Dialogue mediated information retrieval, automatic keyphrase assignment, and identification clouds

by

Michiel Hazewinkel
CWI
POBox 94079
1090GB Amsterdam
The Netherlands

Abstract. The central concept in this paper is that of an identification cloud of a keyphrase. Very roughly this is (textual) context information that indicates whether a standard keyphrase is present, or, better, should be present, whether it is linguistically recognizable or not (or even totally absent). Identification clouds capture a certain amount of expert information for a given field. Applications include automatic keyphrase assignment and dialogue mediated information retrieval (as discussed in this paper). The problem arises how to generate (semi-)automatically identification clouds and a corresponding weak enriched thesaurus for a given field. A possible (updatable and adaptive) solution is described.

MSC2000: 68T35, 68U35, 91F20

Key words and key phrases. Identification cloud, information retrieval, information space, disambiguization, automatic indexing, weak enriched thesaurus, thesaurus, standard keyphrase, dialogue search, neighborhood search.

1. Introduction.

Information retrieval and automatic indexing seem to have reached a certain plateau. As I have argued at some length elsewhere, see e.g. [3, 4, 7, 8] there is only so much that can be done with linguistic and statistical means only. To go beyond, it could be necessary to build in some expert knowledge into search engines and the like. This has led to the idea of identification clouds, which is the main topic of this paper.

The same idea grew out of a rather different (though related) concern. It is known and widely acknowledged, that a thesaurus for a given field of inquiry is a very valuable something to have. However, a classical thesaurus according to ISO standard 2788, see [1], and various national and international multilingual standards is not an easily incrementally updatable structure. This has lead to the idea of an enriched weak thesaurus, loc. cit., and identification clouds are a central part of that kind of structure.

In this paper I try to give some idea of what ID clouds are and how they can be used. More applications can be found in the papers quoted. The idea has meanwhile evolved, largely because of the use of ID clouds in the EC project TRIAL SOLUTION, and in this paper I also sketch the refinements that have emerged.

2. Identifcation clouds.

Basically the "*identification cloud*" of an item from a controlled list of standardized key phrases is a list of words and possibly other phrases that are more or less likely to be found near that key phrase in a scientific text treating of the topic described by the key phrase under consideration.

For instance the key phrase

Darboux transformation

could have as part of its identification cloud the list

soliton
dressing transformation
completely integrable
Hamiltonian system
inverse spectral transform
Bäcklund transformation
KdV equation
KP equation
Toda lattice
conservation law
inverse spectral method
exactly solvable
...
(37J35, 37K (the two MSC2000 classification codes for this area of mathematics)
...

Of course the definition as given here is still pretty vague: both the terms 'more or less likely' and 'near' need to be made more precise. Before saying something to that point let us look at a concrete example.

In a record that I saw some five years ago now there occurred the phrase:

" ... using the Darboux process the complete structure of the solutions of the equation can be obtained."

At first sight it looks like there is here a natural key phrase, viz. "Darboux process", to be extracted. Presumably, some sort of stochastic process like "Cox process", "Dirichlet process", or "Poisson process". The context made that rather doubtfull; the surrounding sentences did not have in them the kind of words one expects in a paper on stochastic matters. The proper name "Darboux" is also not sufficient to identify what is meant; there are two many terms with "Darboux" in them: "Darboux surface", "Darboux Baire 1 function", "Darboux property", "Darboux function", "Darboux transformation", "Darboux theorem", "Darboux equation",....(these all come from the indexes of [2]).

The various words and phrases occurring in the surrounding sentences settled the matter These included such words as 'soliton' and others from the example above and are typical for the surrounding words of the term "Darboux transformation" and typical for the area classified by 37K (and 37J35) (one of the classifications—indeed the main one—of "Darboux transformation"). Thus the '*identification cloud*' of the term "Darboux transformation" made it

possible to extract the right term. What the authors meant is that repeated use of the process 'apply a Darboux transformation' should give all solutions.

A human mathematician, more or less expert in the area of completely integrable systems of differential equations, would have no difficulty in recognizing the phrase "Darboux process". Thus what identification clouds do is to add some human expertise to the thesaurus (list of key phrases) used by an automatic system.

The idea of an identification cloud is part of the concept of an enriched weak thesaurus as defined and discussed in [4, 5, 8].

3. Application 1: automatic key phrase assignment.

A first application of the idea of identification clouds is the automatic assignment of key phrases to scientific documents of suitable chunks of scientific texts.

It is simply a fact that it often happens that in an abstract or chunk of text a perfectly good key phrase for the matter being discussed is simply not present or so well hidden that linguistic and/or statistical techniques do not suffice to recognize it automatically.

The idea here is simple. If enough of the identification cloud of a term (= standard keyphrase) is present than that key phrase is at least a good candidate for being assigned to the document under consideration.

A C-program that takes as input a keyphrase list with identification clouds and a suitably prepared corpus of documents (chunks of text or abstracts) and that gives as output the same corpus with each item enriched with automatically assigned keyphrases has been written in the contect of the EC project "TRIAL SOLUTION" (Febr. 2000 - Febr. 2003). It also outputs an html file for human use which can used to check how well the program worked. This validation test is currently (April 2002) under way.

It is already clear, that the idea of identification clouds needs refinements; certainly when used on rather elementary material (as in TRIAL SOLUTION). Two of these will be briefly touched on below.

4. Application 2: dialogue mediated information retrieval

Given a keyphrase list with identification clouds, or, better, a weak enriched thesaurus it is possible to use a dialogue with the machine to refine and sharpen queries. Here is an example of how part of such a dialogue could look:

Query: I am interested in spectral analysis of transformations?

Answer: I have:

- spectral decompositions of operators in Hilbert space (in domain 47, operator theory, 201 hits)
- spectral analysis (in domain 46, functional analysis, 26 hits)
- spectrum of a map (in domain 28, measure theory, 62 hits)
- spectral transform (in domain 58, global analysis, 42 hits)
- inverse spectral transform (in domain 58, global analysis, 405 hits)

Please indicate which are of interest to you by selecting up to five of the above and indicating, if desired, other additional words or key phrases.

The way this works is that the machine scans the query against the available identification clouds (using some (approximate) string matching algorithm, e.g. Boyer-Moore) and returns those keyphrases whose ID clouds match best, together with some additional information to help the querier to make up his mind.

5. Application 3: distances in information spaces.

As it is the collection of standard keyphrases is just a set. It is a good idea to have a notion of distance on this set: are two selected standard key phrases near, i.e. closely related, or are they quite far from each other. Identification clouds provide one way to get at this idea: two phrases which have large overlap in their identification clouds are near to each other.

A use, again dialogue mediated, of this is as follows.

Query: I am interested in something related to <StandardKeyPhrase 1>. Please give me all standard keyphrases that are within distance x of this one.

For other ways to define distances on information spaces (such as the space of standard key phrases) and other potential uses of distance, see [8].

6. Application 4: disambiguization.

Ambiguous terms are a perennial problem in (automatic) indexing and thesaurus building.

Identification clouds can serve to distinguish linguistically identical terms from very different areas of the field of inquiry in question. E.g. "regular ring" in mathematics, or the technical term "net" which has at least five completely different meanings in various parts of mathematics and theoretical computer science.

Identification clouds also serve to distinguish rather different instances of the same basic idea in different specializations. E.g. *spectrum* of a commutative algebra in mathematics, *spectrum* of an operator in a different part of mathematics, and *spectrum* (of a substance) in physics or chemistry are distantly related and ultimately based on the same idea but are in practice completely different terms.

Possibly an even worse problem is caused by phrases and words which have very specific technical meanings but also occur in scientific texts in everyday language meanings. A nice example is the technical concept "end" as it occurs in group theory, topology and complex function theory (three technically different though related concepts). Searching for "end" in a large database such as MATH of FIZ/STN (Berlin, Karlsruhe) is completely hopeless. Searching for "end" together with its ID cloud for its technical meaning in group theory would be a completely different matter. Note that specifying group theory as well in the query would not help much; there are simply too many ways in which the word 'end' occurs (end of a section, to this end, end of the argument, end of proof, ..). There are many more words like this; also phrases. For more about the 'story of ends', see [7].

7. Automatic generation of identification clouds.

Take a large enough, well indexed corpus, and divide it into suitable chunks called documents. For instance take the 700 000 abstracts of articles in the STN/FIZ database Math (ZMG data)[1], or take as documents the sections or pages of a large handbook or encyclopaedia such as the Handbook of Theoretical Computer Science, [10] or the Encyclopaedia of Mathematics, [2] or an index like [6, 9] Now use a parser for prepositional noun phrases (PNP's) (or an automaton recognizing PNP's) or a software indexing program like KEA, TExTract or CLARIT, to generate from these documents a list of key phrases, keeping track of what phrases come from what document. Now assign, as ID clouds, to the items of the list of keyphrases those words and phrases found by, say, the software indexing program, which occur in the same document as the key phrase under consideration.

---

[1] Though this one is not really well indexed in the sencse that the key phrases assigned are not from a controlled list. However, if the intention would be to generate the controlled list at the same time as the correponding ID clouds, this material would be most suitable.

8. Weights.

One thing that emerged out of the use of identification clouds in the project TRIAL SOLUTION was that it is wise to give weights (numbers between 0 and 1 adding up to 1) to the elements making up an identification cloud. How to assign weights optimally is a large problem. Obviously this cannot be done by hand: a more or less adequate list of standard key phrases needs at least 150 000 terms. I propose to use something like the following adaptive procedure.

Suppose one has an identification cloud of a term consisting of items $1,\cdots,n$ with weights $p_1,p_2,\cdots,p_n$ adding up to 1. Let a subset $S\subset\{1,2,\cdots,n\}$ be successful in identifying the phrase involved. Then the new weights are:

$$\text{For } i\in S,\quad p_i' = p_i\left(\frac{\sum_{i\in S} p_i + r(1-\sum_{i\in S} p_i)}{\sum_{i\in S} p_i}\right)$$

$$\text{For } i\notin S,\quad p_i' = p_i - rp_i$$

where r is a fixed number to be chosen, $0<r<1$. (Note that the new weights again add up to 1; note also that the $i\in S$ increase in relative importance and the $i\notin S$ decrease in relative importance; if $S=\{1,\cdots,n\}$ nothing happens.) This is an adaptation of a reasonably well known algorithm for communication (telephone call) routing that works well in practice but is otherwise still quite fairly mysterious.

9. Further refinements and issues.

Another refinement that came out of the experiences with the TRIAL SOLUTION project is that it could be a very good idea to allow negative weights. Let's look at an example.

"The next topic to be discussed is that of the Fibonacci *numbers*. The generating formula is very simple. But all in all these numbers and their surprisingly many applications are sufficiently *complex* to make the topic very interesting. Similar things happen in the study of fractals."

Both 'complex' and 'numbers' occur in this fragment of text (italized). But, obviously it would be totally inappropriate to assign the technical keyphrase 'complex numbers' to this fragment. A negative weight on 'Fibonacci' in the ID cloud of 'complex numbers' will prevent that.

The concrete example of section 2 above also illustrates the possible value of negative information.

There are a good many other issues to be addressed. Here is one. It is more or less obvious that making one keyphrase list with ID clouds for all of science and technology is a hopeless task. What one aims at is instead an Atlas of Science and Technology consisting of many weak thesauri that partially overlap and may have different levels of detail. Here the problem arises of how to match the different 'charts'.

Another issue is how to adapt the adaptive scheme of the previous section to a situation with negative weights and how to handle insertion and deletion of ID cloud members.

Probably the most crucial issue to be addressed at this stage is the formulation of a good probabilistic model of ID clouds complete with statistical estimators. A project in this direction has been started by the CWI, Amsterdam together with the IMI, Lithuanian Acad. of Sciences, Vilnius.

References.

1. Jean Aitchison, Alan Gilchrist, *Thesaurus construction*, Aslib, 2-nd Edition, 1990.

2. M Hazewinkel (ed.), *Encyclopaedia of mathematics; 13 volumes including three supplements*, KAP, 1988-2001.

3. Michiel Hazewinkel, *Classification in mathematics, discrete metric spaces, and approximation by trees*, Nieuw Archief voor Wiskunde **13** (1995), 325-361.

4. Michiel Hazewinkel, *Enriched thesauri and their uses in information storage and retrieval.* In: C Thanos (ed.), Proceedings of the first DELOS workshop, Sophia Antipolis, March 1996, INRIA, 1997, 27-32.

5. Michiel Hazewinkel, *Topologies and metrics on information spaces.* In: J Plümer R Schwänzl (ed.), Proceedings of the workshop: "Metadata: qualifying web objects", http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html, 1997,

6. Michiel Hazewinkel, *Index "Theoretical Computer Science", Volumes 1-200* 68, Theoretical Computer Science **213/214**(1999), 1-699.

7. Michiel Hazewinkel, *Key words and key phrases in scientific databases. Aspects of guaranteeing output quality for databases of information.* In: Proceedings of the ISI conference on Statistical Publishing, Warsaw, August 1999, ISI, 1999, 44-48.

8. Michiel Hazewinkel, *Topologies and metrics on information spaces*, CWI Quarterly **12**:2 (1999), 93-110. Preliminary version: http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html.

9. Michiel Hazewinkel, *Index Discrete Mathematics Vols 1-200* 68, Discrete Mathematics **227/228** (2001), 1-648.

10. Jan van Leeuwen (ed.), *Handbook of theoretical computer science*, Elsevier, 1990.