A Contrastive Indexing Method and its Integration in Aquarelle Folders

P. Bonhomme & L. Romary

CRIN/CNRS & INRIA Lorraine - FRANCE.

The Aquarelle[1] architecture is designed and built in a user oriented perspective. Folder archiving, indexing and retrieval must thus be designed for data access through «user friendly», yet expeditious, methods. This is a major challenge to the Information Systems in the Aquarelle organisation. This document addresses the role of advanced full text indexing in delivering data access and archiving applications. Our aim is to investigate the effective use of structured folders using automatic or assisted indexing in a multimedia and multilingual environment.

1.0 General Presentation

One of the main objectives of the Aquarelle Project is to provide a user with the technical facilities supporting information retrieval by querying. And our specific task at the CRIN-CNRS & INRIA Lorraine (within Work Package 6 [6]) is to design a Full Text Retrieval System, including mechanisms for automatic indexing of documents using the information available within texts dealing with these documents. The prototype we are developing, based upon a new method we called Contrastive Full Text Indexing, has been integrated within the Grif SGML Editor and tested on two different protocols:

- Blind (or free indexing), that is with no reference to a given thesaurus. This has the advantage of providing non technical words which a user could employ within his request.
- Oriented (or controlled) indexing, with an existing thesaurus containing words close to the artwork being indexed. This collocation word method is a ˝contrastive˝ indexing method [2].

This task is a collaboration between INRIA, GRIF and FORTH/ICS

In addition, since Aquarelle is to support Multilingual or Crosslingual Information Retrieval, it is important, and crucial for perpetuating the Aquarelle Project, to supply the users with multilingual thesauri if they exist or to contribute to the translation of the thesaurus. We achieve a tool which automatically aligns a text and its translation [5, 1]. This tool will assist the translation of thesauri resources across language.

This task is a collaboration between INRIA, FORTH and ILSP

2.0 Full Text Indexing and Contrastive Method

Our aim is to facilitate the effective use of structured folders using automatic or assisted indexing of SGML document.

2.1 General Method for an Indexing System

We want to provide a key word extraction to index the textual content of the different documents of a given collection or the sub-documents (from an SGML point of view) of a single document. The following indexing method is based on the number of occurrences of tokens in the documents of a collection:

1. Use of the SGML structure for each document (or sub-document) to locate textual data. Where appropriate content identifiers (tokens/terms) might be found?
   - text, description,...
   - titles, abstracts,...

1. The Aquarelle Project is managed by ERCIM, the European Research Consortium for Informatics and Mathematics. Telematics Application Programmme, Information Engineering Sector IE-2005. http://aqua.inria.fr

2. Extraction of content tokens with or without words stemming or lemmatization.
3. Assignment of term weights depending on term importance using:

- The Inverted Document Frequency (IDF) weights
- The Reduced Deviation (RD) method to supply user with indications about words occurring more than the mean around a given hub word-type, token or term. This probabilistic method is used in corpus linguistic studies, collocation word extraction and more generally in semantics and lexicology.

2.2 Contrastive Full Text Indexing Method (CFTI)

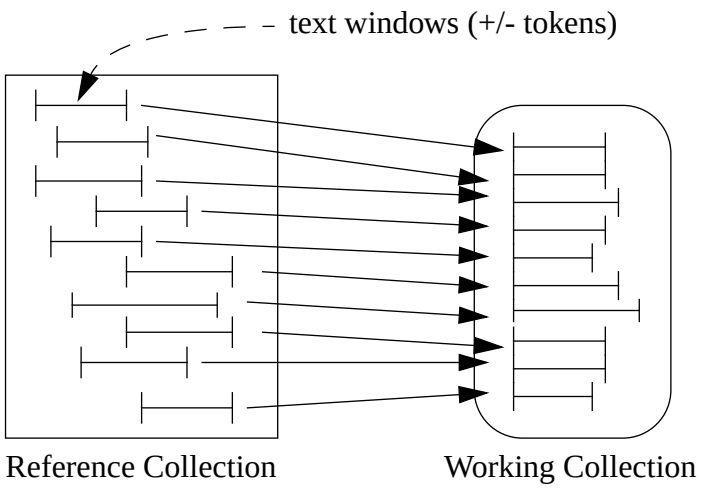
2.2.1 General Framework

The Aquarelle users are supposed to be museum curators, urban planners, commercial publishers and researchers. All users should be able to collect the information relevant to their need or answering their curiosity, wherever the information components are located with the aim of organising exhibitions and producing such information products as books or CD-ROM. The amount of information increase the need, for a professional user, in information retrieval and extraction and in content analysis. It is necessary to get rid of the non relevant information by either filtering with a thesaurus and some authority file supports or with a semantic access to the plain text.

2.2.2 Specification for a Contrastive Approach

The meaning of a word is *decidable* with its linguistic and/or situational context. We use the test of statistical significance to filter the context and to reveal a set of words occurring around a specified word (a hub-word) within a short span in a text (a window). Words co-occurring around a hub-word share some recurring semes organized in structures. We observed that the test of Reduced Deviation could assist the thematic retrieval within a collection of documents or within parts of a document. The main principle of a contrastive approach is to compare (or to contrast) a reference collection of texts with a working collection of textual data. The working collection extracted from the reference collection, is a subset of the whole reference collection. The extraction of the working collection can result from:

- A text windowing procedure,



- A selection of some SGML elements (and their content).

2.2.3 Theoretical Principle

The goal of this method is to supply user with indications about words occurring more than the mean around a given hub-word or term. It is left to the user to decide which among those words are semantically linked with the given hub-word. In this case information uniformly distributed (empty words, more frequently words,...) will not appear in the result of a Reduced Deviation procedure. This kind of non information is of course non relevant for the document and is also language independent. The remaining words will be more relevant compared with other documents (or other parts in the same document).

2.2.4 Probabilist Principle

Our model is based on a probabilistic indexing method which is a statistical assumption about distribution of words in a document. Our hypothesis is that a hub-word (can be a term) has an influence on a given word or another term (a target word). We would be able to determine if a given word (target) co-occur more than the mean around the hub-word. To accept or to reject this hypothesis implies that we take a risk corresponding to the possibility that the target word is not linked to the central word and that its frequent occurrence around the central word is not due to a coincidence. We calculate, to evaluate the risk, what we call: a reduced deviation. We assume, according to our hypothesis, that words occurring around the central word have the same distribution than the whole text. We introduced a normal approximation of the distribution: N(0,1). We calculate the reduced deviation rank *r* of a word *m*:

$$r = \frac{Freq_{observed} - Freq_{theor}}{\sqrt{Freq_{theor} \times (1 - p_m)}} \quad \text{(EQ 1)}$$

$$\text{with: } p_m = \frac{\text{frequency of the word m in the reference collection}}{Freq_{theor}} \quad \text{(EQ 2)}$$

The more important the value (absolute) of the reduced deviation is, the more important the likelihood of the given word will correspond to a semantic link with the hub-word.

Given this, we can provide a contrastive indexing in comparison to other documents in the collection or to sub-document (part) in the document.

2.3 Controlled Indexing

2.3.1 Using a Thesaurus or Authority List

A *hub word* can result from an IDF processing or come from a relative thesaurus. Thesaurus can be used to broaden the indexing term by replacing them with the corresponding thesaurus class identifiers or by adding to them the thesaurus class identifiers:

- choose broader, related or narrowed terms,
- term associations, term classification,
- facets or/and hierarchies.

2.3.2 Expert-user refinement by hand

It is important for the user, to have the possibility *to drive* his own indexing procedure, by changing the parameters of the system:

1. Statistic parameters (size of the windows, threshold, working and reference corpus,...)

2. Linguistic and document parameters (lemmatization or not, stemming words, list of words, definable SGML data structure, working corpus/document, reference corpus/document,...).

2.4 An experiment

A series of experiments are conducted to study the design issues of the Contrastive Indexing system within the Grif SGML Editor. The objectives of this experiment were:

- To implement two different and complementary indexing methods, a global method (IDF) and a local term method (RD);
- To fit statistical thresholds and their influences on results indexing;
- To find the optimum numbers of document and size of documents in number of characters, words, lines,...
- To establish stop list to remove *fluff* words or use IDF to remove high-frequency terms non relevant for indexing.

This experiment have been tested on:

- A set of 30 folders[1] in SGML format (description of the city of Cognac),
- A set of documents extracted from the *Mérimée data base*[2] and converted from HTML into SGML/TEI.

TABLE 1. Example of token extraction in the Collection Cognac

Collection COGNAC - Target Word: cognac				**Collection COGNAC - Element: <DESC>**			
RD	**TOKEN**	**Fq. WIN**	**Fq. TXT**	**RD**	**TOKEN**	**Fq. WIN**	**Fq. TXT**
24.8104	institut	28	93	11.0399	maclou	20	80
21.6968	cognaçais	30	86	10.4327	autel	22	82
15.7159	archéo	36	73	9.96775	église	39	118
15.6222	charentes	19	50	8.57429	paroissiale	21	70
14.0965	municipal	15	40	8.43421	escalier	20	67
11.7747	poitou	17	37	8.4182	saint	101	220
11.3867	musée	17	36	7.8329	périphérie	20	64
11.3519	saint	125	128	7.18366	support	22	65
9.9904	élévation	25	41	7.17207	hermanowicz	34	89
9.5808	périphérie	36	50	5.92809	lambert	27	68
9.37531	antérieure	12	25	5.83096	sud	30	73
9.34375	léger	22	36	5.5313	hôtel	25	62
8.77337	maître	14	26	4.96011	creuse	17	44
8.34589	histoire	14	25	4.79652	figure	19	47
7.91525	détail	19	29	4.76852	distillerie	18	45
7.11915	françois	50	51	4.67325	poitou	13	35
7.11911	élévations	17	25	4.65019	charente	80	146
7.11704	paroissiale	35	40	4.49553	moellon	16	40
6.60507	civat	12	19	4.30795	calcaire	17	41
6.03496	hôtels	35	36	4.2064	paris	36	73
5.82131	église	59	51	4.20493	couvert	14	35

1. coming from the *Inventaire Général* and encoding with the CI DTD provided by *Euroclid*

TABLE 1. Example of token extraction in the Collection Cognac

Collection COGNAC - Target Word: cognac				Collection COGNAC - Element: <DESC>			
RD	**TOKEN**	**Fq. WIN**	**Fq. TXT**	**RD**	**TOKEN**	**Fq. WIN**	**Fq. TXT**
5.15833	prieuré	11	15	4.20493	commerce	14	35
4.92605	commerce	14	17	4.02458	toit	20	45
4.82293	ville	69	52	3.9654	ensemble	14	34
4.39818	portail	13	15	3.92757	retable	13	32
4.20409	maulny	33	28	3.89277	sculpture	12	30
4.00413	léger	23	21	3.82414	martell	20	44
3.7119	hôtel	26	22	3.7681	françois	50	92
3.64024	ensemble	20	18	3.37964	prieure	15	35
3.63245	autel	40	30	3.24144	oeuvre	23	46
3.4955	chapelle	16	15	3.05144	vierge	11	25
3.48562	place	34	26	11.0399	maclou	20	80

3.0 Integration in Aquarelle Folders

The access to the Aquarelle System is gained by means of a User Client. The Aquarelle User Client offers a document authoring environment for professional usage. It provides two kinds of connection to the Aquarelle System:

1. A full access through the Aquarelle Access Server on which the user is registered;
2. A more restricted access through a WWW client.

In the User Client, a user can have access to details of the informations that is available (in his language or not), using query composition and submission facilities, SGML editing and obviously access to Archive and Folder Servers. It's important for a user to have an interface with the Access Server to retrieve what resources are available and what resources are relevant to work on them. In that aim, the User Client provides query and retrieval utilities to specify queries that identify such relevant documents.

Within Aquarelle, a folder is considered as an SGML document. In his User Client environment, a user has an Authority Editor System for creating, editing and modifying their structured documents or folders. The Grif SGML Editor is used for folder editing using some specific DTDs and the material for composing a folder being edited is stored into the Client workspace. At the publishing time, to made available the folder in the Folder Server by the Access Server, workspace folder, sub-folders and links are turned into Aquarelle folders and links. At that time, user has the possibility directly from the SGML Editor to launch a keyword extraction procedure on his folder. The result will be stored in the folder profile itself.

We used the Grif Application Toolkit Environment (GATE) to develop the CFTI system. The GATE API enables:

- To create of integrated SGML authoring environments,
- To customize Grif SGML Editor with new menus and functions,
- To develop Document Oriented Interfaces.

4.0 Multilingual Resources Features and Limitations

Using and producing language resources to support querying and retrieval is a crucial point in the Aquarelle multilingual environments. It includes the use of both trends in language engineering:

2. http://www.culture.fr/cgi-bin/mistral/merimee

1. the statistical trend: allowing for higher precision, it can be used as an automatic start-up in language processing and can be improve by an expert-user refinement.
2. the semantic trend: allowing for a broader coverage using knowledge bases [4], thesaurus and overriding the initial automatic results when needed.

To implement multilingual querying using one of this approach (or both), it is necessary to give the corresponding translation of each thesaural term for each new language recognized. But two main problems remain since:

- concepts expressed by one single term in one language sometime are expressed by distinct terms in another,
- concepts (or specific terms) expressed in one language sometime are not expressed in another.

The difficulty of the multilingual thesaurus construction resides in inter-cultural differences of concepts reflected in the thesaurus structure. Terms at the leaf-level are very difficult to translate. In that case, it is useful to provide an assistant to translate query or thesaurus. This kind of tools is based on a parallel concordancer we developed to generate automatically parallel texts.

As there are more comparable texts [3] than true parallel texts, we want to try our contrastive indexing system

TABLE 2. Example of token extraction with a bilingual parallel collection

l'abbaye de saint-savin (French) with word *Cyprien*				**the abbey of saint-savin (English) with word *Cyprian***			
RD	**Tokens**	**F. Text**	**F. Win**	**RD**	**Tokens**	F. Text	F. Win
11.1239	savin	128	17	11.7426	st	180	19
9.10967	saints	35	7	10.0558	savin	133	14
5.96213	frères	15	3	8.38392	saints	27	5
4.37415	probablement	12	2	8.04809	brothers	11	3
3.98536	vie	14	2	4.5643	life	14	2
3.81915	présence	15	2	4.2158	probably	16	2
3.66779	homme	16	2	3.92405	registers	18	2
3.52909	épisodes	17	2	3.92405	abbot	18	2
3.40129	registres	18	2	3.56263	episodes	21	2
3.17293	autel	20	2	3.35941	altar	23	2

with multilingual comparable texts (see: Table 2, "Example of token extraction with a bilingual parallel collection," on page 6). Are we able to find some multilingual statistic relief using the Reduced Deviation method with multilingual documents talking about the same subject but with no word-for-word translation ?

5.0 Conclusion & Future Works

The system combines the common developments in information retrieval and free text indexing with the recent advances in lexical statistic. The open architecture makes it easier to integrate the CFTI system to other applications such as web server, digital library, multimedia databases and document delivery. The handling of structured and unstructured data types (with the use of the SGML standart) provides a platform for indexing and linking documents and multimedia objects.

The following improvements could be accomplished with a modification of the CFTI system:

- lemmatization and parsing to work with word-type rather than word-token,
- multiple terms, compound names and NPs,

- communication with the thesaurus browser (not only cut and paste).

To test the system, we also planned to evaluate the prototype by the users. The criteria used for evaluating should be:

- The user-friendly use and access,
- The statistical parameters and thresholds,
- Utilisation of a lemmatizer, a stemmer or without any (loss of the information granularity),
- The coverage of the collection, the relevant matter,
- The quantify, the quality of retrieved folders,
- The relevance, precision and recall for retrieval.

6.0 References

[1] P. Bonhomme and L. Romary. Managing multilingual texts for educational purpose. In *Actes des Quinzièmes Journées Internationales IA 95*, Montpellier, 1995. EC2 & Cie.

[2] P. Bonhomme and L. Romary. Design and integration of full-text indexing in aqaurelle folders. Deliverable D6.3, INRIA, 1997. Aquarelle Project.

[3] C. Peters and E. Picchi. Cross language information retrieval: A system for comparable corpus querying. In *Workshop on Croo-linguistic Information Retrieval*, pages 24–33. SIGIR'96, 1996.

[4] R. Richardson, A. Smeaton, and J. Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. In *Proceedings of AICS Conference*. Trinity College, Dublin, September 1994.

[5] L. Romary, N. Mehl, and D. Woolls. The lingua parallel concordancing project. In L. Burnard, editor, *Electronic Texts and the Text Encoding Initiative, A Special Issue of TEXT Technology*, pages 206–220. Oxford University, 1995.

[6] C. Scholl, C. Schmuck, A. Risk, A. Michard, and J. Pascon. Aquarelle annex i - project programme. IE-2005 Information Engineering, 1995.