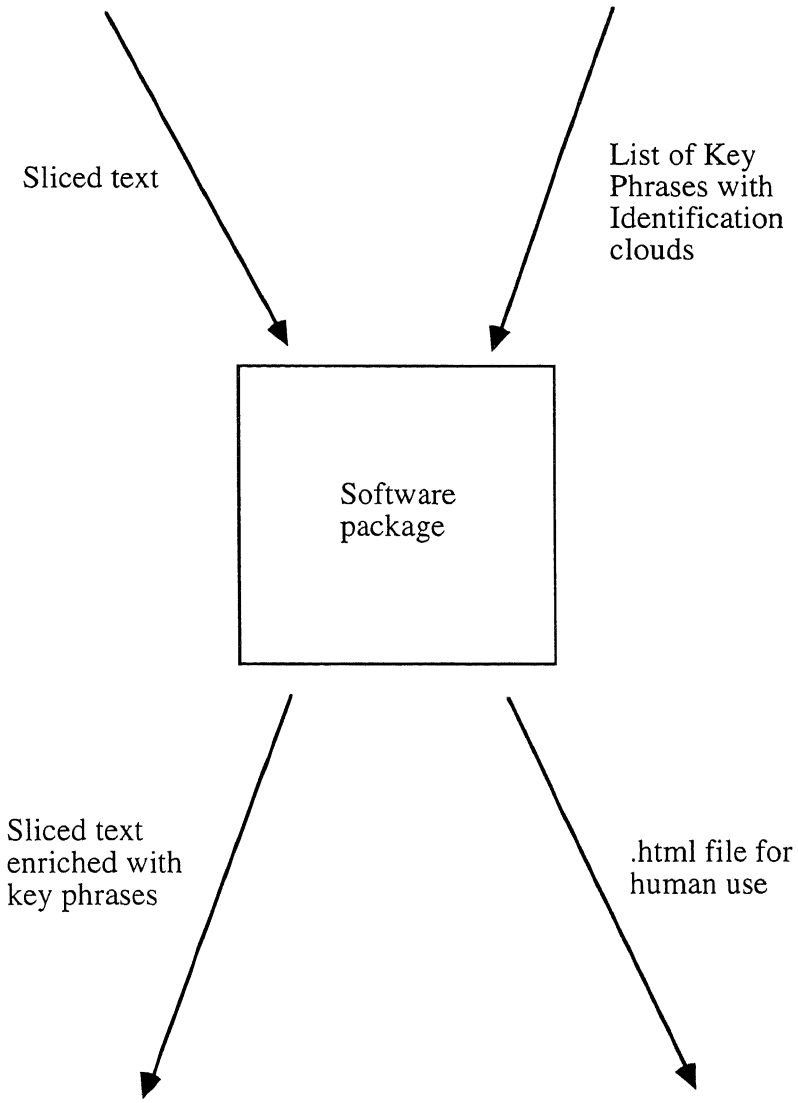
Identification clouds and automatic assignment of key phrases: lessons learned from the TRIAL SOLUTION project

by

Michiel Hazewinkel
CWI
POBox 94079
1090GB Amsterdam
The Netherlands
<mich@cwi.nl>

1. Introduction.

The "Automatic Key Phrase Assigner" deliverable of the TRIAL SOLUTION project (IST-1999-11397: Febr. 2000-May 2003) is a piece of software that pictorially can be depicted as follows



Thus, there are two inputs:

• a 'sliced text'. That means a scientific text, say an undergraduate textbook on analysis (such as were actually used during the project), sliced, that means cut-up in chunks that are small enough to be coherent and reuseable and large enough to have internal meaning. The sliced text comes in a special packaged format that is universal for all TRIAL SOLUTION components.

• a list of key phrases together with for each key phrase its identification cloud. This concept will be described in more detail below in section 2. But, roughly the identification cloud of a standard key phrase is a list of words (and maybe very short other key phrases) that one expects (often) to find in the neighbourhood of the key phrase concept that is being discussed.

There are also two outputs

• the same sliced text, packaged in the same way, enriched with key phrases that have been assigned to each slice.

• a .html file for human use that is to be used for expert validation of the the automatic key phrase assignment. This one, when displayed, highlights the key phrases that have been found and also gives the 'evidence' for assigning them in terms of the items from the identification cloud that were found and the percentage (weight) of the identification cloud that was found.

2. The idea of identification clouds.

The rough idea of an identification cloud (for a key pharse, or concept, or formula) is the following.

It may well happen that a very good key phrase for a given article is simply not present or so linguisitically mangled that even the best NLP techniques (Natural Language Processing) are not good enough to recognize it. Yet professionals in the field in question will have no difficulty in attributing that key phrase to that article. They do that by context.

The basic idea is that a good key phrase can be recognised by (part of) the collection of characteristic words (and short phrases) that one would normally expect to find in its neighbourhood. Thus the idenfification cloud of a (standardized) key phrase is a collection of words and short key phrases that belong to it and can be expected to show up in the immediate surrounding text. The same idea applies to formulas. If, for instance, the surrounding text is all about projective systems and projections and/or inverse ssytems then a 'π' in that neighbourhood is likely to be a projection type mapping; if it is all about groups and representation and compositions the same 'π' is likely to be a permutation (or possibly a composition).

A well known quote is

> "Tell me what company thou keepest, and I will tell thee what thou art"
>
> Miguel de Cervantes (1547-1616),
> Don Quixote, Pt II, Ch. 23

The same applies to formulas and (missing) key phrases.

When the idea was started in 1995, see [11, 12, 13], I thought it was original. As it turns out it was not. An earlier version is from [10]. But there the context is syntactical rather than semantical as in the case of identification clouds.

Perhaps it should be stressed just how important context is. In everyday literary language the meaning of a string of alphabetical characters is absolutely totally dependent on context. Here is an example where the part of the context that is present is insufficient.

> "Many thanks for your book. I shall lose no time in reading it"
>
> Benjamin Disraeli (1804-1881)

There are two meanings, of course, to the second sentence. But, unless the reader is aware of further context, it will not be evident that Disraeli did not intent to waste a further second on that gift.

Here is an example of an identification cloud (in the primitive original sense):

Key Phrase:
Darboux transformation

Id Cloud:
soliton
dressing transformation
Liouville integrable
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))

And, as a matter of fact this ID Cloud did solve a real key phrase assignment problem. See [13, 16] for details.

Potential and actual applications of the general idea of identification clouds include:
- dialogue mediated information retrieval
- distances in information spaces.
- disambiguation
- slicing texts (TRIAL SOLUTION)
- formula recognition
- synonyms
- crosslingual IR
- automatic classification
- automatic key phrase assignment (TRIAL SOLUTION)

The last application was an integral part of TRIAL SOLUTION. The other one marked with this phrase gives a potential answer to the question how to slice a text when it is not a well marked and structured document like a LaTeX document, see the discussion note from 'TRIAL' [14].

3. Lesson 1 from TRIAL: 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.

Here is an example:

```
<KEYPHRASE NAME=<Burgers-Gleichung> THRESHOLD=<0.67>>
    <WORD VALUE=<Burgers-Gleichung> WEIGHT=<0.7>>
    <WORD VALUE=<Burgers> WEIGHT=<0.4>>
    <WORD VALUE=<Gleichung> WEIGHT=<0.2>>
    <WORD VALUE=<Boussinesq> WEIGHT=<0.025>>
    <WORD VALUE=<nichtlinear> WEIGHT=<0.025>>
    <WORD VALUE=<Evolutionsgleichung> WEIGHT=<0.025>>
```

<WORD VALUE=<Solitonlösung> WEIGHT=<0.025>>
<WORD VALUE=<Transformation> WEIGHT=<0.025>>
<WORD VALUE=<KdV> WEIGHT=<0.025>>
<WORD VALUE=<sinh> WEIGHT=<0.025>>
<WORD VALUE=<Gordon> WEIGHT=<0.025>>
<WORD VALUE=<Hirota> WEIGHT=<0.025>>
<WORD VALUE=<Kadomzev> WEIGHT=<0.025>>
<WORD VALUE=<Pedviashwili> WEIGHT=<0.025>>
<WORD VALUE=<Soliton> WEIGHT=<0.025>>
<WORD VALUE=<Bäcklund> WEIGHT=<0.025>>
<WORD VALUE=<inverse spektral> WEIGHT=<0.025>>
<WORD VALUE=<HOPF> WEIGHT=<0.025>>
<WORD VALUE=<COLE> WEIGHT=<0.025>>
<\KEYPHRASE>

Here the "threshold value" in the first line means that the key phrase "Burgers Gleichung" is to assigned to a chunk of text iff enough identification cloud terms are recognized so that their combined weights add up to 0.67 or more.

Of course if the phrase itself occurs that is enough as reflected by the first item in the 'WORD VALUE list'. Note further that the occurrence of "Burgers" and of "equation" is not quite enough. There is a good reason for that. For one thing there is also a concept called "Burgers vector" (in connection with torsion in differential geometry); also "Burgers" is a fairly common surname. Further "equation" (= Gleichung) is of such frequent occurrence (in mathematics) that it can turn up just about anywhere. Thus the occurrence of both "Burgers" and "equation" in a chunk of text is not enough to decide that "Burgers equation" is a suitable key phrase for that chunk. But if three or more of the sort of words that belong to completely integrable dynamical systems are also present one can be quite sure that it is indeed a suitable key phrase.

This particular identification cloud is designed to find occurrences of the Burgers equation as it occurs in the area of completely integrable dynamical systems (soliton equations, Liouville integrable systems).

On the other hand the Burgers equation (the same one, not a different object with the same name) also occurs in a quite different area of mathematics. It is the simplest nonlinear diffusion equation and plays a role as such and in discussions of turbulence. To catch those occurrences a rather different set of supporting words and phrases is needed (like diffusion, turbulence, eddy, nonlinearity,...). Thus a second identification cloud is needed. Just combining the two identification clouds would be dangerous.

4. Lesson 2 from TRIAL: need for automated generation tools

Creating a classical thesaurus for a given field according to various international and other standards, [1], is an immense task and generally considerd undoable in the old fashioned expert-by-hand way, [4, 5]. Moreover it is a structure that is not by nature incrementally updatable.

The concept of an enriched weak thesaurus, [12], was evolved in order to have something comparable but which is incrementally updatable. Controlled key phrase lists are a main part of a weak enriched thesaurus, as are identification clouds.

Making an adequate controlled key phrase list for a given scientific field is still a very very large job. Certainly if identification clouds are to be included. For mathematics one would need something like 150 000 key phrases. It is fair to say that at the beginning of the TRIAL SOLUTION project the size of this task was seriously underestimated.

Thus there is a real need for automatized tools to do the job.

One idea is to use one of the first generation key phrase extractors (back of the book

indexing systems), like KEA (freely available, [18]), or TexTract, or CLARIT (both commercial, [2, 6, 7, 8]) to generate a first candidate list, and this also gives a first list of significant words, and, keeping track of what words occur where, these are to be seen as preliminary identification clouds.

Now start an iterative procedure to improve and update both, ideally using weights. There is a good weight updating formula, see below.

Some question which are not settled yet are the insertion and deletion of elements of identification clouds and the use of negative weights.

Some of the tools from TRIAL SOLUTION (IST-1999-11397) can be adapted to this task.

Human intervention will need to be interspersed in the iteration loops to check that things are going well. Thus we are really talking about semi-automated tools (at least for the moment).

This particular idea has been proposed to be carried out (MKM-NoE proposal, FP6).

As already mentioned, it has become clear from the project TRIAL SOLUTION that, at least in some cases, the idea of identification clouds needs refinement, specifically one needs sometimes weights and even negative weights (to rule out spurious assignments of key phrases; see below). Also in a longer text obviously the mere fact that most of the identification cloud of a term is present is not enough; they should occur more or less grouped. Thus one needs (stochastic) models of identification clouds.

A preliminary investigation of this matter is a task within the current network MKMnet (Task 2.2). Gathering of experimental data is currently under way.

The next step is to fit known statistical distributions and to develop statistical estimators. This is needed for the optimal assigment of weights and also for the semi-automatic generation of identification clouds.

There is something like a three-way chicken and egg problem here. To obtain optimal weights for the identification clouds one needs a stochastic model; to get such a model one needs statistical data on identification clouds; and to collect these data one needs the weights.

Still something can be done by starting with a known controlled list of key phrases for instance for discrete mathematics (M Hazewinkel, Index for the Journal 'Discrete Applied Mathematics, Volumes 1-91 (about 20500 terms (not counting linguistic variations); M Hazewinkel, Index for the Journal Discrete Mathematics volumes 1-200 (about 30000 terms) and using iteration. Another smaller starting point could be the thesaurus for commutative algebra (some 1300 terms) which came out of the INTAS project ERETIMA (INTAS 97-0741), [9].

Concrete proposals to carry out some of these ideas are in the INGADIM and CITIZEMS proposals for FP6.

Now that weights have entered the picture, one also needs automatic updating of them. That can be handled by an adaptation of an adaptive algorithm that is successful in the routing of telephone calls, [3, 17]. The adapted algorithm looks as follows:

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:

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

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

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.) As said, 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.

5. Lesson 3 from TRIAL SOLUTION: negative weights.
Another concept for the refinement of the idea of identifications clouds 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."

Or even worse:

"These mixed spectrum solutions must be *numbered* amoung the more *complex* ones of the KdV equation. Still they can be not neglected."

Both 'complex' and 'numbers' occur in the first fragment of text above (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.

For the second text fragment the technique of stemming, which needs to be used, will give "number", and "complex" also occurs. But here also it would be totally inappropriate to assign the key phrase "complex numbers". It is not so easy to see how to avoid this.

There are still other possible sources of difficulties because "complex" is also a technical term in algebraic topology and homological algebra so one can have a fragment like

"The Betti numbers of this cell complex are..."

or still worse:

"The idea of a simplicial complex numbers amoung the most versatile notions that ..."

Here even the exact phrase "complex numbers" occurs and negative weights are a must to avoid a spurious assignment.

Quite generally it seems fairly clear that the presence of the constituents of a standard key phrase in a given chunk of text is by no means sufficient to be sure that that key phrase is indeed appropriate. This is especially the case for concepts that are made up out of frequently occurring words like "complex numbers" or "boundary value formula". But we have also seen this in the case of the "Burgers equation" above. For the case of the phrase "complex numbers" one needs an identification cloud like

<KEYPHRASE NAME=<complex numbers> THRESHOLD=<0.47>>
<WORD VALUE=<complex numbers> WEIGHT=<0.5>>
<WORD VALUE=<complex> WEIGHT=<0.2>>
<WORD VALUE=<numbers> WEIGHT=<0.2>>
<WORD VALUE=<field> WEIGHT=<0.06>>
<WORD VALUE=<imaginary part> WEIGHT=<0.06>>
<WORD VALUE=<real part> WEIGHT=<0.06>>
<WORD VALUE=<absolute value> WEIGHT=<0.06>>
<WORD VALUE=<Gauss> WEIGHT=<0.06>>
<WORD VALUE=<argument> WEIGHT=<0.06>>
<WORD VALUE=<principal value> WEIGHT=<0.06>>

<WORD VALUE=<vector representation> WEIGHT=<0.06>>
<WORD VALUE=<addition> WEIGHT=<0.06>>
<WORD VALUE=<multiplication> WEIGHT=<0.06>>
<WORD VALUE=<Fibonacci> WEIGHT=<-0.5>>
<WORD VALUE=<Betti> WEIGHT=<-0.5>>
<WORD VALUE=<simplicial complex> WEIGHT=<-0.5>>
<\KEYPHRASE>

So that besides "complex" and "number" one needs at least 2 more bits of supporting evidence to have a reasonable chance that the fragment in question is indeed has to do with the field of complex numbers. On the other hand if at least 8 of the last ten positive weight terms of the identifcation cloud above are present one is also rather sure that the fragment in question has to do with the field of complex numbers. The tentative identification cloud given above reflects this. But it is clear that assigning weights properly is a delicate matter; it is also clear that much can be done with weights.

Thus also in the case of occurrences of the same concept in the same part of mathematics, more than one identification cloud may be a good idea, reflecting different styles of presentation and different terminological traditions.

6. Lesson 4 from TRIAL SOLUTION: text levels.

Originally the idea of identification clouds was developed for use in the context of research level texts. I.e. research papers and also abstracts of those.

The (stochastic) model of occurrences in that type of text differs significantly from that of textbooks, especially undergraduate level textbooks, such as the material that the TRIAL SOLUTION project was concerned with.

Thus, for instance, whether negative weights are really needed in research level texts remains something to be tested. The same holds for weights.

Concrete stochastic models, based on real data, such as briefly discussed in section 4 above, can do much to settle these matters.

7. Coda.

To the writer of these pages it is clear that the project TRIAL SOLUTION has been very fruitful in contributing to the ideas surrounding enriched weak thesauri and identification clouds and as such the project has given birth to many other potential activities, see above.

Of course, this was by no means the only concern of TRIAL SOLUTION.

Many of the things mentioned here are more fully discussed in [16] and there is overlap between this discussion note and that survey paper. Some other related matters are discussed in [15]. Both these papers own their existence in a certain measure to the TRIAL SOLUTION project.

References.

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

2. H Bego, *TExtraxt. Back-of-the-book index creation system*, 1997.

3. G Bel, P Chemouil, J M Garsia, F Le Gall, J Bernusso, *Adaptive traffic routing in telephone networks*, Large Scale Systems **8** (1985), 267-282.

4. Ian Crowlesmith, *Creating a treasure trove of words*, Elsevier Science World. 14-15, 1993.

5. Ian Crowlesmith, *The development of a biomedical thesaurus*, NBBI Thesaurus Seminar. 1993.

6. David A Evans, *Snapshots of the Clarit text retrieval*, Preprint, copies of slides, Carnegie Mellon university, 1994.

7. D A Evans, K Ginther-Webster, M Hart, R G Lefferts, I A Monarch, *Automatic indexing using selective NLP and first-order thesauri.* In: A Lichnérowicz (ed.), Intelligent text and image handling, Elsevier, 1991, 524-643.

8. David M Evans, Robert C Lefferts, *Clarit-Trec experiments*, Preprint, Carnegie Mellon, 1994.

9. R V Gamkrelidze, F Guenthner, M Hazewinkel, A I Onishchik, *Eretima: English Russian bilingual thesaurus for Invariant theory, Lie groups, Algebraic geometry, Dynamical systems, Optimal control, Commutative algebra. INTAS project 96-0741*, 2001.

10. Gregory Grefenstette, *Explorations in automatic thesaurus discovery*, KAP, 1994.

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

12. 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.

13. 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

14. Michiel Hazewinkel, *Statistics of information clouds*, Discussion paper, CWI, Amsterdam, 2001.

15. Michiel Hazewinkel, *Dialogue mediated information retrieval, automatic key phrase assignment and identifications clouds.* In: Proceedings 'Crimea 2002'. Nine-th international conference: libraries and associations in a transient world, new technologies and new forms of cooperation, 2002, 212-221.

16. Michiel Hazewinkel, *Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage.* In: H Gzyl (ed.), Surveys from IWAP 2002, KAP, 2003,

17. P R Srikantakumar, K S Narendra, *A learning model for routing in telephone networks*, SIAM J. Control and Optimization **20**:1 (1982), 34-57.

18. Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, Craig C Nevill-Manning, *KEA: practical automatic keyphrase extraction*, Univ. of Waikato, 1999.