# Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages

Mathias Géry, Jean-Pierre Chevallet

# Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages

**Mathias Géry, Jean-Pierre Chevallet**

Equipe MRIM (Modélisation et Recherche d'Information Multimédia)
Laboratoire CLIPS-IMAG, B.P. 53, 38041 Grenoble Cedex 9, France
E-mail : {Mathias.Gery, Jean-Pierre.Chevallet}@imag.fr

**Abstract :** The World Wide Web is a distributed, heterogeneous and semi-structured information space. With the growth of available data, retrieving interesting information is becoming quite difficult and classical search engines give often very poor results. The Web is changing very quickly, and search engines mainly use old and well-known IR techniques. One of the main problems is the lack of explicit HTML page structure, and more generally the lack of explicit Web sites structure. We show in this paper that it is possible to extract such a structure, which can be explicit or implicit: hypertext links between pages, the implicit relations between pages, the HTML tags describing structure, etc. We present some preliminary results of a Web sample analysis extracting several levels of structure (a hierarchical tree structure, a graph-like structure).

**Keywords :** Web Information Retrieval, Web Pages Analysis, Structure Extraction, Statistics

## 1 Introduction

The task of an *Information Retrieval System* (IRS) is to process a whole set of electronic documents (corpus), with an aim of making it possible to retrieve those matching with their information need. On the contrary of *Databases Management Systems* (DBMS), the user expresses with a query the semantic content of the documents that he seeks. We distinguish two principal tasks:

**Indexing:** The extraction and storage of the documents semantic content. This phase requires a representation model of these contents, called document model.

**Querying:** The representation of the user's information need, generally in a query form. It is followed by the retrieval task, and the presentation of the results. This phase requires a representation model called query model, and a matching function to evaluate documents relevance.

IRS were classically used for textual documents databases, or multimedia databases like medical corpora. The Web growth constitutes a new applicability field for IR. The number of users on the Web has been estimated at 119 millions in 1998 (NUA Ltd Internet survey, July 1998), 333 millions in 2000 (NUA Ltd Internet survey, June 2000). The number of accessible pages has been estimated in December 1997 at 320 millions [20], in February 1999 at 800 millions [21] and in July 2000 at more than 2 billions [26].

The Web is a huge and sometimes chaotic information space without central authority. In this context, and in spite of standardization efforts, these documents are very heterogeneous in their contents as in their presentations: HTML standard is respected in less than 7% of HTML pages [4]. We can expect to find almost everything there, but retrieving relevant information seems to be like *Finding the Needle in the Haystack*...

Nowadays, the great challenge for research in IR is to help people to profit of the huge amount of resources existing on the Web. But it exists yet no approach that satisfies this information need in a both effective [1] and efficient [2] way. For assisting the user in his task, some search engines (like Altavista, AllTheWeb, or Google [3]) are available on the Web. They are able to process huge documents volumes with several tens

[1] measure of IR-tool quality, evaluated classically using precision and recall measures
[2] measure of system resources use: memory usage, network load, etc...
[3] http://www.altavista.com, http://www.alltheweb.com, http://www.google.com

of million indexed pages. They are nevertheless very fast and are able to solve several thousands of queries per second. In spite of all their efforts, the answers provided by these systems are generally not very satisfactory. Preliminary results obtained with a test collection of the TREC conference Web Track has showed the poor results quality of 5 well known engines of the Web, compared to those of 6 systems taking part to TREC [17].

In fact, most of existing search engines use well-known techniques like those described by Salton 30 years ago [30]. Most of them prefer a wide coverage of Web with a low indexing quality to a better indexing on a smaller part of the Web. In particular, they consider generally HTML pages as atomic and independent documents, without taking into account relations existing between them. The notion of document for a search engine is reduced to its physical appearance, a HTML page. But Web's structure is used in few of Web search engines like Google [6].

With an aim of Structured IR, we wanted to determine which structure exists on the Web, and which structure it is possible to extract. This paper is organized as follow: after presentation of related works in section 2 (i.e. IR with structured documents, hypertexts and Web), we will present our hypotheses about what will be an ideal structure on the Web in section 3.1. Then we will propose our approach to validate our hypothesis and check if this kind of structure exists on the Web in section 3.2. Finally we will introduce the Web sample that we have analysed in section 4.1 and some preliminary results of our experimentations in sections 4.2, 4.3 and 4.4, while section 6 gives a conclusion about this work-in-progress and some future directions of our works.

# 2 IR and structure on the Web

The Web is not only a simple set of atomic documents. The HTML standard allows description of structured multimedia documents, it is widely used to publish on Web. Furthermore, Web is an hypertext, with URL's use (Uniform Resource Locator) for the description of links. This structure was used for IR, as well in the context of structured documents as in the context of classical hypertexts. We distinguish 3 main approaches proposing techniques of information access using structure: navigation, DBMS and IR approach.

## 2.1 Navigation approach

Navigation is based on links, used for finding and consulting some interesting information. In the case of a navigation within a hypertext composed by several hundreds of nodes, this solution can be useful. This task is more difficult to achieve on larger hypertext, mainly because of disorientation and cognitive overload problems. Furthermore, it is necessary to have the right links at the right place. A solution is proposed by "Web Directories" as Yahoo or Open Directory Project [4] which propose an organized hierarchy of several millions of sites. These hierarchies are built and verified manually, and thus it is expensive and difficult to keep them up-to-date. Furthermore, exhaustiveness is impossible to reach.

## 2.2 DBMS approach

Documents are represented using a data schema encapsulated in a relational or Object Oriented [10] data schema. It allows an interrogation using a declarative query language based on an exact matching and forced by the data schema structure. The hypertext structure integration in the database schema has been much studied, for example by [25], [3] (ARANEUS project), [16] (TSIMMIS project), etc. Integration attempts at the level of query language can be found in *hyperpaths* [1] or POQL [10]. In fact, these approaches are extensions of the proposed solution for the documents structure integration.

## 2.3 IR approach

IR approach deals with structured documents, promoting a hierarchical indexing : during the indexing process, information is propagated from document sections to top of document, along composition relations.

[4] http://www.yahoo.com, http://www.dmoz.org

This method is refined by Lee [22] who distinguishes several strategies of ascent. Paradis in [28] distinguishes several structure types, data ascent depending on different link types.

Hypertext structure has been taken into account at the indexing step. For example, the hypertext graph can be incorporated into a global indexing schema using conceptual graph model [9] or using inference networks [11]. World Wide Web Worm [24] enables the indexing of multimedia documents by the use of the anchor's surrounding text. Amitay [2] promotes also document's context use. Marchiori [23] adds the "navigation cost notion" that expresses the navigation effort for reaching a given page.

SmartWeb [13] considers the accessible information of a Web page at indexing step, so page relevance is evaluated considering the page content but also the page's neighbors content. Kleinberg (HITS [18]) promotes the use of both links directions: he introduces the hub page[5] and authority page [6] concepts. For automatic ressources compilation, the CLEVER system [8] based on the same idea, obtains good results against manually generated compilation (Yahoo![7]). Gurrin [15] has tried to improve Kleinberg's approach. He distinguishes 2 links types (structural and functional) and uses only structural ones. The well-known Google search engine [6] uses textual anchors to describe pages referenced by links from these anchors.

### 2.4 Related Works : Discussion

We think that navigation approach is well adapted to manually manageable collections, but the Web is too big to be acceded only with navigation. Navigation can be an interesting help to other techniques, for example to consult search results.

About DBMS approaches, we think that a declarative query language is not adapted to the Web heterogeneity. Moreover, these approaches rely on the underlying data base schema, and Web pages have to be expressed following this schema or following predefined templates. According to Nestorov [27] we think that even if some Web pages are strongly structured, this structure is too irregular to be modeled efficiently with structured models like relational or object.

IR approach enables natural language querying, and considers relevance in a hypertext context. At present, most of the IR approaches are based on pages connectivity use, with the notion of relevance propagation along links. The drawback is the bad use of this information because of the fact that relations (links) and nodes (documents) are not typed on the Web.

We think that these approaches are interesting and useful. The lack of explicit Web structure to improve them encourages us to work on Web structure extraction. Several works have focused on statistics studies [5], [4], [31], dealing with the use of HTML tags or with the links distribution which leads for example to the notion of hub and authority pages. Pirolli [29] has categorized Web pages following 5 predefined categories which are related to site structure, based on usage, site connectivity and content data. Broder [7] has studied the Web connectivity and has extracted a macroscopic Web structure. But none of these works deals with Web structure (structured documents and hypertexts) extraction related to IR objectives.

## 3 Is the Web well structured?

The main objectives of our Web sample analysis are to identify the Web explicit structure, and to extract the Web implicit structure. Obviously, the Web is clearly not structured in the DataBase sense of the term. But HTML allows people to publish structured sites. Thus we will talk about hierarchically structured Web as well as structured in the hypertext sense. The question is : "Is the Web sufficiently structured (especially hierarchically) to index it following a structured IR model?". This main objective leads us to some other interesting questions like : "What is a Document on the Web" or "How can I classify a Web link?".

We present our approach to answer these questions. Firstly, in section 3.1 we present the kind of structure that we wanted to identify/extract from the Web. We hypothesize that this *ideal structure* for the Web exists. The underlying problematic is about a structured IR model: our final goal is to develop an IR model adapted to Web.

[5] A page that references a lot of authorities pages.

[6] A page that is referenced by a lot of hubs pages.

[7] www.yahoo.com

Secondly we will present in section 3.2 our interpretation of HTML use related to our hypothesis. Web's structure depends mainly of the HTML use by sites authors, thus finally we will present in section 4 some preliminary results of an Web sample analysis.

### 3.1 Hypotheses

We will try to check the following assumptions, directly related to the concept of information in a semi-structured and heterogeneous context:

**Hypothesis 1: information granularity.** We think that information units on the Web can have different granularities. By assembling these various constituents, one can build entities of more important size. We distinguish at least 6 main granularities, and hypothesis 1 is detailed following these 6 granularities:

- **H1.1: elementary constituent.** We think that on the Web there is the notion of elementary constituents, which can be composed using a morpheme, a word or a sentence. In our approach, elementary constituent is at the sentence level.
- **H1.2: paragraph.** By assembling sentences, one can build paragraph-sized entities. This is our first level of structure. This structure is a list, reflecting the logical sequence existing in order to constitute an understandable unit.
- **H1.3: document section.** This second level includes all the elements that composes a "classical" document, like sub-sections, sections, chapters, etc. All of them are built using paragraphs. They include also some other attributes like title, author, etc.
- **H1.4: document.** This third level is the first one that introduces a tree like structure, based on document sections. Moreover, reader must follow a reading direction for a better understanding. For example, people generally read "introduction" before "conclusion".
- **H1.5: hyper-document.** This level loses the reading organization when gluing documents. This level can be associated with parts of hypertext, where a reading direction is not obligatory any more.
- **H1.6: clusters of hyper-document.** This last level is useful to glue the hyper-documents that have some characteristics in common, like the theme or the authors. This can be seen as the library shelf metaphor.

**Hypothesis 2: relations.** There are various relations between documents, whatever their granularity. We distinguish at least 3 main relations types, and hypothesis 2 is detailed following these 3 types:

- **H2.1: composition.** This relation expresses the hierarchical (tree-like) build of higher granularity entity. This relation is used in the five first levels of the previous granularity description (i.e. paragraphs are composed by sentences). Composition deals with attributes shared along composition relations, for example author name. It also deals with the compound element lifetime: a paragraph doesn't exist any more without its sentences. The composition can be split in weak and strong composition according to the sharing status. The composition is weak if an element can be shared. In this case the relation draws a lattice, otherwise we obtain a tree.
- **H2.2: sequence.** Certain documents parts can be organized by the author in an orderly way: part B precedes part C and follows part A. This order suggests a reading direction to the reader, for a best understanding. This relation only concerns **H1.1** to **H1.4**, it can be modeled using the probability that a part $A$ can be best understood after the reading of a part $B$. This conditional probability value can be the fuzzy value of the sequence from $B$ to $A$.
- **H2.3: reference.** This relation is weak, in the sense that it can link elements at any granularity level because they have something in common. For example, an author can refer to another document for a complementary information or two documents can refer each other because of their similarity.

The next generation of Web search engines will have to consider all these granularities and relations. In the next section, we interpret the HTML usage on the Web, in relation with these hypotheses.

### 3.2 Web analysis to validate assumptions

Our objectives are to study different Web characteristics, with an aim of validating our hypotheses. Without considering under-sentences granularities, we have made the hypothesis that it exists 6 main granularities on the Web (cf section 3.1), from sentence level until cluster of hyper-documents level. To validate hypothesis 1.1, 1.2 and 1.3, we have chosen HTML tags as describing inside-page granularities.

**H1.1** It is possible with HTML to describe elementary constituents, with `<ADDRESS>` or `<CODE>`. Several are at the presentation level, others at the semantics level. We place our analysis at the sentence level, and we do not have found a lot of tags that explicitly isolate sentence like `<CITE>` do. All others tags are internal sentence elements.

**H1.2** We propose to place at this level simple paragraphs and "blocs elements" like `<TABLE>` or `<FORM>`. It exists sub-blocs elements like `<PRE>` that we place also at this level. Of course we propose to use paragraphs separators `<P>`, `<HR>`.

**H1.3** To express document sections, one can use HTML separators `<Hn>`. In fact, we could use the whole Web page as a section.

**H1.4** We propose to consider the physical HTML page as a document. But we could also take a set of interconnected Web pages as document assuming that links between them represent composition.

**H1.5** The first proposition we can do is to consider an hyper-document to be an Internet site which is defined as a set of pages on the same site.

**H1.6** To represent our cluster of hyper-document, we propose the notion of Web domain (i.e. ".imag.fr").

To validate hypothesis 2, we have tried to identify composition and sequence links. All unidentified links are categorized as reference links. Implicit similarity and reference relations are not extracted.

**H2.1** Composition can be identified by inside-pages **H1.3** tags, representing strong composition. Also, inside-sites links can be identified as hierarchical, representing strong or weak composition.

**H2.2** The sequence can be found by looking at the implicit position of a fragment relatively to the following text segment (inside-pages). Also, some inside-sites links from a page to one of its sisters can be considered.

**H2.3** All the remaining links are classified in this category. This type of relation can be represented on the Web using hypertext links. But it can also be implicit, like quotations for example.

It is possible to describe a structure, but is it a reality on the Web? We have to verify if these sub-page granularity tags are used by authors (**H1.1** to **H1.3**), and we have to check if the concept of page, site and domain are relevant on the Web (**H1.4** to **H1.6**). For each page, we have to rebuild hierarchical tree structure, and to identify a structured documents hierarchy between HTML pages.

## 4 Experiments results

We will present in this section some preliminary results about a Web sample analysis, and particularly statistics used to validate our assumptions.

### 4.1 Web pages sample: IMAG collection

We have collected an "October 5 2000 snapshot" Web sample, using our Web crawler "CLIPS-Index" (cf section 5). We have chosen to restrict our experiment to the Web pages of the IMAG domain[8], which are browsable starting from URL "http://www.imag.fr". These pages deal with several topics, but most of them

[8] Institut d'Informatique et de Mathématiques Appliquées de Grenoble : hosts which name is ended by .imag.fr

are scientific documents, particularly in computer science field. Main characteristics of this collection are summarized in figure 1.

| | #per page | # in coll |
|---|---|---|
| Hosts | | 39 |
| Pages | | 38’994 |
| French language | | 5’649 |
| English language | | 23’819 |
| Others language | | 9’068 |
| Distinct terms | | 241’000 |
| Size (HTML) | 11,6 Ko | 443 Mb |
| Size (text) | 3,7 Ko | 141 Mb |
| Lines | 207 | 8’079’676 |
| Links | 37,8 | 1’475’096 |

Figure 1: IMAG sample: main characteristics

| Extension | #pages | % |
|---|---|---|
| .html | 25’665 | 65,82 |
| .htm | 2’530 | 6,49 |
| .java | 1’021 | 2,62 |
| .cgi | 219 | 0,56 |
| .txt | 82 | 0,21 |
| .php3 | 71 | 0,18 |
| No extension | 8’134 | 20,86 |
| Directory | 933 | 2,39 |
| Others | 339 | 0,87 |
| Total | 38’994 | 100 |

Figure 2: Pages format

Our spider has collected, taking less than 2 hours, almost 39.000 pages which are identified by their URL from 39 hosts, for a size of 443 Mb. It is not surprising that most of the pages are in HTML format (72 % of .html and .htm, cf figure 2). After analysis and textual extraction, it remains about 140 Mb of textual data containing more than 241.000 distinct terms.

### 4.2 Granularity analysis

We have extracted statistics related to entities granularities described in section 3.1. It appears that HTML ability to represent different inside-page granularities as described in section 3.2 is widely used: each page contains on average 17 level 1 objects, 17 blocs elements, 29 paragraphs separators and 3,3 section (cf figure 3). **Hypothesis 1.1, 1.2 and 1.3** seem to be correct, but need manual experiments to be validated.

| Level | Object | #objects |
|---|---|---|
| H1.1 | Level 1 objects | 663’000 |
| H1.2 | Blocs elements | 659’000 |
| | Paragraphs separators | 1’142’000 |
| H1.3 | HTML separators | 130’000 |
| H1.4 | Pages | 38’994 |
| H1.5 | Sites | 39 |
| H1.6 | Domains | 1 |

Figure 3: Inside and outside-page granularity

Pages average size is 3,3 sections or 11,65 Ko (cf figure 1). This is greater than other studies results (almost 7 Ko [31], [5]). Textual pages size (pages without HTML tags) is on average of 3,69 Ko. But these statistics are related to physical aspects of documents. We have to consider entities linkages to conclude something about logical aspects. It exists on average 37 links per page in our collection (cf figure 4): if we don’t consider redundant links (same source and same destination), it remains only 550’000 distinct links: on average 14,11 per page, which is not far from other studies (13,9/page [31], 16,1/page [4]).

| Links | #links | % | Per page | Per site | Distinct | % | Per page | Per site |
|---|---|---|---|---|---|---|---|---|
| Inside-pages | 118’248 | 8 | 2,97 | 3’128 | 13’897 | 2,53 | 0,36 | 356 |
| Outside-pages | 1’318’490 | 89,38 | 33,81 | 33’807 | 500’472 | 90,96 | 12,83 | 12’832 |
| Outside-sites | 2’093 | 0,14 | 0,05 | 57,67 | 1’708 | 0,31 | 0,04 | 43,79 |
| Outside-domain | 36’265 | 2,46 | 0,93 | 930 | 34’130 | 6,20 | 0,87 | 875 |
| Total | 1’475’096 | 100 | 37,12 | 39’118 | 550’207 | 100 | 14,11 | 14’108 |

Figure 4: Links analysis: all/distincts links

Web pages are heavily linked together, but without link categorization it is difficult to distinguish which pages are hyper-documents, which are structured documents or which are sections (**Hypothesis 1.4**). Especially, we can't confirm that a Web document is represented by an HTML page.

| #links/page | #pages | % |
|---|---|---|
| 0 | 38'275 | 97,63 |
| 1 | 396 | 1,52 |
| 2 | 106 | 0,36 |
| 3 | 85 | 0,18 |
| 4 | 39 | 0,1 |
| 5 + | 93 | 0,21 |

Figure 5: Outside-sites links per page

There are a few outside-sites links: only 2,6% of them, contained by 2,4% of pages. Thus, we think that the site compactness validate **Hypothesis 1.5**: hyper-documents are represented by sites. Only 5,4% of these outside-sites links are inside-domain links: most of sites are connected with outside-domain sites, we conclude that a cluster of hyper-documents is not represented by a Web domain (**Hypothesis 1.6**).

### 4.3 Internal HTML page structure extraction

We have identified several internal pages levels (cf section 4.2): section, paragraph, sentence and even internal sentence. These levels are defined by HTML pages writers. With these structure elements (cf figure 3), we are able to rebuild hierarchical tree structure which are relatively large (cf figure 6).
Particularly, we have to extract most of the composition and sequence relations. Composition relations are implicit, from page to all its sections, but also from sections to all their paragraphs, etc. Sequence relations are also implicit, from each page element (except the last one) to its physical successor. Hypertext links which do not correspond to an extracted composition or sequence relation are supposed representing reference relations.

| Depth | #pages | % |
|---|---|---|
| 0 | 2'995 | 7,68 |
| 1 | 1'963 | 5,03 |
| 2 | 2'227 | 5,71 |
| 3 | 5'354 | 13,73 |
| 4 | 12'159 | 31,18 |
| 5 | 4'101 | 10,52 |
| 6 | 8'345 | 21,4 |
| 7 | 1'083 | 2,78 |
| 8 et + | 767 | 1,97 |

Figure 6: Internal structure extraction

### 4.4 External HTML page structure extraction

We make the assumption that Web site directory structure includes some semantics that can be automatically extracted. This semantics is proposed "a priori", because we suppose the manner that pages are placed in the directory hierarchy follows the "principle of least effort". We assume the directory hierarchy reflects the composition relation. It must of course be validated by experimentation using manual validation. We examine in this part all ways links are joining pages across the directory hierarchy and we propose the following links categories: **internal** (inside-page), **hierarchical** and **transversal** (inside-site), **cross** (outside-site) and **out** (outside-domain) (cf figure 7).

| Types | #links | % |
|---|---|---|
| Internal | 118'248 | 8,02 |
| Hierarchical | 880'421 | 59,69 |
| Transversal | 438'069 | 29,70 |
| Cross | 2'093 | 0,14 |
| Out | 36'265 | 2,46 |

Figure 7: Links types

We are interested by categorizing the relations represented by these links. We have interpreted each link type to type the relation that it expresses, in the following way:

**Internal** 8% of links stay in the same page. We have no proposition for their category.

**Hierarchical** We call hierarchical links those whose the source and target are in the same directory path. These links are the most common in our sample with 60%. If these links reflect the composition

structure, we can deduce that this sample is strongly structured.

**Transversal** The target of the link is neither in the ascendant directories nor in the descendant directories but is in the same site. There are 30% of links in this category. We can probably classify them in the weak composition or in reference links.

**Cross site** The target is on an other site: only 0,1% are concerned. They are candidates to be reference.

**Outside IMAG** Target is outside IMAG domain: only 2,5%, they are also candidates to be reference.

We detail hierarchical links in 3 categories: **horizontal, up and down**.

| #up-levels | #links | % |
|---|---|---|
| same directory | 432'359 | 29,31 |
| +1 | 223'717 | 15,17 |
| +2 | 51'330 | 3,48 |
| +3 | 48'232 | 3,27 |
| +4 | 37'185 | 2,52 |
| +5 | 11'573 | 0,78 |
| +6 | 450 | 0,03 |
| +7 | 2 | 0 |
| Total | 804'848 | 54,56 |

Figure 8: Hierarchical links: up/horizontal

| #down-levels | #links | % |
|---|---|---|
| -8 | 8 | 0 |
| -7 | 548 | 0,04 |
| -6 | 137 | 0,01 |
| -5 | 764 | 0,05 |
| -4 | 4'455 | 0,30 |
| -3 | 8'583 | 0,58 |
| -2 | 10'118 | 0,7 |
| -1 | 50'960 | 3,45 |
| Total | 75'573 | 5,12 |

Figure 9: Hierarchical links: down

**Horizontal** Target is in the same directory. These links are candidate to express a **sequence relation**. In our experiment 29% of links are in this category (cf figure 8).

**Up** Source is deeper in the directory path. These links go up in site hierarchy. This is the second ranking value with 25% (cf figure 8). It exists more links going up than going down, we think that this is caused by a lot of *"Back to the Top"* links.

**Down:** Target is deeper in the directory path. These links are less frequent (5%) (cf figure 9) than other hierarchicals. We think that they could belong to the **composition** hierarchy.

We conclude that the directory path is not built in a random manner: the majority of the links follow it. The site also seems to have a strong consistency: 98% of the links are inside-sites links.

# 5 Technical details: CLIPS-Index and Web pages analysis

We have developed a robot called CLIPS-Index[9] in collaboration with Dominique Vaufreydaz (from GEOD team), with the aim of creating Web corpora. This spider crawl the Web, collecting and storing pages. CLIPS-Index tries to collect the bigger amount of information in this heterogeneous context which is not respectful of the existing standard. It is an interesting problem to collect the Web correctly. In spite of this, our spider is quite efficient: for example, we have collected (October 5 2000) 38'994 pages on the **.imag.fr** domain, comparatively to Altavista which index 24.859 pages (October 24 2000) on the same domain and AllTheWeb which index 21.208 pages (October 24 2000). 3,5 millions pages from french-speaking Web domains where collected during 4 days, using a 600Mhz PC with 1 Gb RAM. CLIPS-Index crawls this huge hypertext without considering non-textual pages, and respects the robot exclusion protocol [19]. It does not overload distant Web servers, despite the launching of several hundred HTTP queries simultaneous. CLIPS-Index, running on an ordinary 333Mhz PC with 128Mo RAM which cost less than 1.000 dollars, is able to find, load, analyze and stock something like $\frac{1}{2}$ millions pages per day.

[9] http://CLIPS-Index.imag.fr

We have also developed several analysis tools (23'000 lines) using PERL (Practical Extraction and Report Language) for HTML extraction, links analysis and typing, topology analysis, statistics extraction, text indexing, language extraction, etc.

## 6 Conclusion and future works

We think that it is interesting and useful to use Web structure for IR. Because of the lack of Web explicit structure, we have to identify explicit structure and extract implicit one. We have proposed a framework composed by 6 entities granularities and 3 main relations types. We have proposed some rules to extract these granularities and relations, based mainly on HTML possibilities to describe structured elements, and on study of relations that exist between hypertext links and Web server directories hierarchy.

Our first experiments show that, in a first hand the hypotheses H1.1, H1.2, H1.3 (internal structure level) and H1.5 (site granularity) seem to be correct, and in a second hand that hypotheses H1.4 (page granularity) and H1.6 (cluster of sites granularity as internet sub-domains) seem to be false.

It is possible to identify and extract structure from the Web: several granularities and several types of relations. But we have to continue these experiments. Firstly, we have to improve our relations categorization and our hierarchical structure extraction. Secondly, we need to check extracted informations manually, to validate our hypotheses. Thirdly, we have to analyze bigger collections: several domains, more heterogeneous pages. IMAG collection is undoubtedly not very representative of the Web, because of its small size compared to French Web. Moreover it represents only a single Web domain. And finally, our main objective is to propose a structured IR model, based on these 6 granularity levels and 3 relations types. An Information Retrieval System based on this model will use some IR methods used in the context of structured documents and hypertexts [12]. It will actually use Web structure for IR, and thus will be able to help facing the IR problem on the Web. We are also working on the use of DataMining techniques for extracting useful knowledge for improving IR results [14].

## References


[1] B. Amann. *Interrogation d'Hypertextes*. PhD thesis, Conservatoire National des Arts et Métiers de Paris, 1994.

[2] E. Amitay. Using common hypertext links to identify the best phrasal description of target Web document. In *Conference on Research and Development in IR (SIGIR'98)*, Melbourne, Australia, 1998.

[3] P. Atzeni, G. Mecca, and P. Merialdo. Semistructured and structured data in the Web : Going back and forth. In *Workshop on Management of Semistructured Data*, Tucson, 1997.

[4] D. Beckett. 30 % accessible - a survey to the UK Wide Web. In *World Wide Web Conference (WWW'97)*, Santa Clara, California, 1997.

[5] T. Bray. Measuring the Web. In *World Wide Web Conference (WWW'96)*, Paris, France, May 1996.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *World Wide Web Conference (WWW'98)*, Brisbane, Australia, 1998.

[7] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. In *World Wide Web Conference (WWW'00)*, Amsterdam, Netherlands, 2000.

[8] S. Chakrabarti, B. E. Dom, R. K. David Gibson, P. Raghavan, S. Rajagopalan, and A. Tomkins. Spectral filtering for resource discovery. In *Conference on Research and Development in IR (SIGIR'98)*, Melbourne, Australia, 1998.

[9] Y. Chiaramella and A. Kheirbek. An integrated model for hypermedia and information retrieval. In M. Agosti and A. F. Smeaton, editors, *Information Retrieval and Hypertext*. Kluwer Academic Publisher, 1996.

[10] V. Christophidès and A. Rizk. Querying structured documents with hypertext links using OODBMS. In *European Conference on Hypertext Technology (ECHT'94)*, Edinburgh, Scotland, 1994.

[11] W. B. Croft and H. Turtle. A retrieval model for incorporating hypertext links. In *ACM Conference on Hypertext (HT'89)*, Pittsburg, USA, 1989.

[12] F. Fourel, P. Mulhem, and M.-F. Bruandet. A generic framework for structured document access. In *Database and Expert Systems Applications (DEXA'98), LNCS 1460*, Vienna, Austria, 1998.

[13] M. Géry. Smartweb : Recherche de zones de pertinence sur le world wide web. In *Congrès INFORSID'99*, La Garde, France, 1999.

[14] M. Géry and M. H. Haddad. Knowledge discovery for automatic query expansion on the World Wide Web. In *Workshop on the World-Wide Web and Conceptual Modeling (WWWCM'99), LNCS 1727*, Paris, France, 1999.

[15] C. Gurrin and A. F. Smeaton. A connectivity analysis approach to increasing precision in retrieval from hyperlinked documents. In *Text REtrieval Conference (TREC'99)*, Gaithersburg, Maryland, 1999.

[16] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting semistructured information from the Web. In *Workshop on Management of Semistructured Data*, Tucson, 1997.

[17] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman. Results and challenges in Web search evaluation. In *World Wide Web Conference (WWW'99)*, Toronto, Canada, May 1999.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environnement. In *ACM-SIAM Symposium on Discrete Algorithms (SODA'98)*, San Francisco, California, 1998.

[19] M. Koster. A method for Web robots control. Technical report, Internet Engineering Task Force (IETF), 1996.

[20] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280, April 1998.

[21] S. Lawrence and C. L. Giles. Accessibility of information on the Web. *Nature*, July 1999.

[22] Y. K. Lee, S.-J. Yoo, and K. Yoon. Index structures for structured documents. In *ACM Conference on Digital Libraries (DL'96)*, Bethesda, Maryland, 1996.

[23] M. Marchiori. The quest for correct information on the Web : Hyper search engines. In *World Wide Web Conference (WWW'97)*, Santa Clara, California, 1997.

[24] O. A. McBryan. GENVL and WWWW: Tools for taming the Web. In *World Wide Web Conference (WWW'94)*, Geneva, Switzerland, 1994.

[25] A. Mendelzon, G. Mihaila, and T. Milo. Querying the World Wide Web. In *Conference on Parallel and Distributed Information Systems (PDIS'96)*, 1996.

[26] A. Moore and B. H. Murray. Sizing the internet. Technical report, Cyveillance, 2000.

[27] S. Nestorov, S. Abiteboul, and R. Motwani. Inferring structure in semistructured data. In *Workshop on Management of Semistructured Data*, Tucson, 1997.

[28] F. Paradis. Using linguistic and discourse structures to derive topics. In *Conference on Information and Knowledge Management (CIKM'95)*, Baltimore, Maryland, 1995.

[29] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear : extracting usable structures from the Web. In *Conference on Human Factors in Computing Systems (CHI'96)*, Vancouver, Canada, 1996.

[30] G. Salton. *The SMART retrieval system : experiments in automatic document processing*. Prentice Hall, 1971.

[31] A. Woodruff, P. M. Aoki, E. Brewer, P. Gauthier, and L. A. Rowe. An investigation of documents from the World Wide Web. *Computer Networks and ISDN Systems*, 28, 1996.