

基於欄位填充機制的 XML 文件檢索方法

- (Reducing the semantic gap in XML Retrieval : A Slot Filling Approach)

(以蝴蝶與蛋白質領域為案例)

指導教授 : 項潔教授

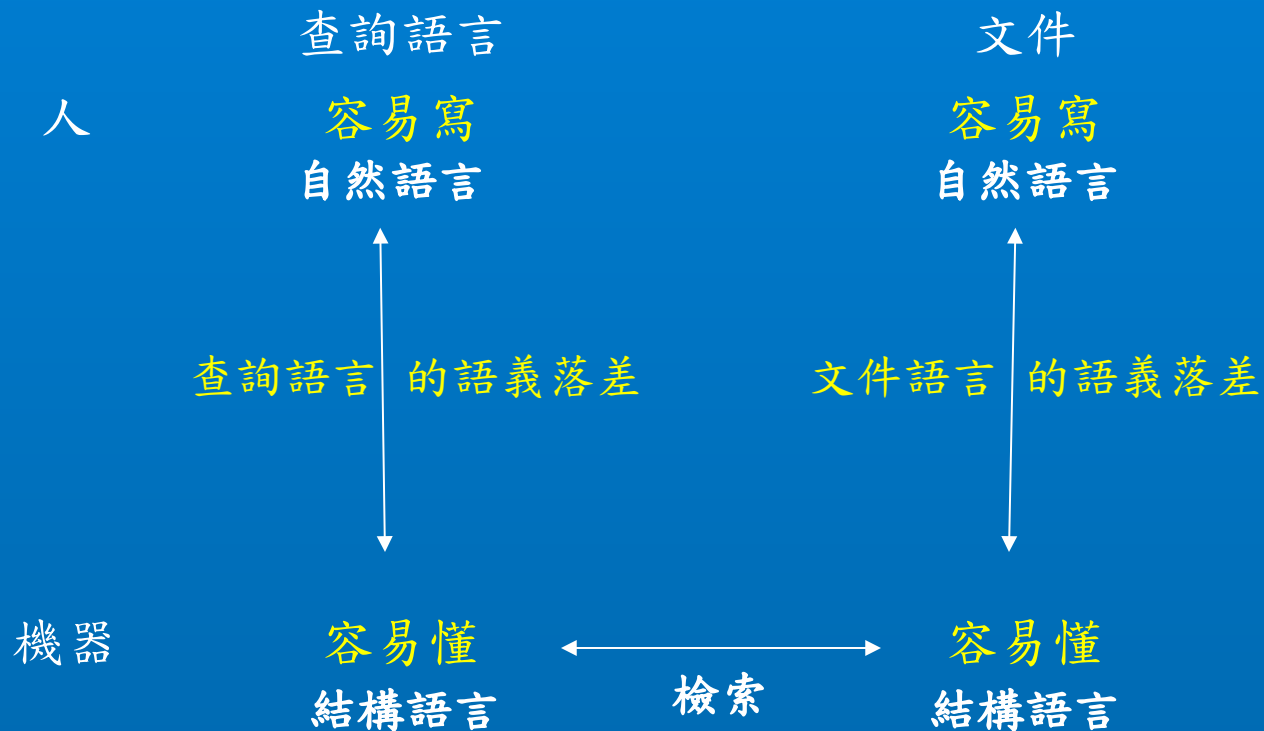
報告人 : 陳鍾誠

Outline

- Motivation
- Problem Statement
- Research Approach
 - Data : document, ontology and query.
 - Method : querying, mapping, and mining.
- Contribution
- Comparison
- Conclusion

Motivation

- 人與機器 之間有語義落差 (Semantic gap).



Motivation

- 事件：XML 出現在 1997 年
 - XML 的一個主要目的是用來降低人與機器之間的語義落差。
- 問題：XML 是否降低了人與機器間的語義落差？
 - 1. XML 查詢語言容易寫嗎？
 - 2. 機器容易讀懂 XML 文件嗎？

Problem Statement

➤ 問題 1 : XML 查詢語言容易寫嗎 ?

- 解釋 : XML 的查詢語言很強, 但很難學習與使用 .
 - 機器易讀, 人卻不容易寫得出來 .
- 範例 : XML 查詢語言
 - For \$b in //butterfly
 - Let \$c=?b//adult//color
 - Where ?c = “green”
 - Return ?b
- 意義 : 找出顏色為綠色的蝴蝶

Problem Statement

➤ 問題 2：機器容易讀懂 XML 文件嗎？

- 解釋：目前沒有已知有效的方法讓機器讀懂 XML 文件。

- A. 目前的檢索方法不考慮 tag 的語義資訊。
- B. 人很難寫出非常詳細的 tag, tag 不夠詳細時，機器又很難讀懂。

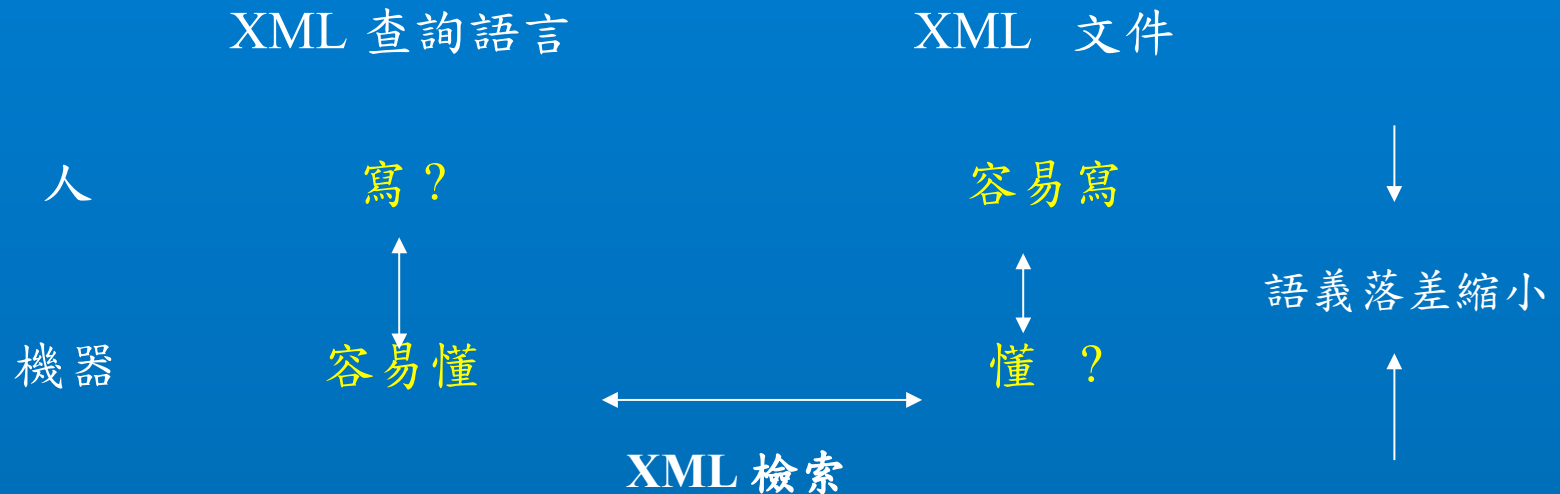
- 範例：XML 文件

- - <butterfly>
 - <cname>拉拉山三線蝶 </cname>
 - <egg><color>淡綠 </color></egg>
 - <larva><color>終齡幼蟲頭部褐色，體呈翠綠色 </color></larva> <pupa><color><color>蛹體底色呈黃褐色 </color></pupa>
 - <adult><color>雄蝶前、後翅表底色為黑色，前翅中室內有一枚長形白斑 </color><adult>
- </butterfly>

- 解釋：查詢“綠色 蝴蝶”時，上述文件會被檢索出來，但語義並不符合。

Problem Statement

- 觀察：單憑 XML 與 查詢語言無法有效縮小語義落差。



Goal

➤ 研究目標

- 降低人與機器之間在 XML 上的語意落差 .

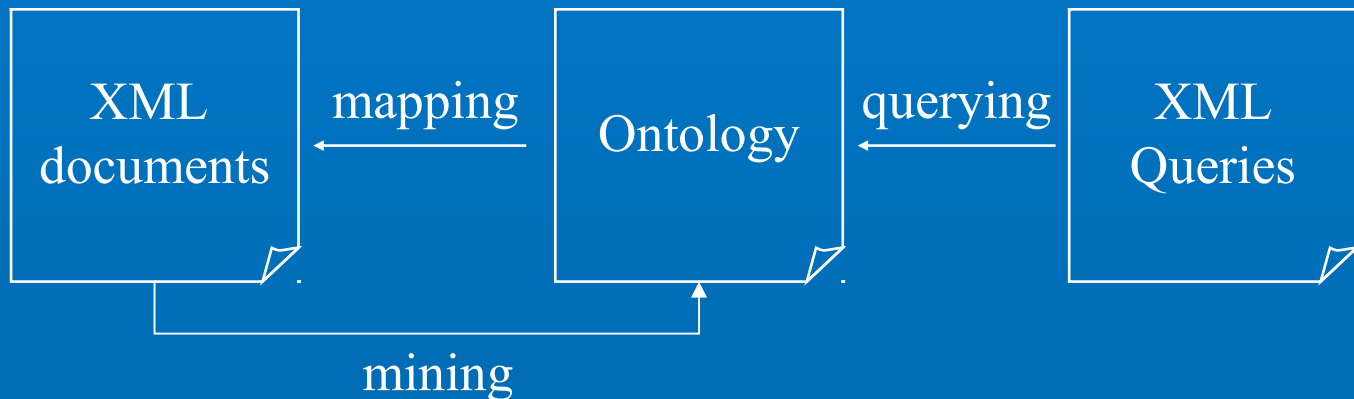
➤ 子目標

- 降低人與機器之間在 “ XML 查詢語言 ” 上的語義落差 .
 - 讓人很容易的寫出 XML 查詢語言 .
- 降低人與機器之間在 “ XML 文件理解 ” 上的語義落差 .
 - 讓機器很容易的讀懂 XML 文件 .

Research Approach

Data : Document, Ontology and Query

Method : querying, mapping, and mining



Our Approach

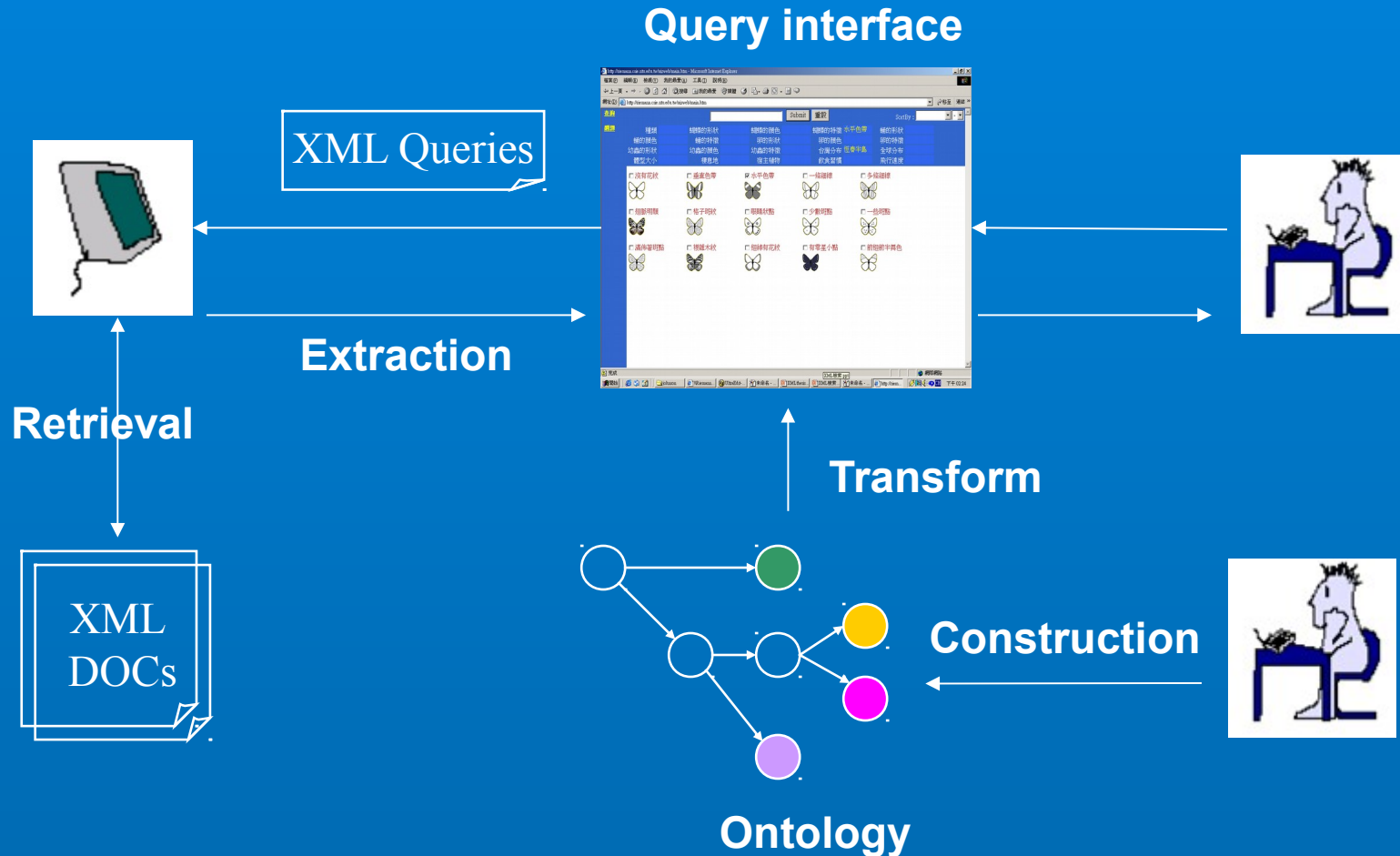
➤ 方法

- 利用 Ontology 作為人與機器的中介，用以降低語意落差。

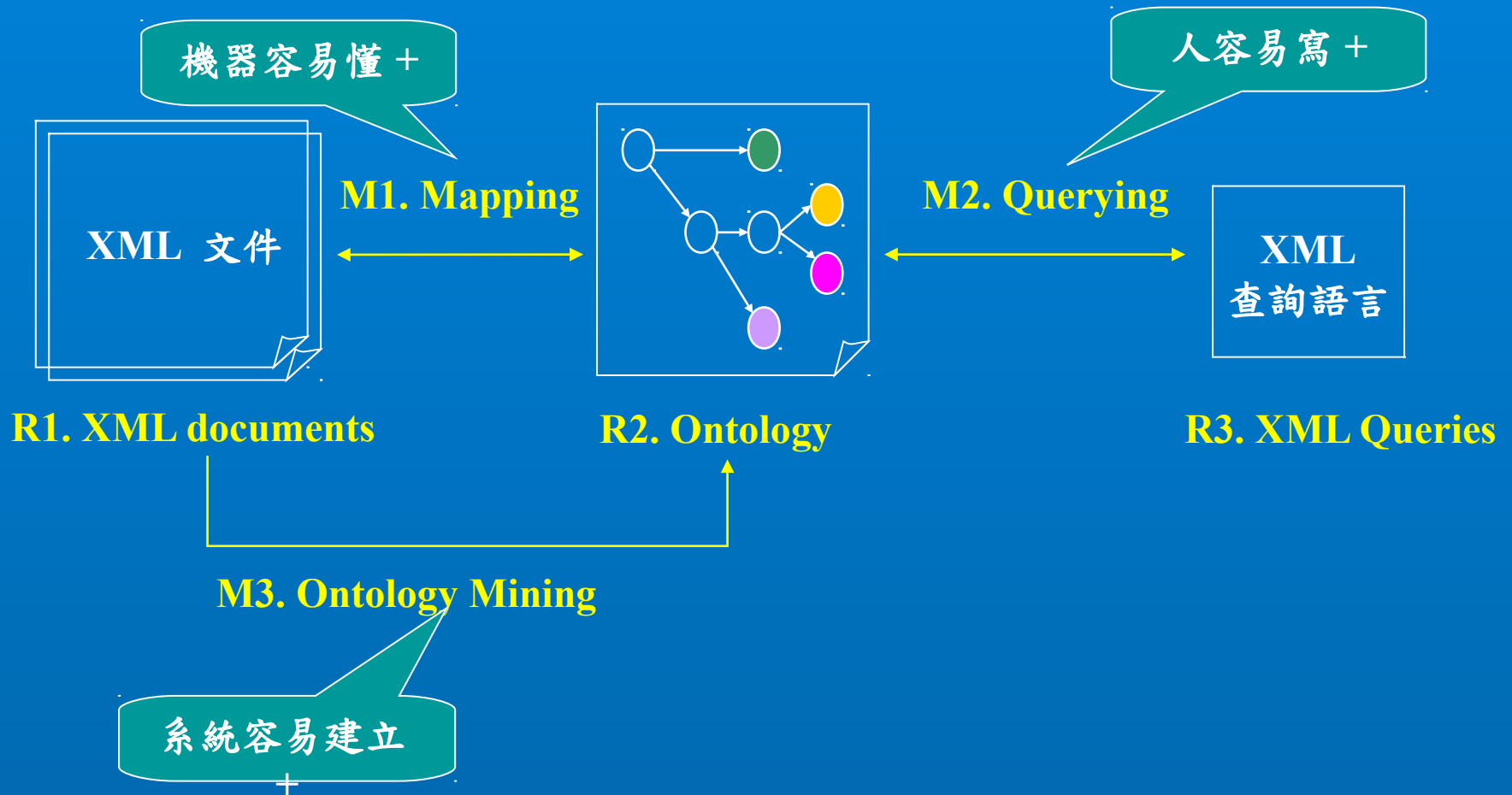
➤ 子方法

- XML 查詢語言：利用 Ontology 幫助人查詢 XML 文件。
 - 透過 Ontology 建立查詢介面，讓人“容易寫出” XML 查詢語言。
- XML 文件語言：利用 Ontology 幫助機器理解 XML 文件。
 - 將 XML 文件映射到 Ontology 中，使機器“讀懂” XML 文件。
- Ontology：自動建立 Ontology，降低系統建立成本。
 - 從 XML 文件中統計出每個 tag 的重要詞彙，建立 Ontology。

An XML Retrieval Scenario



Architecture



Components

➤ Data

- R1. XML :
 - How to represent XML documents ?
- R2. Ontology :
 - How to represent ontologies ?
- R3. Query :
 - How to represent XML queries ?

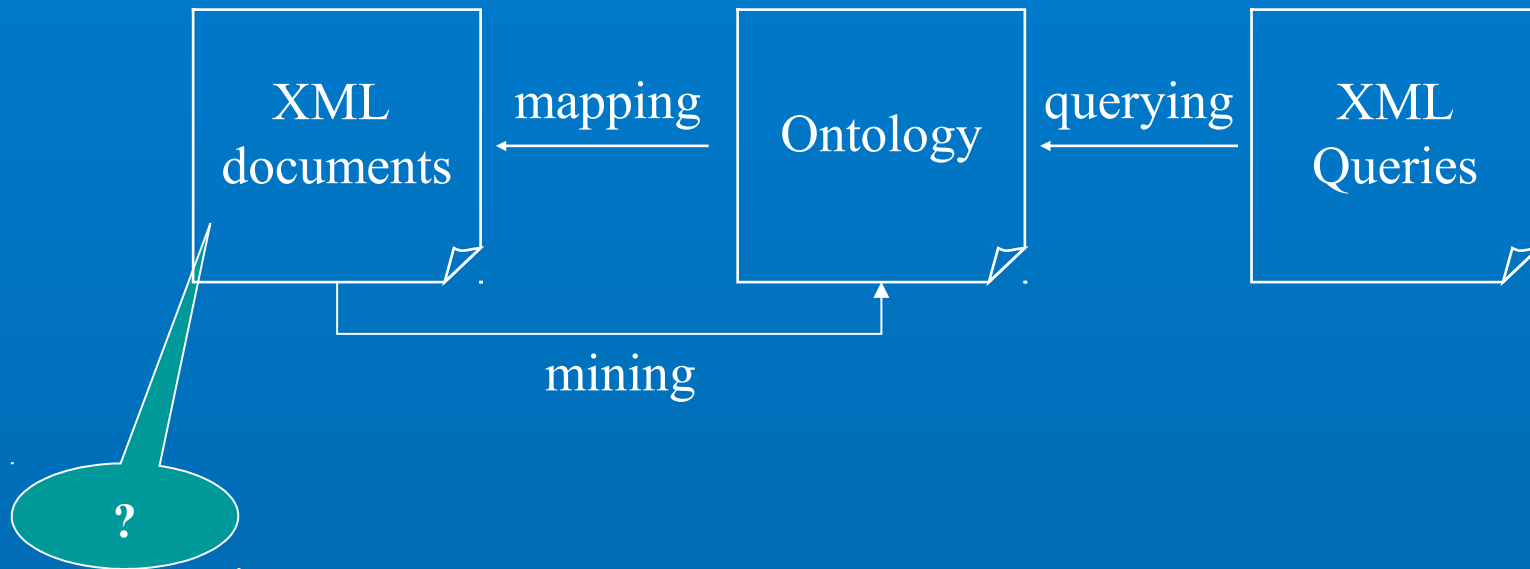
➤ Methods

- M1. Querying :
 - How to use ontology to help user build XML queries ?
- M2. Mapping :
 - How to map XML documents into an ontology ?
- M3. Ontology Mining :
 - How to mine ontology from XML documents ?

Data 1 : XML

➤ Question

- How to represent XML documents ?



XML

➤ XML Document

```
<butterfly>
  <cname> 阿里山小灰蛺蝶 </cname>
  <adult>
    <color> 雄蝶前、後翅表底色為茶褐色，前翅外緣各翅室有一銀色細帶紋 </color>
    <feature> 後翅外緣呈輕微鋸齒狀 </feature>
    <size> 本種為中小型蝶種，展翅約為 40-50mm </size>
  </adult>
</butterfly>
```

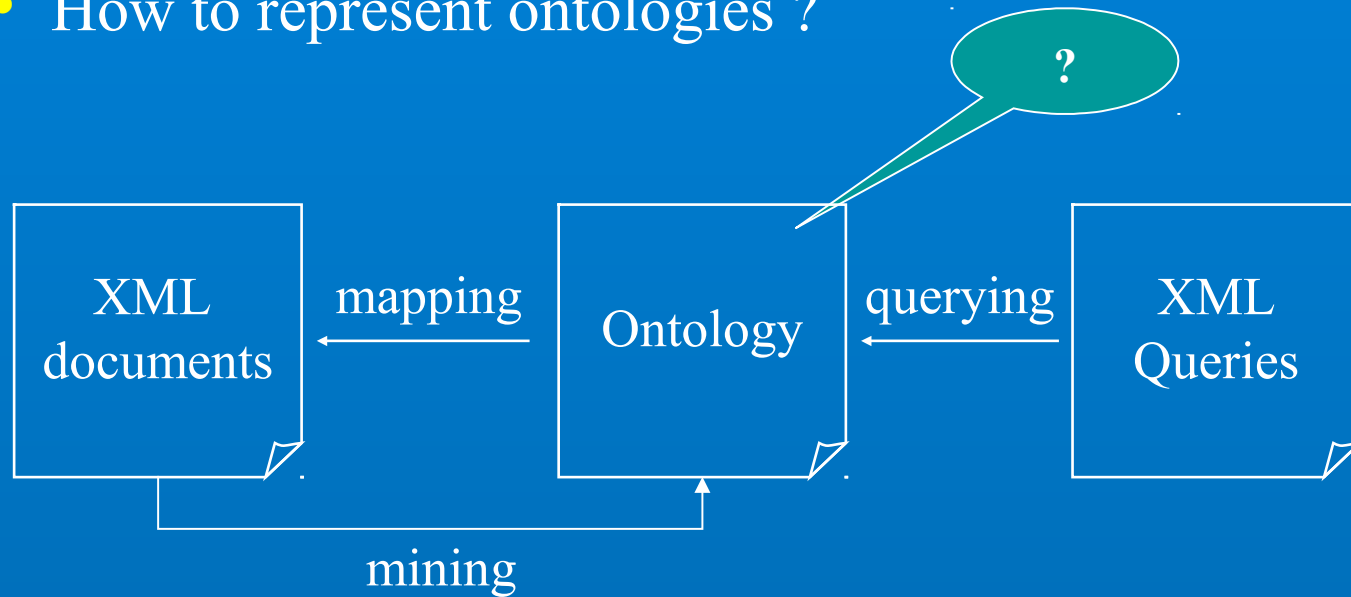
➤ Pair Representation : $R(d) = \{ (p, c) \}$

```
{
  (butterfly/cname, 阿里山小灰蛺蝶 ),
  (butterfly/adult/color, 雄蝶前、後翅表底色為茶褐色，前翅外緣各翅室有一銀色細帶紋 ),
  (butterfly/adult/feature, 後翅外緣呈輕微鋸齒狀 ),
  (butterfly/adult/size, 本種為中小型蝶種，展翅約為 40-50mm)
}
```

Data 2 : Ontology

➤ Question

- How to represent ontologies ?



Ontology : Slot-Tree

Slot-Tree

➤ <s slot=" 蝴蝶 " path="//butterfly">

- <s slot=" 學名 " path="//butterfly//cname"/>

- <s slot=" 成蟲 " path="//butterfly//adult">

- <s slot=" 顏色 " path="//butterfly//adult//color">

- <v value=" 黑色 "/>

- <v value=" 淺棕色 " keys=" 褐色 " />

- <v value=" 黑白相間 " match=" 黑色 & 白色 " /></s>

- <s slot=" 形狀 " path="//butterfly//adult//feature">

- <v value=" 有尾突 " keys=" 尾突 , 突出 " />

- <v value=" 翅緣破裂 " keys=" 鋸齒 , 波浪 " /></s>

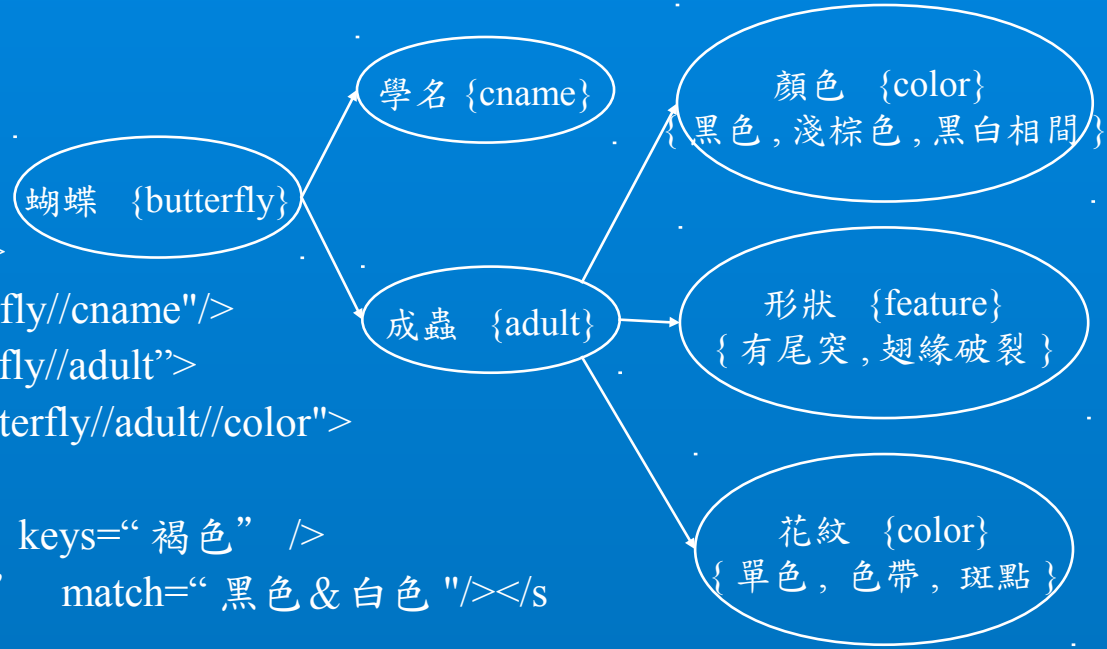
- <s slot=" 花紋 " path="//butterfly//adult//color">

- <v value=" 單色 " keys=" 無 .. 花紋 " />

- <v value=" 色帶 " keys=" 條紋 , 帶紋 " />

- <v value=" 斑點 " keys=" 圓班 , 圓點 " /></s></s>

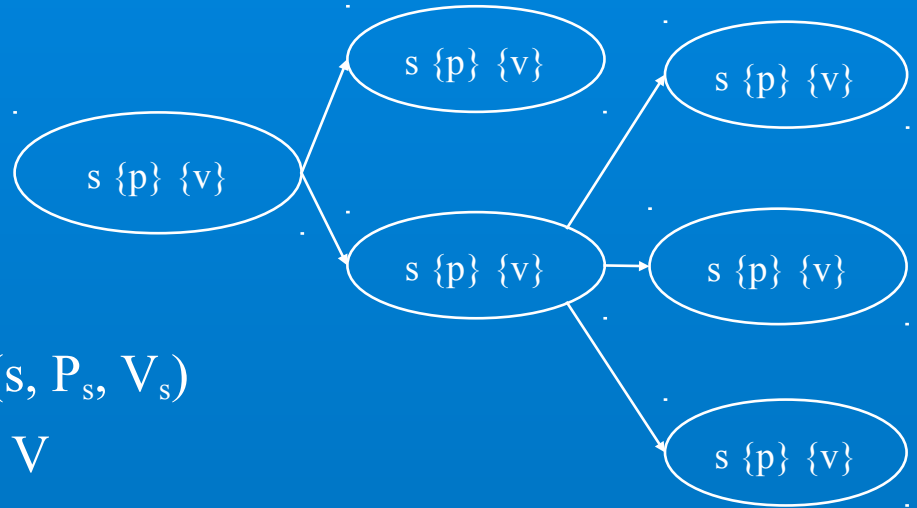
➤ </s>



Ontology : Slot-Tree

➤ Definition : Slot-Tree

- $ST = (T, S, P, V)$
- Slot-Tree is a tree T
 - Each node of T is a tuple (s, P_s, V_s)
 - Where $s \in S, P_s \subseteq P, V_s \subseteq V$



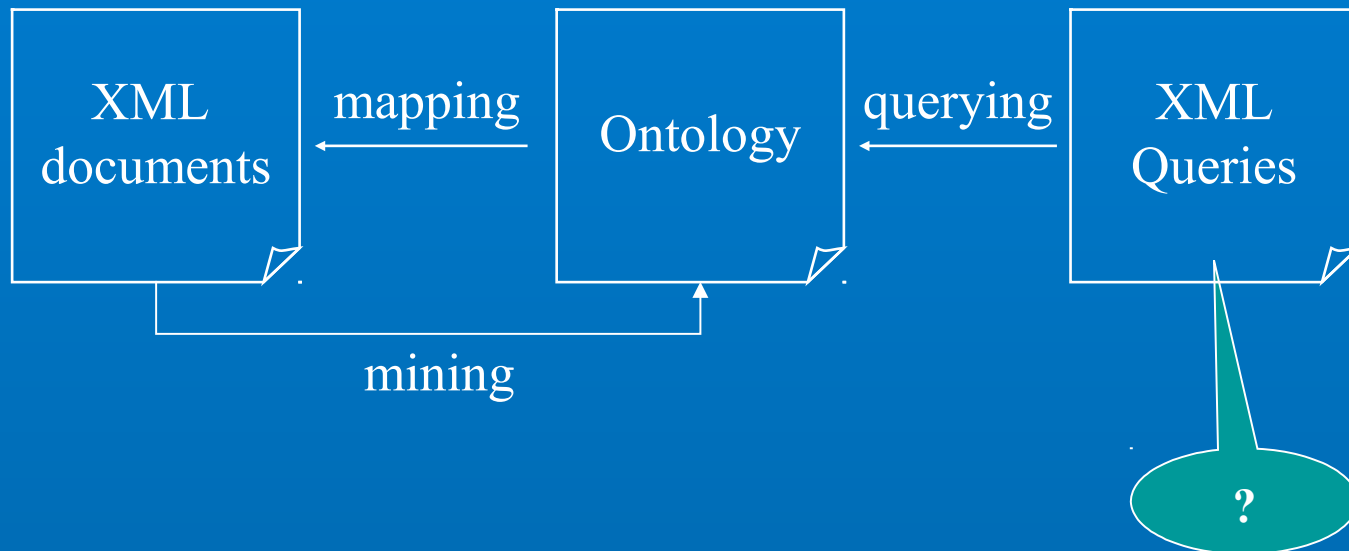
➤ Syntax

- $S \rightarrow \langle s \text{ slot}=\text{"L"} \text{ path}=\text{"XP*"} \rangle V^* S^* \langle /s \rangle$
- $V \rightarrow \langle v \text{ value}=\text{"L"} \text{ keys}=\text{"K"} \text{ match}=\text{"R"} \rangle$
- $K \rightarrow L^*$

Data 3 : Query

➤ Question

- How to represent XML queries ?



Query : Path

➤ Syntax : PATH

- $XP \rightarrow PART^*$;
- $PART \rightarrow TAG \{ COND \}$
- $PART \rightarrow / TAG \{ COND \}$
- $PART \rightarrow // TAG \{ COND \}$
- $COND \rightarrow [@ ATT = 'L']$
- $TAG \rightarrow L$
- $ATT \rightarrow L$

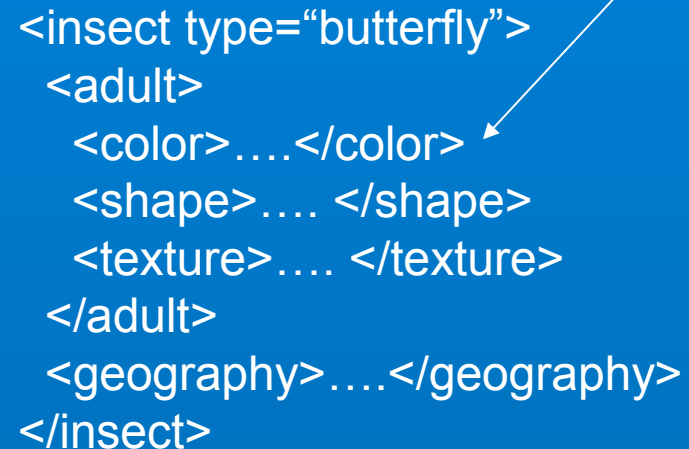
➤ Example

- /butterfly/adult/color
- //insect//color
- //insect[@type='butterfly']//color,

➤ Semantics

- $match(node, xp) = \{true, false\}$

//insect[@type='butterfly']//color,



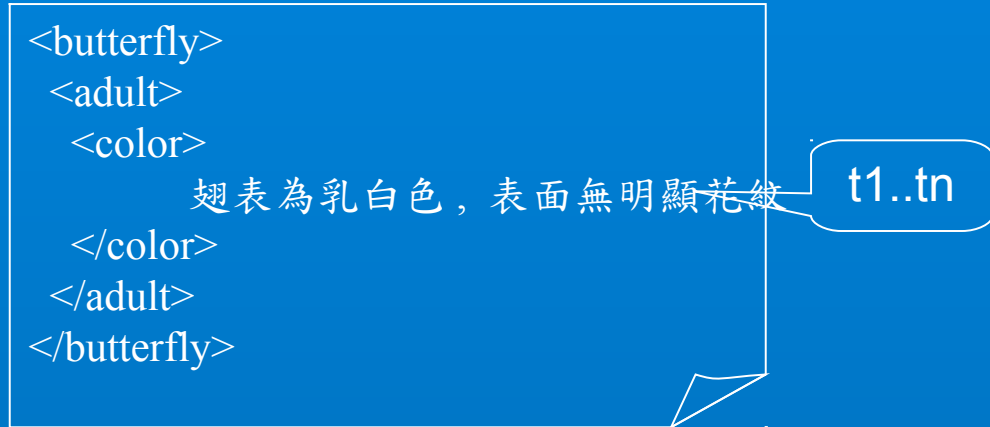
```
<insect type="butterfly">
  <adult>
    <color>....</color>
    <shape>.... </shape>
    <texture>.... </texture>
  </adult>
  <geography>....</geography>
</insect>
```

The diagram illustrates the semantics of the XPath query //insect[@type='butterfly']//color. It shows an XML document structure where the query selects the 'color' element within an 'adult' element of an 'insect' element whose 'type' attribute is 'butterfly'. An arrow points from the query text to the corresponding 'color' element in the XML tree.

Query : Rule

➤ Syntax : Rule

- $R \rightarrow (R \ \& \ R)$
- $R \rightarrow (R \ | \ R)$
- $R \rightarrow E$
- $R \rightarrow -E$
- $E \rightarrow L \{..L\}$



➤ Example

- $R = \text{“黑色 \& 白色”}$
- $R = \text{“白色 \& - 乳白色”}$
- $R = \text{“無 .. 花紋”}$

$\text{match}(t1..tn, \text{“白色 \& - 乳白色”}) = \text{false}$

$\text{match}(t1..tn, \text{“無 .. 花紋”}) = \text{true}$

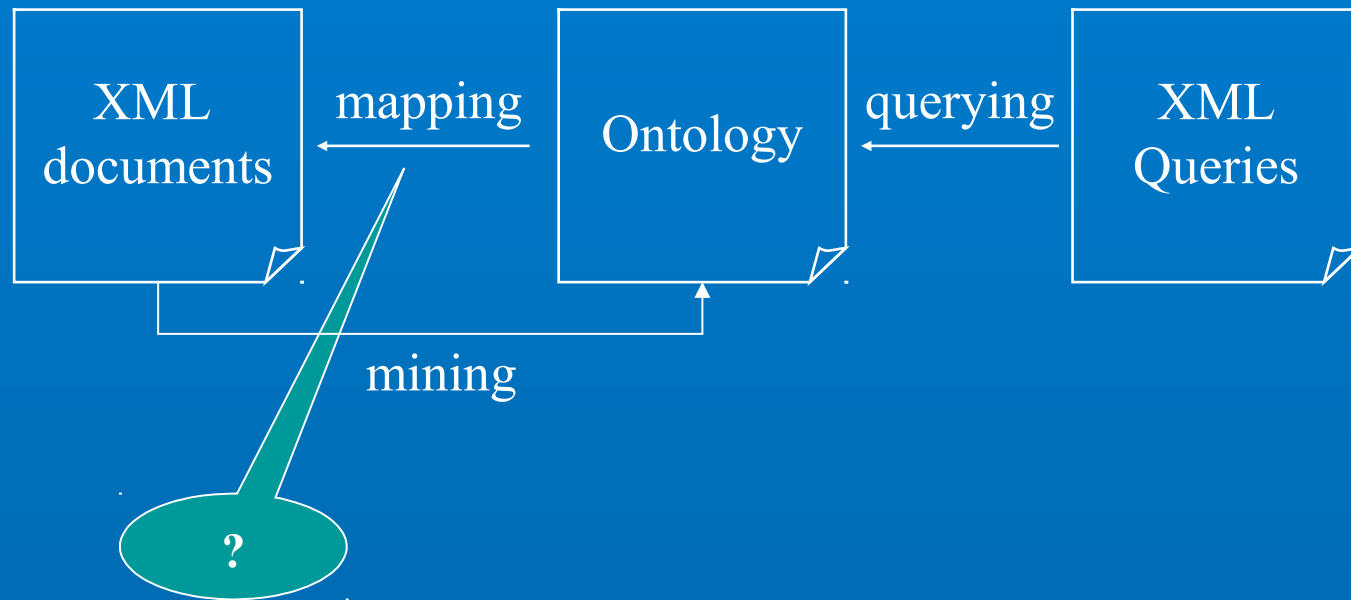
➤ Semantics

- $\text{match}(t1..tn, R) = \{\text{true}, \text{false}\}$

Method 1 : Mapping

➤ Question

- How to map XML documents into an ontology ?



Mapping : Slot-Filling

d

```
<butterfly>
  <cname> 阿里山小灰蛺蝶 </cname>
  <adult>
    <color> 雄蝶前、後翅表底色為茶褐色，
            前翅外緣各翅室有一銀色細帶紋 </color>
    <feature> 後翅外緣呈輕微鋸齒狀 </feature>
    <size> 本種為中小型蝶種，
            展翅約為 40-50mm </size>
  </adult>
</butterfly>
```

```
<s slot=" 蝴蝶" path="//butterfly">
  <s slot=" 成蟲" path="//butterfly//adult">
    <s slot=" 顏色 " path="//butterfly//adult//color">
      <v value=" 黑色 "/>
      <v value=" 淺棕色" keys=" 褐色" />
      <v value=" 黑白相間" match=" 黑色 & 白色" /></s>
    <s slot=" 形狀" path="//butterfly//adult//feature">
      <v value=" 有尾突" keys=" 尾突, 突出" />
      <v value=" 翅緣破裂" keys=" 鋸齒, 波浪" /></s>
    <s slot=" 花紋 " path="//butterfly//adult//color">
      <v value=" 單色" keys=" 無 .. 花紋" />
      <v value=" 色帶" keys=" 條紋, 帶紋" />
      <v value=" 斑點" keys=" 圓班, 圓點" /></s></s>
  </s>
```

蝴蝶

成蟲

顏色，淺棕色：

1

形狀，翅緣破裂

: 1

花紋，色帶：1

Mapping : Algorithm

$$B(d/T) = \{ (s,v) \mid \forall_{v \in V_s, t \in d_s} w(v, d_s) > \varepsilon \}$$

Algorithm Slot-Filling(d, T)

SV = {}

for each s in T

$d_s = \{c \mid (s, p) \in M(T), (p, c) \in d\}$

for each v in s

if $w(v, d_s) > \varepsilon$ then put $(s,v : w(v, d_s))$ into SV

end for

end for

return SV

$O(|V_s|)$

$O(|d_s|)$

Time Complexity = $\sum_s |d_s| * |V_s|$ Worst Case $\rightarrow O(|d| * |T|)$

$|d_s|$: size of blocks that match s .

$|V_s|$: size of nodes that match s .

$|T|$: size of slot-tree T .

$|d|$: size of document d .

Mapping : Extraction

➤ Slot-Filling

- $(d/T)_{s,v} : (s,v)$ 這一格共被填入多少分數

➤ Extraction Algorithm

- $E(d/T, \varepsilon) = \{(s,v) \mid (d/T)_{s,v} \geq \varepsilon\}$

➤ Example

- $E(d/T, 0.5) = \{ (\text{蝴蝶} / \text{成蟲} / \text{顏色}, \text{淺棕色}),$
 $(\text{蝴蝶} / \text{成蟲} / \text{形狀}, \text{翅緣破裂}),$
 $(\text{蝴蝶} / \text{成蟲} / \text{花紋}, \text{色帶}) \}$

Mapping : IR Model

➤ Slot Vector Space Model (SVSM)

- $V(d/T) = (d/T_{s1,v1} \dots d/T_{s1,vx} \dots d/T_{sn,vy})$

➤ Similarity($d1, d2 \mid T$)

- $S(d1, d2 \mid T) = V(d1/T) \bullet V(d2/T)$

蝴蝶 / 成蟲 / 顏色 , 淺棕色 : 1
蝴蝶 / 成蟲 / 形狀 , 翅緣破裂 : 1
蝴蝶 / 成蟲 / 花紋 , 色帶 : 1

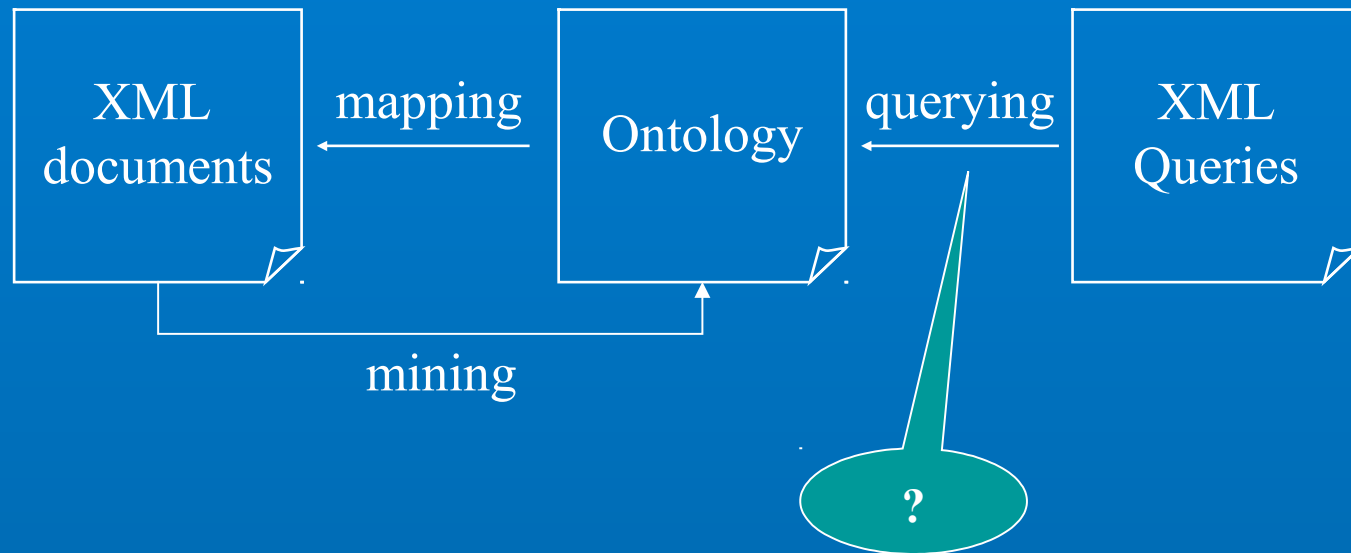
Multidimensional
similarity

蝴蝶 / 成蟲 / 顏色 , 淺棕色 : 1
蝴蝶 / 成蟲 / 形狀 , 有尾突 : 1
蝴蝶 / 成蟲 / 花紋 , 色帶 : 1

Method 2 : Querying

➤ Question

- How to use ontology to help user build XML queries ?



Querying : Interface

Slot

http://riemann.csie.ntu.edu.tw:8080/xir/frame.htm - Microsoft Internet Explorer

File Edit View Favorites Tools Help







Back Forward Stop Home Search Media Print

Address http://riemann.csie.ntu.edu.tw:8080/xir/frame.htm Go

Submit Reset SortBy : 翅的長度

種類	蝴蝶的形狀	翅緣破裂	蝴蝶的顏色	大致淺棕色	蝴蝶的特徵	蛹的形狀
蛹的顏色	蛹的特徵		卵的形狀		卵的顏色	卵的特徵
幼蟲的形狀	幼蟲的顏色		幼蟲的特徵		台灣分布	全球分布
體型大小	棲息地		宿主植物		飲食習慣	飛行速度

☐ 類似燕尾 ☐ 細小尾突 ☒ 翅緣破裂 ☐ 翅緣波浪狀 ☐ 似蛾狀 ☐ 似枯葉狀

Value

Done Internet

開始 Profiles - Enter... PHD Microsoft Powe... HiNet - Micros... http://riemann.c... 下午 02:37

Querying : Language

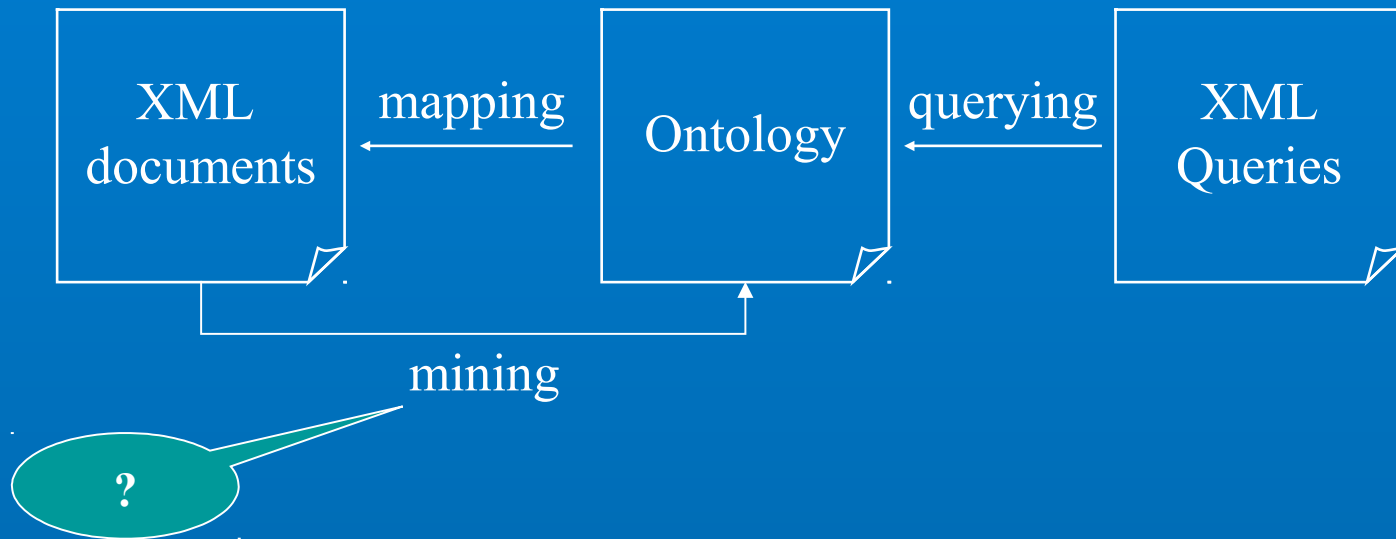
➤ Query

- `<s slot=“ 蝴蝶” >`
 - `<s slot=“ 成蟲” />`
 - `<s slot=“ 顏色” values=“ 淺棕色” />`
 - `<s slot=“ 形狀” values=“ 翅緣破裂” />`
 - `</s>`
- `</s>`

Method 3 : Ontology Mining

➤ Question

- How to mine ontology from XML documents ?



Ontology Mining

➤ Question :

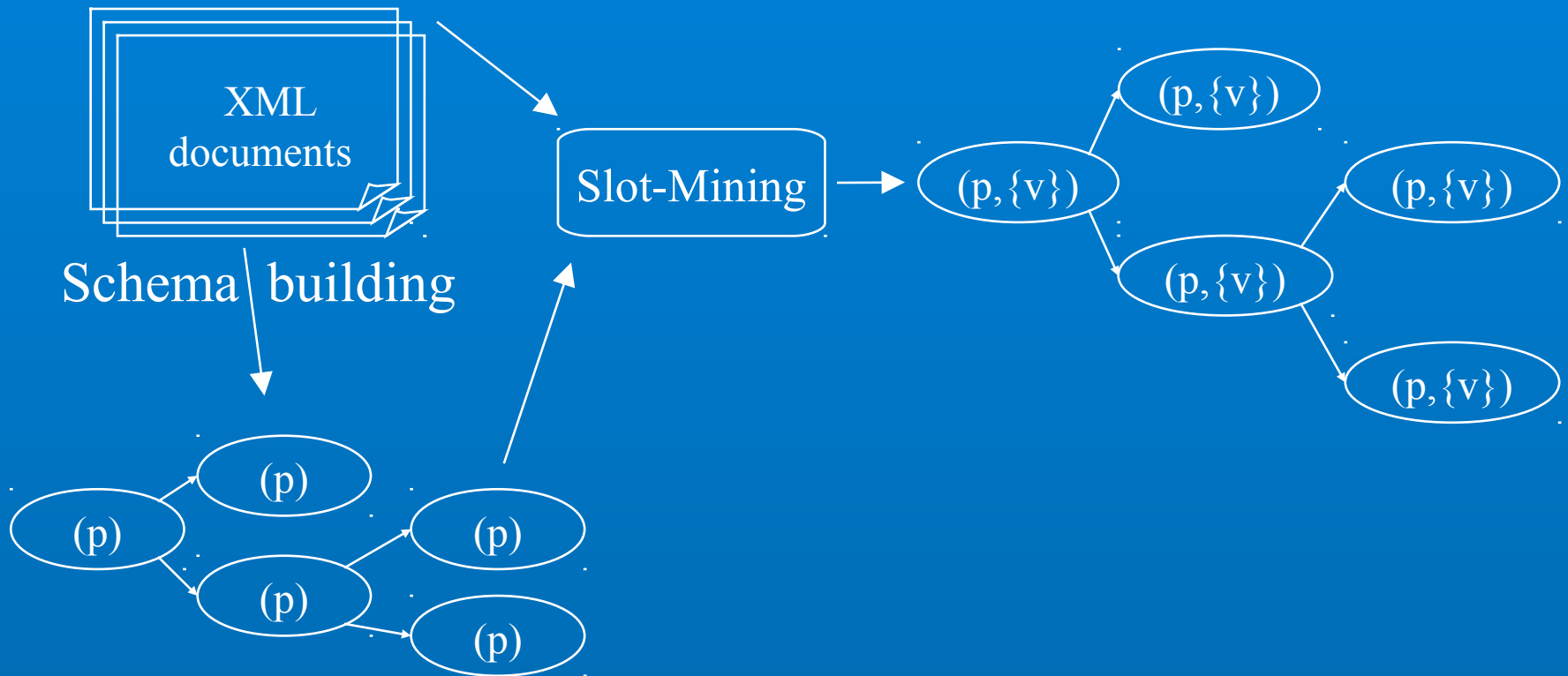
- Mining ontology from XML documents.

➤ Method

- Mining the relation between tag and value.
- Using correlation analysis to mining the (p, v) pairs.

- $V_p = \{ t \mid \text{Cor}(p, t) \geq r \}$

Ontology Mining : Process



Ontology Mining : Example

■ 自動建立 Slot-Tree

■ Mining(D) $\rightarrow \{(p,v)\}$

台灣
高山
褐色
鋸齒狀
...

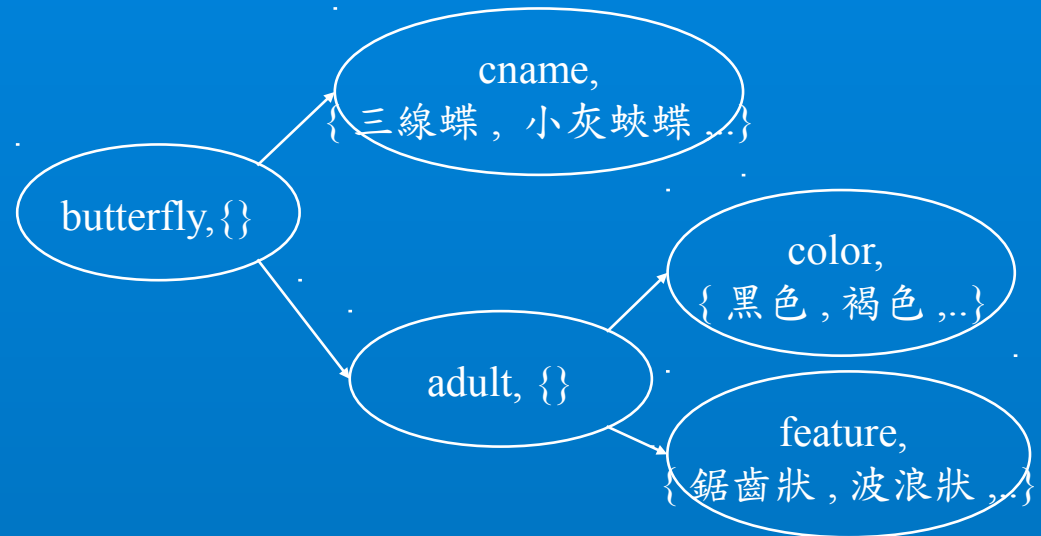
<feature>.. 邊緣呈鋸齒狀 </feature>
<feature>.. 翅緣為鋸齒狀 ..</feature>
<feature>.. 有鋸齒狀翅膀... </feature>
<feature>.. 呈現鋸齒狀 ..</feature>

鋸齒狀 : 7
(feature : 6)
(color : 1)

$\text{Cor}(p, t) \geq r$

組合成 tree

(feature, 鋸齒狀)



Ontology Mining : Model

➤ XML 文件集合 D 的向量表示法

- $V(D) = (D_{p1,t1}, \dots, D_{p1,tk}, \dots, D_{pn,t1}, \dots, D_{pn,tk})$
- 簡寫：
 - $|D_p| = \sum_t D_{p,t} \quad |D_t| = \sum_p D_{p,t} \quad |D| = \sum_p \sum_t D_{p,t}$

➤ Slot-Mining Algorithm

- $\text{Cor}(p, t) = P(t | p) / P(t) = (D_{p,t} / |D_p|) / (|D_t| / |D|)$
- $\{v\} = \text{mining}(D, p) = \{ t \mid \text{Cor}(p, t) \geq r \}$

Ontology Mining : Algorithm

Mining(D) = { (p, t) }

Algorithm Slot-Mining (D)

P = {p | p is a path in D}

for each (p,t) in D

$|D_{p,t}| = |D_{p,t}| + 1$

$|D_p| = |D_p| + 1$

$|D_t| = |D_t| + 1$

$|D| = |D| + 1$

end for

for each (p,t) in PT

$p(t | p) = |D_{p,t}| / |D_p|$

$p(t) = |D_t| / |D|$

if $p(t|p)/p(t) > r$ then put (p,t) into SV

end for

return SV

$|D|$: D 所包含的詞數
 $|D_t|$: t 在 D 中出現的次數
 $|D_p|$: D_p 所包含的詞數
 $|D_{p,t}|$: t 在 D_p 中出現的次數

Time Complexity : $O(|D| * \log|D|)$

Ontology Mining : Results

Analysis

\butterfly\classification\cfamily	鳳蝶科，蛱蝶科，蛇目蝶科，粉蝶科，斑蝶科，弄蝶科，小灰蝶科
\butterfly\classification\family	Satyridae, Pieridae, Papilionidae, Papilio, Nymphalidae, Lycaenidae, Hesperidae, Danaidae
\butterfly\footnote	高冷蔬菜區，非常，開發，開墾，長達，近年來，種經，種族群，破壞，生活史，...
\butterfly\geographic\global	馬來半島，非洲，錫金 西部，蘇門達臘，蘇門答臘 蘇門，群島，美洲，緬甸北部，緬甸，琉球群島，琉球，爪哇，...
\butterfly\honeyplant\	馬櫻丹，馬利筋，馬利，金露花，野花，豐草，菊科野花，菊科，菊科，花蜜，腐熟，繁星花，繁星，紫花霍香薊，...
\butterfly\life_stage\adult\predator	鳥類，青蛙，螳螂，蜻蜓，蜥蜴，蜘蛛，捕食性天敵，捕食，性天敵，天敵，
\butterfly\life_stage\egg\feature	高饅頭形，饅頭，頂點，頂部微凸，頂部，角形，表面，著生，菱形，花紋，縱脊，細長刺毛，細長，細小突起，細小，精孔
\butterfly\life_stage\adult\color	黑褐色，黑褐，黑色細帶紋，黑色斑點，黑色斑紋，黑色性徵，黑色帶紋，黑色小斑，黑色小圓斑，黑色外框，...

Domain : Protein

Protein

- Collection : Protein Information Resource
 - <http://pir.georgetown.edu/>
- Information : Fields
 - ID, name, source organism, function, classification, feature, length, type, sequence
 - create_date, keyword, reference (author, citation), access information

XML Data

```

<ProteinEntry id="S35333">
  <created_date>03Feb1994</created_date>
  <protein><name>steroid receptor protein svp44</name></protein>
  <organism><source>zebra fish</source><formal>Brachydanio rerio</formal></organism>
  <reference>
    <authors><author>Fjose, A.</author><author>Nornes, S.</author>...</authors>
    <citation>EMBO J.</citation>
    <volume>12</volume><year>1993</year><pages>14031414</pages>
    <title>Functional conservation of vertebrate sevenup related genes in neurogenesis ...
    <xrefs><xref><db>MUID</db><uid>93223680</uid></xref></xrefs></reference>
  <accinfo label="FJO">
    <accession>S35333</accession><moltype>mRNA</moltype><seqspec>1411</seqspec>
    <xrefs><xref><db>EMBL</db><uid>X70299</uid></xref>...</accinfo></reference>
  <genetics><gene><uid>svp44</uid></gene></genetics>
    <classification><superfamily>unassigned erbArelated proteins</superfamily>...
    <keywords>...DNA binding...zinc finger,...</keywords>
  <feature label="ERBA"><featuretype>domain</featuretype>
    <description>erbA transforming protein homology</description>
    <seqspec>74320</seqspec></feature>...
  <summary><length>411</length><type>complete</type></summary>
<sequence>MAMVVSVWRDPQED.... </sequence> ....

```

Domain Knowledge

<frame>

<s slot=" 分子種類 " path="/ProteinEntry/reference/accinfo/mol-type">

<v value="protein" /><v value="DNA" /><v value="RNA" /></s>

<s slot=" 分子形狀 " path="/" menu="yes">

<v value=" 螺旋 =Alpha" keys="Helix"/><v value=" 平板 =Beta" keys="Sheet"/>

<v value="Alpha+Beta" /><v value="Parallel-Beta" /><v value="AntiParallel-Beta" /></s>

<s slot=" 分子來源 " path="/">

<v value=" 人 =Human" /><v value=" 動物 =Animal" /><v value=" 植物 =Plants" />

<v value=" 細菌 =Bacteria" /><v value=" 病毒 =Virus" /><v value=" 酵母 =Yeast" />

<v value=" 魚 =Fish" /><v value=" 蟲 =Insects" /><v value=" 鳥 =bird" /><v value=" 獸 " /></s>

<s slot=" 身體部位 " path="/">

<v value=" 心臟 =Heart"/><v value=" 肺臟 =Lung"/><v value=" 肝臟 =Liver"/>...

<v value=" 血液 =Blood"/><v value=" 骨骼 =Bone"/><v value=" 荷爾蒙 =pheromone"/>...

<v value=" 根 =Root"/><v value=" 莖 =Stem+Trunk"/><v value=" 葉 =Leaf"/>...</s>

<s slot=" 細胞部位 " path="/">

<v value=" 細胞核 =Nucleus"/><v value=" 細胞質 =Cytoplasm"/>...

<v value=" 內質網 =Endoplasmic_reticulum"/><v value=" 高基氏體 =Golgi_Bodies"/>...</s>

<s slot=" 身體功能 " path="/">

<v value=" 消化 =Digestion"/><v value=" 運動 =Motion"/><v value=" 感覺 =Perception"/>...

XML Retrieval

D:\dege\xir\web\protein\frame.htm - Microsoft Internet Explorer

File Edit View Favorites Tools Help

← Back → Search Media

Address D:\dege\xir\web\protein\frame.htm Go

Submit Reset

SortBy : 蛋白質大小 -

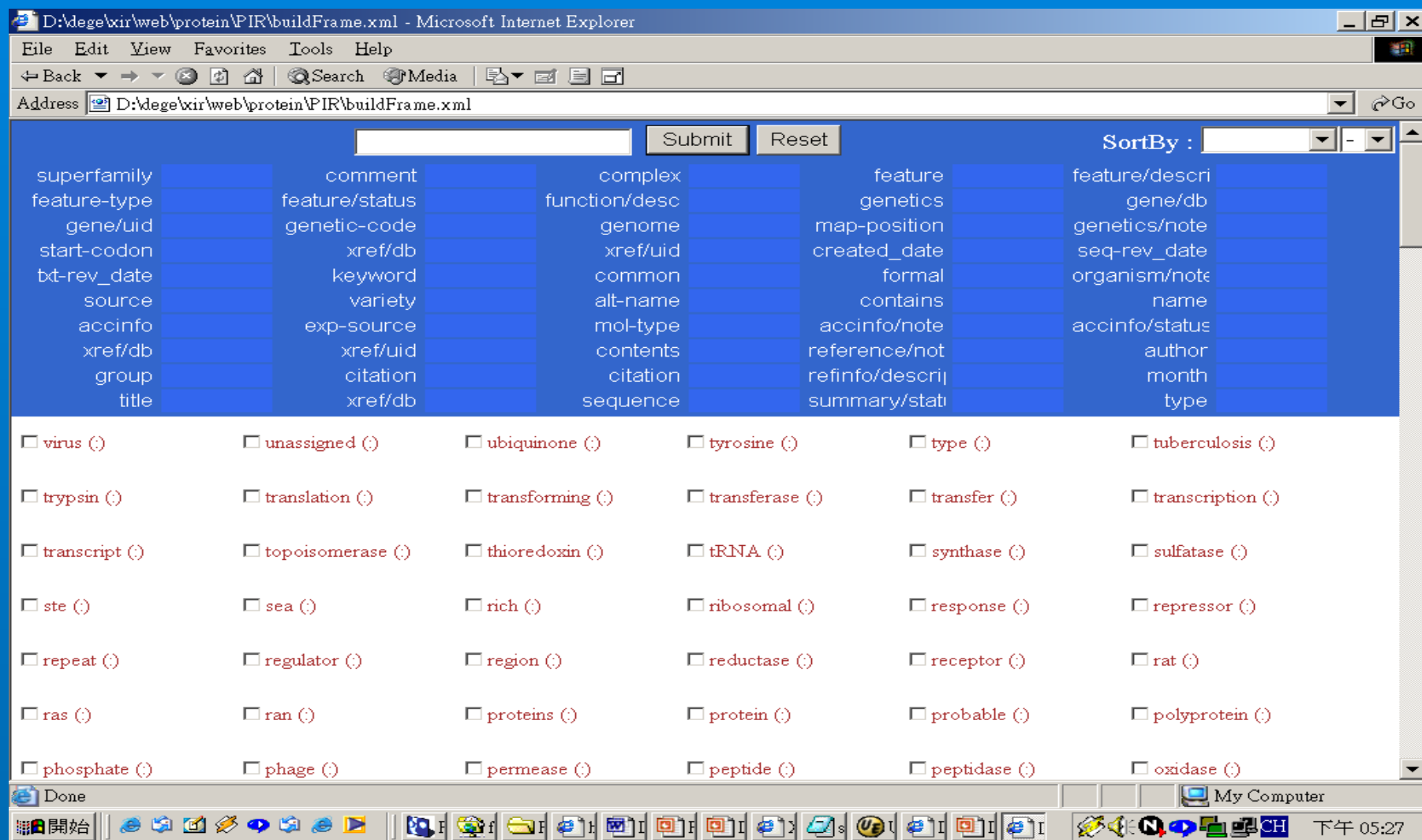
分子種類	分子形狀 平板=Beta	分子來源 人=Human	身體部位	細胞部位
身體功能	細胞功能	相關元素 維生素=vit	特性	蛋白質大小
全部/片斷	參照資料庫	記錄資料庫	論文資料庫	出版日期

<input type="checkbox"/> DNA	<input type="checkbox"/> RNA	<input type="checkbox"/> 酸=Acid	<input type="checkbox"/> 鹼=Base	<input type="checkbox"/> 鹽=Salt	<input type="checkbox"/> 醣=Carbohydrate
<input type="checkbox"/> 酶=Enzyme	<input type="checkbox"/> 核酸=Nucleotides	<input type="checkbox"/> 脂肪=Lipid	<input checked="" type="checkbox"/> 維生素=vitamin	<input type="checkbox"/> 離子=Anion/Cation	<input type="checkbox"/> 碳=Carbon
<input type="checkbox"/> 磷=Phosphatase	<input type="checkbox"/> 能量=ATP	<input type="checkbox"/> 金屬	<input type="checkbox"/> 鈉	<input type="checkbox"/> 鉀=Potassium	<input type="checkbox"/> 鈣=calcium
<input type="checkbox"/> 鐵=iron	<input type="checkbox"/> 銅=copper	<input type="checkbox"/> 鋁=aluminum	<input type="checkbox"/> 鎂=phosphatase	<input type="checkbox"/> 重金屬=heavy_metal	

Done My Computer

開始 下午 05:22

XML Text Mining



XML Text Mining

Analysis

/ProteinEntry/classification/superfamily	virus, unassigned, ubiquinone, tyrosine, type, tuberculosis, trypsin, translation, transforming, transferase, transfer, transcription, transcript, topoisomerase, thioredoxin, tRNA, ...
/ProteinEntry/comment	ste,protein,phosphorylation,phosphorylated,phosphorylase,phospho,phosphate,non, ...
/ProteinEntry/complex	tet,phosphorylase,phospho,mer,homotetramer
/ProteinEntry/feature	TMM,SIG,RRH,MAT,KIN,IMM,HOX,FOX,ERBA,ACP,ABC ...
/ProteinEntry/header/created_date	Sep,Oct,Nov,May,Mar,Jun,Jul,Jan,Feb,Dec,Aug, Apr ...
/ProteinEntry/keywords/keyword	zinc,transmembrane,transferase,transfer,transcription,transcript,tet,ste,ribosome,regulation,reductase,receptor,rat,ras,ran,pyridoxal,proteinase ...
/ProteinEntry/genetics/xrefs/xref/db	SGD,OMIM,MIPS,MIP,GDB
...	...

Domain : Butterflies

Source

➤ Collection：台灣蝴蝶數位博物館

- 暨南大學
- 台灣大學
- 國立自然科學博物館

➤ Information：記錄訊息

- 名稱，科別，種類，宿主植物，地理分布
- 卵，幼蟲，蛹，成蟲
 - 顏色，形狀，特徵，成長期，天敵

XML Data

```
<butterfly>
  <cname> 拉拉山三線蝶 </cname>
  <classification>
    <cfamily> 蛱蝶科 </cfamily>
    <family>Nymphalidae</family><genus>Athyma</genus> <species>fortuna</species>
    <sub_species>kodairai</sub_species></classification>
  <hostplant> 忍冬科 (Caprifoliaceae) 的松田氏紅子仔 </hostplant>
  <honeyplant> 成蝶喜吸食腐熟水果汁液或樹幹流出汁液。 </honeyplant>
  <geographic>
    <taiwan> 分布於台灣中北部地區，海拔 1000-2000 公尺間山區均有分布 </taiwan>
    <global> 中國大陸中部有原名亞種分布。 </global></geographic>
  <adult>
    <feature> 成蟲前翅外觀大致呈現三角形，翅形稍微橫長。後翅卵圓形，外觀接近
      三角形。雌蝶翅型較為寬圓。 </feature>
    <color> 雄蝶前、後翅表底色為黑色，前翅中室內有一枚長形白斑，各翅室中橫線
      部位有一大型白色橢圓斑，前翅端有兩枚小型白斑。後翅有兩條明顯白色橫帶
      紋，前後翅緣皆有不明顯小白紋。雌蟲翅表色澤花紋與雄蟲相似。 </color>
    <size> 本種為中型蝶種，展翅約為 50-60mm。 </size>
    <characteristic> 前翅中室內有一枚長形白斑。 </characteristic>
```

...

Domain Knowledge

<butterfly>

<family slot=" 種類 " path="//butterfly//cfamily//">

<v value=" 弄蝶 " keys="Hesperiidae" /><v value=" 小灰蝶 " keys="ycaenidae" /> ...</family>

<adult slot=" 蝴蝶成蟲 " keys="Adult" path="//butterfly//adult//">

<shape slot=" 蝴蝶的形狀 " keys="Adult:Shape" path="//butterfly//adult//shape//">

<v value=" 類似燕尾 " image="swallowtail.gif"/> <v value=" 翅緣波浪狀 " .../>...</shape>

<color slot=" 蝴蝶的顏色 " keys="Adult:Color" path="//butterfly//adult//color//">

<v value=" 黑色 " keys="Black" />... <v value=" 黑白相間 " keys="Black_White"/>...</color>

<texture slot=" 蝴蝶的特徵 " keys="Adult:Texture" path="//butterfly//adult//color//; //butterfly//adult//texture// ">

<v value=" 沒有花紋 " image="mono.gif" /><v value=" 少數斑點 " image="spot.gif" /> ...</texture></adult>

<pupa slot=" 蝴蝶的蛹 " keys="Pupa" path="//butterfly//pupa//">...

<s slot=" 蛹的特徵 " keys="Pupa:Feature" path="//butterfly//pupa//feature//">

<v value=" 帶蛹 " keys="Laying_Pupa"/><v value=" 垂蛹 " keys="Hanging_Pupa"/> </s></pupa>

<egg slot=" 蝴蝶的卵 " keys="Egg" path="//butterfly//egg//">

<s slot=" 卵的形狀 " keys="Egg:Shape" path="//butterfly//egg//feature//">...

<s slot=" 台灣分布 " keys="Taiwan" path="//butterfly//geographic//taiwan//">

<v value=" 台灣北部 " keys="North_Taiwan+ 北 " /> ...</s>

<s slot=" 全球分布 " path="//butterfly//geographic//global//">

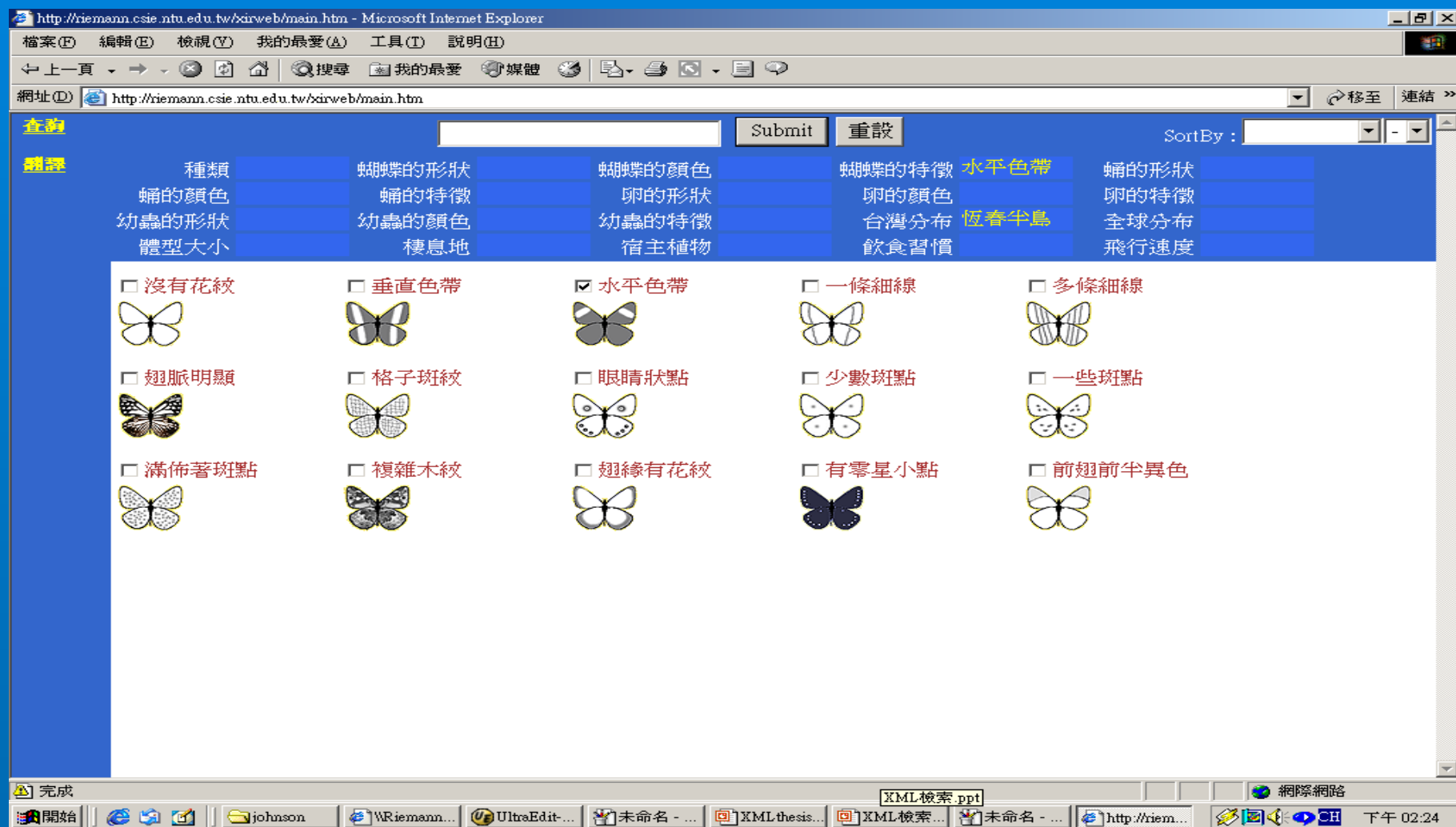
<v value=" 東南亞 " keys="South_Asia " /><v value=" 中國大陸 " keys="China" /></s>

<s slot=" 飲食習慣 " keys="Eat Food" path="//butterfly//adult//behavior//; //butterfly//honeyplant//">

<v value=" 食花蜜 " keys="Nectar" /><v value=" 食腐汁 " keys="Juice " />...</s>

</butterfly>

XML Retrieval



XML Browsing

http://riemann.csie.ntu.edu.tw/~xdrweb/main.htm - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

← 上一頁 → 搜尋 我的最愛 媒體

網址(D) http://riemann.csie.ntu.edu.tw/~xdrweb/main.htm 移至 連結 >>

查詢

引擎

查詢條件=//butterfly//cname;%//butterfly//adult//texture/*%水平色帶,h_Band;//butterfly//geographic//taiwan%恆春半島,HunChan 排序=in_links : decrease

[butterfly/xml/Appias_albina_semperi.xml](#) - 1.0 - [全文](#), [關連](#), [摘要](#),


hasFeature name=h_band rank=56 base=60

cname 尖翅粉蝶

taiwan 族群分布於台灣南部中低海拔山區，恆春半島、台東縣較易見到，台南縣亦有分布，除此之外，蘭嶼、龜山島及東北角亦有少量族群分布。通常可見成蟲活動於平地至海拔 800 公尺間山區。

Abstract

種類：粉蝶	宿主植物：大戟科	台灣分布：台灣北部
台灣分布：台灣東部	台灣分布：台灣南部	台灣分布：恆春半島
台灣分布：蘭嶼	全球分布：東南亞	全球分布：中國大陸
全球分布：新幾內亞	全球分布：澳洲	卵的形狀：梭子形
卵的顏色：淡綠	卵的顏色：光澤	幼蟲的形狀：細長
幼蟲的顏色：翠綠色	幼蟲的顏色：黃綠色	幼蟲的顏色：淡黃色
幼蟲的顏色：白色	蛹的特徵：帶蛹	蛹的顏色：黃綠色
蛹的顏色：褐色	蛹的顏色：灰色	蛹的顏色：白色
體型大小：中型	體型大小：大型	棲息地：平地
棲息地：低海拔山區	飛行速度：飛行迅速	飛行速度：飛行迅速



[butterfly/xml/Salatura_melanippus_edmondii.xml](#) - 1.0 - [全文](#), [關連](#), [摘要](#),

hasFeature name=h_band rank=50 base=60


hasFeature name=h_band rank=51 base=60

cname 黑脈白斑蝶

taiwan 主要產於蘭嶼，台灣恆春半島海拔 500 公尺以下低山地區亦偶爾可見。

Abstract

種類：斑蝶	台灣分布：恆春半島	台灣分布：蘭嶼
全球分布：東南亞	卵的形狀：砲彈形	卵的顏色：乳白
卵的顏色：淡黃	卵的顏色：光澤	幼蟲的形狀：細長
幼蟲的顏色：白色	幼蟲的顏色：黑色	蛹的特徵：垂蛹
蛹的顏色：黃綠色	蛹的顏色：白色	體型大小：中型

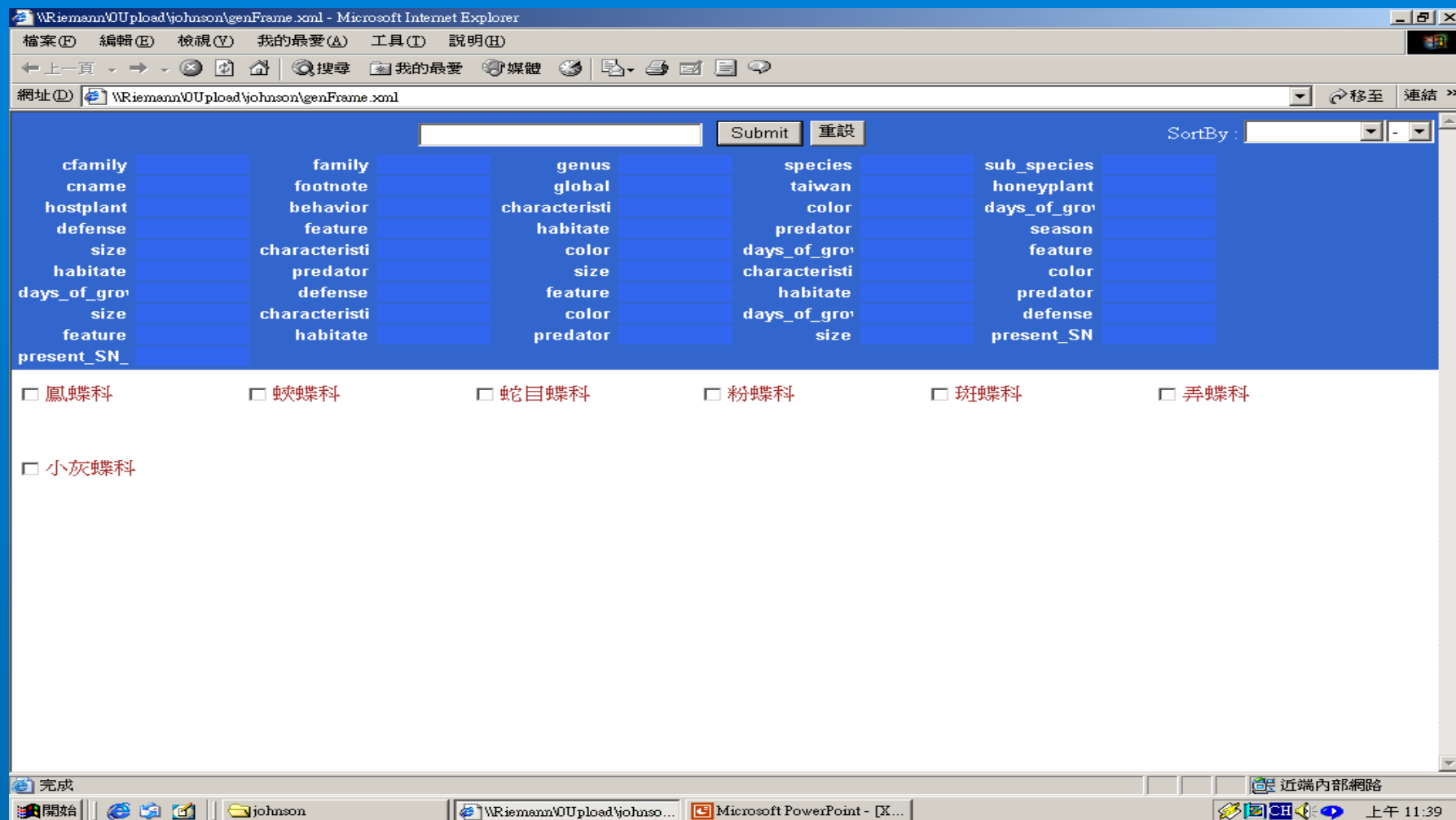


完成

開始 | johnson | \Riemann... | UltraEdit... | 未命名 - ... | XMLthesis... | XML檢索... | 未命名 - ... | http://riem... | 國際網路

下午 02:22

XML Text Mining



XML Text Mining

Analysis

\butterfly\classification\cfamily	鳳蝶科，蛱蝶科，蛇目蝶科，粉蝶科，斑蝶科，弄蝶科，小灰蝶科
\butterfly\classification\family	Satyridae, Pieridae, Papilionidae, Papilio, Nymphalidae, Lycaenidae, Hesperidae, Danaidae
\butterfly\footnote	高冷蔬菜區，非常，開發，開墾，長達，近年來，種經，種族群，破壞，生活史，...
\butterfly\geographic\global	馬來半島，非洲，錫金 西部，蘇門達臘，蘇門答臘 蘇門，群島，美洲，緬甸北部，緬甸，琉球群島，琉球，爪哇，...
\butterfly\honeyplant\	馬櫻丹，馬利筋，馬利，金露花，野花，豐草，菊科野花，菊科，菊科，花蜜，腐熟，繁星花，繁星，紫花霍香薊，...
\butterfly\life_stage\adult\predator	鳥類，青蛙，螳螂，蜻蜓，蜥蜴，蜘蛛，捕食性天敵，捕食，性天敵，天敵，
\butterfly\life_stage\egg\characteristic	表面平滑，表面，縱脊，細微，精孔，突起，條縱脊，條細微縱脊，條細微縱脊，明顯縱脊，明顯，數條，平滑，刻點，...
\butterfly\life_stage\egg\feature	高饅頭形，饅頭，頂點，頂部微凸，頂部，角形，表面，著生，菱形，花紋，縱脊，細長刺毛，細長，細小突起，

Contributions

- 降低人與機器之間在 XML 上的語意落差 .
 - 使人更容易使用 XML 查詢語言
 - 使機器更了解 XML 文件的語義
 - 整合 XML 檢索的語義，包含 tag 整合與詞彙整合 .
 - 使人更容易瀏覽 XML 文件
 - 使人更容易建立 XML 文件的 ontology

Contribution 1

➤ 使人更容易使用 XML 查詢語言

- 簡單的 Slot 介面可以下出精確複雜的 Query ；
 - 降低了 XML 查詢語言 上的語義落差，可以提昇檢索的 Precision
- Open Problem : [SIGIR 2000]
 - XML 的查詢語言很強，可提升 precision, 但使用者不會用 ？

Contribution 2

➤ 使機器更了解 XML 文件的語義

- 方法 : Mapping
 - Slot-Filling 準確的將 XML 文件映射到 ontology 中
- 優點 : 利用 tag 可以降低 Slot-Filling 的困難度
 - 大而分散的範圍 → 小而集中的範圍
 - 任意的文件 d → 單一領域的 p

Contribution 3

➤ Mapping → 使人更容易瀏覽 XML 文件

- 利用 Slot-Filling 機制，可對檢索結果進行摘要，以便瀏覽；
- Open Question [SIGIR 2000]
 - 如何組織檢索結果？(Doc ? Tree ? Graph ?)

(s , v)

成蟲的顏色，淺棕色
成蟲的形狀，翅緣破
裂

成蟲的花紋，色帶

Contribution 4

- 使人更容易建立 XML 文件的 ontology
 - 利用 XML 的 tag，可統計出 path 與 value 之間的關係；
 - Mining 的結果可以組合成樹狀結構；
 - 很容易能 mapping 到 frame (Slot-Tree) 中；

Comparison : Approaches

	查詢語言	文件
人	<p>easy $\xrightarrow{\text{寫}}$ hard</p> <p>NL DB Logic XML</p> <p>←</p>	<p>easy $\xrightarrow{\text{寫}}$ hard</p> <p>NL DB Logic XML</p>
機器	<p>easy $\xrightarrow{\text{懂}}$ hard</p> <p>NL DB Logic XML</p>	<p>easy $\xrightarrow{\text{懂}}$ hard</p> <p>NL DB Logic XML</p> <p>←</p>

Comparison : XML systems

- XML-GL
 - 適合用圖形的方式描述的一種 XML 查詢語言。
- XYZfind
 - 兩層式的檢索方法，先檢索 tag, 再檢索 XML 文件。
- Lore+Data Guider
 - 結合樹狀式的圖形介面與 OODB 的 XML 檢索系統。
- RDF
 - 採取物件式表達法的 XML 文件規格。
- DAML
 - 加入邏輯表達法的 XML 文件規格。

Comparison : XML systems

	XML 查詢語言	XML 文件
人	<p>easy $\xrightarrow{\text{寫}}$ hard</p> <p>X-GL DAML</p> <p>Lore</p> <p>XYZ</p> <p>Slot $\xleftarrow{\text{GUI}}$ XML</p>	<p>easy $\xrightarrow{\text{寫}}$ hard</p> <p>X-GL</p> <p>Lore DAML</p> <p>XYZ RDF</p> <p>Slot</p> <p>XML</p>
機器	<p>easy $\xrightarrow{\text{懂}}$ hard</p> <p>XYZ DAML</p> <p>X-GL RDF</p> <p>Lore</p> <p>Slot</p> <p>XML</p>	<p>easy $\xrightarrow{\text{懂}}$ hard</p> <p>DAML</p> <p>X-GL</p> <p>RDF Lore</p> <p>XYZ</p> <p>Slot $\xleftarrow{\hspace{1cm}}$ XML</p>

Discussion

➤ 貢獻

- 貢獻：降低人與機器之間在 XML 上的語意落差。
 - A. 降低人與機器之間在 XML queries 上的語意落差。
 - 利用 slot-tree 建立 query interface, 讓人容易下 XML queries
 - B. 降低人與機器之間在 XML documents 上的語意落差。
 - 利用 slot-filling, 讓機器容易了解 XML documents.
- 限制：必須將檢索範圍限定在單一領域上。

Future Work

➤ 整合數個同領域的文件集合

- 方法：Slot Vector Space Model (SVSM)

- 整合相同領域的異質性的 XML 文件

- 例如：蝴蝶 + 昆蟲

- `<s slot="蝴蝶" path="//butterfly; //insect[@type='butterfly']">`

蝴蝶

昆蟲中的蝴蝶

- 整合相同意義的詞彙，有利進行多國語言檢索

- 例如：顏色的中英文檢索

- `<v value="黑色" keys="黑色, Black"/>`

- Open Question [SIGIR 2000]

- 1. 如何處理異質性的 XML 文件？
 - 2. 如何進行多國語言檢索？

Future Work

➤ 待研究的問題

- 關於本方法的量化實驗研究
 - 對本方法的結果進行 Recall / precision 衡量 .
- 數個相關領域的 XML 檢索如何進行 ?
 - 可否利用各種型態的 Link ?
 - 可否利用 Ontology , Taxonomy, 或 Thesaurus ?
- 不限定領域的 XML 檢索如何進行 ?
 - Two level searching – XYZfind 。
 - 加入 Domain Search 的步驟就能有效進行不限領域的 XML 檢索嗎 ?

Thank you .