A Framework for Capturing Domain Knowledge via the Web

Chao Wang, Faculty of Information Technology [HREF1], University of Technology, Sydney [HREF2], PO Box 123, Broadway, NSW 2007, Australia. cwang[AT]it.uts.edu.au

Jie Lu, Faculty of Information Technology [HREF1], University of Technology, Sydney [HREF2], PO Box 123, Broadway, NSW 2007, Australia. jielu[AT]it.uts.edu.au

Guangquan Zhang, Faculty of Information Technology [HREF1], University of Technology, Sydney [HREF2], PO Box 123, Broadway, NSW 2007, Australia. zhangg[AT]it.uts.edu.au

Abstract

Domain knowledge can be formalized and represented by ontologies, which play an important role in the realization of the Semantic Web. However, since the acquisition of knowledge from certain domains usually requires deep involvement of qualified domain experts, construction of such ontologies is difficult and costly, even with the availability of dedicated languages and ontology editing tools. Some effect has been made to reduce this involvement by introducing a general paradigm of automatic domain knowledge learning from various sources. To make this paradigm more specific and practical, this paper proposes a framework for capturing domain knowledge through raw domain data available over the Web. This framework consists of three dedicated parts: data collection, pre-processing and mining, where mining part performs core task of the framework. Each part can be designed with specific optimized methods. The preliminary implementation of certain parts has shown it is able to capture the knowledge of electronic product taxonomy via the Web.

1. Introduction

The aim of Semantic web (Berners-Lee, Hendler & Lassila 2001) is to provide a globally shared platform so that contents published over it can be understood and intelligently processed by computers themselves. To approach this aim, Berners-Lee, who invented the notion of Semantic web, suggests a possible architecture (Berners-Lee 2000) which consists of seven layers. Among those layers, ontology layer plays an important role. An ontology is capable of describing relationships between types of things, maintaining a set of concepts and their hierarchy, a set of relations, and certain axioms (Maedche 2002). It provides a mechanism for knowledge share and exchanging. Without well-defined ontologies, Semantic Web applications or software agents can not communicate with each other in an effective way.

The importance of ontologies has led to proliferation of several languages like OWL ([HREF3]), OIL (Fensel et al. 2000) and edit tools such as OntoEdit ([HREF4]), Protégé ([HREF5]), OilEd (Bechhofer et al. 2001) that help to construct them. Refer to (Denny 2002) for a more detailed survey on those tools. However, as ontologies shall reflect the relations of things in the real word appropriately, languages and tools alone are not enough. Creation of ontologies requires the involvement of human experts, especially when the target ontologies are supposed to describe certain domain knowledge. This makes the process costly.

To minimize the involvement of human experts in observing domain knowledge for creating ontologies, related works have been conducted. Maedche and Staab (2001) propose a paradigm of ontology learning for semantic web. Its general process consists of several sub-processes: import and reuse, extract and prune, refine. Their paradigm is quite general, providing guidelines for related applications; however, for extraction of domain knowledge from the Web, it offers neither more specific architecture nor optimized methods. "OntoMiner" (Davalcu et al. 2003) is a system that extracts taxonomy from domain specific web sites. The domain knowledge it extracts is limited to certain web site and has little integration. To address the above issues, this paper proposes a practical framework for capturing domain knowledge via the Web. It consists of several dedicated parts, each of which can be designed with specific optimized methods. This framework also allows integration of domain knowledge extracted from different web sites, provided those web sites belong to the same domain.

2. Overview of the Framework

Figure 1 gives the overview of our framework for capturing domain knowledge.

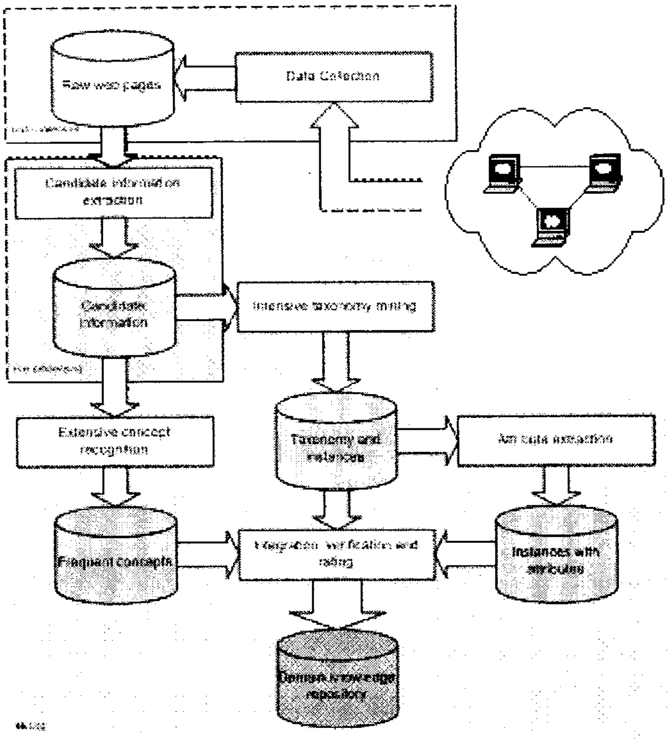

Raw web pages
Data Collection
Candidate information extraction
Candidate information
Intensive taxonomy mining
Extensive concept recognition
Taxonomy and instances
Attribute extraction
Frequent concepts
Integration, verification and rating
Instances with attributes
Domain knowledge repository

Figure 1: Framework for capturing domain knowledge via the Web

The framework can be divided into three main parts: data collection part, pre-processing part, and mining part. Given the instruction from a user, data collection part obtains related web pages from the Web in various ways and builds a repository of raw web pages. The pre-processing part performs filtering tasks for the raw web pages, extracting candidate information useful for late mining. Mining part is the core part of the framework. It consists of four major components: intensive taxonomy mining, attribute extraction, extensive concept recognition, and integration, verification and rating.

The four components of the mining part work in collaboration to mine domain knowledge from the candidate information. Component of intensive taxonomy mining works on candidate information extracted from one same web site. It employs novel technology to rebuild the taxonomy knowledge embedded in those web pages. Meanwhile, it is also able to associate nodes in taxonomy (which we call concepts) with appropriate web contents. We call those web contents the instances of concepts. Those instances are then further processed by the component of attribute extraction. This component extracts the attribute list for the instances, making instances more expressive. While intensive taxonomy mining works in a vertical way, focusing on information from one certain web site, component of extensive concept recognition deals with information from various related web sites. It presents the popularity of concepts that have been found from various web sites. The results of the above three components are finally feed into Integration verification and rating, component that synthesizes and ranks them and put them into the domain knowledge repository.

3. Data collection

To produce desired result, the framework requires adequate input. Therefore, it is essential that raw data collection part provide web pages that contain rich information for the desired domain.

It is assumed that users working on this framework already know certain web sites that contain rich domain knowledge they want to capture. For example, if a user wants to capture domain knowledge related to computer and electronic products, they may know certain web sites such as PC Magazine [HREF6], PC World [HREF7], CNet [HREF8], and etc. Thus they can input URLs of such web sites. The data correction part then uses these URLs as feeds to collect web pages from those web sites. However, users may sometimes have limited information for the domain knowledge they want to capture. In this case, web search engines and web directories can help solve this problem. For example, from the Google directory [HREF9], we can find out lots of web sites related to computer and electronic products via the entry of "Computers and Internet" and "Electronics".

After the seed URLs are input, the data collection part starts to download web pages from those web sites and pools them into the repository of raw web pages for further process.

4. Web page pre-processing

The collected web pages may contain desired domain knowledge; however, not all the contents of the web pages are useful for building domain knowledge, even the web pages are rich of them. Thus pre-processing is required to extract potential useful information from the raw web pages. However, the diversity of web page design raises the difficulty in pre-processing. A dedicated processing method for one web page / web site is always output undesirable results for other web pages / web sites. Capturing the common features of web pages is essential for pre-processing.

4.1. Common features of web pages

Due to the flexible use of the HTML tags, it is not practical to rely only on the types of tags to extract the wanted information from the web pages. For example, the tag <H1> may indicate the text enclosed by it would be a level one heading; however, most of the commercial web pages don't use it as they have their own schemes to express headings. More over, Tag <TABLE> is supposed to be used to arrange a group of structured data, but in practise, it is also widely used as a mechanism of layout design.

As the types of HTML tags are not reliable for automatic information extraction, what else should we rely on? After observations of about 100 web pages from professionally designed websites, the following three features have been identified:

1. Use of information block. Web information is grouped into blocks based on its content to make user ease of use. Within one block, the content is logically related.
2. Use of menus and navigation indicators. Menus and navigation indicators (also called "breadcrumb navigation") are widely used in web sites. They always locate within a block and are rich of domain knowledge. Automatic recognition and exaction of them would be beneficial for our task.
3. Similarity of word length of items in menus/navigation indicators. Though the schemes of menus (the tags used, the styles rendered) differ from site to site, two features are common: word length of the menu items is same or similar; and word length of each item is quite small (around 2 or 3 words). We believe the selection of words for menus is always well considered by the web site constructors so that the resulted menu can be informative in content but concise and balanced in its appearance.

4.2. Extraction method

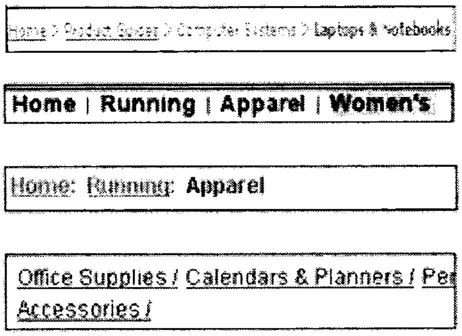
It is effective for information extraction to treat a web page as a tree, i.e., DOM tree [HREF10], based on the HTML tags it contains. During the implementation, nekohtml [HREF11], an open source html parser, is used to parse and create a DOM tree for processing. Given the above analysis of common features of web pages, an algorithm is designed to extract menus and navigation indicators from web pages as candidate information for domain knowledge. The algorithm is presented in another paper (Wang, Lu & Zhang 2005) and here we will not discuss it in detail.

5. Mining of the domain knowledge

Domain knowledge can be mined in two directions: Intensive (vertical) and Extensive (horizontal). Intensive mining aims at reconstruction of concrete taxonomy from certain web sites if available whereas extensive mining is supposed to discover common knowledge shared by various web sites.

5.1. Intensive taxonomy mining

It is common that a web site maintains a logic and consistent way to present its contents. The arrangement of the contents in most cases is reflected by the navigation indicators. Such navigation indicator gives the logical position of the page that contains it within the whole web site context and provides a category or concept of contents of this page. Figure 2 presents four examples of navigation indicators with different styles ([HREF6], [HREF12]). It is obvious that information that navigation indicators carry reflects the domain knowledge we want to extract. Extracting such information and arrange it in a taxonomy is the task of Intensive taxonomy mining.



Home > Product Guides > Computer Systems > Laptops & Notebooks

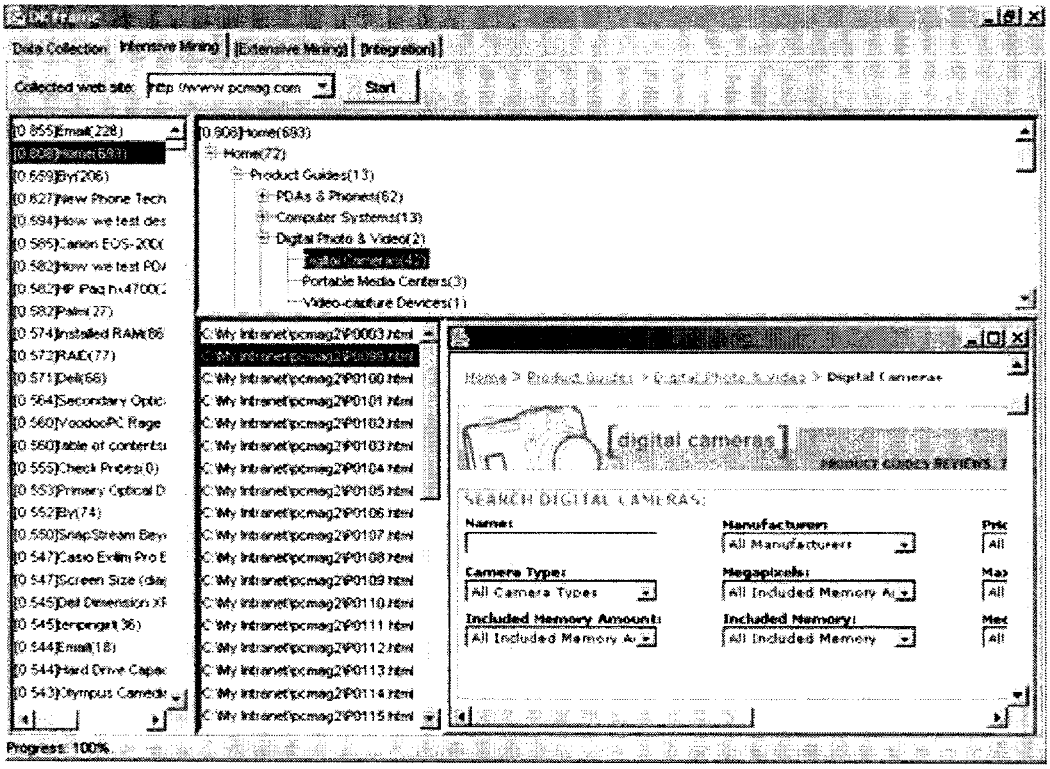
Home | Running | Apparel | Women's

Home: Running: Apparel

Office Supplies / Calendars & Planners / Pe
Accessories /

Figure 2: Different styles of navigation indicators

Pre-processing part has already extracted candidate information including navigation indicators. However, even the candidate information contains unwanted information. In practice, the pieces of candidate information extracted from web pages from one web site can amount to hundreds and thousands. Extracting and rebuilding the taxonomy manually from that large amount of candidate information would be tedious and time-consuming. To overcome this difficulty, we design the entropy-based evaluation, an approach which is discussed in detail in (Wang, Lu & Zhang 2005). With this approach, the right taxonomy can be easily extracted from lots of candidate information. Besides the construction of the taxonomy, contents of web pages can also be assigned as instances to the nodes in the taxonomy. Figure 3 presents the mining result from the web site PCMag. The list at the left shows the candidate information sorted by their entropy values. The higher the value, the more likely the piece of candidate information contains the taxonomy. The tree at the top right displays the captured taxonomy from the web site. Some nodes in the taxonomy tree are associated with several web pages as instances. This example shows a list of pages associated with the node of digital camera.



Collected web site: http://www.pcmag.com

[0.855]Email(228)
[0.808]Home(693)
[0.659]By(206)
[0.627]New Phone Tech
[0.594]How we test des
[0.585]Canon EOS-20D(
[0.582]How we test PD/
[0.582]HP iPaq hx4700(
[0.582]Palm(27)
[0.574]Installed RAM(86
[0.572]RAID(77)
[0.571]Dell(66)
[0.564]Secondary Optic
[0.560]VoodooPC Rage
[0.560]Table of contents
[0.555]Check Prices(0)
[0.553]Primary Optical D
[0.552]By(74)
[0.550]SnapStream Bey
[0.547]Casio Exilim Pro E
[0.547]Screen Size (diag
[0.545]Dell Dimension XF
[0.545]lenspeople(36)
[0.544]Email(18)
[0.544]Hard Drive Capac
[0.543]Olympus Camedia

[0.808]Home(693)
Home(72)
Product Guides(13)
PDAs & Phones(62)
Computer Systems(13)
Digital Photo & Video(2)
Digital Cameras(45)
Portable Media Centers(3)
Video-capture Devices(1)

C:\My Intranet\pcmag2\P0003.html
C:\My Intranet\pcmag2\P0099.html
C:\My Intranet\pcmag2\P0100.html
C:\My Intranet\pcmag2\P0101.html
C:\My Intranet\pcmag2\P0102.html
C:\My Intranet\pcmag2\P0103.html
C:\My Intranet\pcmag2\P0104.html
C:\My Intranet\pcmag2\P0105.html
C:\My Intranet\pcmag2\P0106.html
C:\My Intranet\pcmag2\P0107.html
C:\My Intranet\pcmag2\P0108.html
C:\My Intranet\pcmag2\P0109.html
C:\My Intranet\pcmag2\P0110.html
C:\My Intranet\pcmag2\P0111.html
C:\My Intranet\pcmag2\P0112.html
C:\My Intranet\pcmag2\P0113.html
C:\My Intranet\pcmag2\P0114.html
C:\My Intranet\pcmag2\P0115.html

Home > Product Guides > Digital Photo & Video > Digital Cameras

Name:
Manufacturer: All Manufacturers
Camera Type: All Camera Types
Megapixels: All Included Memory A
Included Memory Amount: All Included Memory A
Included Memory: All Included Memory

Progress: 100%

Figure 3: Results of mining pages from PC Magazine

5.2. Attribute extraction

Attribute extraction is an extension of intensive taxonomy mining. The generated taxonomy has its nodes (concepts) associated with a number of web pages as instances. Within one concept, its instances always have some common features and those features are described in the corresponding web pages. Thus, it is desirable to extract those common features as the attributes of the concept. For example, a node of "digital camera" in the taxonomy is associated exactly with many web page instances discussing topics about digital camera. It is obvious that, lens, resolution, pixels, and some other features are always covered in those web pages. They construct the attributes of digital camera, and are parts of the domain knowledge we want to capture.

Due to the diversity and noisy nature of web page contents, it is difficult to extract the precise attributes out of the web pages. However, some heuristic rules can be applied. An evident rule is that attributes are often accompanied with values and they are usually separated by ":" or are formatted in a table. The dual phenomenon of attribute-value can also help to find out more rules from the web pages. (Sundaresan & Yi 2000) propose a technique that we think can be adopted to solve this problem.

5.3. Extensive concept recognition

Each web site has its own view of the domain it belongs to. Intensive mining captures such view of domain knowledge presented by each web site. As for many web sites belonging to one domain, it is interesting to find out the common view of the domain from all these web sites. Extensive concept recognition is the component that performs the task of discovering commonly used concepts among the web sites belonging to one certain domain.

Same with the intensive taxonomy mining, the input of extensive concept recognition is the candidate information generated by pre-processing part. It is assumed that all the important concepts of domain knowledge are preserved in the candidate information. However, those pieces of information that are used to build taxonomy are ignored in this task, as the whole candidate information is abundant in domain knowledge. Inclusion of the used information for extensive concept recognition may reduce the chance of discovering more other useful concepts.

For the discovery of frequent patterns, association mining is promising. Association rule mining algorithms, such as Apriori (Agrawal & Srikant 1994) and its variant, is quite effective in finding frequent item sets out of a large number of item sets. We regard each piece of candidate information as a "transaction". This transaction consists of one or more concepts (in form of short words or phrases) presented together. This builds up a database of "concept transactions" obtained from various web sites. Association rule mining algorithm can be applied to this database and as a result, frequent concept sets across those web sites can be discovered.

5.4. Integration, verification and rating

The framework captures the domain knowledge in several ways. An integration process is indispensable as different ways provide us different views of the target domain knowledge. Through integration, those different views can be connected with each other, thus provides a more comprehensive perspective of the domain knowledge. Integration also means redundancy reducing. It is inevitable that domain knowledge captured from different web sites overlaps, especially their taxonomies. Thus merging those taxonomies is necessary to reduce the redundancy.

As for taxonomy integration or mapping, certain research works have been done, aiming at finding automatic and effective ways to accomplish this task. Those research works can be divided in to two categories according to their focuses. Approaches proposed by (Agrawal & Srikant 2001 ; Sarawagi, Chakrabarti & Godbole 2003 ; Zhang & Lee 2004 -a, 2004 -b) focuses on integration of instances from two taxonomy with optimized methods, whereas (Doan et al. 2002) focuses on improving algorithms for mapping labels between ontologies. Despite the difference of focuses, their methods are worth studying. It is promising to adopt and improve those methods to accomplish our specific task of integrating mined taxonomies.

As several techniques are used to automatically capture the target domain knowledge, there are chances that those automatic techniques produce some unwanted results due to our input, the web pages that always diverse in both styles and contents and also contain lots of noisy information (advertisement, etc). Verification is necessary to assure that correct results are recorded and stored to the final domain knowledge repository and inappropriate results are discarded.

When users refer to the captured domain knowledge, it is advisable to provide them with information such as which concepts or concept patterns are used more frequently, which attributes are mentioned more than others. Rating is the process to provide such statistical information. It relies on the information that previous process like extensive concept mining and taxonomy integration produced.

6. Conclusion and future work

This paper proposes a practical framework for capturing domain knowledge via the Web. This framework is made up of three parts: data collection, pre-processing and mining. Mining is the core part of the framework, consisting of four major components. Each part or component has been assigned a distinct task. Under this framework, domain knowledge embedded in raw web pages can be captured in both intensive and extensive ways and finally integrated and rated. Preliminary implementation of certain parts has shown this framework is able to capture the knowledge of electronic product taxonomy.

Our future work is to further implement and improve methods designed for the parts/components so that better results can be achieved. During this course, issues such as robustness of the framework, tolerance to noises, and extensibility for growing domain knowledge will be addressed. It is expected that knowledge of a certain domain captured though this framework will accelerate the construction of corresponding ontology.

References

Agrawal, R. & Srikant, R. 1994, 'Fast Algorithms for Mining Association Rules in Large Databases', in, *Proceedings of 20th International Conference on Very Large Data Bases(VLDB'94)*, Morgan Kaufmann, Santiago de, Chile, pp. 487-499.

Agrawal, R. & Srikant, R. 2001 'On integrating catalogs ', in, *Proceedings of the tenth international conference on World Wide Web*, ACM Press, Hong Kong, pp. 603-612.

Bechhofer, S., Horrocks, I., Goble, C. & Stevens, R. 2001, 'OilEd: A Reason-able Ontology Editor for the Semantic Web', *Joint German/Austrian Conference on AI*, Vienna, Austria, p. 396.

Berners-Lee, T. 2000, Semantic Web - talk at XML2000, [HREF13]

Berners-Lee, T., Hendler, J. & Lassila, O. 2001, 'The Semantic Web', *Scientific American*, pp. 34-43.

Davalcu, H., Vadrevu, S., Nagarajan, S. & Ramakrishnan, I.V. 2003, 'OntoMiner: bootstrapping and populating ontologies from domain-specific Web sites', *IEEE Intelligent Systems*, vol. 18, no. 5, pp. 24-33.

Denny, M. 2002, Ontology Building: A Survey of Editing Tools, [HREF14].

Doan, A., Madhavan, J., Domingos, P. & Halevy, A. 2002 'Learning to map between ontologies on the semantic web ', in, *Proceedings of the eleventh international conference on World Wide Web*, ACM Press, Honolulu, Hawaii, USA pp. 662-673.

Fensel, D., Horrocks, I., Harmelen, F.V., Decker, S., Erdmann, M. & Klein, M. 2000, 'OIL in a Nutshell', Knowledge Acquisition, Modeling, and Management, *Proceedings of the European Knowledge Acquisition Conference(EKAW-2000)*.

Maedche, A. 2002, *Ontology learning for the semantic Web*, Kluwer Academic, Boston.

Maedche, A. & Staab, S. 2001, 'Ontology Learning for the Semantic Web', *IEEE Intelligent Systems*, vol. 16, no. 2, pp. 72-79.

Sarawagi, S., Chakrabarti, S. & Godbole, S. 2003 'Cross-training: learning probabilistic mappings between topics ', in, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, Washington, D.C. , pp. 177-186.

Sundaresan, N. & Yi, J. 2000, 'Mining the Web for relations', in, *Proceedings of the 9th international World Wide Web conference*, North-Holland Publishing Co., Amsterdam, The Netherlands, pp. 699-711.

Wang, C., Lu, J. & Zhang, G. 2005, 'Mining key information of web pages', in, *1st International Workshop on E-Service Intelligence in conjunction with 8th Joint Conference on Information Sciences*. (accepted).

Zhang, D. & Lee, W.S. 2004 -a, 'Web taxonomy integration through co-bootstrapping ', in, *Proceedings of the 27th annual international conference on Research and development in information retrieval*, ACM Press, Sheffield, United Kingdom pp. 410-417.

Zhang, D. & Lee, W.S. 2004 -b, 'Web taxonomy integration using support vector machines ', in, *Proceedings of the 13th international conference on World Wide Web*, ACM Press, New York, NY, USA pp. 472-481.

Hypertext References

HREF1
    http://it.uts.edu.au
HREF2
    http://www.uts.edu.au
HREF3
    http://www.w3.org/2004/OWL/
HREF4
    http://www.ontoprise.de/products/ontoedit_en
HREF5
    http://protege.stanford.edu
HREF6
    http://www.pcmag.com
HREF7
    http://www.pcworld.com
HREF8
    http://www.cnet.com
HREF9
    http://directory.google.com
HREF10
    http://www.w3.org/DOM/
HREF11
    http://www.apache.org/~andyc/neko/doc/html/
HREF12
    http://www.webdesignpractices.com/navigation/breadcrumb.html
HREF13
    http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html
HREF14
    http://www.xml.com/pub/a/2002/11/06/ontologies.html

Copyright