Key words and key phrases in scientific databases.
Aspects of guaranteeing output quality for databases of information.

by
Michiel Hazewinkel
CWI
POBox 94079
1090GB Amsterdam
The Netherlands

Abstract. This paper discusses the well known problem of information retrieval in the setting of large scientific databases of published papers. The paper focusses on 'key-phrases' as a potentially very powerful tool and addresses the problem of the semi-automatic assignment of key-phrases to a record (usually consisting only of an abstract, authors' names, and title plus some standaard bibiographical information).

MSCS: 68P20

Key words and key phrases: information retrieval, information storage, controlled list, controlled vocabulary, key phrases, key words, thesaurus, weak thesaurus, entiched weak thesaurus, dynamic updating, incremental updating, indexing, scientific database, MATH, MATHSCI, linguistics, uncontrolled information, unorganized information, classification scheme, identification cloud, ends.

1. Introduction.

Let me start by stating my main concern.

Main concern. Getting the desired information out of a scientific database (knowing that it is in there). And, preferably, without retrieving so much irrelevant junk that the real information desired is swamped in a sea of nonsense.

This is an old IR topic (Information Retrieval) and in my own modest opinion we are getting nowhere. In particular, I am going to argue that linguistic and statistical techniques are not sufficient.

It makes but little sense to make large amounts of funds available for scientific research if one does not also make sure that the results of that research will be refindable if and when needed. Whether that is currently the case is doubtfull at least. In my own community (mathematics) I have heard it said that it is often easier to re-investigate a (not too complicated) question, rather then trying to find the relevant literature on it. And in [8] Naisbitt writes:

> "Uncontrolled and unorganized information is no longer a resource in an information society. Instead it becomes an enemy of the information worker. Scientists who are overwhelmed with technical data complain of information pollution and charge that it takes less time to do an experiment than to find out whether or not it has already been done."

This is not only a matter of concern for the information retriever (searcher). Authors, desirous to have their work recognized and noticed face the same sort of problem, and, as is argued cogently in [1] they are well advised to supply their articles with good key-phrases.

Additional concern. An average user must be able to get the desired information out; not only (super) experts.

Context. The context of the remarks here is that of largish scientific databases such as MATH (FIZ/STN, Karlsruhe/Berlin), COMPUSCI (FIZ/STN, Karlsruhe/Berlin), and MATHSCI (American Mathematical Society, Ann Arbor, USA). These have something like a million records consisting of various fields like 'Author(s)', 'Title', 'Abstract', 'Classification' and a variety of bibliographic fields that are of less importance for IR tasks. In addition MATH and COMPUSCI have an uncontrolled Key Phrase field; MATHSCI has no key phrases.

Focus of this paper. The focus of this paper (and the original lecture in Warsaw, August 1999) is on

- The importance of a 'standard key-phrase field'. I.e. a key-phrase and key-word field with a controlled vocabulary.
- The idea of an enriched weak thesaurus.

Background. I am writing these remarks with as background and sources of inspiration my own experiences with such projects as:

- Encyclopaedia of Mathematics, 10 volumes, two update volumes (and two more to come), some 42000 key phrases (not counting inversions), see [2, 3, 5].
- Index of the journal 'Artificial Intelligence', volumes 1-89 (some 12000 phrases), see [4]
- Index of the journal 'Theoretical Computer Science', volumes 1-200 (some 48000 key phrases), see [6].

2. Linguistic problems.

Part of the problems encountered when searching a scientific database are linguistic in nature. Different linguistic versions of the same phrase: plurals versus singulars, inversions, misspellings, different transcriptions of proper names from a non-English language.

To give examples, I know of 28 variants (different transcriptions) of the proper name Chebyshev (П Λ Чебышев). Also some 22% of the more important papers on the well-known numerical method 'Crank-Nicolson' are hiding under the misspelt version 'Crank-Nicholson".These are nontrivial problems but can, it seems, be handled by linguistic means, in particular by the use of standard lists of phrases and names and recognition software that can spot and correct mild deviations in a query using availabe lists of standard phrases. The AMS with their database MATHSCI has done valuable and useful work here by means of a standard list of known names of mathematicians.

These matters are not my main topic here, beyond noting that 'standard lists', i.e. a controlled vocabulary, also here provide a solution.

3. The story of "ends".

'End', plural 'ends' is a technical concept in mathematics. It refers to some kind of point at infinity. Depending on the particular part of mathematics involved the concept is slightly different. The task is to collect a comprehensive bibliography on this topic. The problem is of course that the word 'end' or 'ends' also has its normal linguistic meanings as in

- **end** of a rod
- **ends** of a beam

- near the **end** of this chapter
- the book **ends** with

All these phrases (and several more) are quite likely to occur in an abstract and, thus, it is very dificult to select just those mathematical papers that deal with the mathematical concept of 'ends'. Currently it is not possible to search for the concept just using the keywords field. In the case of MATHSCI because there is no keywords field; in the case of MATH because the words in the keywords field, in the title and in the abstract are lumped all together in one grand search field. I consider myself not inexpert in using scientific databases but have so far been unsuccessfull in collecting a good bibiographic survey of papers dealing with the mathematical concept 'end'. The problem is aggravated because the concept does not just occur in one mathematical superspecialism (so that one can use the classification scheme) but in several (with slightly different definitions). To be precise the concept occurs in General topology (54), Manifold theory (57), Complex function theory (30), Group theory (20), Combinatorics (05), and very possibly in a few other fields.

'End' is not the only word which can cause this kind of difficulty. Mathematics and more generally, science, is full of words which have both normal lingusitic and highly technical meanings which can conspire to cause this kind of difficulties. Potential examples are: 'regular', 'sound', 'complete', 'filter', 'net', 'control', ...

4. The concept of an enriched weak thesaurus

The remarks above, hopefully, have indicated that there is a real need for a separately searcheable 'key-phrase field'. This should be a field with entries coming from a controlled vocabulary. As said, the database MATH has a key phrase field. But it is uncontrolled and the entries are provided by authors, editors, others, ... A disturbing 31% of the 3.5 Million key phrases refer just to one article and as such have no information-retrieval-value whatever.

This is nothing new. The value of a controlled vocabulary, in the form of a thesaurus, is enormous. This is well recognized and the value of such a thesaurus is further attested to by the considerable resources that are devoted to maintaining one in the few fields of science, like medicine, that are fortunate enough to have an adequate thesaurus.

Unfortunately, a thesaurus, as defined by various ISO, national, and international standards, for a given field of science, is very expensive to create and, once created, nearly impossible to maintain dynamically, i.e. incrementally, as a field of science evolves. This has led me to define the concept of an *enriched weak thesaurus.* This concept is described and discussed at length in [7]. For the present purposes a main aspect is that in such a thesaurus each key-phrase comes with an socalled '*identification cloud*', a set of nouns that in the published literature are likely to be found in the neighborhood of the phrase in question. Perhaps not all in all cases that the phrase in question would be appropriate for an article, but enough that one can have a founded suspicion that that phrase should be attached to the paper being examined.

This provides a mechanism for the automatic assignment of useful key-phrases to documents. The examples below are meant to illustrate this.

5. Examples (from the real world)

It seems clear that the value of a scietific database can be enhanced by providing each record with an adequate list of key-phrases. The question arises whether that can be done on the basis of the data that are available. As a rule these are: title, authors and abstract and sometimes author provided (uncontrolled) key-phrases. To see whether these data suffice let us look at some real world examples. These examples all come from the data which were used to generate [4], [6].

Example 1.

a **complete axiomatic characterization of first-order temporal logic of linear time**.

As shown in (**Szalas**, 1986, 1986, 1987) there is no finitistic and **complete axiomatization** of

First-Order Temporal Logic of linear and discrete time. In this paper we give an **infinitary proof system** for the logic. We prove that the *proof system is sound and complete*. We also show that any **syntactically consistent temporal theory** has a model. As a corollary we obtain that the Downward Theorem of **Skolem**, **Löwenheim** and **Tarski** holds in the case of considered logic.

KEYWORDS: **algebra of Lindenbaum and Tarski**, **Boolean algebra**, **completeness**, **consistency**, **first-order temporal logic**, model, proof system, **semantic consequence**, soundness, **syntactic consequence**.

sound and complete proof system

first order temporal logic

axiomatization of temporal logic

downward theorem

finitistic axiomatization

downward Löwenheim-Skolem theorem

Kripke structure

Here the available data consisted of an abstract and a list of key-phrases. In bold are indicated the index (thesaurus) phrases which can be picked-out directly from the text. Below the original text are five more phrases, that can be obtained from the available data by relatively simple linguistic means, assuming that one has an adequate list of standard key phrases available. For instance "first order temporal logic" results from "First-Order Temporal Logic" by a simple cleaning up, and "sound and complete proof system" is linguistically close enough to a phrase from the available text: "proof system is sound and complete" (indicated in italics).

Then, in shadow, there is the term "downward Löwenheim-Skolem theorem". This one is a bit more complicated to find. But, again given an adequate standard list, and with "downward theorem", "Löwenheim" en "Skolem" all in the available text it is recognizable as a term that belongs to to this document.

Finally, in bold-shadow, there is the term "Kripke structure". There is no linguistic hint that this term belongs here. However, the identification cloud of this term, would contain many of the key words that occur in this document and that thus strongly suggests that "Kripke structure" could be an important term to assign to this document. (As turned out to be the case by checking the full paper later.)

Example 2.

two-dimensional iterative arrays: characterizations and applications.

We analyse some properties of two-dimensional iterative and **cellular arrays**. For example, we show that **arrays** operating in $T(n)$ time can be sped up to operate in time $n + (T(n) - n)/k$.

....

computation. Unlike previous approaches, we carry out our analyses using sequential machine characterizations of the iterative and cellular arrays. Consequently, we are able to prove our results on the much simpler **sequential machine models**.

iterative arrays

sequential characterizations of cellular arrays

sequential characterizations of iterative arrays

characterizations of cellular arrays

characterizations of iterative arrays

arrays of processors

The style coding of terms is the same as in example 1 above. Here clearly the term “array” is very central. Given that, the term “arrays of processors” in a standard list, and an identification cloud for that phrase, this term can be recognized as belonging to this document.

Example 3.

A *safe* approach to **parallel combinator reduction**.

In this paper we present the results of two pieces of work which, when combined, allow us to take a program text in a **functional language** and produce a **parallel implementation** of that program. We present techniques for discovering **sources of parallelism** in a program at **compile time**, and then show how this parallelism is naturally mapped into a **parallel combinator set** that we will define. To discover sources of **parallelism** in a program, we use **abstract interpretation**. Abstract interpretation is a compile-time technique which is used to gain information about a program that may then be used to optimize the execution of the program. A particular use of abstract interpretation is in **strictness analysis of functional programs**. In a language that has **lazy semantics**, the main **potential for parallelism** arises in the evaluation of operands of strict operators. A function is strict

...

Having identified the sources of **parallelism** at compile-time it is necessary to communicate these to the **run-time system**. In the ...

safe evaluation in parallel

functional programs

optimizing the execution of a program

evaluation in parallel

parallelizing functional programs
safe parallelization

In this example the words and phrases “safe”, “functional program” and “parallel(ization)” are clearly central. Given identification clouds and standard lists of key phrases this leads to the extra two phrases in shadow.

Example 4.

sequential and **concurrent behaviour in Petri net theory**.

Two ways of describing the **behaviour of concurrent systems** have widely been suggested: arbitrary **interleaving** and **partial orders**. Sometimes the latter has been claimed superior because **concurrency** is represented in a `true' way; on the other hand, some authors have claimed that the former is sufficient for all practical purposes. **Petri net** theory offers a framework in which both kinds of **semantics** can be defined formally and hence compared with each other. Occurrence sequences correspond to **interleaved behaviour** while the notion of a process is used to capture **partial-order semantics**. This paper aims at obtaining formal results about the

...

more powerful than **inductive semantics** using

...

of **nets** which are of **finite synchronization** and **1-safe**.

sequential behaviour in Petri net theory
Petri net theory
axiomatic definition of processes

interleaving semantics
1-safe nets

Here, the constituents "1-safe" and "nets" of "1-safe nets" actually occur in the text. But they are so far apart that without standard lists and identification clouds the phrase would probably not be picked up.

The four examples above all come from [4] and [6]. They are the same examples as discussed in [7]. They are not complete; in particular, parts of index phrases that are themselves also suitable index phrases have not been indicated.

6. Conclusions

In this paper I have discussed the value of a controlled list of key-phrases for information retrieval purposes.in the form of (part of) a enriched weak thesaurus The arguments have been illustrated with a number of real life examples.

Probably there should also be a field for free uncontrolled key-phrases that are author supplied. These can be used as part of a mechanism for the dynamic and incremental updating of the (controlled) enriched weak thesaurus.

References

1. Edward E Gbur, Jr, Bruce E Trumbo, *Key words and phrases—the key to scholarly visibility and efficiency in an information explosion*, American Statistician **49**:1(1995), 29-33.

2. Michiel Hazewinkel (ed.), *Encyclopaedia of mathematics; 10 volumes*, KAP, 1988-1994.

3. Michiel Hazewinkel (ed.), *Encyclopaedia of mathematics volume 11 (first supplementary volume)*, KAP, 1997.

4. Michiel Hazewinkel, *Index "Artificial Intelligence', Volumes 1-89*, Artificial Intelligence **96:1** (1997), 1-302.

5. Michiel Hazewinkel (ed.), *Encyclopaedia of mathematics volume 12 (second supplementary volume)*, KAP, 1999.

6. Michiel Hazewinkel, *Index "Theoretical Computer Science", Volumes 1-200*, Theoretical Computer Science **213/214**(1999), 1-699.

7. Michiel Hazewinkel, *Topologies and metrics on information spaces*, CWI Quarterly **12**:2(1999), Preliminary version: http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html

8. J Naisbitt, *Megatrends. Ten new directions transforming our lives*, Warner, 1982.