A Document Browsing Tool Based on Book Indexes

Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Lyne.Da.Sylva@UMontreal.CA

1. Introduction

1.1 Context and objectives

This research project is a contribution to the global field of information retrieval, specifically, to develop tools to enable information access in digital documents. We recognize the need to provide the user with flexible access to the contents of large, potentially complex digital documents, with means other than a search function or a handful of metadata elements.

The goal is to produce a text browsing tool offering a maximum of information based on a fairly superficial linguistic analysis. We are concerned with a type of extensive single-document indexing, and not indexing by a set of keywords (see Klement, 2002, for a clear distinction between the two). The desired browsing tool would not only give at a glance the main topics discussed in the document, but would also present relationships between these topics. It would also give direct access to the text (via hypertext links to specific passages).

The present paper, after reviewing previous research on this and similar topics, discusses the methodology and the main characteristics of a prototype we have devised. Experimental results are presented, as well as an analysis of remaining hurdles and potential applications.

1.2 Example

Figure 1 shows an example of the type of browsing structure sought; underlined numbers signify hypertext links to corresponding numbered passages, paragraphs or sentences.

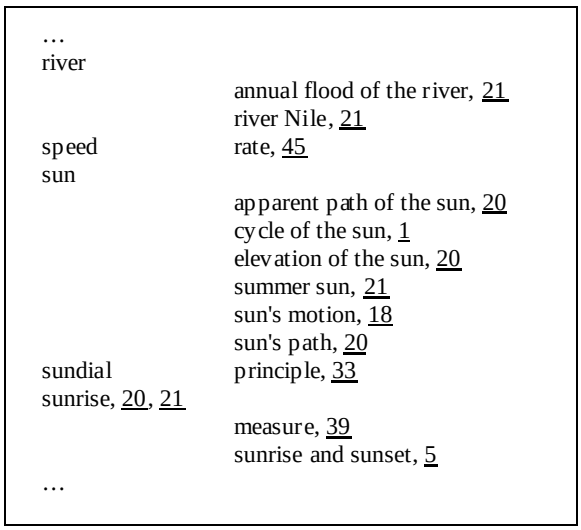

Figure 1: Sample result sought

The traditional tool which corresponds to this presentation is the "back-of-the-book" style index (see Fetters, 1994 or Mulvany, 1994). Its main characteristics are as follows: it presents an inventory of the main concepts in the document; the *entries* are structured: under general *main headings* are grouped a number of *subheadings* which

From : *Proceedings of Computational Linguistics in the North East Workshop (CliNE'04)*, Université Concordia, Montréal, 30 août 2004, pp. 45-52.

specify different aspects of the main headings discussed in a particular passage; and all important discussions (or passages) are indexed, thus covering all material in the document

The *reference* (the hyperlink) provides a direct and easy access to the passage in question. In a digital search environment, where a user's query may yield a number of long documents, building such a tool automatically would be quite helpful to browse a document's content quickly (it would also be useful for current awareness uses).

2. Previous approaches

Initial research in automatic book indexing (namely Artandi, 1963 and Earl, 1970) was met with limited success. Recently, from the field of information retrieval have come text browsing tools (also referred to as "phrase browsing" or "text exploration tools", etc.) with innovative solutions to a similar problem. Research in this direction is thus timely, and can benefit from insights drawn from information organization tradition. We now examine briefly previous approaches to automatic book indexing, their basic principles and methodological difficulties.

Initial work consisted essentially in extracting words (or, better, phrases) from documents, counting occurrences and alphabetizing a list of high-frequency candidates. This approach has some important inherent flaws. One of the most important is its inability to tease apart an important discussion of a concept, from a mere mention of it. Anyone who has used an index obviously made by such a method has been confronted with numerous references for a single entry, where only a fraction pointed to actual, useful passages on the topic. Indeed, the absolute most frequent phrase in a document will appear in very many passages – on almost every page (or a digital equivalent thereof); only a few will be useful. Conversely, quite a few fairly rare phrases will constitute very interesting index entries. Hence, information derived from frequency (used in this way) is quite dubious.

Another difficulty lies in the grouping of phrases, to form coherent entries of significantly linked topics. Such semantically-based groupings prove very difficult indeed to derive automatically (see more details below). In addition, faced with a number of references on a given topic, it proves very difficult to automatically identify the ways in which the corresponding passages differ.

Also, the simple extraction of explicit phrases in the document will not provide much more than a search function would. It seems thus that much of the "value-added" characteristics of a back-of-the-book style index lies in the semantic structure provided by the entries groupings and subdivisions. It is this structure that has been most lacking in implementations, due to the obvious difficulty of the task.

A note on more recent advances from information retrieval: various systems allow some type of document browsing. They typically try to extract useful phrases and construct a representation which enables the user to navigate among related terms (where relatedness is defined in a wide variety of ways) or related documents. Among others, we find systems for "interactive information seeking" (Anick and Tipirneni, 1999), "phrase browsing" (Wacholder et al., 2001, Nevill-Manning et al., 1999), "text exploration" (Yaari, 1997, 2000), "text structure presentation" (Hernandez and Grau, 2003, etc.) and "hierarchical summarization" (Lawrie et al., 2001). The problem they are tackling is quite similar to the one defined here, and they use techniques similar to ours (and from which we can benefit), but we note that these applications largely disregard the methodology used by human indexers. Although we make no claim to cognitively model the thought processes of human indexers, we believe there is much in their analysis of the document that can be usefully incorporated in an implementation.

Finally we note the work of Aït El-Mekki and Nazarenko (2002, 2003, 2004), which is quite similar to our own, except in their treatment of the structure relating index entries.

3. A new approach

Our implemented approach is still very much in the prototype stage and will benefit from many improvements. However, not only does it already present interesting results, but also we view it as indicating a number of fruitful research directions. We sketch the general approach now.

3.1 Methodology inspired by human indexing

The approach we have devised aims to replicate a certain conception of document indexing, inspired by human methods. For indexing, it is crucial of course to identify the key concepts in the document (this requires efficient term recognition and extraction techniques). This must however be done in tandem with a certain type of text segmentation: indexers index passages, not words; a given passage, discussing a given topic (or topics), is delineated in the text, and this "text region" must be identified and labeled with the said topic(s). This point of view – indexing passages and not concepts – allows the system to solve the problem of discarding all but significant mentions of a

given topic. This leaves the problem of establishing links between topics, whether within a given passage or between different passages. The methodology described below explains how this latter goal is achieved.

3.2 Three insights

The processing hinges on three insights.

1. The first regards the text structure and human indexing: a document is a collection of passages, which an indexer identifies and describes briefly. In other words, each passage (when delineated) can be treated as an independent "document" of a thematically-linked "corpus" (i.e. the original document). Techniques used for single document indexing (based on local frequency) can help identify the main topic(s) of the passage. Subsequently, techniques applicable to a collection may be adapted to the document as a whole. Specifically, statistics regarding term frequency and dispersion in the "collection" can be used to identify topics which help discriminate one passage from the others.
2. The second addresses the problem of grouping references into entries and distinguishing references to the same topic. Although semantic inference on a list of phrases is difficult, it is fairly easy to spot, in the source document, pairs of phrases that are linked somehow. They may be linked by a semantic relation such as hypernymy or meronymy (part-whole relationship), or because one is the (syntactic or semantic) argument of the other. They may be linked because statistical analysis reveals an important correlation between the two. When examining some of the index entries in the example above (such as "*speed, rate*"), it should not be surprising that the main heading and the subheading actually cooccurred in the text. Hence in addition to spotting words and phrases that describe the topic of a passage, it should be fairly easy to spot pairs of words or phrases that will become candidates for two-level index terms.
3. The third is a linguistic observation: that, for indexing, not all words are created equal. Some expressions are extremely useful; namely, those belonging to the specialized vocabulary of the document. Some others are not very good indicators of the thematic content; words such as "theory", "introduction", "development", "value", "example", etc. This is true independently of their frequency of occurrence (at least to a great extent). This linguistic observation will allow an interesting, yet easily-implementable type of main/subheading pairs.

3.3 More on relations and terms

A number of relation types are used by indexers in building book indexes. A quantitative analysis of some book indexes, presented in Da Sylva (in press), identifies the main relationships as the following (corresponding examples are shown in Figure 2):

a. Hypernymy (general vs. specific term)
b. Meronymy (part-whole relationships)
c. Synonymy (i.e. semantic equivalence in the document)
d. Specialized vocabulary / basic scientific vocabulary – see explanations below
e. Syntactic argument
f. Term factorization
g. Statistical cooccurrence

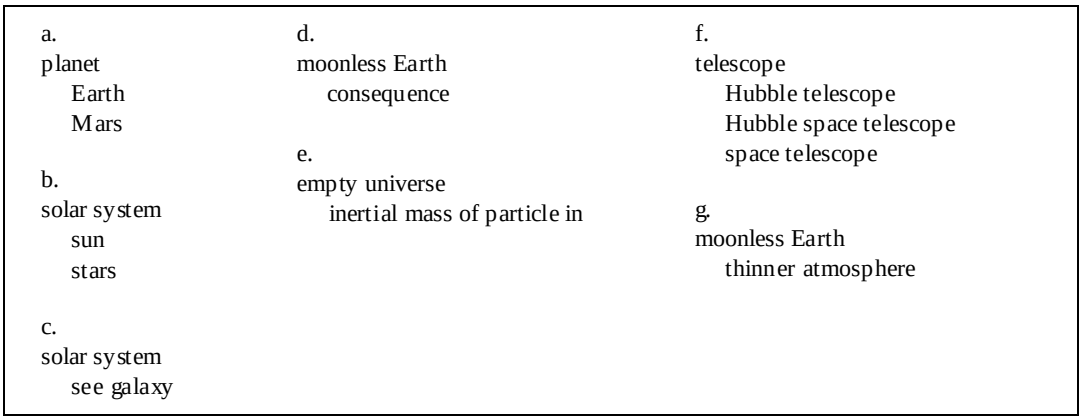

Figure 2: Sample main heading-subheading pairs illustrating types of relationships

We review now these relationships by general type.

3.3.1 Thesaural relationships

Some of these relationships require external semantic or lexical resources. Namely, the first three (a. to c. above) are relations typically expressed in a thesaurus. Given a comprehensive thesaurus appropriate for the document's thematic content, one could hope to spot related pairs of terms in the document and produce structured index entries (ex. planet, Mars). Even better, single specific terms could be spotted and paired with their hypernym taken from the thesaurus, thus producing highly informative indexes which exhibit groupings that are impossible to do automatically otherwise. The same can be said for part-whole term pairs. For synonyms, a thesaurus can yield two useful results. One is to allow the identification of variants of terms to be grouped in a single entry; the other is to provide cross-references, in the index, from one variant to the other, allowing the user to find all references to a single concept, no matter what starting point in the index is chosen.

These are highly desirable characteristics of an automatically-derived index. For the moment, they are slightly out of reach, given the difficulty of acquiring a good domain-specific thesaurus, and the limited use of general-language thesauri. But it constitutes a future goal of our research. We concentrate mostly, for the time being, on other types of relations.

3.3.2 Relationships based on term structure

The structure of extracted terms can be exploited to derive types e. and f. above, which are of a syntagmatic nature (as opposed to other relationships examined in this paper, which are paradigmatic). See Figure 3.

These types are fairly simple to implement. Term factorization implies simply to identify recurring words within phrases and factor them; although an extremely simple technique for semantic groupings, it is not always implemented in indexing systems. Certain syntactic predicate-argument relationships can be extracted from complex phrases containing prepositions. In essence, extracted phrases are split either side of a preposition. Initially, the first half becomes the main heading and the second half is treated as subheading (ex. "*life, in space*"). A second entry may also be produced by splitting after the preposition and using the second part as a main heading (ex. "*space, life in*").

SYNTACTIC ARGUMENT	FACTORISATION
length	billion
of shoreline	billion to one chance
measurements	billion years
of the moon	life
moon	advanced life
measurement of	life in space
life	lunar
in space	lunar samples
shoreline	lunar surface
length of	moon
space	earth and its moon
life in	moon's interior
	star
	single star
	star cluster
	system
	solar system
	trinary systems
	binary and trinary systems
	single star systems

Figure 3: Syntagmatic relationships based on term structure

3.3.3 Specialized vs. basic scientific terminology

As mentioned above, some very general words are poor indexing candidates. They include words such as "theory", "introduction", "development", "value", "example", etc., which would be discarded in most indexing contexts. Waller (1999) refers to them as "basic scientific vocabulary" (BSV). However, when used as subheading, paired with

a word or phrase belonging to the specialized terminology used in the document (specialized vocabulary, or SV), they produce quite insightful descriptions of passages, as Figure 4 illustrates (left column).

Our research on BSV suggests that it consists of a fairly stable list of words, used in all domains of scholarly writing, including pure and applied science, social sciences, humanities, arts, etc. The words are general and abstract in nature. The BSV may be defined on linguistic grounds, by a semantic characterization, or it may be derived automatically by extraction from a suitably balanced corpus. Our prototype uses a list which has been manually constructed and validated to produce output such as is presented in examples in this paper.

SV is identified essentially in terms of high document frequency. The pairing of SV and BSV produces useful index entries using very simple means; it would be quite difficult to produce something equivalent by a proper semantic analysis.

SV/BSV	PHRASE COOCCURRENCE PAIRS
assumptions that we live on an average planet	consequence for a moonless earth
information	interval before advanced life developed
asteroid	static crust
model	thinner atmosphere
earth	Hubble space telescope
analysis	important telescopes ever built
consequence	search for advanced life in space
development	moonless earth
information	static crust
measure	thinner atmosphere
model	static crust
question	thinner atmosphere

Figure 4: Sample SV/BSV pairs (left) and phrase cooccurrence pairs (right)

3.3.4 Statistical cooccurrence

One type of relation that is more easily obtained from the actual text of the document and without recourse to external linguistic resources it that which may be observed by statistical analysis. Indeed, by measures such as tf•idf or likelihood ratio, one can calculate whether the cooccurrence of two terms is simply due to chance, or whether, on the contrary, it is statistically significant. In the latter case, the pair may be used to propose candidate pairs of terms, such as those shown in Figure 4 (right column).

The actual relation observed by statistical cooccurrence may not be easy to characterize. However, this is not necessary, as the main/subheading pair provides a evocative concept cluster.

4. The prototype

A working prototype has been developed, which implements the three insights described above.

4.1 Overview of processing

Processing involves four main steps: concept extraction (words and phrases, with frequency counts); document segmentation; candidate term weighting; index compilation. A brief description of each step is outlined below.

1. Term extraction and frequency counts: limited lemmatization is performed (essentially restricted to nouns, which represent the most useful candidates); phrases are recognized by a simple algorithm which spots sequences of up to four non stop words (breaking on punctuation marks). A manually constructed stop list is used (our prototype has both a French and an English stop list, and can handle both languages). A frequency count is kept of all lemmatized words and phrases. The resulting words and phrases are the basis for subsequent processing.
2. Document segmentation: automatic segmentation is performed, based on lexical cohesion. The algorithm is similar to that presented in Hearst (1997) or others, but is different in that it considers sentences as

indivisible elements; it considers lemmatized content words only; and it handles some types of sentence-initial anaphors. The result is a complete segmentation into disjoint thematic units.

3. Candidate term weighting: candidate terms receive a weight which is a function of their frequency and of their form (phrases are weighted more highly than single words, for instance). The distribution of words and phrases throughout the document is also a factor; the weighting thus uses a version of tf•idf. Each thematic unit is assigned a number of these candidates, and only the most highly weighted ones will be retained in the final index. This corresponds to keeping only the most salient or discriminative concepts for each segment, thus discarding uninformative mentions of topics.
4. Index compilation: a limited number of candidates for each segment is transferred to the final index, distributed unequally among candidate types (approximately 30% are drawn each from SV-BSV pairs, cooccurrence pairs and compound terms, whereas 10% are single words). Candidate entries from different segments may share the same main heading, thus performing some types of groupings as a side effect of gathering pairs of terms (and lemmatization increases the chance of such collapsing). Specifically, phrases are factored by their initial and final word (a current limitation prevents factorization by internal words) and thus topics which are similar but not identical are nonetheless grouped in the index.

Figure 6 presents an overview of the processing implemented in our prototype.

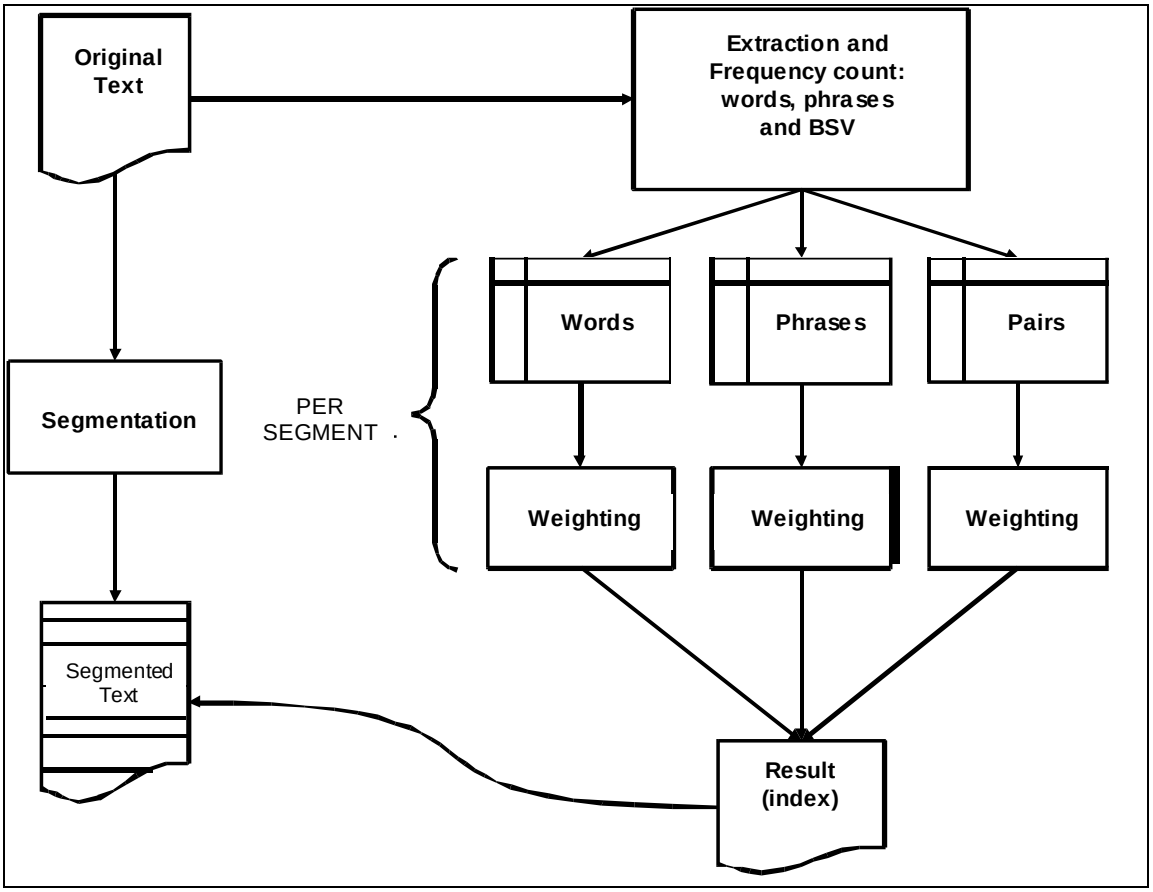

Figure 6: Overview of processing

4.2 Sample results

The examples presented throughout the previous section are output produced by our prototype. We show below a short index for the *Stargazers* text (Baker, 1990) used among others by Hearst (1997) for testing text segmentation. This sample index is probably much too short to be of real use as an index, but is shown for demonstration purposes only (no hyperlinks to passages have been included). It illustrates how candidate terms of different types are brought together in the final index (without, however, explicit markings as to their entry type).

Preferred texts for our experiments are from the “popular science” genre. This genre has been chosen for two reasons: it has a “reasonable” amount of structure (it is more structured than general prose or literary text but not as strict, or predictable, as news text) and it involves a mixture of general language and specialized language, putting in focus the distinction between the BSV and SV.

In the implemented version, each index entry is followed by a number, which is a hyperlink to the document passage (the segment) which contains the heading (and subheading, where applicable). The index thus provides a navigation tool to access specific passages in the document, dealing with a given topic; the structure of the index entries suggests some of the semantic relationships present among the concepts in the document.

5. Future prospects

The current prototype is undergoing a number of improvements. Namely, term identification is presently based simply on recurrence of content words (non-stop words), but will soon be modified to operate on patterns of part-of-speech tags. Also, as there is significant interaction between BSV and SV items, more sophisticated means of delineating which words of a text belong to its SV and which to the BSV are required; specifically, any highly frequent word in a document should be considered part of its specialized terminology, and so would be removed from a working list of BSV items. Thirdly, the algorithm for text segmentation will be tuned to its specific book indexing application; current implementations are generic and do not address all of our concerns..

Among remaining hurdles, we note that no word sense disambiguation is attempted for polysemous BSV items such as “application”. If highly frequent, every occurrence of it will be considered SV; otherwise it will be considered BSV throughout. Obviously, this is not always the case in actual texts, where the two meanings can co-exist.

6. Conclusion

Our approach to single document indexing uses insights from human indexing: the result of text segmentation becomes the structure of the index, the structure to which is attached a limited number of highly weighted index term candidates; the candidates may be isolated words and phrases, but preference is given to pairs of words or phrases, which are perceived to entertain semantic (or other) links within each segment. Relationships holding across segments are obtained by alphabetizing and factoring the entries. The resulting index is a structured list of concepts, which seems more insightful than flat list of words or phrases.

Although the current prototype produces indexes which are intuitively more useful than simple lists (and even simple graphs) of terms, no proper evaluation has been performed. Evaluation presents important methodological difficulties (those associated with evaluating something which even humans have difficulty agreeing on). In any case, a number of significant improvements to the prototype are planned in the immediate future, as mentioned above.

For this research, we foresee applications not only for indexing (such as browsing tool for large documents found on the Web, computer-assisted book indexing, and a test-bed for theories on indexing), but also for producing abstracts, based on the segmentation and on the thematic description of each passage; however, more work needs to be done to transform the index entries into a coherent summary.

This research also provides a stimulating framework for the study of properties of indexes.

6. References

Aït El Mekki T., Nazarenko A. (2002). L'index, une représentation synthétique de document. In *Atelier « Le résumé de texte automatique : solutions et perspectives »*, Paris, 14 décembre 2002. http://www-lipn.univ-paris13.fr/~aitelmekki/atalaresume.ps

Aït El Mekki T., Nazarenko A. (2003). Le réseau terminologique, un élément central pour la construction d'index de documents. In *Actes des cinquièmes rencontres Terminologie et intelligence artificielle*, pp. 1-10.

Aït El Mekki T., Nazarenko A. (2004). Une mesure de pertinence pour le tri de l'information dans un index de "fin de livre". In *TALN 2004*, Fès, April 19-21, 2004. http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/AitElMekki-Nazarenko.pdf (accessed June 15th 2004).

Anick, P. ; Tipirneni, S. (1999). The paraphrase search assistant: Terminological feedback for iterative information seeking. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 153-159.

Artandi S. (1963). *Book indexing by computer*. New Brunswick, N.J., S.S. Artandi.

Baker, David. (1990). Stargazers look for life. *South Magazine*, 117, pp. 76-77.

Da Sylva, Lyne. (In press). Relations sémantiques pour l'indexation automatique. Définir des objectifs pour la détection automatique. *Document numérique, Numéro spécial « Fouille de textes et organisation de documents »*.

Earl L.L. (1970). Experiments in automatic extraction and indexing. In *Information Storage and Retrieval*, 6, pp. 313-334.

Fetters L.K. (1994). *Handbook of Indexing Techniques : a Guide for Beginning Indexers*. Port Aransas, TX : American Society of Indexers.

Hearst M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. In *Computational Linguistics*, 23(1), pp. 33-64.

Hernandez N. ; Grau, B. (2003). What is this text about? Combining topic and meta descriptors for text structure presentation. In *Proceedings of the 21st annual international conference on Documentation (ACM SIGDOC*), San Francisco, 12-15 Oct. 2003, pp. 117-124.

Klement, S. (2002). Open-system versus closed-system indexing. In *The Indexer*, 23(1), 23--31.

Lawrie D., Croft B. (2001). Finding Topic Words for Hierarchical Summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 349 - 357), New Orleans, Louisiana.

Mulvany N. (1994). *Indexing books*. Chicago: University of Chicago Press.

Nevill-Manning, C.G., Witten, I.H., Paynter, G.W. (1999). Lexically-generated subject hierarchies for browsing large collections. *International Journal of Digital Libraries*, 2(2/3), 111--123.

Wacholder, N., Nevill-Manning, C. (2001). Workshop report: The Technology of Browsing Applications, Workshop held in conjunction with JCDL 2001. *SIGIR Forum* 35(1), pp. 16-19. http://www.acm.org/sigir/forum/S2001-TOC.html.

Waller, S. (1999). *L'analyse documentaire. Une approche méthodologique*, Paris, ADBS Éditions.

Yaari, Y. (1997). Segmentation of Expository Texts by Hierarchical Agglomerative Clustering. In *Proceedings of Recent Advances in Natural Language Proceessing* (pp. 59-65), Bulgaria.

Yaari, Y. (2000). *NLP-assisted exploration of texts*. http://citeseer.ist.psu.edu/412683.html.