A Knowledge-Based Approach for Searching Semi-Structured Documents

Xiaoying (Sharon) Gao
Intelligent Agent Laboratory
Department of Computer Science
The University of Melbourne
Parkville 3052, Australia

{xga, leon}@cs.mu.oz.au

Abstract

The amount of information on the Internet is increasing dramatically, making efficient Information Retrieval (IR) very difficult. Most of the search engines are using traditional IR methods based on keyword matching and are already showing a number of limitations, particularly they suffer from low precision. One suggestion for improving their performance is to use Natural Language Processing (NLP) technology to support content-based IR. However, current NLP technology is not mature enough to enable free text parsing and understanding. This research focuses on searching semi-structured documents, such as Web pages generated from databases and online services. Shallow NLP and a knowledge-based approach is used to parse both users' query and semi-structured documents to a set of knowledge units (concepts) and the knowledge units (concepts) are used as the basis for matching instead of keywords. An online Classified Advertisement Search Agent (CASA) has been built as an example. It can automatically search through online real estate advertisements to help users find accommodation. This paper focuses on the knowledge unit representation and text parsing algorithm which are used for parsing semi-structured documents and users' query.

1. Introduction

Efficient information retrieval (IR) from the Internet has become an important research issue recently. Traditional keyword search based IR methods, which treat text as bags of unordered words, have two major problems. One is that one word may have many meanings and this reduces the precision rate. The other is that many different words may have the same meaning and this results in lower recall rate. Ideally, IR should be based on concept (meaning) matching instead of keyword matching. However, current Natural Language Processing (NLP) technology is not mature enough to support natural language understanding.

This research uses Information Extraction (IE) technology, which is more limited than NLP and "full text understanding", to extract knowledge units (concepts) from the text and to use the knowledge units for further information retrieval. A Knowledge unit is defined as word groups or phrases with an independent and specific meaning. For example, *"$200 perweek"* is a knowledge unit *price* with the value *200*. In the search process, both the query and the documents are parsed to obtain a set of knowledge units. Then the knowledge units are used as the basis for matching instead of keywords.

As an application, an online Classified Advertisement Search Agent (CASA) has been built to read online real estate advertisements to help users find accommodation. In this domain, the main basic knowledge units are *suburb, price, size, type, furniture, transport, facility, bond, available time, common words* and *abbreviations*. Basic knowledge units can be clustered to form larger knowledge units such *as real estate ad, paragraph* and *document*.

The main challenge is to build a text parser to extract knowledge units from text. Related previous work dates back to Schank's primitive-act frame system in the 1970s (Winston 1984). Also, this work is related to research on Information Extraction (SCIE-97 1997) and Message Understanding (MUC 1993). However, this research differs from others in that it uses a knowledge-based approach and introduces the knowledge unit representation using three kinds of constraints and the parsing algorithm consisting of three steps. It not only concerns the text but also the HTML tags and other structures. The method is particularly good at parsing semi-structured documents in HTML such as web pages generated by Web services.

The paper is organized as follows. Section 2 presents the knowledge unit representation. Section 3 outlines the text parsing algorithm. The parsing performance is evaluated in Section 4. Conclusions and future work is summarized in Section 5.

2. Knowledge Unit Representation

This section aims to present a way to represent every knowledge unit, that is, to find a way to match a group of words (including their synonyms and abbreviations) to a certain concept. If knowledge units could be represented as a set of rules and facts, then the rules and facts could be used in the parsing algorithm to extract knowledge units from online advertisements.

Towards this aim, text is viewed in three perspectives: structure, length and content, each providing a kind of constraint on the text that gives a

Xiaoying(Sharon) Gao (1998) **A Knowledge-Based Approach for Searching Semi-Structured Documents.** In D. Estival (ed.) *Abstracts for the ANLP Post Graduate Workshop,* pp 21-24.

clue to its meaning (i.e. its matching concept). Structure is the boundary constraint for a group of words, which is usually expressed with HTML tags, character template, punctuation and special characters. Length is the number of characters in all the words. Content is the words, or their synonyms and abbreviations, which represent word meaning, word order, and relations between the words. An example of the structure, length and content of the text *"<B> Hello World </B>"* is given in Figure 1.

Text
"<B> Hello World </B>"
Structure
Begin with HTML tag *"<B>"*
End with HTML tag *"</B>"*
Character template
"clll cllll"
(*'c'* represents capital letter;
'l' represents lower case letter;
'n' represents number)
Length
11
Content
"hello world"

Figure 1: An example of the structure, length and content of text

For most knowledge units, one kind of constraint is sufficient. For example, knowledge units such as *price, size, type* can be defined by their content constraints. Other knowledge units such as *paragraph* can be defined by their structure constraints (e.g. *paragraph* begins with tag *<p>* and ends with tag *</p>*). Some knowledge units need all three constraints. Smaller knowledge units can be used to define larger knowledge units.

2.1 Content Constraints Representation

Definite Clauses Grammar (DCG) rules (Sterling and Shapiro 1994) are used to represent content. Here is one example of representing the knowledge unit *price* in DCG rules:

```
price(X) --> [$], number(X1), timeunit(N),{X is X1*N}.
number(X) --> [X], {integer(X)}.
timeunit(1) --> [perweek]; [pw]; [pwk].
timeunit(7/30) --> [permonth]; [pcm]; [p,c,m].
timeunit(1) --> [].
/* The knowledge represented:
There are three components in price: $, an integer number and a time unit.
There are two kinds of time units: per week and per month
Synonyms and abbreviations for per week and per month
The translation rate between the two is 1:7/30.
The default time unit is per week. */
```

DCG rules are good at representing ordered and continuous content. To represent free order content with discontinuous constituents, a set of predicates such as include_freeorder/1, include_any_of/1, exclude/1 are created. One example is:

```
content(real_estate_ad,
include_freeorder([ku(suburb), ku(price), ku(size), ku(type)])).
/*A real estate advertisement consists of knowledge units suburb, price, size and type. The four knowledge units may come in any order and any extra information may appear among them */
```

2.2 Structure Constraints Representation

A set of predicates such as begin_with/1, begin_after/1, end_with/1, end_before/1 are created to describe the structure constraints. HTML tags (tags/1), character template (c_t/1), and knowledge units (k_u/1) can be used to specify the structure.

```
structure(paragraph, end_with([tags("<p"), tags("<hr"), tags("</p")])).
/* A paragraph ends with one of the paragraph tags or line tag */

structure(suburb, begin_with([c_t( "ccc")])).
structure(suburb, end_before([c_t("*l"),tags("<")])).
/* A suburb name consists of upper case letters */
```

2.3 Length Constraints Representation

Two predicates max_length/1, min_length/1 are created to represent length constraints.

```
length(suburb, max_length(20)).
/* A suburb name has less than 20 characters */
```

3. Text parsing

Text is fragmented to paragraphs. For every paragraph, basic knowledge units such as *suburb*, *price, size* and *type* are extracted. Then the basic knowledge units are clustered to groups to form the knowledge unit *real estate ad*.

For every knowledge unit, the extraction consists of three major steps: structure parsing, length parsing and content parsing. For knowledge units represented with three constraints, the process is to first locate the required text and extract a fragment according to the structure constraints. Then the length of the fragment is tested and its content checked. The parsing algorithm is given below. If other knowledge units are used in structure or content constraints, this algorithm is recursively called.

a) If the knowledge unit has structure constraints, the structure parser is started. Structure parser scans the text and extracts a fragment according to the structure constraints.

- If the structure constraint is specified using a character template (e.g. the structure constraint of *suburb)*, then the text string is parsed character by character into a structure string consisting only of tags, character labels (*"c"* for capital letters, *"l"* for lower case letters, *"n"* for numbers), punctuation and special characters. The text string and the structure string have the same length.
- Set the begin and the end pointer by checking the structure constraints on either the structure string (for character template parsing only) or the text string.
- Use the begin and the end pointer to extract a fragment from the text string. The structure string is only used to set the pointers. The output text is always obtained from the text string.

b) If the knowledge unit has length constraints, the length parser is triggered to check maximum and minimum length.

c) If the knowledge unit has content constraints, the text is parsed to a word list that only consists of words (without tags and punctuation). Then one of the two content parsers is triggered. One is a top down parser for knowledge units represented in DCG rules. The other is a parser for free order content with discontinuous constituents which parses every constituents in turn skipping over unrecognized words using the rule:

unrecognized(AnyWord)--> [Anyword].

d) After the three steps, the knowledge unit is extracted and stored into the database.

4. Evaluation

The text parsing performance is evaluated using traditional IR evaluation standards. It is tested on a static collection of Web pages and precision and recall are calculated by comparing CASA's responses with manual parsing results. The text parser was tested on six Web pages downloaded from two Web sites. The six pages consist of "Houses to let" advertisements from Victoria. The first three pages are from *Newsclassifieds* (Newsclassifieds 1997) and most of the advertisements are from the *Leader Newspaper Group*. The other three are from *Fairfax Market* (Fairfax 1997) and all advertisements are from *The Age*. The parsing results on four major knowledge units (*suburb, size, price* and *type*) are shown in Table 1. It can be seen that the overall precision is 96% and the recall is 78%.

The main problem is that the text parser has low precision and especially low recall for suburb parsing. The main reason for this is that the text parser is not sufficiently flexible to parse varying advertisement structures. For example, it fails when an advertisement begins with the full property address instead of a suburb name, and when the suburb is presented in lower case letters. Another problem is that the text parser gets confused when the alternatives of a knowledge unit are given in a single advertisement, for example, phrases such *as "2/3 br", "$400-$450 per week"*, and *"house, unit style"* are not parsed correctly.

Table 1: Test parsing results

Knowledge Units	**Precision***	**Recall ***
Suburb	90	63
Price	99	88
Size	97	80
Type	96	78
Overall	96	78

* Precision=$N_{Correct}/N_{Response}$; Recall = $N_{Correct}/N_{Key}$ in which $N_{Response}$ is the number of knowledge units returned, N_{key} is the total number of knowledge units in the document and $N_{Correct}$ is the number of correct knowledge units returned.

To evaluate CASA's retrieval performance, a comparison between CASA and the search engine of *Newsclassifieds* is shown in Table 2. CASA's search was restricted in this site and its query refining function was not used. The search scope for both systems was limited to online advertisements from "Victoria" state and in the "Houses to let" category on one day (21/11/97). The data was obtained by running 17 queries on both systems. Since a relatively small amount of advertisements were retrieved for each query (about 0-10), instead of the average precision, the sum of the numbers of advertisements retrieved, the numbers of correct advertisements and mistakes were calculated.

Table 2: Comparison between *Newsclassifieds* search engine and CASA

System	***Newsclassifieds***	**CASA**
Sum of $N_{Response}$	186	44
Sum of $N_{Correct}$	51	40
Sum of $N_{Mistake}$	135	4

The results show that CASA makes much fewer mistakes than the *Newsclassifieds* search engine. The most common problem of *Newsclassifieds* search engine is that the search fails when a fragment (or paragraph) consists of a set of advertisements for different real estate properties. When a user specifies a suburb name and a price, the suburb name matches that of one property but the price matches that of another. In contrast, CASA's search, based on knowledge units matching, is particularly good at separating advertisements so that CASA shows a big advantage especially when there is a set of advertisements for different properties within a single paragraph. Another main problem of the *Newsclassifieds* search engine is it performs very poorly on prices. CASA's impressive accuracy proves that the search strategy based on knowledge unit matching has been successful in this particular domain.

5. Conclusions and Future Work

This paper presents a knowledge-based approach which uses information extraction technology to support IR based on knowledge unit matching. In the search process, both query and document are parsed through a text parser, which extracts knowledge units from text, then the extracted knowledge units are used as the basis for further information retrieval. A knowledge-based information agent CASA has been built to search online real estate advertisements from the Internet to help users find accommodation. CASA has shown better performance than the advertisement search engine at *Newsclassifieds.*

This research automates the online real estate advertisement search task and has proved that knowledge-based approach has been successful in this domain. The methods are particularly useful for monitoring Internet online services that provide semi-structured information.

Future work is needed to test knowledge-based architecture, knowledge representation and algorithms in other domains such as car advertisements, job advertisements and other Web services presenting semi-structured documents.

6. Acknowledgements

This work is carried out under the supervision of Prof. Leon Sterling. Also I would like to thank Liz Sonenberg and Dominique Estival for their helpful comments and suggestions.

7. References

Fairfax. (1997). http://www.fairfax.com.au/market.

MUC. (1993). *Fifth Message Understanding Conference (MUC-5) : proceedings of a conference held in Baltimore, Maryland, August 25-27, 1992*, San Mateo, Calif.

Newsclassifieds. (1997). http://www.newsclassifieds.com.au.

SCIE-97. (1997). *Information extraction : a multidisciplinary approach to an emerging information technolology*, Springer, New York.

Sterling, L., and Shapiro, E. (1994). *The Art of Prolog*, The MIT Press.

Winston, P. H. (1984). *Artificial Intelligence*, Addison-Wesley Publishing Company.