CWI Quarterly Volume 12 (2) 1999, pp. 93 – 110

Topologies and metrics on information spaces

Michiel Hazewinkel
CWI
P.O. Box 94079, 1090GB Amsterdam, The Netherlands
e-mail: mich@cwi.nl

This paper is concerned with information retrieval from large scientific data bases of scientific literature. The central idea is to define metrics on the information space of terms (key phrases) and the information space of documents. This leads naturally to the idea of an enriched weak thesaurus and the semi- automatic incremental generation of such a tool for information retrieval. Quite a large number of unsolved (mathematical) problems turn up in this context. Some of these are described and discussed. They mostly have to do with missing information and classification and clustering issues.

NOTE. The present text is a revised and expanded version of the write-up of a talk presented at the workshop on "Metadata: qualifying webobjects" that took place at the University of Osnabrück, 12–15 October, 1997, organized by Roland Schwänzl and Judith Plümer [1].

1. INTRODUCTION

This paper is concerned with the matter of finding information in science. More precisely it is concerned with finding information about something using a (rather large) bibliographic database such as the data base MATH of FIZ/STN (Karlsruhe) which has records of basically the whole mathematical literature from 1931 to the present.

Most scientists seem to feel that this is not (yet) a very serious matter. They are well acquainted with what is going on in their own superspecialism and that, together with the informal network of friends and colleagues who can be asked questions, seems to them to work well enough.

[1] See http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html for the write-up of the Osnabrück talk

I disagree. Mainly on the basis of my experiences with [4, 9] (which forced me to find information also outside my own immediate fields of expertise). I think the problem is orders of magnitude worse than one generally realises and I consider it odd that enormous financial investments are made towards new research while almost nothing is devoted to the matter of (re)finding an existing bit of knowledge if and when needed.

Given the size of the information spaces involved (about a million records, about 4 million key phrases for the case of mathematics) it becomes necessary to give some additional structure to these sets. After all, typing in a key phrase and than getting some 6000 hits is not very useful. In this paper I explore the idea of providing these spaces with a metric (and related topological notions) so that it becomes possible to localize one's searches and so that one can thus vastly increases the chance of finding more relevant information and less garbage. This paper almost completely concentrates on these matters and by and large neglects the more traditional (linguistic) concerns in the world of indexes, thesauri and linguistics-based information retrieval.

The key concerns in this paper are the structure of an enriched weak thesaurus (and how to generate such a structure semi- automatically and incrementally), various problems of missing and contaminated (dirty) data, and the mathematical problems (and some results) that arise from these.

2. Enriched weak thesauri and their uses

2.1. Enriched weak thesaurus

Very roughly, a classical thesaurus, according to ISO standard 2788 and various national and international multilingual standards, consists of a (large) list of words and phrases provided with additional structure as follows:

Terms come provided with extra information in the form of broader terms (more general terms), narrower terms (more specific terms), and related terms (terms that are close (in some sense) but are not broader of narrower). There is an implied suggestion that the whole thesaurus has more or less a tree structure.

In addition there is often additional structure in the form of preferred terms, information on synonyms etc. See e.g. [1] for more detailed information. There is no doubt that a thesaurus for a given field of interest is an immensely valuable thing to have. Thesauri are also very expensive to construct; indeed presently almost impossibly expensive. Apart from that, such a structure is very difficult to update incrementally (dynamically).

Thus it is natural to look for structures that can at least be semi- automatically generated and for which the updating problem is less severe. One candidate for such a structure is what I call a enriched weak thesaurus. The definition is as follows. It consists of:

- a (large) list of terms (key words and key phrases), which, together, are adequate to describe a given field of interest, e.g. mathematics.
- a distance function on this set of terms turning it into a semi-metric space.

The phrase semi-metric (instead of metric) here refers to the fact that two different terms can have distance zero; then they are synonyms. The metric replaces the ideas of 'broader -', 'narrower -', and 'related terms' in a classical thesaurus. What is lost is the local partial order [2] given by 'narrower' and 'broader', whence the term 'weak thesaurus'; on the other hand the metric gives (some) quantitative information as opposed to the purely qualitative information of 'broader', 'narrower' and 'related'. There arises the (mathematical) problem, whether perhaps the partial order information can be recovered from the metric information. This is briefly discussed in Section 5.

There are some more bits of structure to incorporate. In addition there is a classification scheme (such as the MSCS (Mathematics Subject Classification Scheme) for mathematics, or PAC for physics and astronomy. Each term of the thesaurus is given one or more classification numbers from the classification scheme, and, conversely, each node in the classification scheme has attached to it all terms from the thesaurus that are linked to that classification node. Thus the classification scheme also gets enriched: the nodes are given content and meaning. This enrichment part roughly consists of:

- links to a classification scheme.

Finally, partly for dealing with problems of missing and inaccurate data, each term is assigned a collection of terms that in actual documents in the field concerned are likely to occur the neighbourhood of that term. I call these

- identification clouds,

and they form the last part of the enrichment structure. There is overlap between this part of the enrichment structure and the metric structure. Conceivably, in very large enriched weak thesauri, for the identification clouds one could simply take all terms that are within a given, to be determined, distance of the term concerned. See Sections 4.4 and 4.5 below for a discussion of how identification clouds are to be used.

2.2. Potential uses of the metric on the information spaces of terms and documents

Let us suppose that we have defined an adequate metric on the information space of documents and the information space of terms. Here are some of the (potential) uses one can make of those.

- Automatic assignment of key phrases and classifications to documents. Given an adequate list of standard key phrases, all provided with classification numbers, it is a straightforward (though not trivial) matter to write a program that can scan a document against these and can come up with a list of suggested suitable key phrases and classification numbers.

[2] See the footnote to Section 5.5 below for a definition of this term

– Dialogue mediated search. Given 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.`

– Local search. There are several possible versions of this. Here is one. It might easily happen that a user looking for information has a good example of a relevant document in his mind (for instance one of his own papers). Using the metric on the document information space it is now possible to formulate a query like: "find all documents with one or more of the following key phrases and which is within distance y of the following document".

There are many more kinds of local search using also the metric on the term information space (or both metrics). For instance, if no example document is available to localize the search, a virtual document can be introduced by specifying a set of terms. After that the query can be as before. Or, suppose one is interested in interrelations between two different parts of mathematics A and B. Describe both A and B by sets of terms, and then ask for documents whose sets of terms come close to both A and B in some way that can be specified by the metric on term information space. (E.g. for both A and B the document must contain terms that are within distance 2 of A and B).

There are basically two kinds of things that one can do using the metrics on the two information spaces. On the one hand, localize searches by specifying a centre and limiting the search to a neighborhood of that centre; on the other hand one can give oneself some more latitude by requiring only that some nearby terms (to a given term) occur in the document instead of a precisely specified one (which may easily not occur in an ideal document for the search concerned for a large variety of reasons).

3. Generation of weak enriched thesauri

The first step in (semi-)automatically generating a weak enriched thesaurus is to generate a list of key phrases (terms) from the available data. This is a well-known problem and there exist (first generation) computer programs to assist one in doing this; for instance, the program Textract, which I made use of for [12]. As said, these are only first generation programs and much remains to be done, but that is not the topic of this paper. So, assume that there is a clean and complete [3] list T of key phrases and that we know for each phrase in which documents it occurs. This gives a bipartite graph (such as depicted in Section 5 below), with as vertices the set of terms T (depicted on the left) and the set of documents D (depicted on the right) with an edge between $t \in T$ and $d \in D$ if and only if the term t occurs in (better, has been assigned to) the document d.

Given these data there are very natural metrics on both T and D. The distance$m(t, t')$ between two terms is simply the number of documents which have t but not t' plus the number of documents that have t' but not t (Hamming distance). And similarly for D: the distance $m(d, d')$ between two documents is the number of terms that are assigned to the first document but not to the second plus the number of terms assigned to the second but not to the first. For real life applications one needs more sophisticated versions of this construction. For instance by assigning weights to documents. (If two terms are assigned to the same very long document this implies grosso modo rather less of a relation than if this is the case for a short document.)

Note that this is an easily updatable structure. Also, note that it would be silly to store the metrics obtained. This would take very much storage place. Each term has a rather short list of documents to which it has been assigned. That is the inverted list of the list that gives for each document its set of key phrases. And from these two lists it is practically trivial to calculate the distance between terms, or the distance between documents, if and when needed.

4. Imperfect data: dirty data, hidden terms, and missing terms

As indicated above, given clean and complete data it is a simple matter to generate suitable metrics on the given information spaces of terms and documents; indeed the definition is explicitly given above. Unfortunately the data as currently available are far from clean and far from complete. Here I will briefly discuss three aspects of this matter.

All examples come from the two large databases of mathematics STN/FIZ MATH (*Zentralblatt für Mathematik*) and MATHSCI (*Mathematical Reviews*) and the classification scheme MSCS-1991 (Mathematical Subject Classification Scheme, 1991) used in these two, and my experiences with compiling the three large indices [10, 11, 12], and the indexing and classification work that went into [4, 7, 9].

[3] I.e., no dirty data and no missing data as discussed in Section 4 below

Ч	еб	ы	ш	ев
Ch	eb	y	sh	ev
Tch		i	sch	ef
Tsch		(j)	ch	eff
				ew
				ow (etc.)
				(iev etc.)
				(jev etc.)

TABLE 1.

Chebyshev	4103	Tchebys**ch**ev	2
Chebyshef	1	Tchebyschef	2
Chebysheff	8	Tchebyscheff	6
Cheby**ch**ev	89	**Tsch**ebyshev	2
Chebychef	1	Tschebyshef	1
Chebycheff	11	Tschebysheff	3
Chebyschev	13	Tschebyshew	1
Cheb**i**shev	1	Tschebyc**h**ev	2
Tchebyshev	34	Tschebychef	2
Tchebyshef	1	Tschebycheff	11
Tchebysheff	12	Tschebys**ch**ev	3
Tcheby**ch**ev	45	Tschebyschef	5
Tchebychef	13	Tschebyscheff	139
Tchebycheff	216	Tschebyschew	4
Tschebyschow	0*		

TABLE 2.

Note that the two databases mentioned have records consisting of (at best): author(s), title, abstract, source (bibliographic data), key phrases, classification. The full texts of the papers are not available; nor will that change in the foreseeable future. As to size, they consist of somewhat less than a million records with some 4 million key phrases in the case of FIZ/STN MATH (MATHSCI has no key phrases).

*but occurs thus in the "Kleine Encyclopaedie der Mathematik"

4.1. Dirty data

The first aspect is a familiar and obvious one: dirty data. For instance, misspellings (in particular of proper names), linguistic variants of terms (inversions, morphological variants, to some extent synonym trouble) and unusable key phrases (to long and/or too complicated).

For instance, I know of 29 ways in which the name of the Russian mathematician P.L. Chebyshev (П.Л. Чебышев) is rendered in the published mathematical literature.

Here is the example in detail. In Table 1 each (group of) symbol(s) are indicated the various transcriptions that can and have been used. Between brackets are very rare transcription occurrences.

Fortunately, not all possible combinations arise (it seems); but enough of them do in fact occur. The ones I have seen are indicated in Table 2, together with their frequency in the MATH database (a few years ago).

This type of difficulty causes two kinds of problems. On the one hand, a user, typing in a wrong spelling, is likely to find very little or nothing of what he is looking for. On the other hand if he types in the correct spelling he may well miss an important segment of the available literature. This happens e.g. with the Crank-Nicolson method from numerical analysis; quite a few papers on the topic (some 22%) have Crank-Nicholson instead.

Next, the matter of linguistic variants (morphological variants like singular-plural, composite words written with or without a dash, inversions, ...) has had a great deal of attention in the literature, and I shall say nothing about it.

Thirdly, there is the matter of unusable key phrases, usually rather long ones, which are so specific that they often apply to one document only. For instance "algebraic and differential invariants of smooth four dimensional manifolds with boundary". This one actually would be relevant for quite a few documents but still it needs to be broken up into several parts.

Much can be done to handle these matters of linguistically dirty data by linguistic means and standard lists of names and phrases. Indeed for the matter of proper names some impressive work has been done for the MATHSCI database.

There are some pitfalls, for instance "topological algebra" and "algebraic topology" refer to two very different parts of mathematics, and here the same 'identification cloud' idea that I will try to describe in Section 4.4 below might come in useful.

4.2. Hidden terms

A matter which can not be handled by purely linguistic means is that of hidden key phrases (terms). Let me describe a rather simple example.

In a record that I saw recently there occurs 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. "Dar-

boux process", to be extracted. Presumably, some sort of stochastic process like "Cox process", "Dirichlet process", or "Poisson process". The context made that rather doubtful; 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". .... The various words occurring in the surrounding sentences settled the matter. These were typical for the surrounding words of the term "Darboux transformation" and typical for the area classified by 58F07 (one of the classifications - indeed the main one - of "Darboux transformation"). Thus the *'identification cloud'* (see Section 4.4 below) 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.

This is a rather simple example. It may very well happen that various parts of a good key phrase for a paper are scattered over several (two or three) sentences, and/or that only some parts of it are present, or even that no part of an ideal key-phrase is present in the data at hand. Several examples are described in detail in Section 4.5.

4.3. Missing terms 1

In addition to hidden terms there are frequently completely missing terms. Especially terms rather more general (in level of specialization) than the subject treated in the paper may not get mentioned at all. Thus, for example, a paper dealing with the "Dyer-Lashof algebra of cohomology operations" may not have in its record any mention of "algebraic topology" (which is the field to which this subject belongs).

This also can be dealt with (semi-)automatically by considering the identification cloud of (likely) more general terms and comparing these with the set of words and short phrases occurring in the available material.

There are also other ways in which one may gain insight in the matter of missing terms, see Section 5.3 below.

One can even do much about missing terms at the same level of specialization as the paper at hand.

These matters, hidden terms and missing terms, are of importance both for completing the data of existing records and for new records (automatic assignment of key phrases and classifications). They are also most important for determining the correct metrics on both term space and document space

4.4. The identification cloud of a term

The above, in particular the example of Section 4.2, will already have made it more or less clear what is to be understood by the phrase *'identification cloud'*. Each term in the standard list of terms (key phrases) should come together

with a cloud (list) of words and phrases that are likely to be found in a text in the neighborhood of an occurrence of the term in question. This cloud will include the terms from the standard list closest to the given one in the metric of Section 3 on the information space of terms.

These identification clouds also 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. The identification cloud also serves 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.

4.5. Examples of missing terms and the uses of identification clouds

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 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 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 phrases that occur in this document and that thus strongly suggests that "Kripke structure" could be an important term to assign to this document.

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 [10, 11]. They are not complete; in particular, parts of index phrases that are themselves also suitable index phrases have not been indicated.

I should stress, that these examples are not automatically generated. Adequate lists of standard phrases for this area do not yet exist; nor are there identification clouds for these terms. These index jobs were done by hand. However, I believe that I work this way myself. I am primarily a mathematician and not really an expert in the areas of computer science from which these examples come. However, through long experience with abstracts and indices in this area, I do know which groups and phrases sort of belong together; i.e. I have some sort of identification clouds in my head and those are what I use. Afterwards, I checked whether the 'new' phrases did really fit. They did.

5. Some mathematical problems

In this section I will discuss and describe some mathematical problems and results that come out of the information retrieval issues at hand. All these problems have much to do with the (semi-)automatic generation of thesauri. Indeed, they need to be solved to be able to do just that.

5.1. Missing terms 2: missing centres of clusters[4]

The field of mathematics is fortunate in that it has a universally used, quite detailed, and very valuable classification scheme, called MSCS. The current version is that of 1991; an update is in the making (2000). This classification scheme reflects the history of mathematics and how the field split historically into various more specialized parts; it is essentially a top-down construction. Whether it fits very well with the metric structure of the information space of terms that comes from the actual present day collection of documents, i.e. some such metric as defined in Section 3 above, is another matter. I rather suspect that it does not do that very well.

Thus, it is certainly interesting to apply clustering techniques ([13, 14, 15, 16] to the metric information space of terms of Section 3 and to see what kind of hierarchical structure is suggested by the results. I call this bottom-up classification and one project in that direction (BUC'M, Bottom-Up Classification in Mathematics) is under way at CWI, Amsterdam. Another bottom up effort is the INTAS sponsored project ERETIMA involving Yaroslav University

4 Clustering in the sense of clustering theory; not to be confused with identification clouds, though the two are not unrelated; for instance, a cluster containing a given term can serve as a first approximation of an identification cloud for that term. However, as I see it, identification clouds will often contain single word terms that are useless as key phrases (for humans, because of great generality) but can very well serve to help pinpoint what subdomain one is dealing with.

and the Russian Academy of Sciences in Russia, CIS in München and CWI, Amsterdam.

Suppose then that one level of clustering has been carried out and that the space of terms has been divided into a number of (possibly overlapping) subsets, called (first level) clusters. These clusters need names and very possibly one of the terms in the cluster is that name (or a new term (new in the sense that it is not as yet in the information space, not in the sense that a new term needs to be invented) may be needed (missing terms again). In both cases we need to find the centre of the cluster in terms of the its zero one coordinates labelled by the elements of the document space (just like the original terms themselves have a zero or one coordinate for each document; see Section 3 above). A rather simple mathematical argument shows that these centres are determined by majority vote. I.e. a centre has coordinate 1 for document d if the majority of terms in the cluster has a 1 there. Ties are broken arbitrarily and there may be several (candidate) centres of a cluster. (A centre is a point that minimizes the summed squared distances of the elements of the cluster to that point.)

5.2. Missing terms 3: other sources of metric information

In the case we are considering, there is, in principle, more information that can be used (though not in the records discussed so far). What I am alluding to is a technique called 'cocitation analysis' ([17]). The basic idea is that papers on more or less the same topic will have a tendency to refer to the same source papers. Thus there is a second way to determine a metric on the document information space.

It will in any case be interesting to compare this metric with the one defined in Section 3 above. But there are also other uses one can make of it. One such I will describe here and it deals again with the missing term problem. consider again the information spaces of documents and terms, this time depicted as a bipartite graph, as shown on the left, with on the left terms and on the right documents with a term linked to a document if and only if that term has been assigned to that document. As in Section 3 this defines a metric on the space of documents. Now remove part of the terms (as indicated in Figure 1), together with all the links from these terms. The remaining bipartitie graph defines a new smaller metric. Suppose we still know the original metric on the space of documents (say from co-citation analysis). Can the removed terms be recovered? This amounts to a question of deciding when a metric is a cut metric and finding a cut representation of it. This matter has been studied a great deal but is still very far from solved, see [3]. Actually, one needs to find the best cut metric approximation to the function that is obtained by subtracting the new metric (after removal of some terms) from the original metric.

5.3. Transfer problem

Related to the above is the transfer problem: give a bipartite graph (as on the left) and a metric on one of the spaces, what is the best metric on the other space

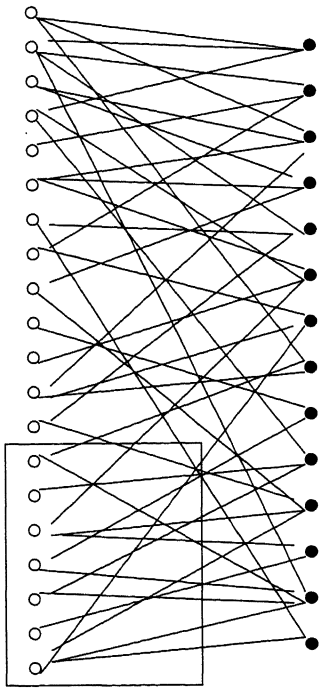

FIGURE 1.

given the bipartite relation. This problem has been discussed in some detail in [8]. Briefly the problem is to transfer metric information from document space, obtained e.g. through cocitation analysis as a clustering technique, to the term information space.

5.4. Classification trees and networks

There are many different clustering methods in the literature. All of them have some drawbacks and probably there is no such thing as a best clustering method. One of the simplest ones to understand and implement is *single link clustering.* An axiomatic analysis, [14], indicates that is also one of the better ones.

Every clustering method, applied to the metric information space of terms, yields a (classification) tree. From this the question arises how to compare such trees and which one is the best. Each such tree determines a metric: the distance between two nodes, in particular two leafs, is the minimal number of steps needed to get from one to the other. More generally, the edges (steps) can have weights with all weights issueing from the same parent node equal. Thus a classification tree defines a second metric on the information space of terms. And, given a notion of distance between metric spaces, one can speak

about a *best* classification tree. One good notion of distance between different metrics on the same set is *Lipshits distance*. Here there is a rather nice theorem available: single link clustering is best under Lipshits distance [5, 6].

This does not imply that, after all, single link clustering is a good clustering method. The drawbacks of single link techniques, particularly the phenomenon known as chaining, are well known. It rather points to the idea that trees are not really a particularly good way to describe the structure of a field of science.

In line with that, personally, I am rather sceptical about trees (= classification schemes) as a way of adequately describing a field of science. Trees are nice and there is a nice characterization of them within the class of metric spaces (the four point condition, cf. e.g. [2]). but they are also not flexible enough. For instance, I know of no good way to integrate two different hierarchical tree structures on the same set of terms (such as might be given by two different classification schemes). Something more general is needed like directed networks with a height function. These are discussed in [5].

5.5. Local partial order

The main loss of information of a weak thesaurus as defined in Section 2 above compared to a classical thesaurus is the loss of the local partial order[5] implied by the notions of broader (more general) and narrower (more special) terms. (It is not realistic to hope for a global partial order even though most literature on the topic sort of implies, or even assumes, that there is such a thing).

Given complete and clean data, it may be possible to find the local partial order structure. The idea is that statistically the more general terms (with respect to a given term) should turn up more often in the same document than more specialistic terms. No research has so far been done on this matter (as far as I know).

Of course applying clustering techniques, together with the identification of the centers (= names) of clusters also defines a partial order (given by the resulting hierarchy or tree).

6. Statistical dynamics of indexes and thesauri

The problem considered here, in this final section, is how a global index, a list of terms supposed to describe a given field of enquiry, evolves as indexing proceeds and, simultaneously, the field develops (at a far from trivial pace). To fix ideas let us think about theoretical computer science or artificial intelligence. In both cases an attempt was made to generate an adequate index of terms on the basis of a subset of the available literature [10, 11, 12]. The question arises how does such an index evolve chronologically (assuming, for simplicity, that

[5] Given a set X, a relation R on it and a covering $\mathcal{U} = \{U_i\}_{i \in I}$ of X, it can easily happen that the relation R restricted to each U_i is a partial ordering while R itself is not a partial order on the whole set X. Then we have a local partial ordering (relative to the covering $\mathcal{U}$). A local partial ordering that is not a global partial ordering has intransitivity cycles but none of these lies entirely in one of the sets U_i.

the indexing is also done chronologically), and, most important, how does one judge on the basis of these data whether the index generated is adequate for the field in question or not.

Here is a very simple (naive) stochastic model for this situation and a preliminary analysis of it. At starting time (time zero) there is an (unknown) collection, $K(0)$, of key phrases that is adequate for the field in question. In addition there is an infinite universe of potential terms that can be dreamed up by authors and others of new (important) key phrases. Thus, from the point of view of indexing and thesauri the field grows as:

$$K(t+1) = K(t) \coprod B(t),$$

where $\coprod$ stands for disjoint union and $B(t)$ is the collection of new terms generated in period t. Now indexing starts. At time zero no terms have been identified. Let $X(t)$ stand for the set of terms selected (found) at time t, $X(t) \subset K(t)$. Hence $X(0) = \emptyset$. A generalization would be that one starts with an existing thesaurus and tries to bring it up-to-date; then $X(0)$ is a known subset of $K(0)$.

The indexing proceeds as follows. At time t a set of terms $S(t)$ is selected and added to $X(t)$. This set $S(t)$ consists of two parts, $S(t) = A(t) \cup C(t)$, $A(t) \subset K(t)$, $C(t) \subset B(t)$, $A(t) \cup C(t) = \emptyset$. Thus

$$X(t+1) = X(t) \cup S(t) \subset K(t+1).$$

As a rule, of course, part of $A(t)$ is already in $X(t)$. The main problem is to have criteria or estimates to decide whether eventually $X(t)$ exhausts $K(t)$ or not. For instance in the form

$$y(t) = \frac{x(t)}{k(t)} \to 1, \quad \text{as} \quad t \to \infty,$$

where $x(t)$ is the cardinality of $X(t)$ and similarly for $k(t)$. The (only) basic observable is $S(t)$ and deriving from that $X(t)$.

Let us do some rather crude average reasoning. First, let us assume linear growth of the field of science in question:

$$k(t) = k(0) + tv$$

for some constant v. Also on average u terms are selected (per period) with a fraction $\frac{x(t)}{k(t)}$ coming from known stuff, and a fraction $\frac{k(t)-x(t)}{k(t)}$ new terms. There results a recursion equation for $x(t)$:

$$x(t+1) = x(t) + u\left[1 - \frac{x(t)}{k(t)}\right].$$

Let $y(t) = \frac{x(t)}{k(t)}$ be the fraction of terms covered by the thesaurus at this time. Then

$$y(t+1) - y(t) = \frac{u}{k(t+1)} - \frac{(u+v)y(t)}{k(t+1)}.$$

Assume that the differential equation

$$y' = \frac{u}{k(t+1)} - \frac{(u+v)y(t)}{k(t+1)}$$

approximates the difference equation above well enough (which is certainly the case). This differential equation is actually explicitly solvable and the solution is:

$$y(t) = \frac{u}{u+v} - \frac{u(k+v)^{1+(u/v)}}{(u+v)[k+(t+1)v]^{1+(u/v)}},$$

where $k = k(0)$. So

$$\lim_{t\to\infty} = \frac{u}{u+v}$$

and $y(t)$ grows monotonically from 0 to the asymptotic limit value $u/(u+v)$.

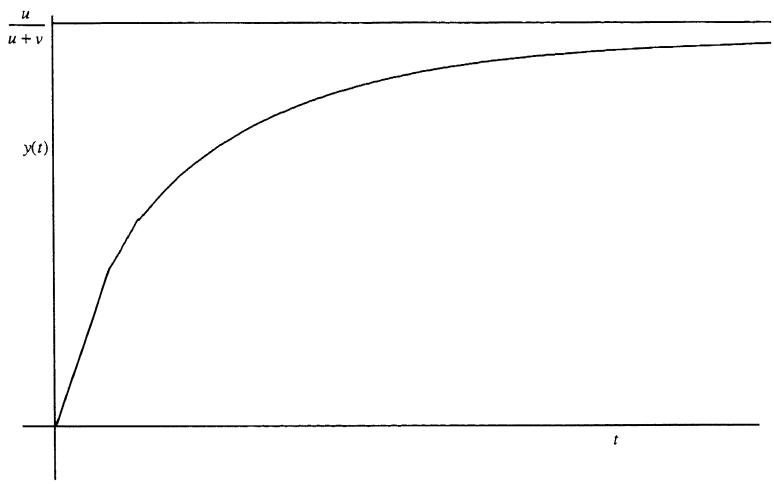

FIGURE 2.

Of course this is quite primitive. Frequently replacing stochastic phenomena with averages (in a nonlinear case) does not work. So a more sophisticated analysis of this kind of stochastic processes – apparently a new kind – is needed, as well as simulations. Research on this matter is currently carried out at the Inst. of Mathematics and Informatics of the Lithuanian Academy of Sciences in Vilnius.

REFERENCES

1. JEAN AITCHISON, ALAN GILCHRIST (1990). *Thesaurus construction*, Aslib, 2-nd Edition.

2. PETER BUNEMAN (1971). The recovery of trees from measures of dissimilarity. F.R. HODSON, D.G. KENDALL and P. TAUTU (ed.), *Mathematics in the archeological and historical sciences*, Edinburgh Univ. Press, 387–395.
3. MICHEL MARIE DEZA, MONIQUE LAURENT (1997). *Geometry of cuts and metrics*, Springer.
4. M. HAZEWINKEL (ed.) (1988–1994). *Encyclopaedia of mathematics*; 10 volumes, KAP.
5. MICHIEL HAZEWINKEL (1995). Classification in mathematics, discrete metric spaces, and approximation by trees, *Nieuw Archief voor Wiskunde* **13**, 325–361.
6. MICHIEL HAZEWINKEL. Lipshitz distance and hierarchical clustering, *J. Classification*, to appear.
7. MICHIEL HAZEWINKEL (ed.) (1995-...). *Handbook of algebra*, Elsevier.
8. MICHIEL HAZEWINKEL, (1996). Tree-tree matrices and other combinatorial problems from taxonomy, *European J. Combinatorics* **17**, 191–208.
9. M. HAZEWINKEL (ed.) (1997). *Encyclopaedia of mathematics volume* **11** (first supplementary volume), KAP.
10. MICHIEL HAZEWINKEL. (1997), *Index "Artificial Intelligence"*, Volumes 1–89, Elsevier.
11. MICHIEL HAZEWINKEL, (1999). *Index "Theoretical Computer Science"*, Volumes 1–200, Theoretical Computer Science 213/214, 1–699.
12. M. HAZEWINKEL, H. BEGO, S. VAN DONGEN, (1995). Index for volumes 101–150 of TCS, *Theoretical Computer Science* **150**:2.
13. M. JAMBU, M-O. LEBEAUX (1983). *Cluster analysis and data analysis*, North Holland.
14. NICHOLAS JARDINE, ROBIN SIBSON (1971). *Mathematical taxonomy*, Wiley.
15. BORIS MIRKIN (1996). *Mathematical classification and clustering*, KAP.
16. BORIS MIRKIN, et al. (ed.) (1997). *Mathematical hierarchies and biology*, AMS.
17. H. SMALL, E. SWEENEY (1985). Clustering the science citation index using co-citations I: a comparison of methods, *Scientometrics* **7**, 393–404.