Keyphrase assignment.
Comparision between 'identification cloud' based automatic methods and human experts.
Deliverable D2.5 of MKMnet.

by

Michiel Hazewinkel
CWI, Amsterdam

1. Introduction.

For the idea of 'identification clouds' and what can conceivably be done with them, see M Hazewinkel, Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage, In H Gzyl (ed.) Applied probability 2002. Survey papers arising from IWAP2002, Caracas, Venezuela, Jan. 2002, KAP, 2004 (which is also deleverable D2.2 of MKMnet), and M Hazewinkel, Dialogue mediated information retrieval, automatic keyphrase assignement and identification clouds, In: Proc, Crimea 2002, nine-th international conference: libraries and associations in a transient word, new technologies and new forms of cooperation.

One of the potential applications of identification clouds is the automatic assignment of keyphrases to (abstracts of) documents. The basic idea here is to incorporate some human expert knowledge in the automatic system on the basis of the fact that a human expert can pretty quickly tell what a given scientific paper is about by running his eye over it and noting a number of characteristic words (and formulas). This is the essential idea of identification clouds: a potential key phrase comes with a 'cloud' of words (and possibly other data) that one expects (to some extent) to see in its neighbourhood. If enough of the cloud is present the key phrase in question is probably a good one to attach to the document examined.

When it is attempted to implement the idea a number of problems surface. Many of these have occurred before, such as inguistic ones (linguistic variations, stemming, ...), programming ones (standard formatting, packaging, ...) and a variety of techniques exist to deal with them (more or less adequately). A new problem that surfaces is of a statistical nature. Obviously if the items of an identification clouds are spread out widely it means far less than if they are nicely custered together. So to have something like optimal identification clouds one needs to know something about the statistics of key phrases and key words as they occur in documents.

This now appears to be something like a major obstruction. There is simply not enough readily accessibe material to work with. For instance, there is a handmade index of the major field 'discrete mathematics' on the basis of the 6000 or so articles that appeared in the journal of that title (the first 200 volumes).[1] There are some 30000 key phrases and on average they occur just twice. That is utterly insufficient to get any idea of the statistical distributions involved.

Still it is possible to select a few keyphrases, concoct identification clouds by hand, and see how the automatic machinery works on these data. That has been done and the present document reports on this experiment.

2. Set-up..

To test the idea of identification clouds 16 key phrases were selected from the index of R L Graham a.o. , Handbook of combinatorics, Elsevier, 1995, 2303 pp. The corresponding identification clouds were composed by reading in loc. cit. and noting which significant sounding words occurred in the surrounding text. It is unlikely that these ID clouds are anywhere near optimal.

The 16 key phrases selected are the following:

[1] See M Hazewinkel, Subject index volumes1-200 of the journal 'Dsicrete Mathematics, discrete Mathematics,**227/228** (2001), 1-648.

transitive group
induced subgraph
module of an arithmetic sequence
packing problem
Schur algebra
asymptotic property (in discrete mathematics)
chromatic polynomial
domination problem (in discrete mathematics)
domination problem (in topology)
complement of a graph
Erdos-Szekeres conjecture
Sierpinski gasket
Cayley graph
Bruck-Chowla-Ryser theorem
shellable complex
Steiner triple system.

Actually, originally 19 keyphrases were selected. But there was an unfortunate transcription error (2 missing left angled brackets) and this caused three of these to be lumped together to one rather large ID cloud labelled 'Artin braid group'.[2] These three are:

Artin braid group
Eulerian graph
blocking number

This caused the program to find 39 articles to which 'Artin braid group' is assigned. Three of these are completely correct and this fits perfectly, because there are in fact precisely three articles in the control (human made) index with this key phrase (and they are the same). The remaining 36 in great majority fit 'Eulerian graph' (using the circumstance that Eulerian and Hamiltonian graphs are closely related).

These key phrases with their corresponding identification clouds are listed in a separate document that goes herewith. A few comments are in order. The material (corpus) on which these were to be tested were around 6000 abstracts from the journal 'Discrete Mathematics, volumes 1-200'. Given that setting it would have been silly to include one word key phrases like 'graph', 'edge', 'vertex'; within this corpus these carry almost no informational meaning at all.

Also note that 'domination problem' occurs twice. There are two completey different meanings to this phrase and as far as is known there are no relations. One is in combinatorics and has been the subject of a monograph; the other is in low dimensional topology. One would not expect relevant material to this second one in the corpus to be tested.

Apart from the remarks just made the selection of these 19 key phrases was as random as can be (in so far as any body of knowledge can be independent of another in mathematics).

3. The data from the Journal of Discrete Mathematics.

An index was made of the 'Journal of Discrete Mathematics, volumes 1-200' a number of years ago by a humen mathematician (M Hazewinkel, Subject index Volumes 1-200, Discrete Mathematics **227/228** (2001), 1-648. As said the key phrases come fom the Handbook of Combinatorics. It was not even checked that the nineteen key phrases selected occurred in the DISC index just mentioned. And in fact some of them don't; at least not precisely.

Here are the data from the index just mentioned concerning these 19 keyphrases. the numbers refer to the articles pubished in the journal cited as listed in loc. cit.

transitive group, 3379
    of automorphisms 736, 3482
    of bounded automorphisms 5469
induced subgraph 391 ... 5906 (94 listings in total)
module of an arithmetic sequence

[2] This illustrates once more how extraordinary sensitive software programs can be. A human would have spotted this immediately and corrected things. Much more robust programming is needed for these linguistic applications.

(none). But there is ‘modulus of an arithmetic progression, 1998
packing problem 722 ... 5464 (13 lisitings)
packing problem of Lovasz 63
Schur algebra 2885
asymptotic property (in discrete mathematics) 109, 2881, 4635
chromatic polynomial 141 ... 6018 (42 listings)
chromatic plynomial of a graph 1935 ... 5343 (9 listings)
domination problem (in discrete mathematics), 2861, 2864, 2866, 2868
domination problem (in topology), none
complement of a graph, 54, ..., 3820 (16 listings)
Erdos-Szekeres conjecture, 4879
Sierpinski gasket, 5894
Cayley graph, 604 - 5682 (38 listings)
of a cyclic group, 4720
of a finite group, 735
of a group, 1525, 3486
of an Abelian group, 5682
of the group of integers, 5588
on an Abelian group, 4109, 4755
structure, 5588
Artin braid group,
Artin coloured braid group, 4869
Artin pure braid group, 4869
(braid group, 3358, 3360, 4869)
Eulerian graph, 1733, 2215, 3141, 4605, 5222
blocking number,
Bruck-Chowla-Ryser theorem, none
shellable complex,
(shellability of a simplical complex, 2690, 4839)
Steiner triple system, 28, ..., 5597 (47 listings)
(STS, 724, ..., 5402 (13 listings).

4. Description of the automatic keyphrase assignment package.
The sofware package involved, was written by Kees Blom, CWI, Amsterdam. It takes a list of ID clouds (for examples see the separate document alreeady mentiooned that goes herewith) and a list of abstracts and produces an .html file which is the same as the list of abstracts but with those abstracts highlighted (in yellow) for which a candidate keyphrase was found. In addition the relevant ID cloud is added (in front) and the items which constitute the evidence for the assignment of the keyphrase to this abstract are indicated (with weights). The original version of the package only worked for IMS packaged list of abstracts. So now there is a prepocessor that turns a list of abstracts into an IMS packaged collection of chunks of text.

The software package is freely available and can be obtained from Kees Blom <Kees.Blom@cwi.nl>.

5. The results of the test.
All in all the automatic keyphrase assigment problem returned 113 hits from 5826 abstracts (mostly very short abstracts). This is rather less than the sum total of the numbers of listings in section 3 above. The correponding .html files goes herewith as a separate document. The ‘hits’ are highlighteed in yellow and list the corresponding keyphrase and identification clouds.

Non-English language abstracts were excluded (bringing the total from 6034 to 5826).

Here is a list of the hits with from time to time some comments. Note that the 39 hits ‘Artin braid group’ have been retained so that the list before exactly corresponds to the output file aka.html.

69: transitive group. This one does not occur in DISC index (the expert madeindex mentioned in section 3. There is no doubt that this is an appropriated index phrase for this articles.

173: chromatic polynomial. As in the DISC index.

284: Artin braid group. Not in DISC index and not appropriate

309: complement of a graph. Not in the DISC index but appropriate for this article.

310: Steiner triple system. As in the DISC index.

317: Cayley graph. Appropriate. Not in the DISC index as such. But 'Cayley colour graph, 317' occurs in the DISC index.

391: complement of a graph. Not in the DISC index, but (mildly) appropriate.

447: Artin braid group. Not in DISC index and not particularly appropriate.

604: Artin braid group. Not in DISC index and not appropriate.

604: Cayley graph. In DISC index and appropriate.

679: Artin braid group. Not in DISC index and not appropriate

724: Steiner triple system. In DISC index and appropriate

735: Artin braid group. Not in DISC index and not appropriate.

890: Artin braid group. Not in DISC index and not appropriate.

913: Artin braid group. Not in DISC index and not appropriate.

946: Artin braid group. Not in DISC index and not appropriate.

1037: transitive group. Not in DISC index but entirely appropriate.

1039: Steiner triple system. In DISC index and entirely appropriate

1201: Artin braid group. Not in DISC index and not appropriate

1225: Modulus of an arithmetic sequence. Not in DISC index. Mildly appropriate.

1294: chromatic polynomial. In DISC indec and entirely appropriate

1514: Steiner triple system. Not in DISC index as such, but in DISC index under Steiner system (not entirely correctly). Entirely appropriate.

1524: Artin braid group. Not in DISC index and not appropriate

1733: Artin braid group. Not in DISC index and not appropriate

1772: Steiner triple system. In DISC index and entirely appppropriate.

1802: Artin braid group. Not in DISC index and not appropriate.

1843: Steiner triple system. Not in DISC index as such, but does occur as STS(15) (which is not good enough). Entirely appropriate.

1953: chromatic polynomial. Not in DISC index, but entirely appropriate.

2057: Artin braid group. Not in DISC index and not appropriate.

2099: transitive group. Not in DISC index but enirely appropriate. But does occur in the DISC index as 'transitive permuation group'.

2196: Steiner triple system. In DISC index and entrely appropriate.

2206: modulus of an arithmetic sequence. Not in DISC index but more or less appropriate.

2246: Steiner triple system. In DISC index and entrely appropriate.

2396: Artin braid group. Not in DISC index and not appropriate.

2521: Artin braid group. Not in DISC index and not appropriate.

2589: Steiner triple system. In DISC index and entrely appropriate.

2596: Steiner triple system. In DISC index and entrely appropriate.

2619: Steiner triple system. In DISC index and entrely appropriate.

2658: complement of a graph. Not in DISC index and not appropriate. Reason it was picked up nevertheless is that there is a remark in the abstract saying that the algorithms discussed here 'complement' a number of well known ones.

2690: shellable complex. In DISC index (as 'shellability of a simplicial complex') and entirely appropriate.

2707: Artin braid group. Not in DISC index and not appropriate.

2806: Artin braid group. Not in DISC index and not appropriate.

2830: chromatic polynomial. In DISC index and appropriate.

2843: Artin braid group. Not in DISC index and not appropriate.

2859: domination problem (for graphs). Not in DISC index but entirely appropriate. The listing 2859 does occur in the DISC index under 'domination number'.

2861: domination problem (for graphs). In DISC index and entirely appropriate

2875: domination problem (for graphs). Not in DISC index but entirely appropriate. The listing 2875 does occur in the DISC index under 'domination number'.

2856: Cayley graph. Not in DISC index but entirely appropriate.

3058: Steiner triple system. In DISC index and entrely appropriate.

3115: Artin braid group. Not in DISC index and not appropriate.

3141: Artin braid group. Not in DISC index and not appropriate.

3160: Steiner triple system. Entirely appropriate but not in DISC index.

3164: Steiner triple system. Entirely appropriate but in DISC index only under 'Steiner system' (which is not really correct for this item).

3207: Artin braid group. Not in DISC index and not appropriate.

3309: modulus of an arithmetic sequence. Not in DISC index but entirely appropriate.

3356: Artin braid group. Not in DISC index and not appropriate.

3358: Artin braid group. Entirely appropriate and in DISC index

3360: Artin braid group. Entirely appropriate and in DISC index

3387: Artin braid group. Not in DISC index and not appropriate.

3464: Steiner triple system. Not in DISC index and entrely appropriate.

3486: Cayley graph: Entirely appropriate and in DISC index (as Cayley graph of a group which is not entirely appropriate).

3527: Cayley graph. Entirely appropriate and in DISC index.

3548: Cayley graph. Entirely appropriate and in DISC index.

3643: complement of a graph. Appropriate but not in DISC index.

3847: induced subgraph. Appropriate but not in DISC index.

3918: Cayley graph. Entirely appropriate and in DISC index.

3921: Artin braid group. Not in DISC index and not appropriate.

3959: chromatic polynomial. Appropriate and in DISC index.

4009: chromatic polynomial. Appropriate and in DISC index.

4015: Steiner triple system. Not in DISC index and entirely appropriate.

4070: Artin braid group. Not in DISC index and not appropriate.

4210: Artin braid group. Not in DISC index and not appropriate.

4305: domination problem (for graphs): Appropriate and in DISC index (but under 'domination number').

4352: domination problem (for graphs): Appropriate and in DISC index (but under 'domination number').

4363: Artin braid group. Not in DISC index and not appropriate.

4379: Artin braid group. Not in DISC index and not appropriate.

4407: chromatic polynomial. Appropriate and in DISC index.

4429: complement of a graph. Appropriate and not in DISC index.

4515: induced subgraph. Appropriate and in DISC index.

4623: Artin braid group. Not in DISC index and not appropriate.

4720: Cayley graph. Entirely appropriate and in DISC index.

4728: Steiner triple system. In DISC index and entirely appropriate.

4793: Cayley graph. Entirely appropriate and in DISC index.

4805: domination problem (for graphs). Apppropriate and in DISC index (but under domination number).

4869: Artin braid group. Entirely appropriate and in DISC index.

4871: complement of a graph. Appropriate and not in DISC index.

4888: Artin braid group. Not in DISC index and not appropriate.

4898: domination problem (for graphs). Apppropriate and in DISC index (but under domination number).

4935: complement of a graph. Appropriate and in DISC index (but only under complement).

4952: Artin braid group. Not in DISC index and not appropriate.

4992: Artin braid group. Not in DISC index and not appropriate.

5184: Steiner triple system. In DISC index and entirely appropriate.

5250: Artin braid group. Not in DISC index and not appropriate.

5265: Steiner triple system. In DISC index (but only under 'Steiner system') and entirely appropriate.

5301: domination problem (for graphs). Appropriate but not in DISC index.

5303: domination problem (for graphs). Appropriate and in DISC index.

5309: Steiner trile systems: Appropriate and in DISC index.

5315: chromatic polynomial. Appropriate and in DISC index.

5402: Steiner trile systems: Appropriate and in DISC index.

5445: Artin braid group. Not in DISC index and not appropriate.

5602: Cayley graph. Appropriate and in DISC index.

5694: Artin braid group. Not in DISC index and not appropriate.

5724: Cayley graph. Appropriate and in DISC index.

5816: domination problem (for graphs). Appropriate and in DISC index (under domination number)

5821: domination problem (for graphs). Appropriate and in DISC index (under domination number for graphs).

5855: chromatic polynomial. In DISC index and appropriate

5868: domination problem (for graphs). Appropriate and in DISC index (under domination).

5891: Artin braid group. Not in DISC index and not appropriate.

5908: Steiner triple system. Appropriate and in DISC index.

5951: domination problem (for graphs). Appropriate and not in DISC index.

5982: domination problem (for graphs). Appropriate and in DISC index.

6. A few numbers and remarks on the results.

Here is a little table on how many instances of the selected keyphrases are in the human made DISC index, how many were found by the aka program working with the given ID clouds which are in that index, and how many were found by that program which were overlooked by the human indexer.

Keyphrase	# in DISC index	# found by aka	# overlooked by human
transitive group	4	2	1
induced subgraph	94	2	0
modulus of ...	1	0	4
packing problem	14	0	0
Schur algebra	1	0	0
asymptotic ...	3	0	0
chromatic poly...	42	8	1
domination (graphs)	4	4	10
domination (topo)	0	0	0
complement ...	16	3	1
Erdos-Szekeres ...	1	0	0
Sierpinsky ...	1	0	0
Cayley graph	38	10	1
Bruck ...	0	0	0
shellable ...	2	1	0
Steiner triple ...	47	19	3
----------	----------	----------	----------
Artin braid group	3	3	0
blocking number	0	0	0
Eulerian graph	5	5	many

The results recorded below the dotted line come from some postautomatic looking (after it was realized that the three ID clouds in question had been merged).

Here are some comments resulting from detailed inspection of the data.

- induced subgraph. A score of 2 out of 94 is very poor (though perhaps not as bad as 0 out of 1). The problem here is that 'induced subgraph' is a very low information content kyephrase (in the field of discrete mathematics). It occurs all over the place just like 'graph', 'edge', 'vertex'. Thus short of setting the weights in such a way that 'induced' and 'subgraph' together yield the threshold it is difficult to see what to do, though there is no doubt that a better identification cloud can be concocted than the one used. The problem with given 'subgraph' plus 'induced' sufficient weight to reach the threshold is that many spurious assignments will immediately result.

For these and other cases discussed below one should probably admit the phrase itself as a threshold level item from the identification cloud.

- Schur algebra, Erdos-Szekeres conjecture, Bruck-Chowla-Ryser theorem, Sierpinsky gasket. These are such specific keyphrases that on the one hand one should not mess about with ID clouds; meaning that if the full phrase occurs it should have weight enough to carry the day. On the other hand if that were the only possibility one would miss quite a few things. To give an ad hoc example the following phrase might occur: 'we will discuss a conjecture of Erdos as later precisized by Szekeres'. On the other hand giving 'Erdos', 'conjecture' (which means very little) and 'Szekeres' together enough weight for this keyphrase would lead to many spurious assignments. There are hunderds of Erdos conjectures and the name Szekeres is also not unknown. Thus some extra evidence is needed.

It is clear, however, that in the case of the four cases mentioned, and also in the case of 'packing problem' too conservative weights were used.

- It is encouraging to see how even in its current far from perfect state the aka package pickes up quite a few assignements that were missed by the human expert.

7. Conclusions and outlook.

According to S Lawrence, C Lee Giles, Accessability of information on the web, Nature **400** (1999), 107-109, the best one can hope for when searching for information on the web is to find some 16 % of the relevant webpages (using 'Northern Light'). There is no evidence that things have improved since; indeed, things have worsened, because of all kind of commercial interest based tricks (and firms) to fake out the search engines.

Using the numbers above (above the dotted line) the relevant scientiic information retrieval percentage is 21.9% (using a very crude way of estimating things); if the information below the dotted line is added this rises to over 30%.

This gives grounds for optimism for the idea of information clouds. But it is also abundently clear that a lot of work remains to be done, such as:

- what is the appropriate 'informational level' for good search phrases.
- how to optimise 'information clouds'.

Some thoughts about these things can be found in deliveralbe D2.2 of this project (MKMnet). Both these things require some insight in the statistical distribution of keyphrases over texts and the corresponding distributions of ID cloud items.

The data used above and similar collections are utterly insufficient for getting an idea of the kind of statistics involved. One would have to work with far larger collections such as the material of the ZMATH database in a given field (such as statistics and probability, or combinatorics).

A project to do just that was submitted to the EC but was not retained for funding.