Web as Huge Information Source for Noun Phrases Integration in the Information Retrieval Process

Mathias Géry — M. Hatem Haddad — Dominique Vaufreydaz
CLIPS-IMAG Laboratory
MRIM Team (Modeling and Multimedia Information Retrieval)
B.P. 53, 38041 Grenoble Cedex 9, France
{Mathias.Gery, Hatem.Haddad, Dominique.Vaufreydaz }@imag.fr

ABSTRACT: *Web is a rich and diversified source of information. In this article, we propose to benefit from this richness to collect and analyze documents, with the aim of a relational indexation based on noun phrases. Proposed data processing chain includes a spider collecting data to build textual corpora, and a linguistic module analyzing text to extract information. Comparison of obtained corpus with corpus from Amaryllis conference shows the linguistic diversity of collected corpora, and particularly the richness of extracted noun phrases.*

KEYWORDS: *Textual corpora, Web analysis, Noun phrases extraction, Information Retrieval (IR)*

1. Introduction

The Web growth constitutes a new applicability field for IR: we find almost everything there, and retrieving relevant information looks like *Finding the Needle in the Haystack*! New methods must be developed, dealing with heterogeneous context. Natural Language Processing (NLP) was well studied for IR, as mentioned during TREC "NLP tracks" **[TREC]**. NLP can be integrated into several Information Retrieval System (IRS) components: at indexing time (good terms identification to represent content), at querying time (query analysis or interactive query reformulation) and at matching time (dynamic NLP integration). We work on textual content indexing, with the aim to use a smarter NLP than words truncation or stop-words remove. We propose to use noun phrases for documents representation, instead of the restrictive use of simple words. Noun phrases denote generally a specific class of mental objects, which interpretation is usually precise.

To experiment such methods, people use test collections as those proposed during TREC conference **[TREC]** or during French-speaking Amaryllis conference (INIST, OFIL **[AMA]**). These collections are composed by corpus of documents having a common origin: articles of newspaper "*Le monde*" (OFIL) or scientific records (INIST). We hypothesize that the quality of extracted information is directly related to the corpora quality. Thus, experiment new IR methods (linguistic or statistical) requires a great corpus richness. That is not the case in classic corpora, particularly regarding topics diversity. So, experiments are often restricted to specific processes, and restrict information extraction on an unique field.

On the other hand, we have showed that the Web is a very interesting source for spoken language modeling **[VAU01]**. Training such a language model needs a great diversity of words, and the Web is very useful for this task compared to other corpora. Web is a huge and heterogeneous information space: the number of users has been estimated at 119 millions in 1998, 333 millions in 2000 and 500 millions in 2001 **[NUA]**, evolving from simple "readers" to "writers". The number of accessible pages has increased from 320 millions (1997, **[LAW98])** to more than 2 billions (2000, **[MUR00]**). Furthermore, pages are more and more diversified: they handle almost all possible topics, in a lot of languages, and using various forms of expressions. Another interesting Web characteristic is the dynamic aspect of its content, implying a constant evolution of terminology, unlike classic collections that have a fixed vocabulary. So, this huge data amount is very interesting to build rich, diversified and large corpora.

We present in this article the building of textual corpora and noun phrases extraction. We discuss about advantages of the Web regarding corpora like Amaryllis, with the aim of a richer documents representation. We begin by presenting methods of noun phrases extraction

for IR in the 2nde section, especially the method that was used during our study. In the 3rd section, we detail the processing chain, which allows to extract corpora from the Web, to process them and to extract noun phrases. In the 4th section, we present characteristics of collected corpora compared to classic corpora. Finally, we analyze in the 5th section the results obtained by our experiments of knowledge extraction from these corpora.

2. Noun phrases extraction

Both statistical and linguistic approaches are used in noun phrases extraction. The common statistic used is frequency of potential noun phrases and words combinations discovery, according to their appearance regularity **[CHU90], [FAG89]**. These statistical methods allow covering in an exhaustive way all the possible terms combinations, in a window going from bigram to whole document. A drawback is the huge quantity of possible combinations in large corpora: some of them are valid on a statistical point of view but are not semantically correct. Lexter system uses linguistic method **[BOU92]**, arguing that terminological units obey to specific rules of syntactic formation, and for the non-necessity of complete syntactical analysis, replaced by a surface grammatical one. Lexter deals with noun phrases mainly consisting of adjectives and nouns. It analyses and parses a corpus tagged with a part-of-speech tagger. During *analysis*, noun phrases with a maximum length are extracted, regarding potential "terminological frontiers" (pronouns, verbs, etc.). During *parsing*, sub-strings are extracted from extracted noun phrases according to their position within the maximum length noun phrases. **[DAI94]** extracts only 2-word terms according to morpho-syntactic criteria, allowing variations on terms. A statistical score (likelihood ratio) is applied as an additional filter to the candidate terms extracted.

Unlike most of the noun phrases extraction methods, our approach has to be as general as possible to handle any application domain, and particularly the Web. Thus, we based it on the most used morpho-syntactical patterns of a language (French language in our study). Given the huge information amount, an appropriate linguistic treatment should be applied. NLP needs a robust and exhaustive language analysis, too complex for the SRI aimed objective. For this reason, we adopt a superficial analysis, which eliminates the deep structure determination and takes into account only the noun phrases extraction.

3. Data processing chain

We present the processing chain and its 3 main components: the spider (collecting raw data from the Web), corpus analyzer (building standardized textual corpora), and the linguistic analysis module (IRS IOTA **[CHI86]** extracting noun phrases). The outline of the processing chain is presented in Figure 1.

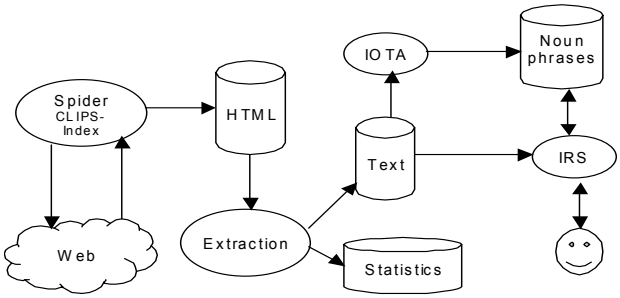

Figure 1: Data processing chain

3.1. CLIPS-Index spider

We have developed a spider called *CLIPS-Index* **[CLI]**, in collaboration with GEOD group (CLIPS laboratory). It crawls the Web, collecting and storing pages, with the aim of creating Web corpora. *CLIPS-Index* tries to collect the larger amount of data on a given domain, in the Web heterogeneous context (the existing standards HTML, HTTP and URL are rarely used properly). Thus, we have to find a compromise between efficient crawl and errors management. *CLIPS-Index* execute the following steps:

- Getting an URL from an existing URL-to-collect repository.
- Collecting the HTML page.
- Analyzing HTML page and extracting a URLs list.
- Storing the HTML page collected.
- Adding new URLs to the URL-to-collect repository.

CLIPS-Index is based on a multithreaded architecture, which allows launching hundreds of HTTP request simultaneously to Web servers. Our spider has several important problems to address. It has to assume the synchronization between all the processes, to avoid collecting the same URL two times. It has also to manage an URL repository accessed hundred times per second, and containing up to several hundreds millions URLs. Despite its fast

collect, *CLIPS-Index* has to be very careful with Web servers. Firstly, it respects the spider control method **[KOS96]** allowing webmasters to choose which parts of their site should be collected. Secondly, it considers a delay between two requests on the same Web server, avoiding to overload Web servers despite the launching of several hundreds of requests per second.

C*LIPS-Index* is fast: running on an ordinary low-cost 500Mhz PC with 128Mo RAM (less than 1.000 dollars), it is able to find, load, analyze and store up to 3 millions pages/day. Its parser is also efficient: for example, we have collected 38'994 pages on the "*.imag.fr*" domain (October 5th 2000), comparatively to AltaVista and AllTheWeb which index 24'859 pages (resp. 21'208) on this domain (October 24th 2000). Tests using GNU "wget" give worst results. *CLIPS-Index* has a robust parser and URL extractor, which are able to deal with the fact that less than 7% of HTML pages are HTML-valid **[BEC97]**.

3.2. Treatment and normalization of Web corpora

This phase consists in normalizing raw data collected, with the aim to obtain files following several formats (for example, TEI format for IOTA or a specific format for the SRI SMART):

- Text extraction from HTML, which should be robust and must give a correctly punctuated text (with the aim of linguistic treatments on the scale of the sentence).
- Mirrors elimination (servers aliases, sites mirrors, etc).
- Lexicon extraction, and calculation of the corpus lexical coverage.
- Statistics extraction, like language or information about Web pages structure **[GER01]**.

3.3. IOTA system

We used the IOTA system for the morpho-syntactic analysis and the noun phrases extraction. The morpho-syntactic analyzer is a surface analyzer which uses a dictionary associated to a morphological model. It is very careful about the non-recognized forms. It allows to extract potential interpretation to a non-recognized form, using a manually-planned resolution corresponding to typical ambiguity cases which the resolution is known. This module output is a labeled corpus. The global word frequency and the word frequency according to a window are calculated. Then, the labeled corpus is used to extract noun phrases, locating syntactical categories borders (we consider for example that a noun phrase begins with a noun or an adjective). A syntactic filter allows to keep only the valid noun phrases regarding the set of morpho-syntactical patterns. These patterns are generic French language patterns ("Noun-Noun", "Noun-Preposition-Noun", etc.).

4. Corpora gathering and characteristics

We have compared classic corpora (INIST, OFIL) with two corpora extracted from Web:

- "***Tunisia***": a relatively small corpus containing pages collected on the ".tn" domain, with a majority of French-speaking documents representative of a country.
- "***Newspapers***" (**NP** and **NP2**): a large French-speaking textual corpus containing pages collected on newspapers Web sites, with a good quality in the use of the French language.

A *CLIPS-Index* parameter is used to filter crawled sites: it is expressed using a regular expression on the sites names. The one used for "*Tunisia*" restrict to sites ending by "*.tn*". The filter used for "*NP*" was built automatically by a "*topical sites names extractor*", which aim is to crawl parts of a directory hierarchy (for example "/News_and_Media/Journals/" from Yahoo!) and to extract a filter from it.

Table 1 shows the main characteristics of these corpora, collected at various dates to analyze their evolution.. They were crawled until the last URLs on the required domains, obtaining low performances (2,87 to 9,44 doc/sec., against almost 30 usually), because of the difficulty to find last URLs on a domain.

Corpora	Crawl date	Crawl time	Number of docs	Docs/sec.
Tunisia	March 16 2001	1 h 08	38'523	9,44
Tunisia 2	August 22 2001	1 h 50	60'787	9,21
Tunisia 3	January 24 2002	7 h 49	109'162	3,88
NP	Nov. 7 2001	17 h 43	244'364	3,83
NP 2	January 11 2002	38 h 29	397'854	2,87

Table 1: Crawls characteristics

Each of these 5 corpora is analyzed and mirroring documents are eliminated to extract textual corpus. Table 2 shows their characteristics compared to classic corpora. The ratio between HTML size and textual size goes from 4,9 to 7,17 because of the heavy use of HTML tags for presentation. Moreover, HTML tools add more various data into pages: a minimal HTML page size is 74 bytes in HTML 4.01, 304 bytes using Netscape Composer and more than 2'000 bytes using Word!

Corpora	Number of docs	HTML/TEI size		Textual size	
		Corpus	Kb/page	Corpus	Kb/page
INIST	163'308	100 Mb	0,63	79 Mb	0,50
OFIL	11'016	33 Mb	3,06	32 Mb	2,93
Tunisia	27'959	161 Mb	5,90	27 Mb	1,00
Tunisia 2	43'651	397 Mb	9,31	55 Mb	1,30
Tunisia 3	79'361	863 Mb	11,13	165 Mb	2,13
NP	198'158	4'391 Mb	22,69	896 Mb	4,63
NP 2	345'860	7'728 Mb	22,88	1'491 Mb	4,41

Table 2: General characteristics of textual corpora

5. Corpora analysis and results

5.1. Languages distribution

Language distribution Language extraction is based on the frequencies of most common words for each language (English, French, Italian, German, Spanish, Dutch, Danish), calculating the proportion of "and, any, by, for, not, of, the, to, etc". These lists of words are extracted from a reference corpus. We notice a great majority of French-speaking pages, particularly in the "*NP*" corpora. The proportion of "Unknown" language extracted from Tunisian corpora is explained by a lot of pages without textual content (replaced by pictures), while pages from "*NP*" are always textual.

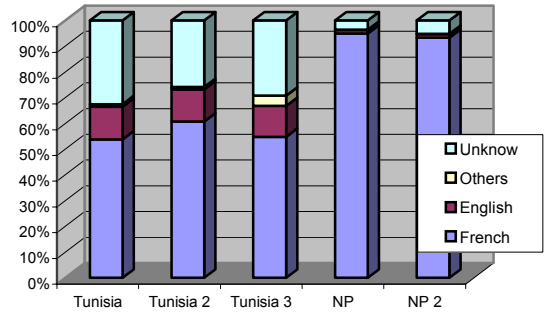

Figure 2: Languages distribution

5.2. Lexicon and French coverage

Table 3 shows distinct terms number for each corpus, total corpus terms number and total document terms number. Indeed, we obtain large corpora: "*NP 2*" is 30 times larger than Amaryllis corpora. Documents from newspapers collections ("*OFIL*", "*NP*", "*NP 2*") are on average larger than others.

Corpus	#terms	#terms/corpus	#terms/doc
INIST	174'659	8,31 millions	50,89
OFIL	119'434	5,15 millions	467,55
Tunisia	113'418	4,21 millions	150,61
Tunisia2	164'569	8,46 millions	193,70
Tunisia3	393'919	25,04 millions	315,57
NP	536'361	133,97 millions	676,07
NP 2	850'659	257,04 millions	743,19

Table 3: Terms number

We estimate corpora variety using lexical coverage, calculating the percentage of French lexical forms appearing in each corpus. A reference lexicon of 400'000 lexical forms was built from 2 lexicons: one containing more than 270.000 lexical forms **[ABU]**, and the other containing more than 300'000 lexical forms derived from BDLex lexicon **[CAL00]**. Figure 3 shows the lexical coverage of each corpus, which is larger for "*NP*" and "*NP 2*" than for classic corpora.

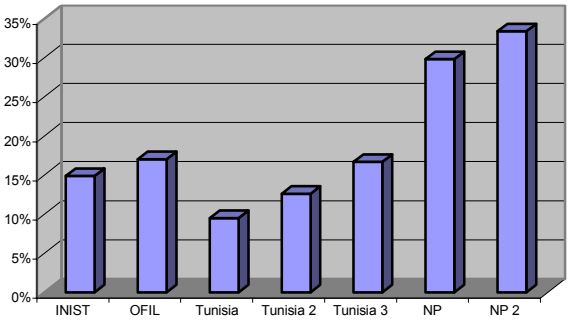

Figure 3: French coverage

	Average	OFIL	INIST	Tunisia	Tunisia 2	Tunisia 3	NP	NP2
Noun	29,75%	30,60%	33,62%	29,21%	29,30%	28,14%	28,88%	28,52%
Adjective	27,71%	27,25%	31,55%	26,48%	27,00%	26,15%	27,84%	27,71%
Noun proper	14,65%	18,38%	11,75%	12,96%	12,89%	13,79%	16,37%	16,42%
Others	27,88%	23,76%	23,09%	31,35%	30,80%	31,91%	26,91%	27,35%

Table 4: Grammatical categories distribution

5.3. Grammatical Categories Distribution

The words grammatical categories distribution is almost identical for every corpora. As shown in Table 4, the dominant categories are "Noun", "Adjective" and "Proper noun". This distribution has been used for the morpho-syntactical patterns selection, that takes into account mainly these three categories.

6. Noun phrases extraction

The average number per document is more important in the OFIL corpus as shown in Figure 5, because of the very high quality of the OFIL corpus. On the other hand, the noun phrases extracted are much more abundant in the Web corpora than in the Amaryllis corpora, as shown in Figure 4. So, Web corpora quality is lower than OFIL quality, but the huge size of available data allows to collect far more data and extract far more noun phrases. We notice that for the same corpus, some frequencies widely increased. For example, the noun phrase "higher education" not present in "*Tunisia*" collection and occurs with a frequency of 1773 in "*Tunisia 3*" corpus against only 992 in "*Tunisia 2*". New noun phrases appear from a collection to the other one, and reflect for example a media event such the noun phrase "Mediterranean games" which appears 972 times in "*Tunisia 2*" and 178 in "*Tunisia*".

7. Conclusion

In this article, we have shown the Web as an excellent source of data to build diversified textual corpora, with the aim to extract information for IR. We have compared qualitatively and quantitatively the corpora obtained with classic corpora from Amaryllis. The huge amount of available information on the Web allows building very large corpora. We have used them to apply linguistic methods of information extraction, but we have also shown that this huge amount of data is also very interesting to apply statistical methods **[GER99]**. However, the main advantage of these corpora comes from the dynamic aspect of Web and the great domains variety. Indeed, it allows to extract information covering many knowledge domains and to follow the vocabulary evolution. Quantity and quality of the extracted information offers many perspectives. We are developing an IR model based on a relational indexation integrating noun phrases into the indexing process. Implementation for the Web requires the use of appropriate corpora, which allows to extract knowledge reflecting a vocabulary used at a given period.

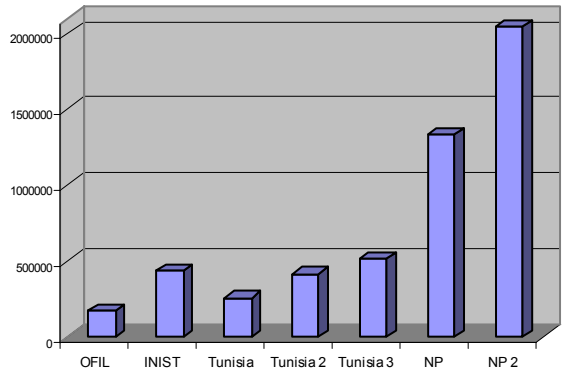

Figure 4: Noun phrases per corpus

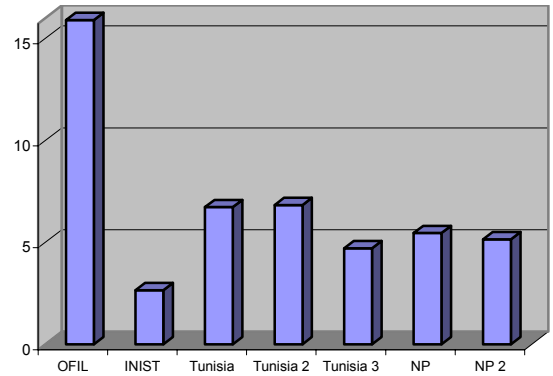

Figure 5: Noun phrases per document

8. Bibliography

[ABU] Association des Bibliophiles Universels, http://abu.cnam.fr

[AMA] Amaryllis conference, http://amaryllis.inist.fr

[BEC97] D. Beckett, 30% accessible - a survey of the UK Wide Web. WWW Conference, Santa Clara, California, 1997.

[BOU92] D. Bourigault, Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases, COLING, Nantes, France, 1992.

[CHI86] Y. Chiaramella, B. Defude, M.F. Bruandet and D. Kerkouba, IOTA: a full test Information Retrieval System, SIGIR, Pisa, Italy, 1986.

[CAL00] M. de Calmès and G. Pérennou, BDLEX: a lexicon for spoken and written French, LREC, Grenade, Spain, 1998.

[CHU90] K.W. Church and P. Hanks, Word association norms mutual information and lexicography, Computational Linguistics, vol. 16, n°1, 1990.

[DAI94] B. Daille, E. Gaussier and J.M. Lange, Towards Automatic Extraction of Monolingual and Bilingual Terminology, COLING, Kyoto, Japan, 1994.

[FAG89] J.L. Fagan, The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval, JASIS, vol. 40, n°2, 1989.

[GER99] M. Géry and H. Haddad, Knowledge discovery for automatic query expansion on the World Wide Web, WWWCM, Paris, France, 1999.

[KOS96] M. Koster, A method for Web robots control, technical report, IETF, 1996.

[LAW98] S. Lawrence and C.L. Giles, Searching the World Wide Web, Science, vol.280, n°5360, 1998.

[MUR00] B.H. Murray and A. Moore, Sizing the Internet, technical report, Cyveillance Inc., 2000.

[NUA] Nua Internet Surveys, July 1998, June 2000, August 2001, http://www.nua.ie/surveys

[TREC] Text REtrieval Conference, http://trec.nist.gov

[VAU01] D. Vaufreydaz and M. Géry, Internet evolution and progress in full automatic French language modelling, ASRU, Madonna di Campiglio, Italie, 2001.