Template-Based Information Mining from HTML Documents

Jane Yung-jen Hsu & Wen-tau Yih
Computer Science and Information Engineering
National Taiwan University

Outline

The Web Information Mining Problem
A Model of Electronic Documents
Document Templates
Template-based Information Extraction
A Case Study: *FAQ Miner*
Conclusion

Web Information Mining

Search for relevant documents
- Search engines
- Web guides
- White & yellow pages

Extract target information from documents
- Document analysis
- Information extraction in resource discovery
- Smart web shopping

The Myth about *Keywords*

Relevant information can be found using *keyword-based methods*. e.g.
- Search for relevant documents
- Filter undesirable information
- Extract useful information

Are keywords sufficient to satisfy most of our informational needs?

Problem with Keywords: Example

Sample FAQ Documents

Semi-Structured Document Hypothesis

A semi-structured document, e.g. an HTML document with tags, provides sufficient structural hints to enable effective extraction of semantically meaningful information.

Machine *readable ƒ -* machine *usable*

Basic Elements of A Document

Content

the actual data in a document

Format

the visual presentation of a document

Structure

the logical elements and their relationships

Content

MEMORANDUM TO: JOHN SMITH, GRADUATE OFFICE FROM: MARK SAM SUBJ: STUDENT APPEALS MEETING DATE: 8 APR, 1997 There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore. Please make every effort to attend. If you cannot attend, please contact Mary Armour, ext. 1234.

MEMORANDUM

TO: JOHN SMITH, GRADUATE OFFICE
FROM: MARK SAM
SUBJ: STUDENT APPEALS MEETING
DATE: 8 APR, 1997

There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore.

Please make every effort to attend. *If you cannot attend, please contact Mary Armour, ext. 1234.*

Format

BLAHBLAHBL

BLA HBLA HBLAHB LBLAHBLA HBLAHB
BLAHA HBLA HBL
BLAHB LBLAHBL ABLAHBL ABLAHBLA
BLAHB L AHBL AHBL

Blahb lahb la h blahbla hb lah Blahblahb la Hblahb Lahbla hb Blahblahbl Ahbl ahb lahb la hblah blah bl ahbl ahbl ah Blah bla Hblahblahb

Blahbl blah blahb lahbla hb lahblahb *Bl ahb lahbla hblahbl ahblah blahbla Hbla Hblahbl ahbl hblah*

MEMORANDUM

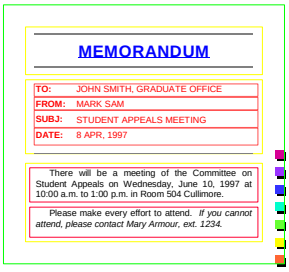
TO: JOHN SMITH, GRADUATE OFFICE
FROM: MARK SAM
SUBJ: STUDENT APPEALS MEETING
DATE: 8 APR, 1997

There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore.

Please make every effort to attend. *If you cannot attend, please contact Mary Armour, ext. 1234.*

Structure

- Memorandum
  - Title
  - Header Block
    - Receiver field
    - Sender field
    - Subject field
    - Date field
  - Memo Body
    - Paragraph 1
    - Paragraph 2



A Model of Electronic Documents

S: a set of structural components
Title, Header Block, Memo Body

C: the sequence of content symbols
MEMORANDUM TO: JOHN SMITH, GRADUATE OFFICE ....

F: format properties of elements in C
Blah blah *blah*

A partial ordering over S

A mapping between C and S

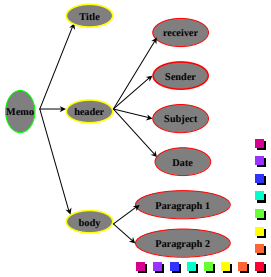
Properties of Document Structure

Context continuity

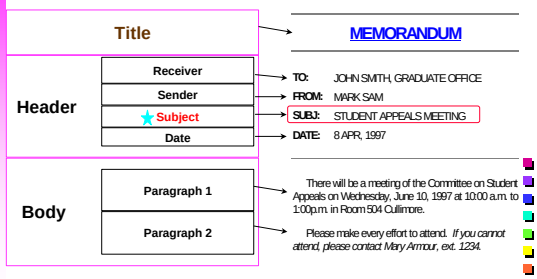
Partial order between levels

Total order within the same level

Order-preserving



Template-based Information Extraction



Memo in SGML

```
<memorandum>
  <title> MEMORANDUM </title>
  <header>
    <rec> TO: JOHN SMITH, GRADUATE OFFICE </rec>
    <send> FROM: MARK SAM </send>
    <subj> SUBJ: STUDENT APPEALS MEETING </sub>
    <date> DATE: 8 APR, 1997 </date>
  </header>
  <body>
    <paragraph>
    There will be a meeting of the Committee on Student
Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00
p.m. in Room 504 Cullimore.
    </paragraph>
    <paragraph>
     Please make every effort to attend.  If you cannot
attend, please contact Mary Armour, ext. 1234.
    </paragraph>
  </body>
</memorandum>
```

MEMORANDUM

TO: JOHN SMITH, GRADUATE OFFICE
FROM: MARK SAM
SUBJ: STUDENT APPEALS MEETING
DATE: 8 APR, 1997

There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore.

Please make every effort to attend. If you cannot attend, please contact Mary Armour, ext. 1234.

Memo in HTML (1/2)

```
<BODY>
  <H1> MEMORANDUM </H1>
  <HR>
  <UL>
    <LI> TO: JOHN SMITH, GRADUATE OFFICE </ LI >
    <LI> FROM: MARK SAM </ LI >
    <LI> SUBJ: STUDENT APPEALS MEETING </ LI >
    <LI > DATE: 8 APR, 1997 </ LI >
  </UL>
  <HR>
  <P>
    There will be a meeting of the Committee on Student
Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00
p.m. in Room 504 Cullimore.
  </P>
  <P>
    Please make every effort to attend.  If you cannot
attend, please contact Mary Armour, ext. 1234.
  </P>
</BODY>
```

MEMORANDUM

TO: JOHN SMITH, GRADUATE OFFICE
FROM: MARK SAM
SUBJ: STUDENT APPEALS MEETING
DATE: 8 APR, 1997

There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore.

Please make every effort to attend. If you cannot attend, please contact Mary Armour, ext. 1234.

Memo in HTML (2/2)

```
<BODY>
  <P> <FONT SIZE=6> <B> MEMORANDUM </B></FONT>
</P>
  <HR>
  <UL>
    <LI> TO: JOHN SMITH, GRADUATE OFFICE </LI >
    <LI> FROM: MARK SAM </LI >
    <LI> SUBJ: STUDENT APPEALS MEETING </LI >
    <LI> DATE: 8 APR, 1997 </LI >
  </UL>
  <HR>
    There will be a meeting of the Committee on Student
Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00
p.m. in Room 504 Cullimore.
  <BR> <BR>
    Please make every effort to attend.  If you cannot
attend , please contact Mary Armour, ext. 1234.
</BODY>
```

MEMORANDUM

TO: JOHN SMITH, GRADUATE OFFICE
FROM: MARK SAM
SUBJ: STUDENT APPEALS MEETING
DATE: 8 APR, 1997

There will be a meeting of the Committee on Student Appeals on Wednesday, June 10, 1997 at 10:00 a.m. to 1:00 p.m. in Room 504 Cullimore.

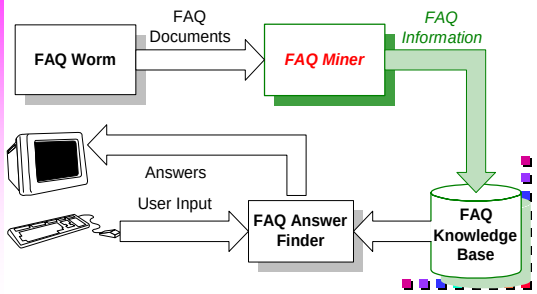
Please make every effort to attend. If you cannot attend, please contact Mary Armour, ext. 1234.

Template: FAQ Documents

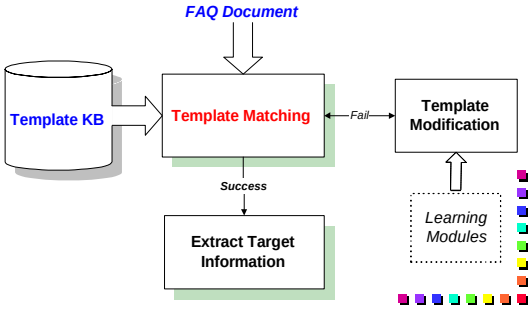
```
Standard_TFAQ
        i
        Title
                <TITLE> TERM_faq_title </TITLE>
        i
        toc
                index_indicator TERM_TOC_indicator
                index_body
                        (ordered_list <OL> list_item* </OL> |
                         unordered_list <UL> list_item* </UL>)
        i
        q_a_pairs
                question_answer_paragraph*

list_item
        <LI> Hyperlink_Anchor   TERM_question </A> </LI>
```


The FAQ Agent



FAQ Miner Architecture



Sample FAQ documents

Experimental Results

Template	# of documents	Success Ratio
Standard_TFAQ	62	56.4%
No_TOC_Indicator	10	9.1%
Near Pass	13	11.8%
Difficult	25	22.7%

Concluding Remarks

Document structure facilitates information extraction.

HTML documents are tree-structured.

HTML tags provide hints for structural elements.

Effective information mining is possible.

What¡s next?

- Tree-structured document templates
- Semantic parsing