



A Search Engine for the Chilean & Korean Webs

Felipe Lalanne

Project Objectives

- Create a Cross Language (Korean/Spanish) search engine
- In other words. Be able to find information in both Chilean and Korean webs, regardless of the language of the query

What do we need?

- To build a search engine we need 3 main steps.
 - Web crawling: I.E download text from the pages of multiple web sites in a short time.
 - Text indexing: I.E to order the text in such a way that it will be easy to find a page related to a user query.
 - Text searching: I.E to find the query in the index and retrieve the matching web sites.

Cross Language Feature

- In the case of the current project, we need to add a fourth phase before the text search.
 - Query translation: Detect the language of the query and translate it.
 - Use both the plain and the translated query as input to the text search.

A First Approach

- As a first attempt, we used standard tools, to accomplish the required 3 steps of the process:
 - WIRE (Web Information Retrieval Environment) developed here in Chile by Carlos Castillo, was used as a web crawler.
 - Swish-e an Open Source tool for indexing and searching.
 - A web front-end to Swish-e was also created to perform the queries through a web page.

Search the Chilean and Korean Web

Search: Whole Web Government Sites

	Whole Web		Government Sites	
	Sites	Pages	Sites	Pages
Chile	53,528	3,092,954	1,413	150,287
Korea	54,895	8,889,419	4,386	563,404

This prototype has been funded by the Chile-Korea IT Center in cooperation with KCU, Daejeon, Korea & Center for Web Research, DCC, Santiago, Chile

Search the Chilean and Korean Web

Search: Whole Web Government Sites

Chilean results



Results 1 to 4 of 4

19 <http://www.netplan.cl/itcc/programa18nov04.html>

1st International IT Conference in Latin America and the Caribbean
OBJETIVOS DEL EVENTO Compartir ideas y experiencias en Estrategias
nacionales de TI de Korea y paÃses

13 <http://www.netplan.cl/itcc/index.html>

InvitaciÃ³n ITCC Programa del Evento Acceda al programa oficial del
evento aquÃ . Los invitados a la Conferencia podrÃn participar (sin costo)
de un AI

11 <http://www.decom-uv.cl/html/modules.php?name=News&file=article&sid=t>

e1c Departamento de Computaci3n - Invitacion a Taller Web, con becas par
alumnos - Enero 2005 Usuarios registrados. [login] 495

8 <http://www.dcc.uchile.cl/~churtado/>

Carlos Hurtado Larrain Carlos Hurtado Larrain (English) e-mail:
churtado@dcc.uchile.cl fono: +56 2 6784363 fax: +56 2 6895531 Inicio (

Korean results



Results 1 to 2 of 2

3 <http://xxx.snu.ac.kr/list/cs/0405> 57.5 Kb

f8f Computer Science authors/titles
May 2004 Computer Science
Authors and titles for May 2004
cs.CY/0405001 [abs , pdf] : Title:
Toward a New Policy for Scientific
and Technical

3 <http://si.gist.ac.kr/publication.htm> 19.4 Kb

Kwangju Institute of Science and
Technology KOREAN KOREAN 2 1
2004-09-14T02:38:00Z
2004-09-14T02:38:00Z 1 2555
14564 KOREA 121 34 17085
11.5606 106 Clean Clean

Results

- The approach worked to make make simple queries, in English or in Spanish. But didn't get any results when trying to search text in Korean
- The problem is character encoding

Character Encoding

- *A character encoding consists of a code that pairs a set of characters (representations of graphemes or grapheme-like units, such as might appear in an alphabet or syllabary for the communication of a natural language) with a set of something else, such as numbers or electrical pulses
(definition retrieved from Wikipedia.org)*

Character encoding

- For instance, every character in the English language, including some control characters like line breaks, can be represented by a single byte, 8 bits, which result in 256 possible combinations
- Spanish and Korean alphabets, can also be represented by a single byte but the representations are different (they use different encodings)

International Standard

- Unicode has the goal of providing the means to encode the text of every document people may want to store in a computer
- UTF-8 is a variable-length character encoding for Unicode, it can represent the alphabets of many of the world's languages

Back to the project

- As said earlier, the main issue with the first approach was character encoding
 - The chosen indexing engine doesn't have support for euc-kr encoding (Korean alphabet)
- However, there are indexing engines that support UTF-8. Therefore it is necessary to convert the text from both the Chilean (mostly iso-8859-1) and Korean (mostly euc-kr) pages to UTF-8, so they can be indexed

Encoding Conversion

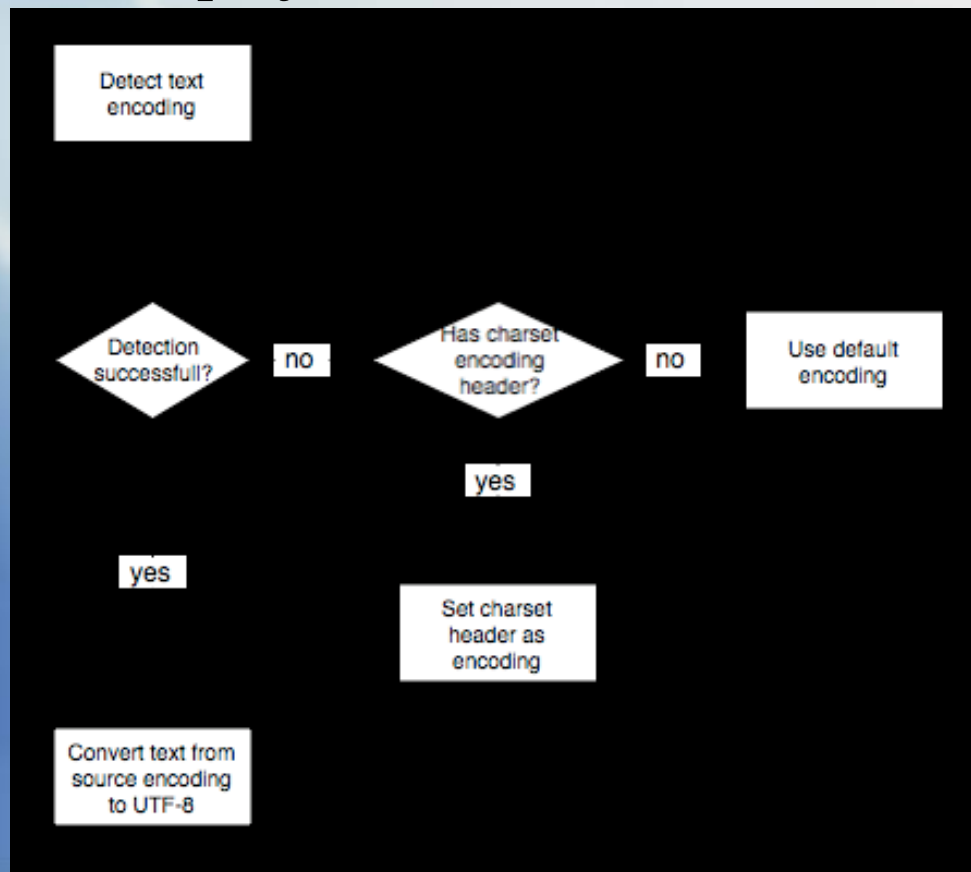
- There are several libraries that allow to convert text between different encodings
- The common issue with them is that the encoding of the origin text needs to be known beforehand. Since in the crawled sites can exist web pages with diferent encodings this is not an easy task

Source encoding detection

- There are three different possibilities:
 - Assume a default encoding.
 - Read the value of the charset from the header returned by the web server.
 - Parse the text and detect automatically the encoding.

Source encoding detection

- A combination of the three was finally used as shown in the following diagram. To detect the encoding a library from the Mozilla project source code was used.



The Crawler

- The WIRE crawler performs four main cyclic tasks.
 - Seeding: Receives URL and adds documents for them to the repository.
 - Managing: Creates batches of documents to harvest.
 - Harvesting: Download the documents from the web.
 - Gathering: Extract urls from the downloaded documents and remove unnecessary markup.

Integration with crawler

- During the harvesting phase the encoding from the header is added to the document information, if there is no such header, the default encoding is used.
- During the gathering phase, after the parsing of the text, the encoding is detected, the conversion is made and the final UTF-8 text is stored.
- This way the text output for the indexing engine will already be Unicode. Therefore there is no need to make changes to this tool.

Text Indexing

- This process consists of creating a relation between the text and the web site.
- The idea is quite similar to a dictionary or a book index which points out the page to look if you want to find information on a certain topic or by a certain keyword.
- There are many open source tools to perform this task. The one chosen for its Unicode support was Lucene.

What's next?

- Create a web front-end to search the text indexed with Lucene.
- Integrate the search engine with the query translation feature.

Some Conclusions

- The differences between Chilean and Korean alphabets make text searching a complex task.
- A common alphabet representation is needed. Unicode can help solve this issues.
- It is necessary to create a consciousness of improving the internet between web developers and hosting providers, to provide such simple information as the web page codification can improve the way text is searched (and found) in the web.