Building Product Image Extraction from the Web

Walaiporn Nakapan[1], Gilles Halin[2], Jean-Claude Bignon[3], Marc Wagner[4], and Pascal Humbert[5]

CRAI (Research Center in Architecture and Engineering) UMR Map, N°694 CNRS

Nancy School of Architecture

Nancy, France.
E-Mail: nakapan[1],halin[2],bignon[3],wagner[4],Humbert[5]@crai.archi.fr

Abstract: HYPERCAT project proposes a digital organization of the technical information relative to the building products and materials. Its application on technical information search by images is one of context-based image search engines. However, a manual construction of an image database for this application can be very costly. The image extracted from the French building product providers' Web sites can solve the problem of acquisition and indexing. The problematic of the Web image extraction for this activity is how can we extract the pertinent images and how to index them? Consequently, this question leads to twofold challenges: image extraction and image indexing. First, an extraction rule is applied to illustrate the extraction process. Second, an indexing process indexes pertinent images with terms in the thesaurus. The application on building product data extraction on the Web is called the Wimex-Bot.
Key words: image, web, data extraction, context-based image indexing

1. Technical information search in the domain of AEC

In the domain of Architecture, Engineering, and Construction (AEC), the technical information is required during the early design phase (conceptual design) and the end of design stage (technical design). The technical information is, for instance, the product's name, the installation, the price, the provider's name, address, telephone number, etc. The technical information comes in various kinds of support, from ordinary paper catalogue to more advanced ones like CD-ROM, or Web site. However, the access to the information is not easy when its quantity is important. The system should provide various search modes adapted to each situation of the user's demand.

The HYPERCAT project [1] presents a digital organisation of the technical information relatives to the building products and materials. The Web site of CRIT[1] serves as the project's platform implementation. It proposes a simple search and a multi-criteria search to the professionals in building industry (architects, constructors, engineers, …). The user types in a product's name of a company's name in order to perform a simple search, while a multi-criteria search mode proposes choices of products by product type, constructive function, material, etc.

The simple search and the multi-criteria search don't seem to be the most suitable search modes for the design activity. This kind of search suits better for the situation where the user knows exactly what he wants. On the contrary, in early design stage, the user doesn't know exactly what he wants. He just wants to take a look at the products available on the market, obtain the illustration of the product, and look for ideas.

As the architect has a capacity to think with the image, the image can be a new support for technical information search for the design activity. The image is a very important working tool for the architect. It helps representing the idea, perceive the problem, and communication his idea to others. Moreover, the image is a very efficient information encoding system. As one may say " a good image is better than a thousand words". The use of image can sometimes replace the use of text, especially the manipulation of specific term of the domain. Therefore, the construction of the technical information search by images would be able to fulfil the classic search modes. Our first application for this purpose is a context-based image search engine for technical information called BATIMAGE [2] (cf. Figure 1).

1 The service of CRIT (Centre de Ressources et Informations Techniques) is the result of collaboration between Nancy School of Architecture and Strasbourg School of Architecture. The Web site can be found at `http://www.crit.archi.fr`

Figure 1: BATIMAGE

An effective Technical information search means the user finds an image of the product that interests him. We need a quality database, not a quantity one. It takes three qualities to make an ideal database for this application: the completeness, the ease of acquisition, and the ease of indexing. A complete image database means that there are images that represent every product in the database. Dealing with thousands of images, it is impossible to scan these images one by one. Moreover, the manual indexing is usually very costly (number of persons and time). For this reason, the indexing mode should be an automatic or a semi-automatic one. These indexing modes eventually suggest the possibility of supplying the image database regularly.

The images from the Web come to our interest as they are used by many image search engines (Yahoo!, Altavista, WebSEEk, etc.). This inexhaustible resource is chosen because the image found there meet the three qualities as mentioned above. The image from the Web are numerous, accessible, and of various subject. It seems that there should be many images of building product on the Web and that there should be a way to extract and index them automatically.

This article presents the Web image extraction in order to construct the image database for the technical information search by images. We discuss the problematic of the image extraction and indexing process, the methods proposed, and the applications.

2. Problematic of building product image extraction from the Web

Building product image extraction needs to be treated specifically for the domain of Architecture, Engineering, and Construction. The problematic of the Web image extraction in the domain is as followed: *how to extract the pertinent images and how to index them*? Consequently, this question leads to twofold challenges: image extraction and image indexing.

On one hand, considering that the size of the Web is very large: *how can we find the image of the French building products?* The choice of Web site is not a problem for us. Since, the HYPERCAT project has already indexed French building product providers; we are only interested the companies' Web site in the database. But Web images come in various forms and content. There are decorative images as well as photos; and there is a photo of a tile in a production line as well as a photo of the same tile on the rooftop of a house. Of course the tile on the rooftop is more interested for the building product search than the tile in a production line. How can we extract only pertinent image? The

pertinent image extraction on the Web is discussed in 3.1.

On the other hand, the context-based image retrieval is chosen for the building product information search. In this kind of search, the user is interested by what product the image represents rather than the similarity between images, as performed by content-based image retrieval. Since the image search engine is a context-based one, these images need to be indexed by texts. It is common that the images in a building product catalogue are accompanied by the text: *can we index an image of a building product using its context?*

But the choice of indexing vocabulary can lead to two problems in the image retrieval process: noise and silence. Free text indexing use an entire piece of text to index an image, thus easy to carry on. But this increases noise and silence during the retrieval. On the contrary, controlled text indexing delimits the number of indexing terms, hence more difficult to carry on but results in less noise and silence. Therefore, we prefer the controlled text indexing, which is represented by a thesaurus. The indexing process has to assure that the text in question (the context of the image) correspond to indexing terms in thesaurus. The process of image indexing using its context and thesaurus terms is discussed in 3.2.

3. The method proposed

Two methods are proposed to overcome the challenges in image extraction and image indexing from the Web resource. The application in paragraph 3 uses the following methods:

- the image extraction rules for building product image extraction form the Web,
- the indexing process for the image indexing with the thesaurus term.

We describe the two methods that are served as a guideline for the application programming afterwards.

3.1. The image extraction rules

For the selection of relevant image, there are four important principles to apply*:*

1. a basic principal,
2. a principle of visual analogy,
3. a principle of legend,
4. a principle of form.

3.1.1. The basic principal

As the images on the Web are on various forms and subjects, the image extracted should be the one that represents the building products. The basic principal allows the extraction of building product images. The pertinent image should correspond to the following criteria:

- the image can contain several elements but it should represent one particular building product,
- the image of building product should come from the page (paper catalogue, web page) that presents the building product.

3.1.2. The principle of visual analogy

A pertinent image should be similar to the real world object and the product should be recognized easily. The principle of visual analogy is applied in order to select images that produce most effects, which means that a very little effort is required for image interpretation. The relevance of an image depends on the following criteria:

- there has to be a similarity of colour between the image and the usually dominant colours of the represented product,
- the object must be represented entirely. The more it is cut, the more it is necessary to interpret the missing parts,
- the scale of representation has to allow the represented object to occupy an important surface area on the image. The less important the surface area, the more of the other image can play the role of main subject,
- it is necessary to maintain elements of an object's usual environment in the object representation. (example: a type of faucet will be better noticed if it is situated near a kitchen sink or near a boiler).

3.1.3. The principle of legend

Sometimes, the information of the image is not entirely encoded in the image itself but in the context of the image. The principle of legend is applied to minimize the effort of interpretation as image alone cannot give sufficient information. Keywords are associated to an image in order to refer its universe of interpretation. These criteria are valid only for our particular approach, in which we give greater importance to the recognition effect than to the suggestion effect. The relevance of an image depends on the following criteria:

- an image selected should have a nearby context or a legend,
- the context of the image should be interesting for the product information search,
- the context of the image is interesting if the keywords found correspond to terms in the thesaurus.

3.1.4. The principle of form

The principle of form helps selecting an image without knowing its content. It eliminates the image that causes interference. The criteria are as follow:

- an image's size (width, height) should be in a limited interval, corresponding to the size of a photograph,

- an image's proportion (width/height) should be in a limited interval, close to the proportion of a square.

All these principles should be applied to the image extraction and indexing process. The basic principal and the principle of form can be applied to an automatic process of image extraction. The principle of legend can be viewed as possibly automated if the image's legend is a text, not an image of a text (alphanumeric image as described by [3]). However, the principle of visual analogy needs the knowledge of automatic form recognition. At this stage it still needs a human intervention.

From these principals, the extraction rules for building product images extraction are derived. The extraction rules are organized in a sequence of questions and a pertinent image should correspond to every question. Some questions are specific for image extraction from the Web but they are still derived from the principals presented above. Each question of the rule can be represented in the following way:

1. *"Is the Web page in a limited distance?"* The distance of the Web page should be verified in order to eliminate the pages that do not concern building products. As the parent page is usually a homepage or an index page of a building product provider, the Web pages that are far from their parent have a large chance to present something else. The distance is set to 0 by default, which means that the URL stays in the same domain (parent URL). We limit the distance to 1, which accept the URL that are referred to by the parent URL and no further.
2. *"Is the page written in French?"* The choice of language used in the page can eliminate all the pages that are written in foreign languages (other than French). As the thesaurus that we use is a French one, images found on that page are impossible to be indexed by the words found there on.
3. *"Is the image found in an interesting page?"* Interesting images seem to be found in appropriate pages, which means the product presentation page. In such way, when we look for an image of building product, its context is usually relevant. On the other hand, the same images that are found in the "presentation of the company" page are irrelevant. This helps to eliminate all the images that are irrelevant such as a map of France, an access plan in the address page, a production chain or a factory in the company presentation page, etc…. The list of forbidden words in <A HREF> tells the robot where the page doesn't worth to be investigated (presentation, history, address, contact, link, etc…).
4. *"Is the form of image OK?"* These criteria examine its physical relevance of the image considering its width, height, and proportion (width/height). It helps eliminating the decorative images commonly found on a Web page, as an image of 3 times larger than its height is a banner, or an image of 15*15 pixels is a puce. Apart form that, it verifies if the image will be readable during the image retrieval process. Since the image will be reduced to small icons (within 160*160 pixels), an image that is reduced 5 times if its original size usually lost most of its information. Then the image file size in kilobyte is verified. If the file size is not greater than a certain size, the image can be too small, or even a bad quality one. Hence, it doesn't worth being investigated. The form criteria are validated as follows:
   - 60 pixels ≤ width ≤ 610 pixels,
   - 60 pixels ≤ height ≤ 660 pixels,
   - 0.58 ≤ proportion ≤ 2.1.
   - 3.2 kilobyte ≤ file size.
5. *"Does the image has a context?"* The image is rejected at this stage even if there is not a nearby context. It means that we cannot know what it is about. However, a context of an image can be a context of another. The calculation of the distance between each image and each context helps determining which context belongs to which image.
6. *"Is the content of image's context interesting?"* Even if the image has a context, it might be useless if the context is not interesting. For our approach, an interesting context means a text describing a building product. Therefore, the content of the context is interesting if we find that it is described by words that has equivalent in the thesaurus of building products.

From these questions, we obtain the following extraction rules:

BEGIN
IF the distance of the Web page ≤ 1
AND the page is written in French
THEN select the page
END IF

FOR every page
    IF the image is in an interesting page
    AND 60 pixels ≤ width ≤ 610 pixels
    AND 60 pixels ≤ height ≤ 660 pixels
    AND 0.58 ≤ proportion ≤ 2.1
    AND image file size ≥ 3.2 kilobyte
    AND image has a context
    THEN select the image
    END IF
END FOR

FOR every context
IF the context contain the thesaurus term
THEN select the context
AND IF
END FOR
END

The image and the context that correspond to every rule are selected. The result is a list of image and its interesting context. Then the indexing process will map these keywords for each image with terms in the thesaurus. We will discuss the context-based indexing using the thesaurus in the following paragraph.

3.2. The indexing process for indexing image with the thesaurus term

In the building product catalogue, we usually find image of the product accompanied by text. This is also the case of the images found on the Web pages of building product providers. Therefore, the idea is to index the image with its context. But in order to fight against noise and silence in the retrieval process, the controlled text is used as indexing vocabulary.

The context will be mapped with the thesaurus terms, which contains specific terms of building products. In order to perform the mapping, The N-gram method (bi-gram and tri-gram) as describe in [4] is used since it gives the best result. The indexing system extracts the motif of context's N-gram closest to those from the thesaurus and presents them according to the weighting obtained via the preliminary analysis of the total frequencies N-grams. The indexing process illustrated in figure 2 can be described as followed:

On one side, each thesaurus term is transformed into common name and stored in a dictionary. For example, the word "Roof window" becomes "roof" and "window". Separately, the process transforms each thesaurus term into the N-gram (bi-gram and tri-gram) and stores them in a list of N-gram. For the same example "Roof window" becomes "ro", "oo", "of", "f✿", "✿w", "wi", "in", "nd", "do", "ow" for bi-gram and "roo", "oof", "of✿", "f✿w", "win", "ind", "ndo", and "dow". This process is to be done only once, unless there is a modification of the thesaurus

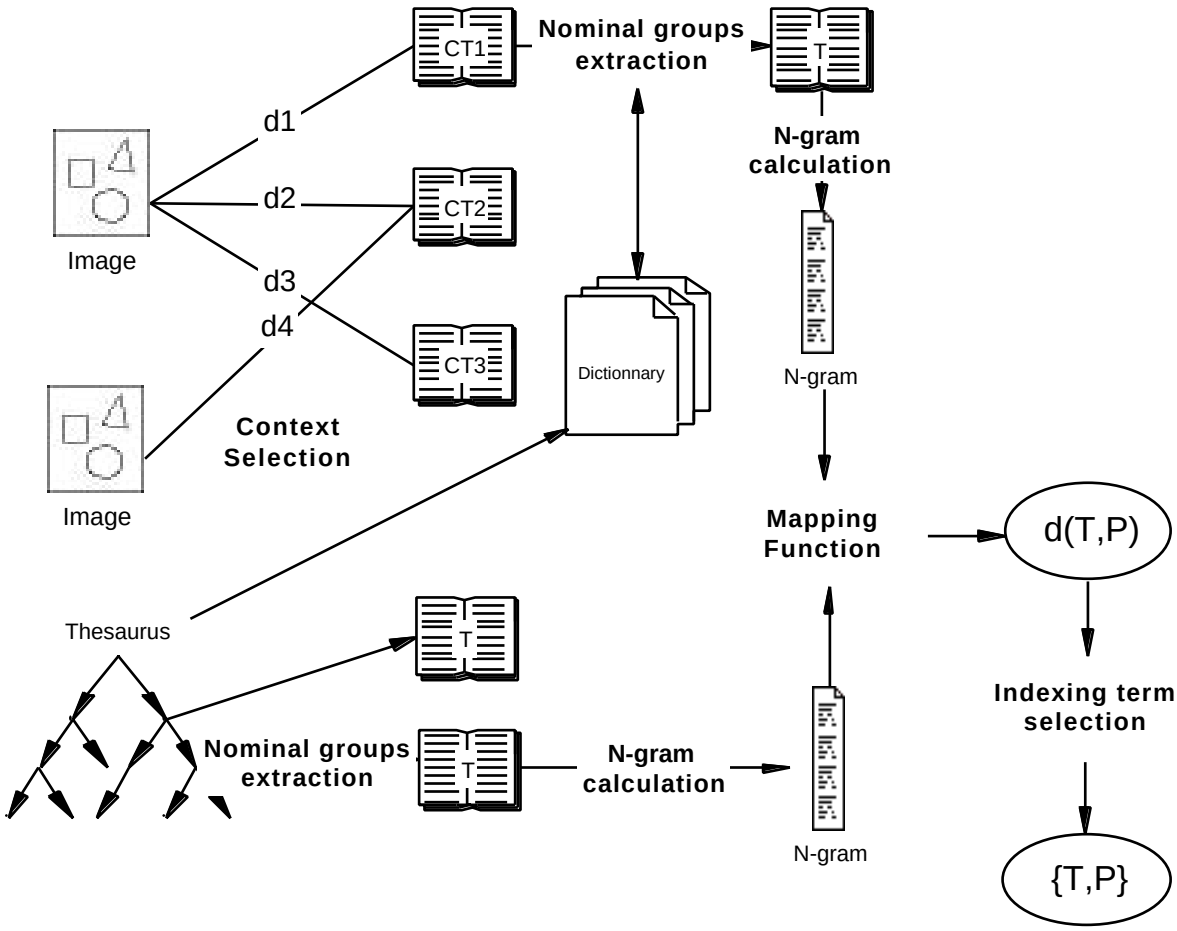

Figure 2: The indexing process

On the other side, the system selects the context (CT) of each image. Notice that a context of an image can be a context of another. Then, each word (W) of the context is compared to the common name in the dictionary. If the word in question is found, the next word (W) is attached to the analysing word becoming a nominal group or a term (T). For example, if the word "roof" is presented in the dictionary, then it is attached to the next word "window" becoming "roof window". Then this nominal group is transformed into N-gram.

Next, the mapping function calculates the distance (d(T,P)) between each context's N-gram and each thesaurus term's N-gram. Then the process ranks terms by weight for each indexing domain. The term with the highest weight of each indexing domain is selected as indexing term for the image.

The indexing terms are used as supports for the interactive and progressive image retrieval. This search method uses the relevance feedback [5] and the vector space model as mapping function [6].

4. Application on the image extraction on the Web: The Wimex-Bot

The Wimex-Bot (cf. figure 3) is the application of image extraction on the Web specializing in building product images. Written in Java, it extracts and indexes building product images in a Web page from their HTML source code. It has been realized in the under the project HYPERCAT in order to facilitate the provision and the revision of the database. The robot goes across the Web beginning from a list of Web pages, which are usually the homepage or the index page of the Web site. Then it analyzes the Web site one by one.

Related work is the *Marie project* of US Naval Postgraduate school [7] which also uses the HTML source code in order to index image, but the system finds the caption of the image and not its entire context.

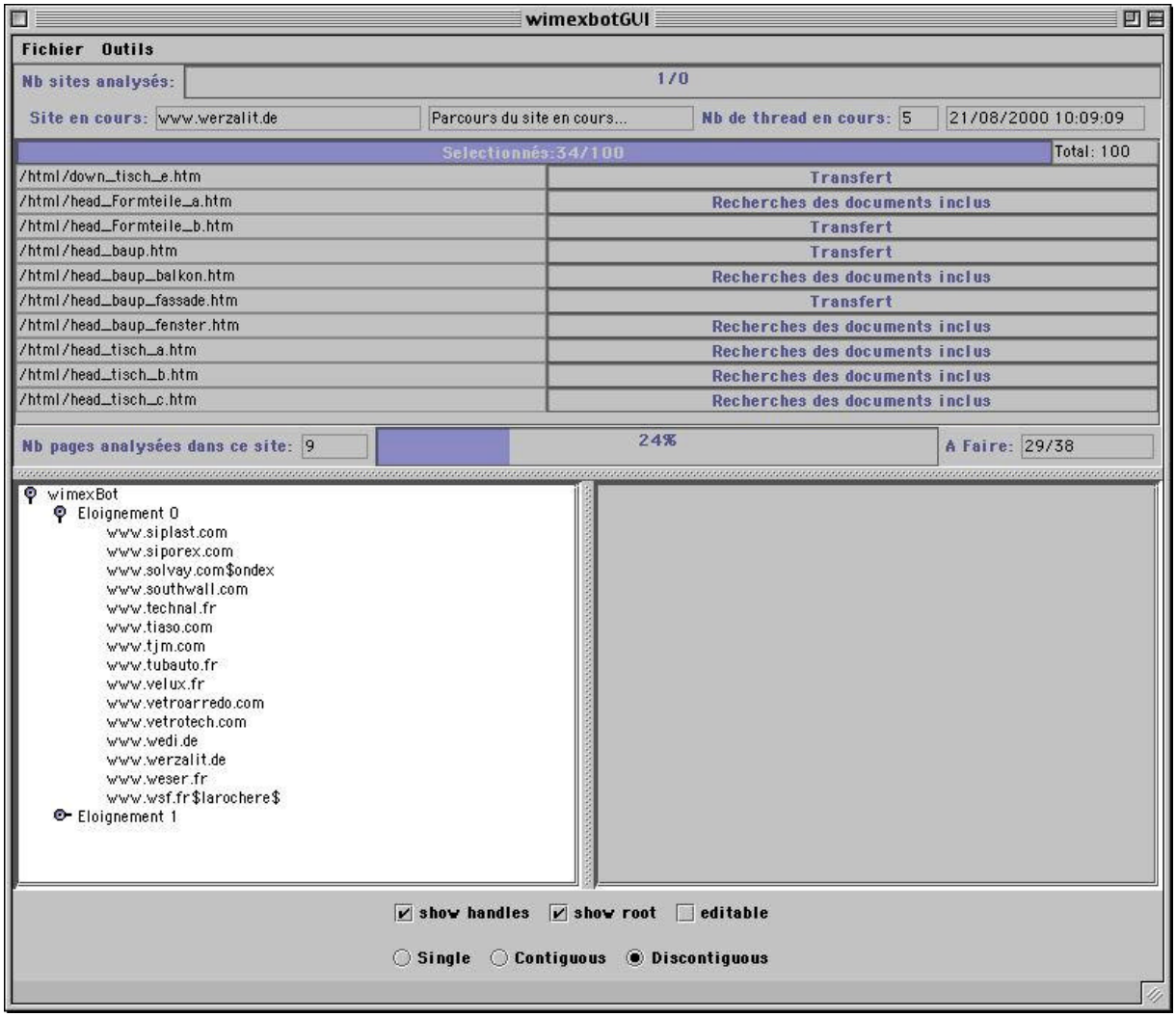

Figure 3 : Wimex-bot

Another early work in image and text extraction on the Web is the *Image Excavator* from School of Computing Science, Simon Fraser University, Canada [8]. This Web agent uses image textual information, like HTML tags in web pages, to derive keywords in English language. By traversing on-line directory structures like Yahoo!'s directory, it is possible to create hierarchies of keywords mapped on the directories in which the image was found.

The Wimex-Bot uses the principle of three spiders: the Transversal spider, the Image extractor, and the Context parser. They have been inspired by the image and video collection process for WebSEEk [9]. The following diagram illustrates the principle of the Wimex-Bot (cf. figure 4).

The Wimex-Bot is composed of two principle parts: the automatic image extraction and indexing and the manual control of pertinence. This makes the application a semi-automatic one.

The three spiders represents the automatic image extraction and indexing part Their function is composed of the following three principle phases:

- page selection and analysis,
- image selection by form,
- image selection by context.

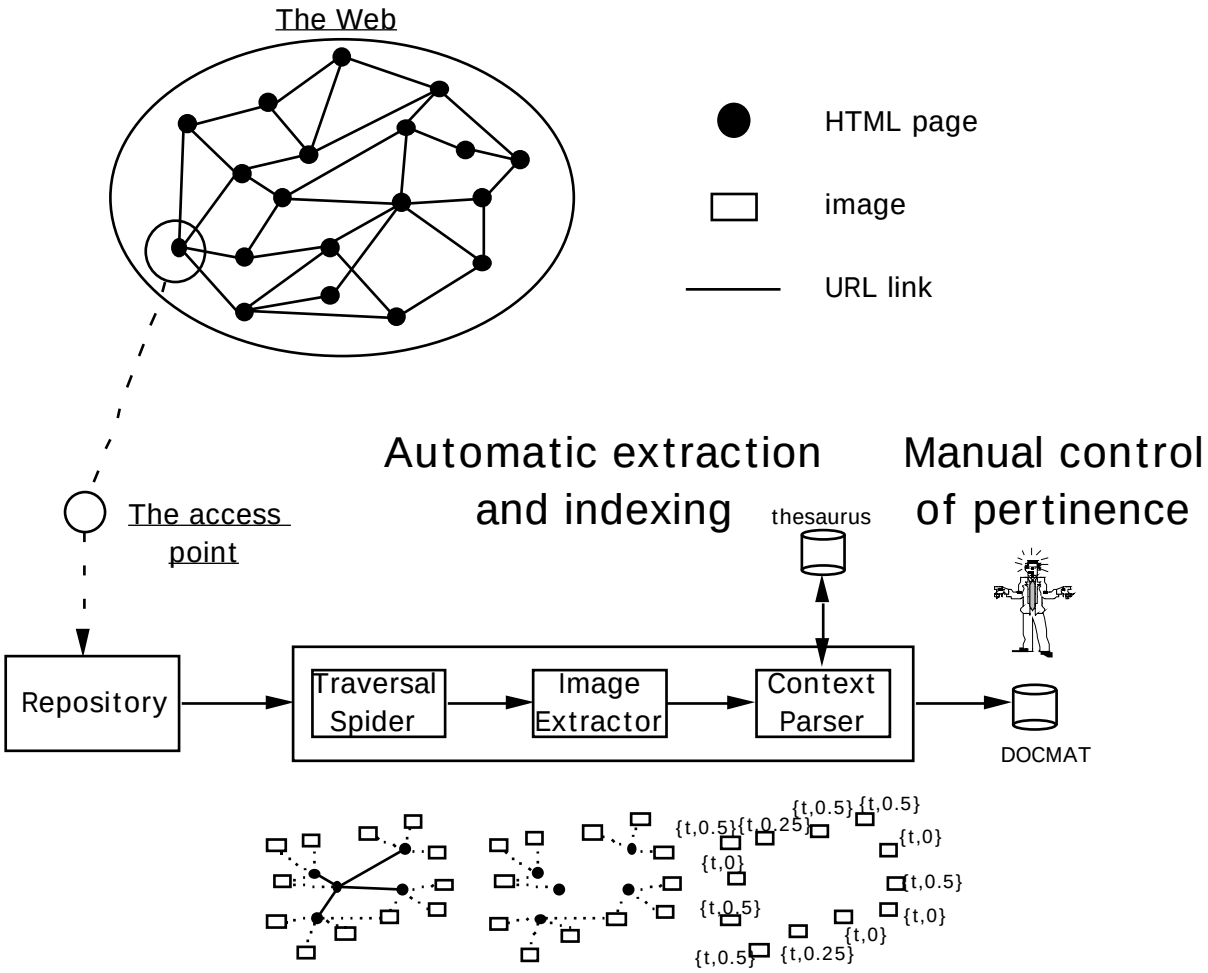

Figure 4: The principle of Wimex-Bot

4.1. Page selection and analysis

The "Transversal spider" begins with a repository and an access point to the Web. The repository usually contains the homepages of the Web sites. These Web pages are selected from the building product database.

The spider travels the Web starting with the beginning page in order to find

- the next pages,
- the image included in each page.

Every HTML pages linked from the beginning page are listed. The distance of the site from the origin page helps to select the page to investigate. Then it analyses each HTML page found. This analysis allows identifying the next page and the image to analyse in the future.

However, the analysis can be carried on only with the page whose context is in French language and the page should not contain the word presented in the list of forbidden words. The forbidden words (address, contact, history, …) are listed in order to reject the pages that doesn't worth being investigated. The reason is that usually the HTML pages containing these forbidden words do not present the building products. The potential pages of each site are selected.

4.2. image selection by form

In order to obtain the readable and pertinent images, we realize that the dimension of the image should be in a certain size.

The "Image extractor" selects only the image from each selected page by its width, height, proportion, and the image file size in byte. These numbers should be in the limited intervals. This stage corresponds to the extraction rules allowing the first filter by form.

4.3. image selection by context

Each context of the image presented in the HTML page that are selected previously is verified in order to the equivalent theme presented in the thesaurus. The thesaurus exploited in used for the building products indexing in the database.

Then the "Context Parser" verifies if an image has a context around it. However, it validates only interesting context. The context is interesting when there are thesaurus terms in the context. At this stage, the indexing process indexes the image with thesaurus terms using its context.

Next, the intervention of administrator seems necessary in order to control the pertinence of extracted data. He validates the image and indexing terms before storing them in the database. The manual validation of the image and its indexing terms is applied. The file of the extraction result containing the URL of each image and its indexing terms. This file will be imported onto the building product database.

Figure 5 : Example of a building product Web page

5. Conclusion

This article presents the building product image extraction from the Web for technical information search by images. The application for this activity, the Wimex-Bot, is composed of three spiders. Two methods are proposed to overcome the challenges in image extraction and image indexing. First, the extraction rules are shown to illustrate the extraction process. Second, an indexing process indexes an image by thesaurus terms using the terms extracted from the image's context.

The first image test set from the application Wimex-Bot is used in the interactive and progressive image search engine for building product information search. It allows the verification of the images extracted from the web and their indexing terms, which will be used in the technical information search process. The result is that the terms and images extracted from the web give a pertinent product search. However, the image extraction process is not absolutely satisfied as it gives some irrelevant images. An intervention of an administrator is still necessary in order to control the pertinence of the image selection. Therefore, we look forward to improve the extraction process by adding other extraction rules, such as the image content (image file type, number of colour used,…).

Apart from that, the main difficulty that the robot has encountered is the massive volume of information to treat. We look forward to improve the algorithm in order to surmount this problem.

Therefore, the next step of the robot development is to develop the manual control of the image extracted and its indexing terms. The administrator will be able to reject the irrelevant images. Then he verifies the indexing term of each chosen image to assure the good operation of the image retrieval process.

Then the robot should allow an evolution of the thesaurus in order to enrich the terms. Manually, the administrator can add the terms that he finds interesting but are not presented there on. The thesaurus used at the beginning is a French language one. However, we can experiment using the thesaurus in other languages. The same extraction process can be applied in order to extract the images from the Web pages written in the respective languages as well.

Lastly, the maintenance of the Web page and the image URL link is necessary. For the time being, the extracted images are downloaded to the local hard disk. And the building product information search by images does not present the links to the original images. But during this process, it is possible that the user wants to visit the original pages where the image is found. If we want to show these original pages, the robot has to revisit them once in a while. First, it verifies if the pages themselves still exist. Second, it verifies if the images are still found there on. With this verification, we can imagine an application to the technological survey that looks for new products in a building product web site.

6. References

[1] Halin, G., Nakapan, W., Bignon, J.C. : Interactive and progressive image retrieval on the WWW. Application on building product search. Presented in International Workshop "Multimedia Databases and Image Communication", Salerno, Italy, Octobre 1999.

[2] Bignon, J.C., Halin, G., Nakapan, W. : Building Product Information Search by Images. Proceedings of the 5th International Conference in Design and Decision Support Systems in Architecture, Nijkerk, The Netherlands, 47-61, August, 2000.

[3] Vendrig J. : Filter Image Browsing. A study to image retrieval in large pictorial databases. Thesis. University of Amsterdam, February, 1997.

[4] Hallab, M., Lelu, A. : Proxilex un outil d'approximation orthographique à partir des fréquences des N-grammes, 5e conférence internationale H2PTM'99, Paris, France (1999) 201-211.

[5] Van Rijs Bergen, C.J.: Information Retrieval. 2nd edn. Butterworths London, 1979.

[6] Halin, G., Créhange, M., Kerekes P.: Machine learning and vectoriel matching for an image retrieval model: EXPRIM and the system RIVAGE. Proceeding of the ACM 13th International Conference on Research and Development in Information Retrieval, Brussel Belgium, 99-114, September, 1990.

[7] Rowe, N. C., Frew, B.: Finding photograph captions multimodally on the World Wide Web. Technical report from the AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora, 45-51, 1997.

[8] Zaïane, O.R., Han, J., Li, Z.-N., Hou, J.: Mining Multimedia Data, Proceeding of CASCON'98: Meeting of Minds, Toronto, Canada, 83-96, November, 1998.

[9] Smith, J. R., Chang, S.-F.: Virtually Searching the Web for Content. IEEE Multimedia Magazine, Vol. 4(3), 12-20, July-September 1997.