

Integrating knowledge from different sources for automatic back-of-the-book indexing

Abstract: The paper reports research on automatic back-of-the-book indexing. It presents a methodology which brings together knowledge from different disciplines. It is inspired by human indexing methodology and the results are more similar to manually-crafted indexes than those produced by previous automatic approaches. Issues of evaluation and applications are addressed.

Résumé : Cette communication présente les résultats de recherche sur l'indexation automatique de livres. L'étude propose une méthodologie qui rassemble des sources de connaissances provenant de disciplines différentes. La méthodologie s'inspire de l'indexation humaine et les résultats se rapprochent plus de l'indexation manuelle que les autres méthodes d'indexation automatique. Sont également touchés les enjeux d'évaluation et d'applicabilité.

1. Introduction

The work reported in this paper deals with the task of automatic back-of-the-book indexing, i.e. automatically producing indexes like those in Figure 1.

river	annual flood of the river, 21
	river Nile, 21
speed	rate, 45
sun	apparent path of the sun, 20
	cycle of the sun, 1
	summer sun, 21
	sun's motion, 18
sundial	principle, 33
sunrise, 20, 21	measure, 39
	sunrise and sunset, 5

Figure 1: Sample book index extract

This constitutes a challenging, little-researched area, which has met with very limited success in the past. We believe that a methodology which brings together knowledge and insights from various sources may help produce more promising results. The objective of the work is to improve on previous approaches to automatic back-of-the-book indexing (ABI) and possibly suggest new applications for this type of resource. In this paper, we will present previous work on this topic, and then outline our methodology for the development of a prototype indexing system. We then present some experimental results. The discussion will address issues of evaluation and applications.

2. Previous approaches

Very little research work has focused on ABI: Artandi (1963), Earl (1970) and Salton (1988) represent the only substantial, reported implementations of this for a period of close to 40 years. These rely heavily on statistics (i.e. frequency of occurrence of extracted words or phrases) and an alphabetical listing of the most frequent ones; this approach, modelled on database indexing methods, does not fare well for book indexes. Here, the goal is not to capture the essential topic of the document as a whole, but to

provide access to specific passages on precise topics, with these topics grouped in a systematic (often semantically motivated) way in the index.

Specific challenges to ABI include, crucially, identifying occurrences truly relevant for indexing (not listing every term occurrence in the index) and structuring the index through semantic and knowledge retrieval principles – this must rely on encyclopaedic knowledge notoriously difficult to impart to automatic systems. Some of these issues were addressed by Nazarenko and Ait El Mekki (2005). Their approach, similar to our, differs in that their index entry structure relies mostly on their complex lexical network (of synonymy and hierarchy relations) whereas ours uses the lexical semantic nature of words and “significant” co-occurrence of concepts in a given text.

3. Methodology

3.1 Overall approach

The novel approach presented here integrates knowledge not only from general computational methods and frequency statistics, but also from lexical semantics, text linguistics and human indexing methodology. The insights gained are the following: recognizing different roles for different types of lexical items; exploiting the context of terms in the document to help structure the index; basing the index on a preliminary step of passage delimitation (or segmentation).

3.2 Indexing algorithm and prototype

We developed a prototype indexing system (Da Sylva and Doll, 2005), based on the processing outlined below (see Figure 2). Each step is detailed in the next section.

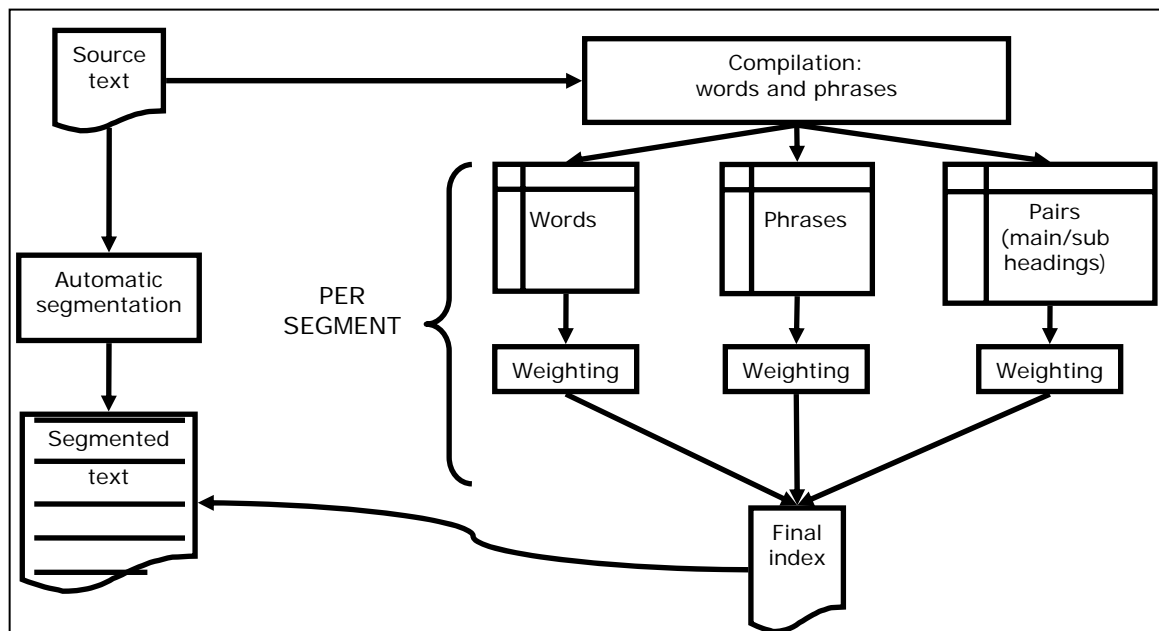


Figure 2: Overview of processing

1. In parallel,
 - i. extract and count all words and multi-word terms;
 - ii. perform automatic segmentation.
2. Identify, within each text segment:
 - i. “salient” words/phrases;
 - ii. “linguistically interesting” words for indexing;

- iii. potential main heading/subheading pairs, to structure the index.
3. Weigh each candidate.
4. Collect only the most highly-weighted candidates for the index; conflate; alphabetize.

Word counting: each word is lemmatized (i.e. plural forms are converted to the singular, comparative adjectives to the base form, verbs to their infinitive...), using a dictionary and word formation rules. Phrases are lemmatized as well (e.g. “sightings of stars” and “sighting of stars” are both converted to “sighting of star”). The most recent version exhibits improved phrase recognition using a part-of-speech tagger.

Automatic segmentation: the segmentation is performed with a lexical cohesion algorithm (see Hearst, 1997). This calculates the number of words in common among successive sentences. A score is calculated for each sentence, reflecting the repetition of words and use of anaphoric pronouns or linking adverbs between successive sentences. When that score is below a certain threshold, a thematic break is posited. In this way, the text is automatically segmented into thematic sections. This provides the basis for subsequent indexing: terms will point to segments rather than actual word occurrences. This mirrors the knowledge that human indexers index passages, and not words; it approximates knowledge we have about text linguistics.

Salience: for each segment, the local topic is determined. The salience of each word/phrase is calculated, with the use of a familiar measure, tf-idf. This is typically used to identify discriminating words in a document, within a collection. For back-of-the-book indexing, each passage is considered a document within the collection represented by the set of passages: thus words which are more (i.e. more frequently) represented in a given segment than in the document as a whole can be identified. This method comes from corpus linguistics.

Lexical distinctions: in our implementation, we make a distinction among words which may have equal frequency, but different uses for indexing. Less useful words are those very general in nature: “application”, “method”, “characteristics”, etc. (generally discarded in database indexing). These arguably occur with equal frequency in any (scholarly) document and are not topical (compared to “single star solar system” or “satellite”). Poor candidates for main headings, they often provide useful subheadings (consider for example “single star solar system, characteristics”). An important part of our research is devoted to identifying this class of so-called “basic scientific vocabulary”. A compiled list of these items is used by the prototype to produce structured index entries like the one in the above example. Thus lexical semantics and indexing practice both concur to suggest the differentiated treatment of different word types in ABI. And they provide a way to structure the index (see below).

Creating potential main heading/subheading pairs: another type of structured index entry consists of the pairing of two words or phrases, which appear (by a tf-idf analysis) to be closely linked in the context, i.e. in a given segment. An example of such a pair might be: “single star solar system, satellite”. Here again, text linguistics helps to produce structured entries.

Weighting: each candidate is weighted by a score combining the tf-idf measure and some of its linguistic features (especially its length – longer terms preferred). Corpus linguistics is at play here.

Index compilation: only the most highly-weighted candidates for each segment make the final index. Some entries are single words, others are phrases, yet others are pairs consisting of a main heading and a subheading. These are all conflated where main headings are identical, and the resulting list is alphabetized. This constitutes the final index, with hypertext pointers to the segments where each term occurs.

4. Results

Results presented here are from a text on the search for life in space. Our system segments the text and identifies a number of candidates for each segment. Only an excerpt is shown here, given space constraints. Recurring words within phrases are extracted, and all related phrases are placed as subheadings of the recurring word (here, “star”, “planet”, “space”, “system”, etc.) to produce the alphabetized version in Figure 4.

Entry type	Segment 1	Segment 7	Segment 8
Pairs	average star, solar system average planet, assumption	star systems, single star solar system planets in binary and trinary systems, data	giant balls of gas, solid surface stars in our galaxy behaviour
Phrases	life in space	trinary systems binary and trinary systems	trinary systems binary and trinary systems single star systems
Word	universe	astronomer	century

Figure 3: Winning candidates from each segment

astronomer, <u>7</u>	space,
ball,	life in space, <u>1</u>
giant balls of gas,	star,
solid surface, <u>8</u>	average star,
century, <u>8</u>	solar system, <u>1</u>
life,	stars in our galaxy,
life in space, <u>1</u>	behaviour, <u>8</u>
galaxy,	star systems,
stars in our galaxy,	single star solar system, <u>7</u>
behaviour, <u>8</u>	system,
gas,	binary and trinary systems, <u>7, 8</u>
giant balls of gas,	planets in binary and trinary systems,
solid surface, <u>8</u>	data, <u>7</u>
planet,	single star systems, <u>8</u>
average planet,	star systems,
assumption, <u>1</u>	single star solar system, <u>7</u>
planets in binary and trinary systems,	trinary systems, <u>7, 8</u>
data, <u>7</u>	universe, <u>1</u>

Figure 4: Resulting index entries

The output resembles a manually-constructed index more than the simple lists produced by past research on ABI. And it includes document-dependent pairs, absent from the Nazarenko & Ait El-Mekki implementation.

5. Discussion

Evaluation of the prototype is an important and quite unresolved issue. Although experimental results seem promising, we have yet to develop a feasible evaluation scenario: manually-produced indexes differ from automatically produced ones, if only because human indexers often reword phrases in the document – actual index entries may be equivalent, but formally different. Our experiments comparing human-built indexes with automatic ones for a book-length document (Darwin’s *Origin of Species*) produced

low agreement on index entry wordings, despite useful segmenting and index entry choices.

Potential applications which we are currently pursuing include: an aid to human back-of-the-book indexing; new presentations of text structure and themes (i.e. elaborate table of contents or “thematic view” into the document); construction of rich, structured metadata for the semantic Web or for better indexing for the “regular” Web (using the additional metadata produced by the ABI).

6. Conclusion

We have devised a method for automatic back-of-the-book indexing which draws from different disciplines, to produce results that are much closer to manually-constructed indexes than previous approaches. Many areas are open for further research but this is an example of the synergy attainable by the diversity of knowledge sources.

7. References

- Artandi, S. *Book indexing by computer*, S.S. Artandi, New Brunswick, N.J., 1963.
- Da Sylva, L; Doll, F. A Document Browsing Tool: Using Lexical Classes to Convey Information. In Lapalme, Guy ; Kégl, Balász, *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005 (Proceedings)*, New York : Springer-Verlag, 2005: 307-318.
- Earl, L.L. Experiments in automatic extraction and indexing. *Information Storage and Retrieval*, 6, 1970: 313-334.
- Hearst, M.: TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1), 1997: 33-64.
- Nazarenko, A.; Aït El Mekki, T. Building back-of-the-book indexes. *Terminology*, 11(1), John Benjamins, 2005: 199-224.
- Salton, G. Syntactic approaches to automatic book indexing. In: *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*: 204-210, 1988.