Eighth International Conference on Web Engineering

WIKE: A Web Information/Knowledge Extraction System for Web Service Generation

Hao Han and Takehiro Tokuda
Department of Computer Science, Tokyo Institute of Technology
Meguro, Tokyo 152-8552, Japan
{han, tokuda}@tt.cs.titech.ac.jp

Abstract

We have a tremendous amount of information/knowledge available on the Web today. We often have a situation in which we need to collect partial contents of a whole page from one or a number of Web sites. Examples are collection of capital city names and population data from country profile sites or collection of company names and their industrial fields from finance sites. We present WIKE, a system for partial information extraction from Web pages without programming. We also give its applications to Web service generation.

Keywords: *information extraction, Web service*

1 Introduction

Today, we have a tremendous amount of information/knowledge available on the Web. We often have a situation, in which we need to collect partial contents of a whole page from one or a number of Web sites. For example, in the BBC Country Profiles [1], there exists a collection of 200 or more country/region information such as capital city, population and latest leader's information. One Web page only has the information of a single country/region. If we would like to collect the information of all the countries/regions, the use of Web browsers would be a time-consuming tedious task. Similar tasks may be collection of disease names and corresponding parts of human body from health/medicine sites or collection of company names and corresponding industrial fields from finance sites.

We present a demo of our system, called **WIKE**, for **W**eb **I**nformation/**K**nowledge **E**xtraction from the general Web applications without programming. The users can pick up the target Web pages, select the desired parts, and extract them by using WIKE. The extraction result is structured data in XML format.

The purpose of this demo is to extract Web information/knowledge from Web sites and generate Web services quickly. Web services are becoming a standard method of sharing information/knowledge, and integrations of Web services are very useful. Unfortunately, many Web sites do not provide the Web services. Web applications are still the main methods for Web information/knowledge sharing.

In this demo, we show how to select the target information from the Web applications, how to compose the structured data, how to configure the Web services, and how to complete all these processes quickly via WIKE.

2 Web Information/Knowledge Extraction

We give an overview of WIKE as shown in Fig. 1. WIKE is based on XML tree approach. Firstly, we get a typical Web page from the target Web application, and define the names and data types for the selected parts to generate an extraction pattern. Then, WIKE uses the extraction pattern to extract the partial information from the ordinary Web pages, and returns a resulting table to show the structured extraction result.

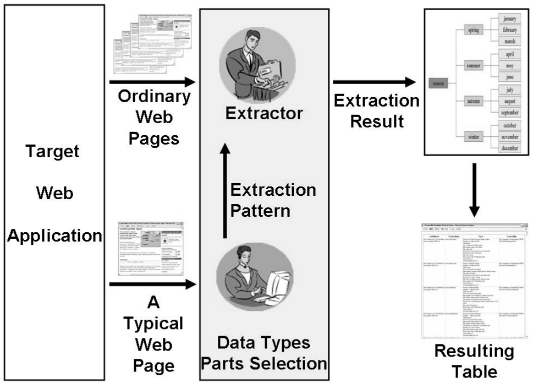


Figure 1. Outline of WIKE

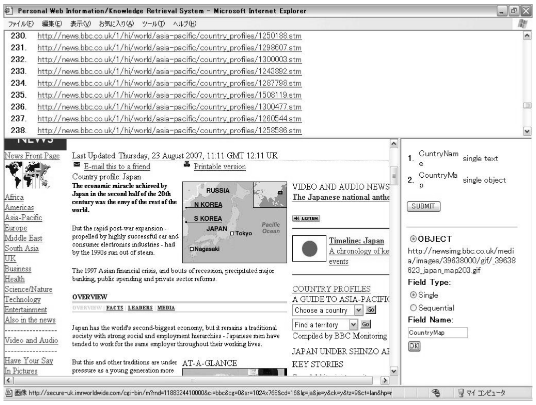
In this section, we demonstrate how to use WIKE to ex-

978-0-7695-3261-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ICWE.2008.30

tract the country information about the listed 200 or more countries/regions from BBC Country Profiles site. The country information includes country name, country map, country facts(capital city, population, and etc.), leader's photo and leader's information.

2.1 Parts Selection and Data Types Definition

Parts selection represents *"Which parts of page are needed?"*, and data type represents *"What kind of information is needed?"*. The data type includes two kinds of information: property and structure. Property is text, object, or link. Text is the character string in Web pages such as an article. Object is one instance of the photo, video and other multimedia file. Link is a reference in a hypertext document to another document or other resource. Structure is single occurrence or continuous occurrence. A single occurrence is a node without similar sibling nodes such as the title of an article. A continuous occurrence is a list of nodes with similar paths such as result list in a search result page. There are six kinds of data types: single text, continuous text, single object, continuous object, single link and continuous link.


Personal Web Information/Knowledge Retrieval System - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

230. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1250188.stm
231. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1298607.stm
232. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1300003.stm
233. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1243892.stm
234. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1287798.stm
235. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1508119.stm
236. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1300477.stm
237. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1260544.stm
238. http://news.bbc.co.uk/1/hi/world/asia-pacific/country_profiles/1258586.stm

Last Updated: Thursday, 23 August 2007, 11:11 GMT 12:11 UK

E-mail this to a friend Printable version

Country profile: Japan

The economic miracle achieved by Japan in the second half of the 20th century was the envy of the rest of the world.

But the rapid post-war expansion - propelled by highly successful car and consumer electronics industries - had by the 1990s run out of steam.

The 1997 Asian financial crisis, and bouts of recession, precipitated major banking, public spending and private sector reforms.

OVERVIEW

OVERVIEW | FACTS | LEADERS | MEDIA

Japan has the world's second-biggest economy, but it remains a traditional society with strong social and employment hierarchies - Japanese men have tended to work for the same employer throughout their working lives.

But this and other traditions are under pressure as a young generation more

1. CuntryName single text
2. CountryMap single object

SUBMIT

◉ OBJECT
http://newsimg.bbc.co.uk/media/images/39638000/gif/_39638623_japan_map203.gif

Field Type:
◉ Single
○ Sequential

Field Name: CountryMap

OK

画像 http://secure-uk.imrworldwide.com/cgi-bin/m?rnd=1188324410000&ci=bbc&cg=0&sr=1024x768&cd=16&lg=ja&je=y&ck=y&tz=9&ct=lan&hp=r

Figure 2. Parts Selection and Data Types Definition

Fig. 2 shows the parts selection and data types definition. We enter the URL of top page of BBC Country Profiles. WIKE lists the URLs of country pages. We select a typical country page. We define the names and data types for each selected part: country name part is *CountryName* of single text type, country map part is *CountryMap* of single object type, country facts part is *Facts* of continuous text type, leader's photo part is *LeaderPicture* of single object type, and leader's information part is *LeaderInfo* of single text type. Each selected part is represented by a node, and each node can be represented by its path from the root. We use the following form to save the path of a selected part:

$body : 0 : ID/N_1 : O_1 : ID_1/N_2 : O_2 : ID_2/N_3 : O_3 : ID_3/....../N_{n-1} : O_{n-1} : ID_{n-1}/N_n : O_n : ID_n$

where, N_n is the name of the n-th node, O_n is the order of the n-th node among the sibling nodes, ID_n is the ID value of the n-th node, and N_{n-1} is the parent node of N_n.

2.2 Information Extraction Method

WIKE uses the defined data types and paths to realize the partial information extraction from the ordinary country pages.

In the tree structure of HTML document, each path represents a node. WIKE extracts the nodes according to the corresponding paths. During the node extraction, if a node can not be found by a path, a similar path would be used to try extracting the node. We give a definition of similar path.

Similar Path: Two paths are similar to each other, if these two paths have the same forms ignoring the difference of orders of nodes among sibling nodes, and the difference of orders is within a given deviation range. The form of path is as follows:

$body : 0 : ID/N_1 : (O_1 - h \sim O_1 + h) : ID_1/N_2 : (O_2 - h \sim O_2 + h) : ID_2/....../N_{n-1} : (O_{n-1} - h \sim O_{n-1} + h) : ID_{n-1}/N_n : (O_n - h \sim O_n + h) : ID_n$

where, h is the deviation value. The ID values are used to choose the most appropriate paths with the minimum deviation value from the deviation range.

If the data type of a part is continuous occurrence, WIKE extracts the list of nodes, and each node of the extracted node list represents a part of continuous parts.

WIKE extracts the partial information from the extracted nodes in text format excluding the HTML tags as described in Table 1. For example, the extracted information of a photo is the value of attribute *src* of node <img>, and the extracted information of a link is the value of attribute *href* of node <a>.

Table 1. Partial Information Extraction

Data Type	Partial Information
single text	node value of single leaf node
single object	attribute value of single node
single link	embedded link value of single node
continuous text	leaf node values of list of nodes
continuous object	attribute values of list of nodes
continuous link	link values of the list of nodes

2.3 Extraction Pattern

In BBC Country Profiles, most of the country pages are similar to each other. The paths of the similar parts of these country pages are similar to each other, too. However, there are still some country pages use the different layout. The defined paths and data types from only one country page can not satisfy the extraction from all the country pages.

WIKE tries extracting the partial information from all the country pages, and lists the country pages from which WIKE can not extract the correct partial information by using the defined paths and data types. The users can define the new paths and data types for one listed country page, and try the extraction again. The users continue this process until the partial information of all the country pages is extracted correctly.

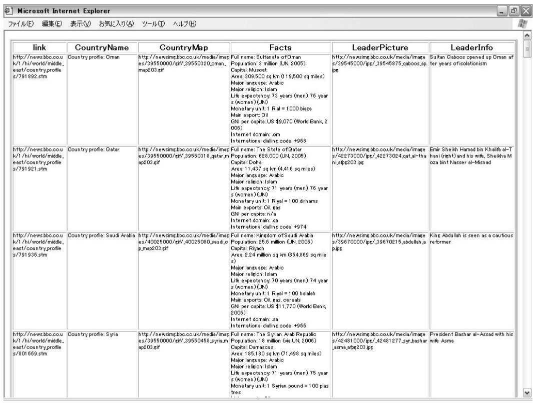
We selected the parts and defined the data types from three typical country pages. All the defined paths and data types comprise the final extraction pattern of BBC Country Profiles.

2.4 Extraction Result

WIKE uses the extraction pattern to extract the partial information from the country pages. The extraction result is an XML-based document. WIKE returns the users a resulting table to show the extracted information.

A resulting table comprises two parts: field name and field value. Field name is the defined name of selected part and field value is the extracted value. In a resulting table, each column is a field and each row shows the extracted results of each country page, except that the first row displays the field names.

Fig. 3 is an extraction resulting table of BBC Country Profiles. The first column *Link* shows the URLs of country pages. The second column *CountryName* shows the country names. The third column *CountryMap* shows the URIs of country map pictures. The fourth column *Facts* shows the country facts. The fifth column *LeaderPicture* shows the URIs of leaders' photos, and the sixth column *LeaderInfo* shows the leaders' information.


Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

link	CountryName	CountryMap	Facts	LeaderPicture	LeaderInfo
http://news.bbc.co.uk/1/hi/world/middle_east/country_profiles/791892.stm	Country profile: Oman	http://newsimg.bbc.co.uk/media/images/39550000/gif/_39550320_oman_map203.gif	Full name: Sultanate of Oman Population: 3 million (UN, 2005) Capital: Muscat Area: 309,500 sq km (119,500 sq miles) Major language: Arabic Major religion: Islam Life expectancy: 73 years (men), 76 years (women) (UN) Monetary unit: 1 Rial = 1000 biaza Main export: Oil GNI per capita: US $9,070 (World Bank, 2006) Internet domain: .om International dialling code: +968	http://newsimg.bbc.co.uk/media/images/39545000/jpg/_39545875_qaboos_ap.jpg	Sultan Qaboos opened up Oman after years of isolationism
http://news.bbc.co.uk/1/hi/world/middle_east/country_profiles/791921.stm	Country profile: Qatar	http://newsimg.bbc.co.uk/media/images/39550000/gif/_39550318_qatar_map203.gif	Full name: The State of Qatar Population: 628,000 (UN, 2005) Capital: Doha Area: 11,437 sq km (4,416 sq miles) Major language: Arabic Major religion: Islam Life expectancy: 71 years (men), 76 years (women) (UN) Monetary unit: 1 Riyal = 100 dirhams Main exports: Oil, gas GNI per capita: n/a Internet domain: .qa International dialling code: +974	http://newsimg.bbc.co.uk/media/images/42273000/jpg/_42273024_qat_al-thani_afpg203.jpg	Emir Sheikh Hamad bin Khalifa al-Thani (right) and his wife, Sheikha Moza bint Nasser al-Misnad
http://news.bbc.co.uk/1/hi/world/middle_east/country_profiles/791936.stm	Country profile: Saudi Arabia	http://newsimg.bbc.co.uk/media/images/40025000/gif/_40025080_saudi_cp_map203.gif	Full name: Kingdom of Saudi Arabia Population: 25.6 million (UN, 2005) Capital: Riyadh Area: 2.24 million sq km (864,869 sq miles) Major language: Arabic Major religion: Islam Life expectancy: 70 years (men), 74 years (women) (UN) Monetary unit: 1 Riyal = 100 halalah Main exports: Oil, gas, cereals GNI per capita: US $11,770 (World Bank, 2006) Internet domain: .sa International dialling code: +966	http://newsimg.bbc.co.uk/media/images/39670000/jpg/_39670215_abdullah_ap.jpg	King Abdullah is seen as a cautious reformer
http://news.bbc.co.uk/1/hi/world/middle_east/country_profiles/801669.stm	Country profile: Syria	http://newsimg.bbc.co.uk/media/images/39550000/gif/_39550458_syria_map203.gif	Full name: The Syrian Arab Republic Population: 18 million (via UN, 2006) Capital: Damascus Area: 185,180 sq km (71,498 sq miles) Major language: Arabic Major religion: Islam Life expectancy: 71 years (men), 75 years (women) (UN) Monetary unit: 1 Syrian pound = 100 piastres	http://newsimg.bbc.co.uk/media/images/42481000/jpg/_42481277_syr_bashar_asma_afpg203.jpg	President Bashar al-Assad with his wife Asma

Figure 3. A Resulting Extracted Table of BBC Country Profiles

3 Web Service Generation

A Web service is a system designed to support interoperable machine to machine interaction over a network, and executed on a remote system hosting the requested service. Usually, a real Web service can access to the server-side database and search for the data in tables to create the response results. A virtual Web service can be generated from the Web application by using the extraction function of WIKE. As shown in Fig. 4, in the proxy server, the extraction patterns are used to extract the partial information, and the resulting tables are generated. By a designed interface, the users send the requests to proxy server and get the responses from the resulting tables.

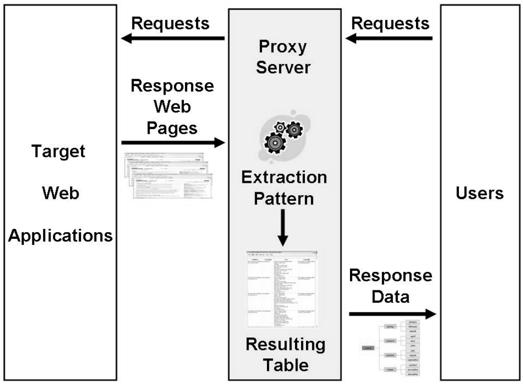


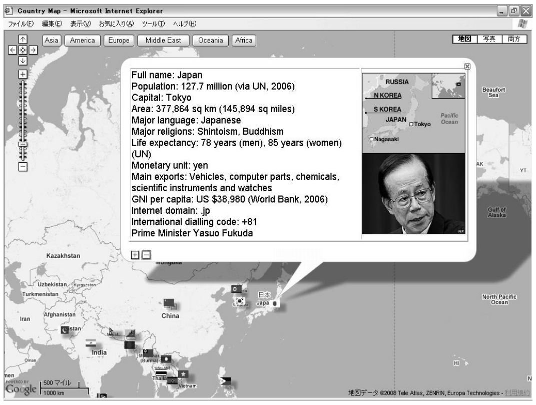
Figure 4. The Web Service Generated by Using Extraction Function of WIKE

In the proxy server, a standard RESTful Web service interfaces is used between the users and the generated Web service. Through the interface, the Web service gets the request from the user, searches for the desired information in the resulting table, and returns the corresponding values to user in XML format. If the structure of desired field is continuous occurrence, the corresponding response data is the value list. The generated Web service works like a real Web service, and can be integrated with other Web services to construct a Mashup system as shown in Fig. 5.

As well as the Web pages with static URLs, the dynamically generated Web pages can also be used to generate the Web services. Fig. 6 shows a response of news search from the generated CNN News [3] search Web service.

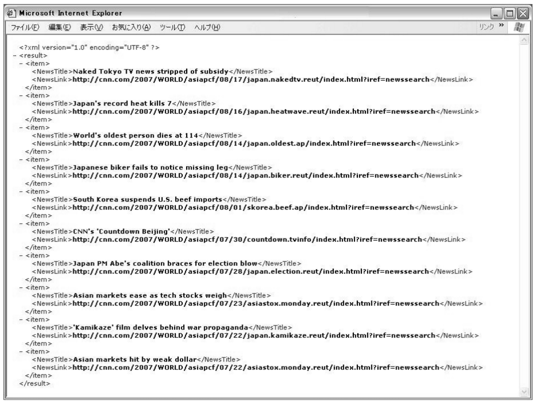
4 Related Work and Evaluation

A number of approaches have been proposed to analyze the Web page structures of Web applications with the purpose of Web information/knowledge extraction and Web service generation.



Full name: Japan
Population: 127.7 million (via UN, 2006)
Capital: Tokyo
Area: 377,864 sq km (145,894 sq miles)
Major language: Japanese
Major religions: Shintoism, Buddhism
Life expectancy: 78 years (men), 85 years (women) (UN)
Monetary unit: yen
Main exports: Vehicles, computer parts, chemicals, scientific instruments and watches
GNI per capita: US $38,980 (World Bank, 2006)
Internet domain: .jp
International dialling code: +81
Prime Minister Yasuo Fukuda

Figure 5. An Integration of Generated Web Service with Google Maps



Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
リンク

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <result>
  - <item>
      <NewsTitle>Naked Tokyo TV news stripped of subsidy</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/08/17/japan.nakedtv.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>Japan's record heat kills 7</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/08/16/japan.heatwave.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>World's oldest person dies at 114</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/08/14/japan.oldest.ap/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>Japanese biker fails to notice missing leg</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/08/14/japan.biker.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>South Korea suspends U.S. beef imports</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/08/01/skorea.beef.ap/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>CNN's 'Countdown Beijing'</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/07/30/countdown.tvinfo/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>Japan PM Abe's coalition braces for election blow</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/07/28/japan.election.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>Asian markets ease as tech stocks weigh</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/07/23/asiastox.monday.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>'Kamikaze' film delves behind war propaganda</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/07/22/japan.kamikaze.reut/index.html?iref=newssearch</NewsLink>
    </item>
  - <item>
      <NewsTitle>Asian markets hit by weak dollar</NewsTitle>
      <NewsLink>http://cnn.com/2007/WORLD/asiapcf/07/22/asiastox.monday.reut/index.html?iref=newssearch</NewsLink>
    </item>
  </result>
```

Figure 6. The Search Results from Generated Web Service of CNN News Search

ANDES [7] and PSO [9] extract the information from Web pages by paths. Crunch [4] is an HTML tag filter to retrieve the contents from the Web pages. These systems need the users to find out the paths or create the tag filters from the HTML documents by hands. IEPAD [2] and Web Service Gateway [5] are difficult for the users to develop the personalized extraction. HTML2RSS [8] is limited to extract the special data structures such as time-series items from blog, BBS, chats and mailing lists. GridXSLT [6] needs the programming of users.

Compared to the existing work, WIKE realizes the extraction of almost all kinds of partial information such as text, object and link. It provides a visual editor for easy parts selection and data types definition. The extraction pattern creation or partial information extraction does not need programming. The simialr paths are used in node extraction, and the users do not need to do parts selection and data types definition for each Web page.

However, this demo still depends on some manual work during the Web service interface configuration, and is not robust enough when the Web applications change the layout of Web pages. If the layout of Web page is changed totally and the nodes can not be found by the similar paths within the deviation range, the extraction would fail partially or completely, and the users have to create a new extraction pattern.

5 Conclusion and Future Work

In this demo, we have presented WIKE, a Web information/knowledge extraction system for Web service generation. WIKE provides a visual editor for easy extraction. The users can select the parts, define the data types, and create the extraction pattern without programming. The typed data is extracted and the extraction result is the XML-based document. The Web service is generated based on the extraction function of WIKE. Through the designed interface, the generated Web service can be integrated with other Web services to construct a Mashup system.

As future work, we will enhance the extraction function of WIKE in order to extract the exact parts from the updated Web pages. Moreover, we will modify WIKE to decrease the manual work during the Web service interface configuration.

References

[1] Country Profiles. http://news.bbc.co.uk/2/hi/country_profiles/.

[2] C.-H. Chang and S.-C. Lui. IEPAD: Web information extraction based on pattern discovery. In *The Proceedings of the 10th International Conference on World Wide Web*, 2001.

[3] CNN. http://www.cnn.com.

[4] S. Gupta and G. Kaiser. Extracting content from accessible Web pages. In *The Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility*, 2005.

[5] H. P. Huy, T. Kawamura, and T. Hasegawa. How to make Web sites talk together - Web service solution. In *The Proceedings of the 14th International Conference on World Wide Web*, 2005.

[6] P. M. Kelly, P. D. Coddington, and A. L. Wendelborn. A simplified approach to Web service development. In *Proceedings of the 2006 Australasian workshops on Grid computing and e-research*, 2006.

[7] J. Myllymaki. Effective Web data extraction with standard XML technologies. In *The Proceedings of the 10th International Conference on World Wide Web*, 2001.

[8] T. Nanno and M. Okumura. HTML2RSS: Automatic generation of RSS feed based on structure analysis of HTML document. In *The Proceedings of the 15th International Conference on World Wide Web*, 2006.

[9] T. Suzuki and T. Tokuda. Path set operations for clipping of parts of Web pages and information extraction from Web pages. In *The 15th International Conference on Software Engineering and Knowledge Engineering*, 2003.