Understanding the Hidden Web

Pierre Senellart

17 June 2005
The Hidden Web

Definition (Hidden Web)
The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?
Definition (Hidden Web)

The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?
Understanding the Hidden Web

Pierre Senellart

Introduction

Process description

Discovery
Wrappers
Semantic Analysis
Indexing and Querying

Summary

The Hidden Web

Definition (Hidden Web)

The set of webpages (which may or may not be dynamically generated) not accessible from the hyperlinked structure of the World Wide Web.

Size estimate (2001) : 500 times larger than the surface Web.

How to understand it and benefit from its content?
Understanding the Hidden Web

Purpose

- **Intensional** indexing of the Hidden Web
- **High-level** queries

⇒ a **semantic** search engine over the Hidden Web

In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).
Purpose

- **Intensional** indexing of the Hidden Web
- **High-level** queries
- ⇒ a **semantic** search engine over the Hidden Web

In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).
Understanding the Hidden Web

Purpose

- Intensional indexing of the Hidden Web
- High-level queries
  \[ \Rightarrow \text{a semantic search engine over the Hidden Web} \]

In a fully, unsupervised, way!

Difficult and broad problem. Possible restriction to some domain (e.g. publications).
Web Service Semantic Interpretation Process

Understanding the Hidden Web
Pierre Senellart

Introduction
Process description
Discovery
Wrappers
Semantic Analysis
Indexing and Querying

Summary
Web Service Semantic Interpretation Process

World Wide Web \(\xrightarrow{\text{discovery}}\) HTML form

WSDL
UDDI

understanding the hidden web
Pierre Senellart

Introduction
Process description
Discovery
Wrappers
Semantic Analysis
Indexing and Querying
Summary
Web Service Semantic Interpretation Process

World Wide Web → discovery → HTML form

- UDDI
- WSDL

wrappers

WSDL
Web Service Semantic Interpretation Process

Understanding the Hidden Web
Pierre Senellart

Introduction
Process description
Discovery
Wrappers
Semantic Analysis
Indexing and Querying
Summary

World Wide Web

HTML form
UDDI
WSDL

discovery

wrappers

Analyzed Web Services

WSDL

analysis
Web Service Semantic Interpretation Process

World Wide Web

- discovery
- HTML form
- UDDI
- WSDL

Wrappers

- analysis
- Analyzed Web Services
- WSDL

- indexing
- Web Services Index

Summary

Process description
- Discovery
- Wrappers
- Semantic Analysis
- Indexing and Querying

Introduction
Web Service Semantic Interpretation Process

World Wide Web → discovery → HTML form

UDDI → wrappers → WSDL

Analyzed Web Services → analysis → WSDL

WSDL → indexing → Web Services Index

Queries → querying → Results
1 Introduction

2 Process description
   - Web Service Discovery
     - Wrapping Web Service Descriptions
     - Web Service Semantic Analysis
     - Web Service Indexing and Querying

3 Summary
Crawling the World Wide Web for:

- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index...)

Only interested in Web Services with no side effects:

Ok
- Yellow Pages
- Publication databases

Not Ok
- Booking services
- Mailing list management
- ...
Crawling the World Wide Web for:

- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index...)

Only interested in Web Services with no side effects:

Ok

- Yellow Pages
- Publication databases
- ...

Not Ok

- Booking services
- Mailing list management
- ...

Web Service Discovery
Web Service Discovery

Crawling the World Wide Web for:

- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index...)

Only interested in Web Services with no side effects:

- Ok
  - Yellow Pages
  - Publication databases
- Not Ok
  - Booking services
  - Mailing list management
  - ...
Crawling the World Wide Web for:
- HTML forms implementing a Web Service
- UDDI registries
- WSDL descriptions
- Other resources (XML, HTML, Web as a full-text index...)

Only interested in Web Services with no side effects:

- Ok
  - Yellow Pages
  - Publication databases
  - ...

- Not Ok
  - Booking services
  - Mailing list management
  - ...
Understanding the Hidden Web

Introduction

Process description
- Web Service Discovery
- Wrapping Web Service Descriptions
- Web Service Semantic Analysis
- Web Service Indexing and Querying

Summary
Analyzing the **structure** of HTML forms.

**Issues**

- What are the **relevant** form fields?
- What is the **concrete** type of each field?
- What is the **label** of each field?
Analyzing HTML forms

Analyzing the **structure** of HTML forms.

**Keyword (one !?!?) search:**

| keyword: | exec | Annuler |

**Issues**

- What are the **relevant** form fields?
- What is the **concrete** type of each field?
- What is the **label** of each field?
Analyzing the **structure** of HTML forms.

**Issues**

- What are the **relevant** form fields?
- What is the **concrete** type of each field?
- What is the **label** of each field?
Probing HTML forms to retrieve sample HTML answer pages:

- **With dictionary words**
  
  ![Image with keyword: information]

- **With nonsense words**
  
  ![Image with keyword: sfsfl:msdf]

- **With domain words**
  
  ![Image with keyword: abiteboul]
Probing HTML forms to retrieve sample HTML answer pages:

- **With dictionary words**
  
  ![Probing HTML forms diagram with dictionary words]

- **With nonsense words**
  
  ![Probing HTML forms diagram with nonsense words]

- **With domain words**
  
  ![Probing HTML forms diagram with domain words]
Probing HTML forms to retrieve sample HTML answer pages:

- With **dictionary** words

![Probing HTML forms](image)

- With **nonsense** words

![Probing HTML forms](image)

- With **domain** words

![Probing HTML forms](image)
Query-answer webpages

Extract data from query-answer webpages.

More about

**Publications with key abiteboul**

**Database Publications**
- Regular Rewriting of Active XML and Unambiguity
  - Gemo Report number 385
  - Authors: Serge Abiteboul, Tova Milo, Omer Benjelloun
  - Reference: pods
  - Year: 2005
  - Abstract
  - Download the paper
- Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue
  - Gemo Report number 384
  - Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
  - Reference: PODS 2005
  - Year: 2000
  - Abstract
  - Download the paper

**Bibliographical Sources**
- WEB
- Virtual Library

**Gemo Report**
- Browse publication server
- Ask your own SQL query
- JASH BibTeX list
- GEMO BibTeX

Issues

- **What part** of the webpage contains the answer?
- How to **extract** structured content?
- How to **label** this structured content?
Query-answer webpages

**Extract** data from query-answer webpages.

**More about**

### Publications with key abiteboul

**Database Publications**
- Regular Rewriting of Active XML and Unambiguity
  - Gemo Report number 385
  - Authors: Serge Abiteboul, Tova Milo, Omer Benjelloun
  - Reference: pods
  - Year: 2005
  - Abstract
  - Download the paper

**Bibliographical Sources**
- Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue
  - Gemo Report number 384
  - Authors: Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
  - Reference: PODS 2005
  - Year: 2000
  - Abstract
  - Download the paper

**Gemo Report**
- Complexity of Answering Queries Using Materialized Views
  - Gemo Report number 383
  - Authors: Serge Abiteboul, Oliver Duschka
  - Reference: Almost published in JCSS - blocked for surrealistic patent reasons
  - Year: 1999
  - Abstract
  - Download the paper

**Issues**

- **What part** of the webpage contains the answer?
- **How to extract** structured content?
- **How to label** this structured content?
Query-answer webpages

Extract data from query-answer webpages.

More about

Publications with key abiteboul

Database Publications

- DBMS & Logic
- Research Index

Bibliographical Sources

- Web
- Virtual Library

Gemo Report

- Browse publication server
- Ask your own SQL query
- JATS BibTeX list
- GEMO BibTeX

Issues

- What part of the webpage contains the answer?
- How to extract structured content?
- How to label this structured content?
### Example: ROADRUNNER information extraction engine

<table>
<thead>
<tr>
<th>Publication of Gemo with key abiteboul</th>
<th>Publications with key abiteboul</th>
<th><em>C</em></th>
<th>Gemo Report number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Regular Rewriting of Active XML and Unambiguity</td>
<td>Gemo Report number 385</td>
</tr>
<tr>
<td>Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!</td>
<td></td>
<td>Diagnosis of Asynchronous Discrete event systems. Datalog to the rescue!</td>
<td>Gemo Report number 384</td>
</tr>
<tr>
<td>Complexity of Answering Queries Using Materialized Views</td>
<td></td>
<td>Complexity of Answering Queries Using Materialized Views</td>
<td>Gemo Report number 383</td>
</tr>
<tr>
<td>Representing and Querying XML with Incomplete Information</td>
<td></td>
<td>Representing and Querying XML with Incomplete Information</td>
<td>Gemo Report number 382</td>
</tr>
<tr>
<td>Active XML: A Data-Centric Perspective on Web Services</td>
<td></td>
<td>Active XML: A Data-Centric Perspective on Web Services</td>
<td>Gemo Report number 381</td>
</tr>
</tbody>
</table>

- **Authors:** Serge Abiteboul, Tova Milo, Omar Benjelloun
- **Authors:** Serge Abiteboul, Z. Abrams, S. Haar, T. Milo
- **Authors:** Serge Abiteboul, Oliver Duschka
- **Authors:** Serge Abiteboul, Luc Segoufin, Victor Vianu
- **Authors:** Serge Abiteboul, Omar Benjelloun, Ioana Manolescu, Tova Milo, Roger Weber
1. Introduction

2. Process description
   - Web Service Discovery
   - Wrapping Web Service Descriptions
   - Web Service Semantic Analysis
   - Web Service Indexing and Querying

3. Summary
IsA **ontology of concepts** (simple DAG)

- Person
  - Man
  - Woman
- Publication
  - Proceedings
  - Article
  - Book

- **n-ary typed roles**
  - `AuthorOf(Person, Publication)`
  - `HasName(Person, Name)`
IsA ontology of concepts (simple DAG)

- Person
  - Man
  - Woman

- Publication
  - Proceedings
  - Article
  - Book

n-ary typed roles

- AuthorOf(Person,Publication)
- HasName(Person,Name)
Semantic representation of a service

What is a service described by?

- A \textit{n-uple} of \textit{typed} input parameters
- A \textit{complex} (= nested) type of its output
- Semantic \textit{relations} between inputs and outputs

\textbf{Definition (Complex types)}

\[ T \leftarrow S|<T, \ldots, T>|T^* \]
Semantic representation of a service

What is a service described by?

- A \textit{n-uple} of \textit{typed} input parameters
- A \textit{complex} (= nested) type of its output
- Semantic \textit{relations} between inputs and outputs

Definition (Complex types)

\[ T \leftarrow S|<T, \ldots, T>|T^* \]
What is a service described by?

- A *n*-uple of *typed* input parameters
- A *complex* (= nested) type of its output
- Semantic *relations* between inputs and outputs

**Definition (Complex types)**

\[ S: \text{set of concepts} \]

\[ T \leftarrow S \mid <T, \ldots, T> \mid T^* \]
What is a service described by?

- A *n*-uple of *typed* input parameters
- A *complex* (= nested) type of its output
- Semantic *relations* between inputs and outputs

**Definition (Complex types)**

\[
T \leftarrow S|<T, \ldots, T>|T^*
\]
Example

**Service** giving authors from publication titles

\[ A* \leftarrow \text{WrittenBy}(P,A), \text{HasTitle}(P,T), \text{Input}(T) \]

Example

**Query:**

\[ <A,T^*> \leftarrow \text{WrittenBy}(P,A), \text{Article}(P), \text{HasTitle}(P,T), \text{KeywordOf}(\text{“xml”},P) \]
**Example**

**Service** giving authors from publication titles

\[ A^* \leftarrow WrittenBy(P,A), HasTitle(P,T), Input(T) \]

**Example**

**Query:**

\[ <A,T^*>^* \leftarrow WrittenBy(P,A), Article(P), HasTitle(P,T), KeywordOf("xml",P) \]
Managing extensional information

How to represent extensional information (i.e. documents) in this formalism?

Definition

A document is a service with no input.

Complex types: natural representation of a DTD.

(Disjunctions a | b simulated by (a?, b?) ).
Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. \d{4} − \d{2} − \d{2} is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.
Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. `\d{4} - \d{2} - \d{2}` is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.
Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. \d{4}-\d{2}-\d{2} is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.
Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. \d{4} - \d{2} - \d{2} is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.
Semantic Interpretation of a Service

How to analyze a Web Service?

- Field labels, variable names, tag names
- Concrete type descriptions
  (e.g. \(d\{4\} - d\{2\} - d\{2\}\) is a date)
- Linguistic analysis of plain text descriptions and pages linking to the service

Note: Extracting concepts is easier than extracting relations between concepts.
Introduction

Process description
- Web Service Discovery
- Wrapping Web Service Descriptions
- Web Service Semantic Analysis
- Web Service Indexing and Querying

Summary
Web Service Indexing and Querying

**Given a query, represented as an Analyzed Web Service, how to know which known web services to query?**

**Issues**

- Subsumption of input/output parameters
- Missing input parameters
- Composition of webservices
Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- Subsumption of input/output parameters
- Missing input parameters
- Composition of webservices
Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

Issues

- **Subsumption** of input/output parameters
- **Missing** input parameters
- Composition of webservices
Given a query, represented as an Analyzed Web Service, how to know which known web services to query?

**Issues**

- **Subsumption** of input/output parameters
- **Missing** input parameters
- **Composition** of webservices
Differences with classical database querying

Three main differences:

- Information can be queried only through views (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!
Differences with classical database querying

Three main differences:

- Information can be queried only through views (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!
Differences with classical database querying

Three main differences:

- Information can be queried only through views (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!
Differences with classical database querying

Three main differences:

- Information can be queried only through views (Local As View)
- Nested types
- Incomplete information

Three sources of complexity!
Differences with classical database querying

Three main differences:

- Information can be queried only through \textit{views} (\textit{Local As View})
- \textbf{Nested} types
- \textbf{Incomplete} information

Three sources of complexity!
Web Service Semantic Interpretation Process

World Wide Web $\xrightarrow{\text{discovery}}$ HTML form

World Wide Web $\xrightarrow{\text{wrappers}}$ Analyzed Web Services

Analyzed Web Services $\xrightarrow{\text{analysis}}$ WSDL

WSDL $\xrightarrow{\text{indexing}}$ Web Services Index

Web Services Index $\xrightarrow{\text{querying}}$ Queries $\rightarrow$ Results