“select * from urai_kalimat_v4(' Clover is one of the typical foods of the city of Surabaya’) as (indeks_kal int2, indeks_kal1 int2, indeks_kal2 int2, group_kata int2, kata text)”

The result of the execution of the function can be shown in table 7.

Table 7: Results of execution function of urai_kalimat_v4.

Index_kal	Index_kal_1	Index_kal2	Group_kata	Words
1	1	1	1	Semanggi
1	1	1	2	Is a
1	1	1	3	one of the typical foods of the city of Surabaya

2.2.4 Translation

Translations are used to retrieve information relevant to be incorporated into the knowledge base of the agent on the basis of the semantic expression of the semantic analysis process (Cahyono & Prihartono, 2013).
From the phrase “one of the typical foods of the city of Surabaya”, the result of semantic analysis of semantic expression is parsed as in table 6, relasi = "is a", object = "semanggi " description = " one of the typical foods of the city of Surabaya "

3 RESULT AND DISCUSSION

To raise knowledge base of culinary tour agent, RDB function of urai_kalimat_v4 becomes very important to parse free texts into relevant information which will be incorporated into the knowledge base of culinary tour agent.
Below is a trial to parse the free text that contains from several sentences using the function urai_kalimat_v4.

Example of sentences:
S=”here are some foods that are typical of the region, clover is a typical food from the city of Surabaya, lontong kupang is a typical food of Sidoarjo, cingur salad is a typical food of the city of Surabaya, bika ambon comes from ambon, gudeg is a typical food of Yogyakarta”

If sentence S is executed by function urai_kalimat_v4.
“Select * from urai_kalimat_v4(”here are some foods that are typical of the region, clover is a typical food from the city of Surabaya, lontong kupang is a typical food of Sidoarjo, cingur salad is a typical food of the city of Surabaya, bika ambon comes from ambon, gudeg is a typical food of Yogyakarta”) as as (indeks_kal int2, indeks_kal1 int2, indeks_kal2 int2, group_kata int2, kata text)”

Then it will generate some semantic expression shown in table 8 below:

Table 8: Semantic expression of test results.

Indeks_kal	Indeks_kal1	Indeks_kal2s	Group_kata	kata
1	1	1	1	here are some foods that are typical of the region
2	2	2	1	Semanggi
2	2	2	2	Is a
2	2	2	3	typical food from the city of Surabaya
3	3	3	1	lontong kupang
3	3	3	2	Adalah
3	3	3	3	typical food of Sidoarjo
4	4	4	1	cingur salad
4	4	4	2	Is a
4	4	4	3	typical food of the city of Surabaya
5	5	5	1	bika ambon
5	5	5	2	from
5	5	5	3	ambon
6	6	6	1	gudeg
6	6	6	2	Is a
6	6	6	3	typical food of Yogyakarta

The word group shows the KB group that will be the knowledge base of the culinary tourism agent of the archipelago stored in the tables of things in the RDBMS (table 5)
Figure 2 show a test using free text file extract application:

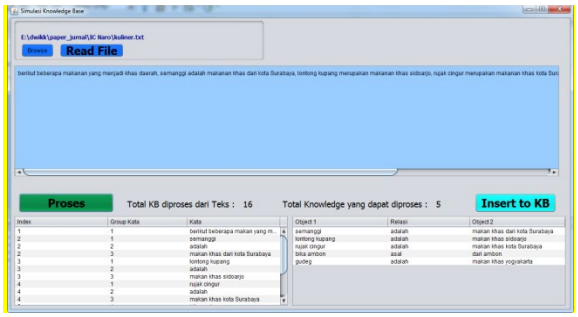

Figure 2: Screenshot of knowledge base generation application using free text.

Figure 2 shows that from the free text input with the culinary text name.txt which contains the text " here are some foods that are typical of the region, clover is a typical food from the city of Surabaya, lontong kupang is a typical food of Sidoarjo, cingur salad is a typical food of the city of Surabaya, bika ambon comes from ambon, gudeg is a typical food of Yogyakarta" we have 6 groups (index) sentences as shown in table 8 , which will be candidates for entry into agent knowledge base, from the 6 groups of sentences, maybe only 5 possible indexes to be included on agent knowledge base. This is because index 1 does not have a relation on the group of words raised, thus index 1 is not candidate to enter agent knowledge base.

From Figure 2 we obtain 5 candidates of knowledge base as shown in table 9 below:

Table 9: Lists of generated knowledge base.

Object 1	relasi	Object 2
Semanggi	Is a	typical food from the city of Surabaya
lontong kupang	Is a	typical food of sidoarjo
rujak cingur	Is a	typical food of Surabaya City
bika ambon	from	Ambon
Gudeg	Is a	typical food of Yogyakarta

4 CONCLUSIONS

processes in generating knowledge base of culinary agent of archipelago uses information retrieval (IR). There are three process in ir including parsing, semantic and translation analysis. In parsing process there are two query function of parshing_kalimat and parsing2, in process of semantic analysis using one query function ie urai_kalimat_v4, while the translational process translates the results from the semantic analysis into agent knowledge base.

Information Retrieval (IR) uses *keyword base parsing* where the determination of keywords for parsing is crucial in the process of retrieving relevant information from the free text given in accordance with the domain of nusantara culinary tourism.

From trials with free texts there are 6 indexes decomposed, from 6 indexes there are 5 indexes that become candidates entered into agent knowledge base, means 83.3% of free text is selected into knowledge base agent.

Free text with bigger sentences in it need to be tested to see the accuracy of the results of this research. Keyword selection in relations becomes important in generating agents knowledge base.

ACKNOWLEDGEMENTS

We would like to thank the Dr. Soetomo University for financial support and friends who always support the completion of this study from beginning to end.

BIOGRAPHIES

DWI CAHYONO

Lecturer of faculty of engineering , dr. soetomo university, surabaya , Indonesia, he graduated from Dr Soetomo University in 1998 and graduated postgraduate at Sepuluh Nopember Institute in 2009, currently also works as a systems analyst at a company in Surabaya, he has 2 registered patents and many research papers. his researc interests include natural language processing, database system, Human Computer interaction and Geographical Information systems. Email at dwik@unitomo.ac.id

VERONIKA NUGRAHENI SRI LESTARI

Lecturer of economic department , dr. soetomo university, surabaya , Indonesia, and still undergoing doctoral education in stiesia, she graduated from Dr Soetomo University and graduated postgraduate at Dr Soetomo University, she has 3 registered paten and more and has a lot of research papers. Her researc interests include transportation economic, urban economiy and industrial economic. Email at venugra@unitomo.ac.id

REFERENCES

Beskow, J., & Mcglashan, S. 1997. Olga - a conversational agent with gestures. *Gesture*.

Cahyono, D., Fadlil, J., Sumpeno, S., Hariadi, M., Sarjana, P., & Elektro, T. 2008. Temu Kembali Informasi Untuk Pembangkitan Basis Pengetahuan Dari Teks Bebas Yang Digunakan Oleh Agen Percakapan Bahasa Alami.

Cahyono, D., & Prihartono, E. 2013. Pembangkitan Basis Pengetahuan Agen Pada Sistem Pembelajaran Cerdas Dari Teks Bebas Dengan Menggunakan Temu Kembali Informasi. *PALIMPSEST (Jurnal Ilmu Informasi Dan Perpustakaan)*, *5*(1), 20–26.

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013).

Chandrasekaran, B. 1986. Generic Tasks in Knowledge-Based Reasoning: High-Level Building Blocks for Expert System Design. *IEEE Expert-Intelligent Systems and Their Applications*, *1*(3), 23–30.

Indradjaja, S. L., & Bressan, S. 2003. Automatic Learning of Stemming Rules for the Indonesian Language. *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, 62–68.

Logan, E. 1993. Information retrieval interaction. *Information Processing & Management*, *29*(6), 794.

Momjian, B., & Bruce. 2001. PostgreSQL : introduction and concepts, 461.

Raghu Ramakrishnan, J. G. 2000. *Data Base Management Systems (2nd Ed.)* (2nd ed.).

Romi Satria Wahono. 2003. *Pengantar Software Agent : Teori dan Aplikasi*, 1–19.